An Integrated Ontology Selection and Ranking Model Based on the Concept and Semantic Web Link

Subhashini R[1], Dr. J. Akilandeswari[2]

[1]Department of Computer Science and Engineering, School of Engineering, Saveetha University, Chennai, India
[2] Department of Information Technology, Sona College of Technology, Salem, India

Abstract- Semantic web is based on knowledge representation which contains large number of ontologies. Ontology should be constructed in such a way that it should meet the requirements of the users. The main difficulty involved in the construction of ontologies is the high cost incurred in building them. Hence reusing ontologies is preferred. Hence, ranking ontologies has become extremely important. In this paper an approach for ranking and reusing the existing ontologies is proposed. This paper describes a novel algorithm called conTive for ranking URIs based on the semantically related content and the internal structure of ontology. Empirical analysis indicates that the integrated algorithm performs better and the ranking results are more promising and convincing that it meets the needs of the user.

Keywords: Semantic web; Ontology; Ontology ranking; ranking measures; Semantic analysis; Swoogle.

I. Introduction

The development of Semantic Web is the extension of World Wide Web, aimed to ensure better understanding of information and knowledge representation. The representation of knowledge in the semantic web is based on Ontologies. Ontology is defined as a formal explicit specification of a shared conceptualization of particular domain [1]. New ontologies can be constructed for the domain or an already existing ontology can be reused [2], [3], [6]. The main problem with the ontology construction is the high cost incurred in for building it. Therefore an approach for reusing the existing ontologies is needed instead of constructing new ontologies for knowledge representation.

The search engines play a vital role in retrieving the information required by the user. Swoogle [8], OntoKhoj [9] are widely used semantic search engines. These approaches perform ranking based on the links or concept structures. However, the retrieved URIs also contains ineffective or irrelevant information because of their poor inter-reference ontologies. This paper proposes an integrated algorithm conTive, which combines the measures of content ranking, where the content rich URIs is retrieved and these URIs are ranked with the modified measures of AKTive, thereby resulting in high quality ontologies. The novelty of this approach lies in the structure and semantic analysis of ontology with respect to a user query.

The paper is organized as follows: Section 2 discusses about the related literature, Section 3 describes the integrated algorithm conTive, which combines the measures of content ranking and the modified measures of AKTive ranking. The experimental results are discussed in section 4 and section 5 draws the conclusion and future work.

II. Related Work

Ontologies are the backbone of semantic web and it represents domain knowledge in the semantic web [10]. Many a times, the cost and effort spent to construct new ontology may not result in the desired quality.

Therefore reusing existing ontology is preferred, for which ranking is of great importance. Google uses the PageRank method to rank documents based on hyperlink analysis. Similarly search engines like Swoogle and OntoKhoj makes use of PageRank method to rank ontology. But the resulted URIs does not satisfy the user because of two reasons: The ontologies are poorly connected and more than half of them are not referred to by any other ontologies [5].

The Content Based Ontology Ranking algorithm was proposed by Matthew Jones and Harith Alani [3]. In order to rank ontologies, it finds a corpus that relates to the domain that the user requires ontology to represent. This algorithm downloads a list of ontologies from a search engine, Swoogle and the retrieved ontologies are ranked by calculating the Class Match Measure and Literal Match Measure. The total of CMS and LMS with the weight values added will be the final rank for the ontology. [3]

AKTiveRank [6] is a technique for ranking ontology based on different analytical measures that assess the ontology in terms of depth of coverage. Users can use ontology search engines (e.g. Swoogle) for searching. There are four ranking measures used to rank ontologies. They are: Class Match Measure (CMM), Density Measure (DEM), Semantic Similarity Measure (SSM) and Betweenness Measure (BEM). Finally a total score is calculated by giving a weightage for each measure [11].

Onto-DSB [10] is a ranking algorithm based on the logical measures to calculate the concept coverage of particular keyword in ontology and to retrieve the best ontology for the users' based on their search. Three logical measures are applied to the structural format. The measures are: Depthness Measure (DTM), Semantic Informative Measure (SIM) and Betweenness Measure (BEM).

