Jurnal Teknologi

Full Paper

ONTOLOGY VALIDATION ALGORITHM ON DATA DRIVEN APPROACH AND VOCABULARY ASPECT

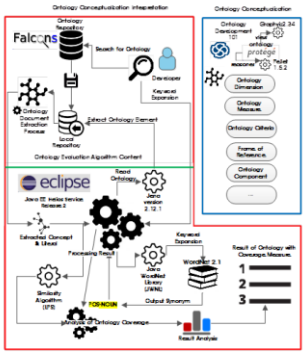
Radziah Mohamad, Nurhamizah Mohd-Hamka*

Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Article history
Received
2 February 2015
Received in revised form
8 October 2015
Accepted
12 October 2015

*Corresponding author
mieza.hamka@gmail.com

Graphical abstract



Abstract

Ontology evaluation is required before using the ontology within applications. Similar with software practice, the purpose of ontology evaluation is to identify the achievement of requirement criteria. Users who require coverage criteria often seeking ontology that contain the terms related to their focused domain knowledge. Users encounter the difficulty to select a suitable ontology from variety of ontology evaluation approaches. Conceptualization of information related to ontology evaluation helps to identify the important component within ontology that helps towards coverage criteria achievement. This work proposes an algorithm to extract ontology documents gained from public ontology repositories like Falcons into its vocabulary parts focused on classes and literals. The algorithm then processes the extracted ontology components with similarity algorithm and later displays the result on the coverage match of ontology with provided terms and the terms that are synonym expanded using WordNet.

Keywords: Data driven, ontology evaluation, similarity, coverage

Abstrak

Pengujian ontologi adalah diperlukan sebelum ontologi digunakan dalam aplikasi. Begitu juga dengan amalan perisian, tujuan penilaian ontologi adalah untuk memastikan kriteria keperluan telah dicapai. Pengguna yang memfokuskan kepada kriteria liputan selalu mencari ontologi yang mengandungi terma yang berkaitan fokus perwakilan domain mereka. Pengguna menghadapi masalah untuk memilih ontologi dari pelbagai variasi pengujian ontologi. Pengkonsepan maklumat berkaitan penilaian ontologi membantu dalam mengenal pasti komponen yang penting dalam ontologi bagi membantu mencapai kriteria liputan. Kajian ini mencadangkan satu algoritma mengekstrak dokumen ontologi dari repositari ontologi awam seperti Falcons ke dalam bahagian perbendaharaan kata seperti kelas dan terjemahan. Algoritma tersebut akan memproses komponen ontologi yang diekstrak melalui algoritma persamaan dan mengeluarkan keputusan persamaan keluasan ontologi dengan terma yang diperoleh dan sinonim bagi terma diperluaskan menggunakan *WordNet*.

Kata kunci: Berasaskan data, penilaian ontologi, persamaan, liputan

1.0 INTRODUCTION

Ontology represents a domain knowledge, thus it requires an ontology to undergo an evaluation process in order to validate the coverage of the ontology. Since various evaluation processes signify different evaluation objectives, this work focuses on gaining the similarity of ontology that represents the selected domain knowledge. When ontology were published in public repositories like Falcons, the

evaluation of the search engine repositories measures the coverage of related search term that matches with the search term given by users.

In order to gain a similarity result, the components within an ontology are extracted and compared with the terms or corpus that are related to the selected domain knowledge known as data driven approach.

Data driven approach consists of calculating the similarity of ontology content from referred set of corpus [1]. This approach is a hybrid of several other approaches from semantic similarity measurement and vocabulary approach. While other works state user profiling as their intermediate data driven approach [2], data driven to text classification [3], Noy *et al*. [4] adds the markup of user based ontology selection history on search terms to draw the analysis of ontology selection from the BioPortal ontology repositories result.

Since there are different aspects of ontology concise into different approach of evaluation [5], the harmonization of the approach need to be taken care to gain better evaluation outcome. The similarity measurement of the ontology with selected corpus has a range of several matching ontology document that contains search keywords provided by users. In addition, several ontology repositories offer the Google-like search engine for semantic Web document like Swoogle, Falcons, Watson and OntoCat. Some of the repositories proposed in the early year 2000 has discontinued publishing ontology document in several repositories that are still currently available like the aforementioned example.

