SOFTWARE METRICS FOR AN EFFICIENT DESIGN OF ONTOLOGIES

Shazia, M. Shoaib, Iqra, K. Kalsoom, S. Majid, and F. Majeed

Department of C S & E, University of Engineering & Technology, Lahore, Pakistan.
[1]Department of Computer Science, LCWU, Pakistan
Corresponding author email: shoaib_uet@hotmail.com

ABSTRACT: Ontology plays an important role in the design of semantic web. Ontologies enable the semantic web with a detailed description of the domain concepts. To reduce the complexity of the ontology an efficient design is required. To propose the metrics for obtaining an efficient design is the main goal of this research. The metrics have been proposed along with their validation. The results are satisfactory. The future research can be carried out for the implementation of semantic web while considering the proposed design and quality metrics. Paper extracted from M.Sc. thesis (Iqra Y., Shzia. A 2009,)

Keywords: Semantic web, Ontology, Relevancy, Ranking, Merging and Domain concepts.

INTRODUCTION

With the rapid growth in knowledge, the popularity and size of ontology also increases and it causes the complexity of ontology to increase. Semantic web contributes to knowledge management and the new technology trying to make it machine-accessible so that searching becomes an easy task and its relevancy percentage can be increased by 30%.

In the semantic web (Shadbolt *et al*., 2006), Ontology is one of the important and crucial layers, so there is a need to propose and study some measures to assess its quality. Web Ontology is defined as "Ontologies are (meta) data schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine process able semantics. The defining of shared and common domain theories; ontologies help both people and machines to communicate concisely supporting not only syntax but also the exchange of semantics" (Maedche and Staab, 2001). Domain ontologies capture knowledge of one particular domain like animals, plants and humans etc. These ontologies provide a detailed description of the domain concepts from a restricted domain (Antoniou and Harmelen, 2004).

Ontology is a schema to the semantic web. Quality of information retrieved as a result of search query depends on the design, construction and quality of ontology (Berners-Lee *et al*., 2001). It provides a way of common understanding of a concept or domain between different people accessing that domain. As an example if ontologies are used in search engine they search for documents that are related semantically not only syntactically (e.g. www.hotbot.com) (Decker *et al*., 2000).

Ontologies actually furnish the semantic web so they should be managed, controlled and standardized by communities and they can be distinguish as deep or shallow ontologies according to the characteristics. Ontologies are used to define parts of data and their interactions and they can be used by anyone to extend it or reuse it according to its own use (Shadbolt *et al*., 2006).

String metric for ontology alignment is proposed in (Stoilos *et al.*, 2005). This string metric makes comparison of different Ontology, their similarities and differences and concluded that it performs better for ontology alignment. Andrew Burton-Jones *et al* (Burton-Jones *et al*., 2005) proposed a suite of metrics that assess the syntactic, semantic, pragmatic, and social aspects of ontology quality.

The main problem with the ontology is that people work only for its structure and don't take its quality and design much under consideration. As there exists a lot of issues related to ontologies, so we have to consider its structure, design and quality for our required information and task. Here we present some quality metrics for the ontology that helps to assess its quality. The Ontologies with lower quality can be improved then for better performance. For better selection of ontologies, we must have some evaluation criteria.

Semantic web is an evolving extension of the World Wide Web in which web content can be expressed not only in natural language, but also in a format that can be read and used by software agents, thus permitting them to find, share and integrate information more easily (w3c, 2009).

Schema, knowledgebase and class metrics are defined in (Tartir *et al., 2005), (*Arshad and Shah, 2007*)*. These metrics serves as mean to evaluate the quality of single ontology or to compare the different ontologies. Coupling and cohesion metrics for ontology has also been defined (Orme *et al*., 2006), (Yao *et al*., 2005). They check the level of interaction between and within the ontologies respectively.

We evaluate Ontology for its quality characteristics using some measures (Arshad and Shah, 2007). Ontology quality depends on several factors that include node similarity, attribute similarity, data usage and Ontology ranking. If the ontology quality is not good then it may brings the unwanted and irrelevant material to the user and reduce the effectiveness of the user. We therefore proposed some quality metrics for the ontology. We have used different measures in order to measure and validate these metrics.

Existing Ontology Metrics: In software development, a metric is the measurement of particular characteristics of the application performance and efficiency. Software metrics are an integral part of the state of measurement. To measure we must first define entity. In our research the entity becomes the ontology layer.