III. Contive: Ranking Algorithm

A. System Design

The proposed algorithm, conTive combines content ranking and the modified measures of AKTive ranking. This paper integrates two algorithms by considering both the internal structure and the semantic web link. It overcomes the limitations of the AKTive Rank algorithm in the semantic web by introducing new and modified measures that result in building high quality ontologies. The semantic similarity measure of the Onto: DSB algorithm handles the link of the semantic web. This result in a better ranking algorithm to put forward the suitable ontology based on their application. The system architecture is depicted in the Fig 1. The user post his/her query to the search engine SWOOGLE. Based on the content, it checks in the database, if found, it retrieves the related ontologies otherwise it will download the required ontologies and save it in the database. The content ranking algorithm is applied to the retrieved URIs and the resultant URIs are content rich and is stored in the database.

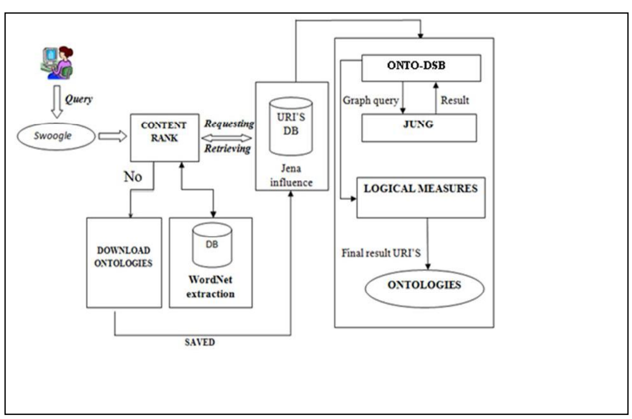

Figure 1: System Architecture of conTive algorithm

The URI's in the database is given as input to the Onto: DSB algorithm (modified measures of AKTive). The measures are applied to the preliminary ranked URIs and as a result the finally ranked ontologies are obtained. These URIs can be reused by the user based on his/her application.

B. Content based Ontology Ranking

In the content based ontology ranking, the system finds a corpus related to the domain that the user requires. It checks for the search term in the database otherwise those ontologies are downloaded from swoogle. WordNet extraction helps the user to expand the search terms more specific to the domain. The clauses used to derive the terms are synset, word sense and word which are collectively named as coordinated terms.

The content-based ontology rank [9] algorithm ranks the ontology by extracting the query terms related words through the WordNet along with two scores: Class match score (CMS) and Literal match score (LMS)

Class Match Score (CMS)

The retrieved/downloaded ontologies are ranked according to how many of the new terms match with the coordinated terms in the WordNet. The following is the formula for determining CMS [10].

$$\mathrm{CMS} = [o \in O] = \sum_{i=1}^{n} I(P_i, o) * 5 \log(n + 2 - i) \qquad (1)$$

Where, O = set of ontologies to be ranked, Pi = set of potential class labels obtained from the corpus, n = number of terms collected from corpus, I (Pi, o) = 1: if o contains a class with label matching Pi ; 0: if Pi does not appear in any of o's class labels.

Literal Match Score (LMS)

The ontologies are also analyzed to see any literal text matches the potential class labels (LMS) using the formula [10]:

$$\mathrm{LMS} = \begin{cases} I(P_i, o) = 1 & \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Total score

The total score of ontology is the combination of the scores which are weighted to emphasis the importance over the other.

$$total = \propto CMS + \beta LMS \qquad (3)$$

Where α and β are the weight factors. Two experiments are performed and the values of α and β is 0.8 and 0.2 in experiment 1 and 0.5 and 0.5 in experiment 2 respectively.

C. Onto: DSB

Onto: DSB is a ranking algorithm [20] based on the logical measures to calculate the concept coverage of particular keyword in ontology and to retrieve the best ontology for the users' based on their search. The steps involved in this algorithm are: The resulting ontologies of content ranking is given as input to the graphical file, which converts RDF files into structural format. Three measures are applied to the structural format. The logical measures are: Depthness Measure (DTM), Semantic Informative Measure (SIM) and Betweenness Measure (BEM). The total rank score is calculated by combining all the measured scores.