The issues arise when the repositories do not show the availability of the ontology document although the semantic Web document is in the top rank of the search engine result. The availability of the ontology is achieved by checking the URI of the ontology document. This has moved towards the evaluation of the vocabulary aspect towards gaining the availability status of the URI namespace of the given ontology [5] and also the imported ontology document that is mapped with the validated ontology.

This work includes the validation of ontology coverage using Letters Pair Similarity algorithm [6] or Dice Correlation. This work aims to help users to select ontology that represents domain knowledge, thus enhances the reuse of ontology. The algorithm focuses on comparing the keywords provided with the classes and literals that are derived from the ontology document.

2.0 MOTIVATION

The objective of this work is to indicate the coverage of domain knowledge that is represented by an ontology. This is done by measuring the numbers of similarity returns from the extracted components of ontology and compare it with the terms that signify their selected domain knowledge. Ontology evaluation approach is performed using corpora of medical abstract, news articles and 19-century English novel [7]. The study suggests the avoidance of the false-positive measures due to matching terms exposed questionable natural language ambiguity and term and concept imperfect relationship. There are various methods involved in data driven approach. Vrandecic [5] suggests referring to ontology grounding [8] as it is helpful during the mentioned approach.

The study by Bouiadjra and Benslimane [9] focus on ontology evaluation from local and by searching ontology via search engine and group ontology lifecycle process into four phases. On this work, the focused phase of evaluation is on reusing the available ontology. Other work proposes the basic of data driven with ontology driven method [10]. While another work proposes an ontology driven method [11].

The approach on data driven ontology evaluation involves text corpora that validates the coverage of ontology on the domain. There is also the use of dice coefficient in Malay corpus retrieval [12] using the stemming technique. The input terms for evaluation are gained from users [4] and expert knowledge [13] in order to widen the keywords into synonym. The synonym gained from WordNet [14] is an English electronic thesauri that is related to the gained input keywords.

3.0 METHODOLOGY

In order to gain the objectives of this study, the similarity gained from ontology components towards corpus of domain knowledge, details of the process are indicated in Figure 1. There are three main parts to develop ontology evaluation approach for this work. It consists of Part A that is Ontology Conceptualization Interpretation and Part C as the Ontology Evaluation Algorithm section that is in red box and Part B as Ontology Conceptualization in blue box.

The above methodology is interpreted from OntoUji ontology that helps to conceptualize the important component and the need to deal with when it comes to ontology evaluation [15]. The three sections describe the following details in general:

1. Ontology Conceptualization Interpretation

This section indicates users upon searching for ontology from public access ontology repository like Falcons. The ontology is then downloaded and store in local repositories and to process for extraction to gain the concepts of situated within the ontology.

2. Ontology Conceptualization

This section indicates the process of ontology to be conceptualized from several literature survey upon ontology evaluation related studies by following the

ontology development methodology and tools to support ontology development like Protégé.

3. Ontology Evaluation Algorithm

This section is the main part where the process of evaluating ontology that is derived from ontology repositories. The extraction uses Jena plugin in Eclipse tools to help gathering the concept within ontology documents. The extracted concept is used during the comparison to gain similarity of corpus that represents domain knowledge with the concept from ontology. The keywords that are being used to search for the ontology are then be compared to WordNet plugin to extend the similarity of the ontology measurement besides comparing with the keywords.

4.0 EVALUATION ALGORITHM

Ontology needs to be validated before it is being used or published for usage. From various search engine identified, users will insert keywords on search input that is related to their interest. Currently, the concern of data driven approach is that, the information of related terms can only be gained once users struck the search button after keyword insertion where related keyword is the important information need to be derived from the user [4].