First we will see some of the existing metrics for the ontology and then we will propose some new metrics. Zhe YANG, Dalu Zhang and Chuan YE (Yang *et al*., 2006) introduced some primitive and complexity metrics as

1. TNOC (Total Number of Concepts): is the sum of concepts in the set C.
   TNOC = C = m.
2. TNOR (Total Number of Relations): is the sum of relations of each concept.
   TNOR $=\Sigma^{m}_{i=1} r_i$
3. TNOP (Total Number of Paths): is the sum of paths of each concept.
   TNOP $= \Sigma^{m}_{i=1} p_i$
4. μ: the average relations per concept. It is the ratio of TNOR to TNOC. It indicates the average connectivity degree of a concept.
5. ρ: the average paths per concept. It is the ratio of TNOP to TNOC.
   Diana Maynard *et al* (Maynard *et al*., 2006) proposed some evaluation metrics based on Precision and Recall measures that are traditionally used for the evaluation of Information extraction and their metrics also include cost components.
6. Augmented Precision: AP = BDM / (BDM + Spurious)
7. Augmented Recall: AR =BDM / (BDM +Missing)
   Where BDM= BR (CP/n0) / [BR (CP/n0) + (DPK/n2) + (DPR/n3)]
   Coupling Metrics (Orme *et al*, 2006) are introduced. Coupling is the degree of interaction between the modules so coupling metrics can be used to improve the quality of system. Given below are some of the coupling metrics.
8. Number of external classes: "NEC is the number of distinct external classes defined outside Oi but used to define new classes and properties in the ontology".

NEC(Oi)= $\Sigma^{m}_{j=1}$ E j for all 1 □□j □ m

9. Reference to external classes: "REC is the number of references to external classes in the ontology Oi".
   REC (O_i)= $\Sigma^{m}_{j=1}$ R j for all 1 □ j □m
10. Referenced includes: "RI is the number of includes at the top of the ontology definition file Oi". RI (Oi)= $\Sigma^{q}_{j=1}$ I j for all 1 □ j □ q
    Cohesion Metrics (Yao *et al.*, 2005) are introduced. Cohesion is the degree of interaction within the modules and they are used for evaluation ontology based application. Ontology cohesion finds the degree of relatedness of properties or attributes of ontology within OWL classes. Cohesion metrics are given by them as
11. Number of Root Class: "Number of Root Classes (NoR) is the number of root classes explicitly defined in the ontology Oi".
    NoR(Oi) = ΣCj for all 1 □□j □□n (number of root classes in Oi)

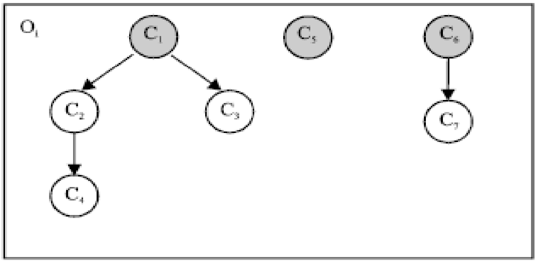


Fig. 1: Ontology O^i with 3 root classes

In the above figure 1, the ontology Oi has 3 root classes, C1, C5, C6 , thus NoR(Oi) = 3.

12. Number of Leaf Classes: "Number of Leaf Classes (NoL) is the number of leaf classes explicitly defined in the ontology Oi".

NoL(Oi) = ΣLj for all 1 □□j □□n (number of leaf classes in Oi)

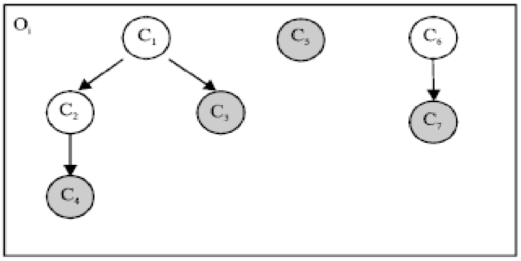


Fig. 2: Ontology O^i with 4 leaf classes

In the above figure 2, the ontology Oi has 4 leaf classes, C4, C3, C5, C7, thus NoL(Oi) = 4.

13. Average Depth of Inheritance Tree of Leaf Nodes: "Average Depth of Inheritance Tree of all Leaf Nodes, ADIT-LN is the sum of depths of all paths divided by the total number of paths".