Depthness Measure

Depthness measure has two main steps. The first step is to count the number of classes matching the search term either partially or exactly and count their total number. The exact match is always considered better than the partial match. In the second step, the ontology classes are classified based on the three criterions follows: Number of subclasses based on the search term, Length of the classification and the relation between the classes in the ontology class set.

From the above steps, the depthness measure is calculated which serves as the input for combined ranking using the following formula:

$$E_a\,[c] = \sum_{i=1}^{3} w_i |A_i| \qquad (4)$$

$$P_a\,[c] = \sum_{i=1}^{3} w_i |A_i| \qquad (5)$$

$$DTM\,[O] = \frac{1}{n+m}\left\{ \sum_{1}^{n} E_a * \mu + \sum_{1}^{m} P_a * \beta \right\} \qquad (6)$$

Where $\mu = 0.6$; $\beta = 0.4$ and, C[o] is the set of classes in the ontology o, and t is the set of search terms, E_a (o, t) is the set contains the exact match classes in the particular ontology with the query term, P_a (o, t) is the set contains the partial match classes in the particular ontology with the query term, n is the number of exact match and m is the number of partial match.

Semantic Informative Measure

Semantic Informative measure is the summation of edge weight along the shortest path between them. The link strength is calculated through the conditional probability that occurs between the two nodes.

$$wt(C_c, C_p) = \left(\mu + (1-\mu)\frac{E}{E(C_p)}\right)\left(\frac{d(C_p)+1}{d(C_p)}\right)^{\alpha}\left[IC(C_c) - IC(C_p)\right]T(C_c, C_p) \qquad (7)$$

$$sim(C_i, C_j) = \sum_{c \in \{path(c_i,c_j) - LSuper(c_i,c_j)\}} wt(c, parent(c)) \qquad (8)$$

Where,

P (C_p) & P (C_c) are the probability of instance parent concept and instance of child concept, IC(c) – the information content of concept c, LS – Link strength, E – Average density in entire taxonomy, E (C_p) – No of children of parent class, T(C_c, C_p) – link relation / type factor.

Betweenness Measure

This measure calculates the number of shortest path that pass through each node in the graph. The nodes that are inter bonded with all the shortest paths will be scored higher than others. In the following formula, n is the number of matched class in the ontology 'o'. The formula used to calculate is:

$$bem(c) = \sum_{c_i \neq c_j \neq c \in C[o]} \frac{\sigma_{c_i c_j}(c)}{\sigma_{c_i c_j}} \qquad (9)$$

$$BEM(o) = \frac{1}{n}\sum_{k=1}^{n} bem(c_k) \qquad (10)$$

Where,

$\sigma_{c_i c_j}$ = shortest path from c_i to c_j

$\sigma_{c_i c_j}(c)$ = the number of shortest paths from c_i to c_j that passes through c

n is the number of matched classes in ontology o,

BEM (o) is the average Betweenness value for ontology o.

D. Combined Score of Analytical Measures

The overall ranking measure of onto: DSB measure is calculated after applying all the measures to the ontology. The final score is calculated using the formula [1]:

$$\text{Total score } (o \in O) = \sum_{i=1}^{3} w_i \frac{M[i]}{max_{1 \le j \le |o|} M[j]} \quad (12)$$

where M = { M[1] , M[2] , M[3]}={DEM, SIM , BEM}, W_i – weight factor, O – The set of ontology to rank.

IV. Experimental Results and Discussions

This section gives the results of the conTive algorithm for the query "Student University" submitted to Swoogle. The list of OWL files is mentioned in Table1. These ontologies are downloaded from the search engine results. Some of the ontologies are duplicated and hence it is dropped and noted as ("-"). An experiment is conducted with different weights for the measures DEM, BEM and SIM. The weightage is set as DTM 0.4, BEM 0.25 and SIM 0.35. In this case ontology i has the total score of 0.7726 marked as the highest and ontology d has 0.25505 which is the lowest. The score is given in table 2.