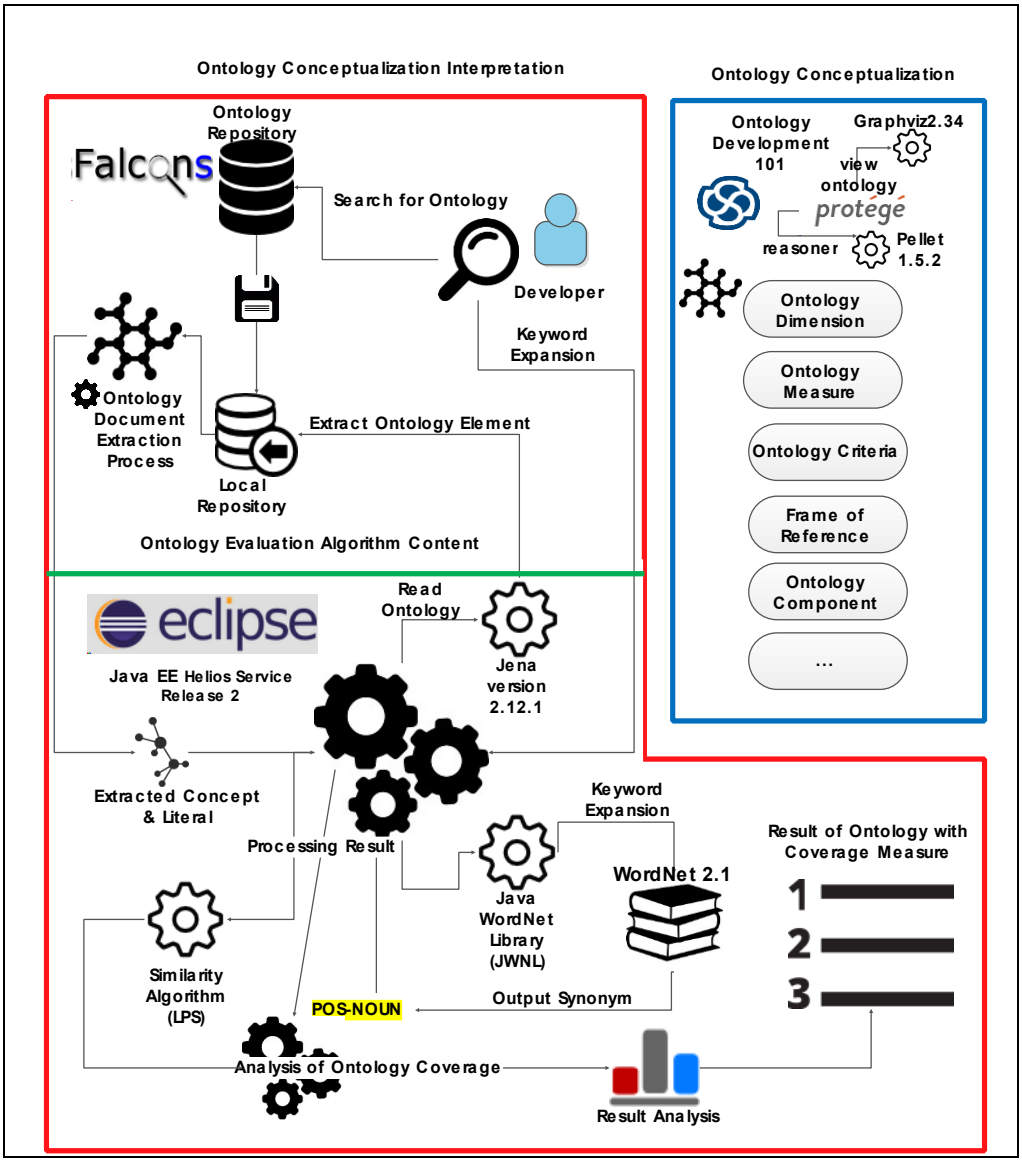


Figure 1 Research Framework

The keywords search is the important source for the evaluation and only users know what they search by the keywords. Figure 2 indicates the summary of the ontology evaluation algorithm that is gained from the process via Section 3.

4.1 Similarity Measure

The derived matching of keywords from input text undergoes calculation process to indicate the measures of ontology similarity. Listing of similarity measures gain from Euzenat and Shvaiko [16] collect eight similarity measures. The measurement gathers the Boolean similarity contains checking of ontology evaluation with Letters Pair Similarity algorithm measurement proposed by White [6] or known as Dice coefficient.