ADIT-LN (Oi) = ΣDj /n for all Dj (Dj is total number of nodes on jth path); 1 □□j □□n (number of paths in Oi)

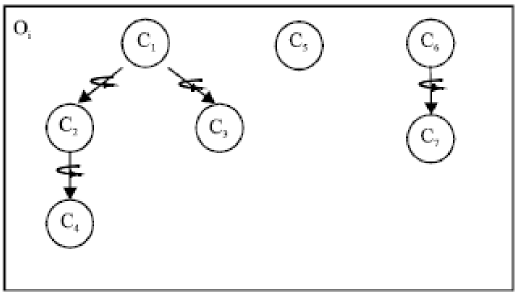


Fig. 3: Ontology O^i with ADIT classes

In the above figure 3, the ontology Oi has ADIT-LN(Oi) = 2.

MATERIALS AND METHODS

In this section we propose design and quality metrics for the ontology. If the ontology quality is not good then it may brings the unwanted and irrelevant material to the user and reduce the effectiveness of the user. We therefore proposed some quality metrics for the ontology. We have used different measures in order to measure and validate these metrics. These metrics will check the quality of the Ontologies and rank them according to their characteristics and values achieve by measuring these ontologies. These proposed metrics will give us information about how much relevant the retrieved data is and the similarity between attributes and nodes of ontology. We characterize our metrics into design and quality metrics. There are given as under.

Ontology Design Metrics: Design metrics measure the different parameters of ontology with respect to design and classify the ontology structure, its storage and its relevancies etc.

Table 1: Ontology Design Metrics

Metric Name	Description
Ontology Relevancy Metric (ORM)	The metric measures the relevancy of ontology at different subsumption hierarchy levels.
Ontology merging Metric (OMM)	It measures the characteristics of ontologies. Positively correlated ontologies can be merged to obtain efficient storing and easiness in searching.
Hierarchical Structure Metric (HSM)	It measures the Precision and Recall to classify different concepts into hierarchical structure.
Domain Specific Metric (DSM)	Use Precision and Recall measure to check whether concepts of particular domain mapped to the domain name.
Ontology Storage Metric (OSM)	Inference support, update support, querying and interfacing support are measured for efficient storage of ontologies.

Ontology Quality Metrics: Ontology quality metrics measure the quality of ontology in order to assess and evaluate it. Following are some of these metrics.

Table 2: Ontology Quality Metrics

Metric Name	Description
Node Similarity Metric (NSM)	It measures the similarity between child node and parent node of ontology tree. It intense to find cohesiveness between parent child relationships.
Attribute Similarity Metric (ASM)	It intense to measure the similarity between characteristics of attributes of different ontologies. Its purpose is to check that no two or more attributes are having same name and so helps in searching.
Data Usage Metric (DUM)	It measures the proportion of usable data coming as a result of search to the user.
Ontology Ranking Metric (ORM)	It measures the quality of ontologies made by different people and ranks it according to the flow and hierarchy of these ontologies. It helps to select the better suitable ontology.

RESULTS AND DISCUSSION

The proposed metrics has been given in the previous chapter. In order to validate our metrics the methodology is consists of two steps, the theoretical validation and empirical validation by mathematical measures as given in the figure 4.

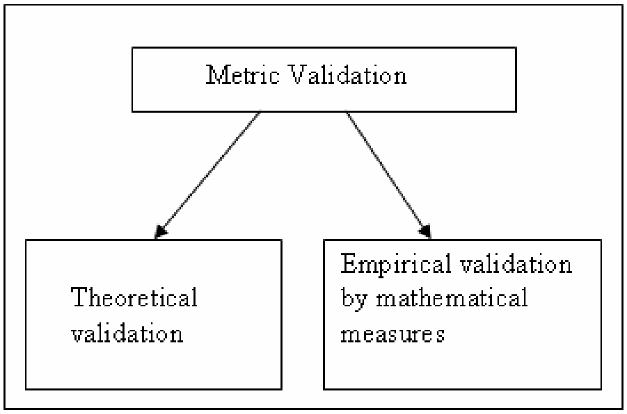


Fig. 4: Metric validation steps

The theoretical validation has been given to check the usefulness of the metrics and empirical validation is used to prove the practical utility of these proposed metrics. The theoretical discussion is as under.

Theoretical Validation

Analysis of Ontology Design Metrics: Ontology researchers have not addressed the issues related to design and quality phase of Ontology from the point of view which we discussed. Under the design phase of Ontology the requirements met by our metrics are relevancy, merging, structure, storage and construction of domain specific Ontology.

The Ontology relevancy metric (ORM) helps the Ontology developers to check how much the specific Ontology is relevant to them. This relevancy can be checked by matching the Ontology entity and its attribute with the retrieved documents. If the relevancy is greater than 50% then the Ontology is marked as relevant.