Table 1- List of owl files from swoogle

Search Result from Swoogle For the query "student university type: OWL"(URI)			
A	http://annotation.semanticweb.org/iswc/iswc.owl	D	http://www.tt.cs.titech.ac.jp/~fukatani/University/HU.owl
_	http://semweb.mcdonaldbradley.com/OWL/Cyc/FreeToGov/060704/FreeToGovCyc.owl	E	http://protege.stanford.edu/plugins/owl/owl-library/koala.owl
B	http://www.openmetadir.org/om2/prim-3.owl	F	http://www.tt.cs.titech.ac.jp/~fukatani/University/TMDU.owl
_	http://counterterror.mindswap.org/2005/terrorism.owl	G	http://www.tt.cs.titech.ac.jp/~fukatani/University/TITech.owl
_	http://www.csd.abdn.ac.uk/~cmckenzi/playpen/rdf/akt_ontology_LITE.owl	H	http://www.informatik.uni-bremen.de/~shi/Lehre/lang-tech-bremen-05/student-work/semantic-rep-assignment/du-liang/semantic_analysis/robot-world.owl
_	http://protege.stanford.edu/plugins/owl/owl-library/ka.owl	I	http://www.mindswap.org/2004/multipleOnt/FactoredOntologies/ItalianUniversities/ita_partition1.owl
C	http://www.mindswap.org/2004/SSSW04/aktive-portal-ontology-latest.owl	J	http://www.cs.toronto.edu/semanticweb/maponto/MapontoExamples/univ-cs.owl

Table 2- Total Score

Onto name	Total score	Onto name	Total Score
A	0.58835	F	0.7156
B	0.629	G	0.5073
C	0.4548	H	0.48775
D	0.25505	I	0.7726
E	0.5816	J	0.2582

Pearson correlation coefficient is a numerical way to quantify the relationship between two variables, e.g. X and Y and it is denoted by the symbol R. The ranks obtained by the proposed algorithm are evaluated using PCC (R) and a value of 0.9718 is obtained. This value is close to +1 which indicates that the algorithm is effective and efficient. The result obtained by the algorithm is in correlation with the results ranked by the experts. Questionnaire was given to the experts and the results are obtained. However, Swoogle obtains a value as -.0144 which clearly depicts that the value is not correlated with the human ranking.

V. Conclusion and Future Work

This paper proposes an integrated algorithm "conTive" which considers both the internal structure and the semantic web link to rank ontologies based on the user query. The URIs are applied to the content ranking and the top ranked ontologies are ranked using the analytical measures of the Onto: DBS algorithm. The experimental results show that the ranked ontologies are of high quality and it is suggested for reusing purpose. The measures of this algorithm can be modified with respect to relevant feedback from the users.

References

1. Thomas R. Gruber, 1995. Towards principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies vol.43:907-928
2. Alani, H., and Brewster, C, "Metrics for Ranking Ontologies", Proceedings of the 4th Int. Workshop on Evaluation of Ontologies for the Web (EON'06), at the 15th Int. World Wide Web Conference (WWW'06). Edinburgh, UK.
3. Alani .H and Brewster .C,"Ontology Ranking based on the Analysis of Concept Structures". In Proceedings of the 3rd International Conference on Knowledge Capture, Banff, Alberta, Canada.
4. Harith Alani,"Ontology Construction from Online Ontologies", In proc. 15th International World Wide Web Conference.
5. Harith Alani, Christoper Brewster and Nigel Shadbolt,"Ranking Ontologies with AkTive Rank".
6. Huang Xiao-xi, Zhou chang-le,"An OWL-based WordNet lexical ontology" journal of Zhejiang university science".
7. http://swoogle.umbc.edu
8. C. Patel, K. Supekar, Y. Lee, and E. Park. Ontokhoj: A semantic web portal for ontology searching, ranking, and classification. In Proc. 5th ACM Int. Workshop on Web Information and Data Management, pages 58–61, New Orleans, Louisiana, USA, 2003
9. R. Subhashini, DR. J. Akilandeswari and V. Sinthuja. Article: A Review on Ontology Ranking Algorithms. *International Journal of Computer Applications* 33(4):6-11, November 2011. Published by Foundation of Computer Science, New York, USA
10. Subhashini R, J. Akilandeswari, " A Novel approach for ranking Ontologies based on the structure and semantics", Journal of Theoretical and Applied Information Technology, Vol 65, Issue 1, 2014.