This is due to the ontology document had import or references to other ontology document gain from the Web. Here we gain the synonym of the keyword and using the technique of Natural Language Processing of Part-of-Speech (POS) tagging. The POS in type of verb were gain from WordNet library from the suggested keyword search.

Step 1: Received keyword from input panel
Step 2: Connect to WordNet and retrieve synonym from keyword
Step 3: Read ontology document
Step 4: Extract concept, literal and URI from the ontology
Step 5: Call Function to Check the availability of the URI using HTTP status return
Step 6: Call Function to Check Similarity of keywords and extracted ontology component in Step 4
Step 7: Return Result

Figure 2 Summary of Ontology Evaluation Algorithm

Figure 2 shows the outline summary of the ontology evaluation algorithm proposed. This works extend the existing works in [17] to include the vocabulary aspect of ontology via checking the availability of the URI methods [5]. The availability of the URI is check via the HTTPS request return status gain from the URI of the ontology document gain from the ontology

The similarity algorithm is based on Letters Pair Similarity algorithm proposed by White [6] and inspired by the Dice Coefficient. The metric involves are situated below.

$$similarity(s1,s2) = \frac{(2 \times |pairs(s1) \cap pairs(s2)|)}{(|pairs(s1)| + |pairs(s2)|)} \quad \text{(Eq. 1)}$$

The algorithm extracts string of keywords into pair of character as in Figure 3. The red box of group of character represents the keywords string gain from search text and the plain box represent the concept extracted from ontology document. The comparison shows the existence of matching of the keywords with the ontology concept by circulating the red thin line on the match occurrence. The matching keywords are then computed by calculating the measures using the Letters Pair Similarity algorithm.

The measures return similarity result as the following calculation. The measures are able to identify the partial matching of keywords with concept written in the ontology documents. The evaluation process will occur to match the keyword given with the concept and literal within the ontology document.

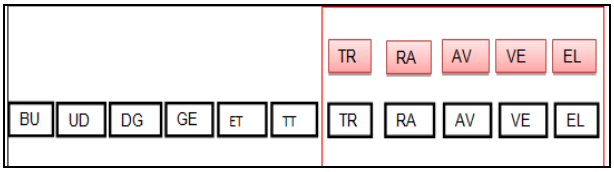


Figure 3 Keyword Similarity Overlap on Letters Pair Similarity

Figure 2 is the algorithm computed for the purpose of automatic evaluation of the ontology document gain from the ontology repository in Web. The ontology documents are gained manually from keywords search from the Falcons repository and run automatically to return the result of similarity with the given keywords.

The extraction of literals is gained from the RDF Node in order to get object property since literals are only limited in the object scope, while concept is gained from the local name that is from the concept without the allocated URI attached.

5.0 RESULT AND DISCUSSION

The evaluation process compares the gained keyword, in this example, 'travel' and compare it with the extraction of concept and literal in the ontology document that is gained from the Falcons repository. The ontology document is stored locally and directly validate thru the algorithm from folder retrieval. We open the Internet access during the evaluation since some of the ontology documents import another Web based ontology document outside the repository.

Table 1 compares the ranking of ontology when it is first searched from the Falcons repositories and the ranking when the ontology were processed using the evaluation steps proposed in this study.

Table 1 Comparison of Rank with Falcons and Proposed Algorithm

Falcons Rank	Proposed Algorithm Rank
1. O_falco3	1. O_falco10
2. O_falco9	2. O_falco15
3. O_falco5	3. O_falco6
4. O_falco6	4. O_falco5
5. O_falco7	5. O_falco4
6. O_falco13	6. O_falco7
7. O_falco4	7. O_falco13
8. O_falco10	8. O_falco16
9. O_falco16	9. O_falco9
10. O_falco15	10. O_falco3

The result shows the similarity of keywords 'travel' with the manually download ontology document from Falcons repository using the similar keyword.

Table **2** shows the synonym of the keyword 'travel' composed from WordNet. Each of the synonyms is compared to the ontology document gain and computed within the Letters Pair Similarity algorithm.