Same is the case with Ontology merging metric (OMM). Ontologies are merged according to its different factors and characteristics. Ontologies with similar characteristics are merged to make it more strong and relevant. It increases the efficiency in storing the Ontology and also easiness in searching, but on the other hand size of Ontology may increase that can cause the retrieval time to increase.

The third aspect of design is the structure of Ontology. In Hierarchical Structure Metric (HSM) after classifying the concepts into hierarchies, identification of correct and not missing concepts has been analyzed. It ensures us how much our classification is good and concepts are classified according to the hierarchy.

The metric domain specific metric (DSM) works by checking the domain name and the concepts defined within the particular Ontology. This metric help to check the domain specification. It finds the best suitable ontology according to the domain given.

The metric Ontology storage metric (OSM) addressed certain issues like inference support, update support, querying and interfacing support, and on the basis of their efficiencies storage system is categorized as efficient or inefficient.

Analysis of Ontology Quality Metrics: Quality metrics proposed quality parameters which have not been considered before. These metrics are NSM, ASM, DUM and ORM.

Node similarity metric (NSM) is used to check the cohesiveness between the parent and child relation that help the developer of Ontology to analyze the quality of the Ontology. The more the cohesion exists between parent and child node, the better is the quality of Ontology.

Attribute similarity metric (ASM) is the quality metric in which similarity between attributes of different Ontologies are checked. If there is more similarity and cohesion exist between attributes of different Ontologies then the Ontology need to be remade otherwise the Ontology will be quality oriented.

Data usage metric (DUM) ensures the amount of data that is correct with respect to the search query. If the amount of correct data is high for the specific query and relevancy is greater than 50%, then Ontology can be classified as strong and retrieved output or document can be considered as relevant.

Ontology ranking metric (ORM) ranks the Onotlogies by checking all the aspects under the category of quality metric. According to the results of all metrics Ontologies are ranked that help the developer to select the best suitable Ontology for its purpose.

Empirical Validation

Node Similarity Metric (NSM): We check the similarity between nodes of same level and also the similarity of nodes across different levels to check the ontology cohesion.

If child node is similar to the parent node and the information between parent and child node is similar than cohesion is high.

NSM $=\sum^{n}_{i=1}$ [Ari + As_i +Ad_i) / n] (for nodes in same level)

NSM $=\sum^{n}_{i=1,j=2}$ [Ar_{ij} + As_{ij} +Ad_{ij}) / n] (for nodes in different level)

Where n = total no. of levels,
i and j are different levels of ontology,
As = same attribute value,
Ar = related attribute value, and
Ad = totally different value.

If cohesion is high for the average level of Ontologies then it means relationship between parent and child is strong in the Ontology.

If NSM is near to 1 then highly cohesive and nodes have strong relation.

Attribute Similarity Metric (ASM): Attribute similarity between two Ontologies are measured by counting the number of Object Attribute Values (OAVs) appeared on two different Ontology belonging to same attribute value (As) or related attribute value (Ar) or totally different value (Ad).

It is represented by

ASM = $\sum^{n}_{i=1}$ (Ar_i + As_i +Ad_i)
i = no. of Ontology,
As = same attribute value,
Ar = related attribute value, and
Ad = totally different value.
We have defined the values as
As=1
Ar=0.5
Ad=0
Threshold = 2/i

If ASM >= threshold value it means attributes are similar and otherwise dissimilar. If attributes are similar then Ontologies need to be remade or can be merged to obtain better results.

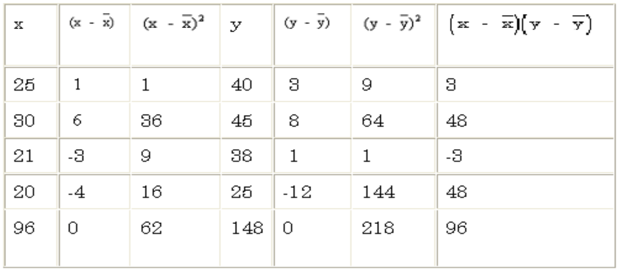
Data Usage Metric (DUM): We find the percentage of related data coming as a result of Ontology search query (information per Ontology) by correlation.

Information per ontology is defined as the information given to the search engine to retrieve the related information from the specific Ontology. If correlation between Ontology search query and retrieved result is near to 1 then it means coming data is usable to the user.

We give the word "Semantic Web" as a query. The retrieved documents contain 25, 30, 21, & 15 matching words respectively.