Table 2 displays the match corpus of keyword 'travel' and the number of class and literals within the ontology documents that match the corpus.

$$\text{Average LPS} = \frac{\sum \text{LPS}}{\sum(\text{Literal Hit} + \text{Concept Hit})} \quad \text{(Eq. 2)}$$

Equation (2) is used to calculate the amounts of hit occur upon the corpus and the literal and concept extracted from the ontology and computed in Table 2. The result helps to identify which of the ontology document have large number of corpus covered within its concept and literal for user own selection.

Figure **4** indicates the result of similarity upon ontology gain from Falcons repositories to corpus that was extended via WordNet. The graph was plotted from result in Table 2. The result indicates the highest measures of literal and concept match from the ontology documents goes to O_falco10 that have 0.6285 matches in Average LPS calculation.

Table 2 Result of Ontology Similarity from Falcons Repository

Ontology ID	Keyword									
	traveling	Average LPS	travelling	Average LPS	change_of_location	Average LPS	travel	Average LPS	locomotion	Average LPS
O_falco3	0	-	0	-	0	-	1	0.0163	0	-
O_falco9	1	0.02496	0	-	0	-	4	0.0301	0	-
O_falco5	0	-	0	-	0	-	3	0.3058	0	-
O_falco6	0	-	0	-	0	-	1	0.3704	0	-
O_falco7	0	-	0	-	0	-	3	0.3007	0	-
O_falco13	0	-	0	-	0	-	3	0.3007	0	-
O_falco4	0	-	0	-	0	-	3	0.3058	0	-
O_falco10	0	-	0	-	0	-	3	0.6285	0	-
O_falco16	0	-	0	-	0	-	2	0.0630	0	-
O_falco15	0	-	0	-	0	-	1	0.4167	0	-

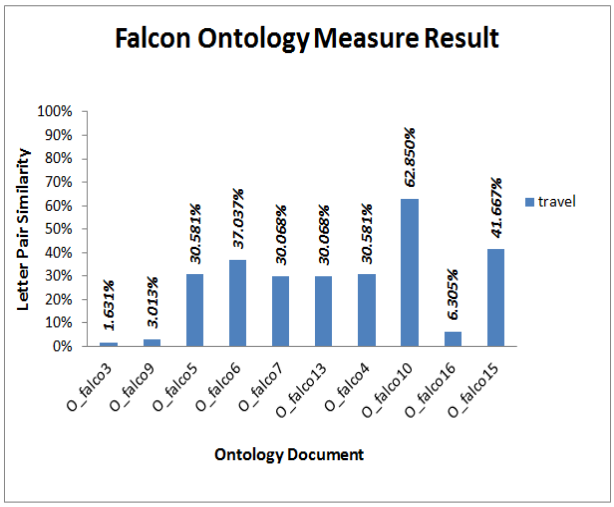


Figure 4 Graph of Falcons Ontology Measure Result

In addition, the graph indicates the lowest Average LPS calculation from ontology O_falco3 and followed by O_falco9. Although the number of match for O_falco9 has result in large number of corpus match in Table 2, the Letters Pairs Similarity calculate to match the corpus based on pairs of alphabetical array which helps to indicate the smaller part until the string pair matches the keywords that are also extracted into pairs of string array.

6.0 CONCLUSION

The increasing numbers of ontology published in public ontology repositories like Falcons, Swoogle and many more have made users upon struggle to select the suitable ontology that would match their preferred domain knowledge. Based on data-driven approach, one of the solutions is to identify ontology that had match corpus connected with their domain knowledge will enhance their ease to select the ontology.

The proposed approach focus on findings the decomposition of string into pair of letters to compare with provided keywords. The keyword also undergoes the similar decomposition process which helps to indicate how much similar does the ontology concept from provided keywords used to search for the ontology and the extension of the keywords match gain from WordNet. From the result, the higher match of concept and literal with keywords and synonym of the keywords shows that the ontology have higher match of pair string thus helps user to select ontology that suitable for their domain knowledge.