According to data usage metric we calculate the information retrieved from ontology. We use co-relation to measure the relevancy of retrieved documents. The calculation has been given (Arshad,.S., Shah, A) in the following figure 5 co-relation value 'r'. Here x is an independent variable & y is a dependent variable.

$$Y= \frac{\text{No of words match}}{\text{Total no of words in document}} *100$$



x	$(x - \bar{x})$	$(x - \bar{x})^2$	y	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
25	1	1	40	3	9	3
30	6	36	45	8	64	48
21	-3	9	38	1	1	-3
20	-4	16	25	-12	144	48
96	0	62	148	0	218	96

Fig. 5: Correlation Computation Steps

$$\bar{x} = 24$$

$$\bar{y} = 37$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}}$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} = \frac{96}{\sqrt{(62)(218)}} = \frac{96}{116.26} = 0.83$$

$r = 0.83$

As the value of r = 0.83 which is close to '1' so we can evaluate that the data retrieved from ontology, that are coming as a result of search query are relevant and hence all of them can be used .If the value of r is less than 0.5 then it means that the retrieved pages are not useable for the user.

This Metric can be used to find % of relevancy for any set of document retrieved against some query.

Ontology Ranking Metric (ORM): We will rank our Ontologies according to the above validated metrics. Ranking depends on the value of the node similarity, attribute similarity and the data usage metric. The higher the cohesion between the nodes of same and different levels and lower the cohesion between attributes of different Ontologies and more the relevancy exist between the search query and its retrieved document , higher is the rank of that Ontology.

CONCLUSION: Semantic web will overcome the problem of WWW by increasing its relevancy percentage and because of its machine understandable processing. Semantic web is more powerful than WWW because it contains the semantics of the words along with the syntax and different ontologies that are domain specific helps to retrieve the relevant document by checking the similarity between nodes and attributes.

In our research, we introduced some design and quality metrics for measuring ontology which can help Ontology developers and users better understand Ontology structure, hierarchy and its quality.

The presented research can be enhanced for the retrieval of domain specific concepts and more related documents in result of search query. The future research can be carried out for the implementation of semantic web while considering the proposed design and quality metrics.

Acknowledgement: This research paper is extracted from the M.Sc thesis of 3rd author which has been published at UET Lahore in 2009.

REFERENCES

Antoniou, G., and F. V Harmelen,. A Semantic Web Primer, The MIT Press Cambridge, Massachusetts London, England (2004).

Yang, Z., D. Zhang,. And C. YE, Evaluation Metrics for Ontology Complexity and Evolution Analysis, In Proceedings of the IEEE international conference on e-business engineering (ICEBE'06), 62-170 (2006).

Maynard, D., W, Peters., and Y. Li, Metrics for Evaluation of Ontology based Information Extraction, 22-26 (2006).

Orme, A. M., H, Yao., and L. Etzkorn, Coupling Metrics for Ontology-Based Systems. *IEEE Software*. 23(2): 102-108 (2006).

Yao, H., Orme, A. M., Etzkorn, L. Cohesion metrics for ontology design and application, *Journal of Computer Science*, 1(1): 107-113 (2005).

Stoilos, G., G. Stamou., and S. Kollias,. A String Metric for Ontology Alignment, ISWC, 624-637 (2005).

Burton-Jones A., V. C., Storey, V., Sugumaran, and P. Ahluwalia1, A Semiotic Metrics Suite for Assessing the Quality of Ontologies, Data and Knowledge Engineering, 55(1): 84-102 (2005).

Berners-Lee T., J. Handler, and O. Lassila The Semantic Web. Scientific American. 5910 (2001).

Shadbolt, N., and W. Hall, Berners-Lee, T. The Semantic Web Revisited, IEEE Intelligent Systems, 21(3): 96–101 (2006).

Decker, S., S. Melnik, F. V., Harmelen, D., Fensel,

M., Klein, Broekstra, J., Erdmann, M., Horrocks, I . The Semantic Web: The Roles of XML and RDF, *IEEE Internet Computing*,4(5): 63-74 (2000).

Arshad, S., and A. Shah, Design Quality Metrics For A Web Page: A Web Application. Department of Computer Science & Engineering, University of Engineering and Technology, Lahore, Pakistan (2007).

Iqra Y., and A., Shzia. Software Metrics For the Design of Ontologies**,** M.Sc Thesis UET Lahore Pakistan (2009).Maedche, M., Staab, S. Ontology Learning for the Semantic Web, (2001).

Tartir, S., I. B., Arpinar, M., Michael P. S., Amit and B. Aleman-Meza, OntoQA: Metric-Based Ontology Quality analysis, IEEE ICDM 2005 Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically heterogeneous Data and knowledge Sources (2005).