In future, the ranking of the ontology are proposed to include the availability of the ontology from the Web sources. The detection of URI ping status defined the availability of the URI of the ontology document. The Green status means that the HTTP reply from the URI gain request code of 200 which is accessible and the Red status represent others than the success code. The ranking of the ontology document is based on the similarity algorithm measurement and does not include the status of the availability of the URI.

Acknowledgement

We would like to thank Universiti Teknologi Malaysia for sponsoring the research through the RUG grant with vote number 05H83 and providing the facilities and support for the research.

References

[1] Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. 2004. Data Driven Ontology Evaluation. In *Proceedings of International Conference on Language Resources and Evaluation*. 2004: 164-169.

[2] Abdullah, N. and Ibrahim, R. 2012. Knowledge Retrieval using Hybrid Semantic Web Search. In *2012 International Conference on Computer & Information Science (ICCIS)*. 61-65.

[3] Netzer, Y., Gabay, D., Adler, M., Goldberg, Y., and Elhadad, M. 2009. Ontology Evaluation through Text Classification. In *Advances in Web and Network Technologies, and Information Management*. vol. 5731, Chen L., Liu C., Zhang X., Wang S., Strasunskas D., Tomassen S. L., Rao J., Li W.-S., Candan K. S., Chiu D. K. W., Zhuang Y., Ellis C. A., and Kim K.-H., Eds. Springer Berlin Heidelberg. 210-221.

[4] Noy, N. F., Alexander, P. R., Harpaz, R., Whetzel, P. L., Fergerson, R. W., and Musen, M. A. 2013. Getting Lucky in Ontology Search: A Data-Driven Evaluation Framework for Ontology Ranking. In International Semantic Web Conference 1. volume 8218 of Lecture Notes in Computer Science. 444-459.

[5] Vrandečić, D. 2010. Ontology Evaluation. Springer.

[6] White, S. 1992. How to Strike a Match.[Online]. Available: http://www.catalysoft.com/articles/strikeamatch.html. [Accessed: 03-Aug-2014].

[7] Yao, L., Divoli, A., Mayzus, I., Evans, J. a, and Rzhetsky, A. 2011. Benchmarking Ontologies: Bigger or Better? *PLoS Comput. Biol.* 7(1): 1-15.

[8] Jakulin, A. and Mladenić, D. 2005. Ontology Grounding. In *Proceedings of 8th Inter- national Multi-Conference Information Society IS-2005*. 170-173.

[9] Bouiadjra, A. B. and Benslimane, S.-M. 2011. FOEval: Full Ontology Evaluation. In *2011 7th International Conference on Natural Language Processing and Knowledge Engineering*. 464-468.

[10] Pivovarov, R. and Elhadad, N. 2012. A Hybrid Knowledge-based and Data-Driven Approach to Identifying Semantically Similar Concepts. *J. Biomed. Inform.* 45(3): 471-481.

[11] Yildiz, B. and Miksch, S. 2007. Ontox-A Method for Ontology-driven Information Extraction. *Comput. Sci. Its Appl.* 4707: 1-14.

[12] Sembok, T. M. T., Bakar, Z. A., and Ahmad, F. 2011. Experiments in Malay Information Retrieval. *Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011*.

[13] Zavitsanos, E., Paliouras, G., and Vouros, G. A. 2011. Gold Standard Evaluation of Ontology Learning Methods through Ontology Transformation and Alignment. *Knowl. Creat. Diffus. Util.* 23(11): 1635-1648.

[14] Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. [Online]. Available: https://wordnet.princeton.edu/. [Accessed: 03-Oct-2014].
[15] Mohd-Hamka N. and Mohamad R. 2014. OntoUji – Ontology to Evaluate Domain Ontology for Semantic Web Services Description. *J. Teknol.* 6(Special Issue on Current and Emerging Trends in Technology, Science and Engineering): 3: 21-26.
[16] Euzenat, J. and Shvaiko, P. 2007.Ontology Matching.
[17] Mohamad, R. and Mohd-Hamka, N. 2014. Similarity Algorithm for Evaluating the Coverage of Domain Ontology for Semantic Web Services. In 2014 8th Malaysian Software Engineering Conference (MySEC). 189-194.