Australian Journal of Basic and Applied Sciences, 9(16) Special 2015, Pages: 312-318

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Semantic Ontology Mapping Using Ensemble Learning

[1]Sangeetha B., [2]Dr. Vidhyapriya R. and [3]Monikasree R.

[1]*Assistant Professor, Department of IT, PSG College of Technology, Coimbatore – 641004.*
[2]*Professor, Department of IT, PSG College of Technology, Coimbatore – 641004.*
[3]*PG Scholar, Department of IT, PSG College of Technology, Coimbatore – 641004.*

ARTICLE INFO

Article history:
Received 20 January 2015
Accepted 02 April 2015
Published 20 May 2015

Keywords:
Semantic web, Ontology Mapping, Similarity measures, Multi layer perceptron and gradient boosting algorithm

ABSTRACT

The semantic web is an extension of the world wide web which enables machines to analyze and respond to complex human requests based on their meaning.The retrieval of the relevant information in the web is done by making the system to understand without any human intervention. In order to automate the search, transistion for semantic web is done which requires semantic data. Ontology provides semantic data which represents a knowledge for a specific domain.It includes concepts and the relationship that naturally exist between the concepts.In various organization, they use different ontologies for the same domain which results in the problem of interoperability.Ontology mapping provides solution for the problem of combining heterogeneous and distributed data sources which is the process of determining the semantic relationship between the concepts used in different ontologies.Mapping is done to find the similarity and difference between the pair of concepts which is automated by machine learning algorithm.In this paper, the relatedness between the concepts using various similarity measures is determined and Multilayer Perceptron , a machine learning algorithm is used to perform classification.In order to improve the base learner by minimizing the error rate ,an ensemble learning method is used. One such learning algorithm is gradient boosting algorithm which is used to optimize the classification. The Performance the system is measured by using benchmark dataset provided by OAEI(Ontology Alignment Evaluation Initiative).

To Cite This Article: Sangeetha B., Dr. Vidhyapriya R. and Monikasree R., Semantic Ontology Mapping Using Ensemble Learning. ***Aust. J. Basic & Appl. Sci.,*** *9(16): 312-318, 2015*

INTRODUCTION

Over the years, the amount of information has been increasing through the world wide web continuously. All these informations are not readily available to the user and therefore searching is difficult because of unstructured and huge volume of data accessible which is scattered throughout the network. People have to connect all the sources of relevant information and interpret manually. Hence retrieval of accurate information is one of the issue in current web.In order to resolve this issue, people are intended to make the transition from current web to semantic web. The Semantic Web is the envision of the World Wide Web which is used to make web search automatically.Semantic data are used for the automated search in the semantic web.Ontology is a framework for arranging the information into an concepts and relating these concepts with each other.Ontology Alignment also known as ontology mapping which is used to map more than one ontology in the same domain to resolve the issue of interoperability. Alignment is done with several similarity measures and classification of the semantic relatedness concept is done using machine learning algorithm.This paper is arranged as follows Section II describes Related works, Section III deals about System Model, Section IV shows the implementation part, Section V depicts the accuracy of the proposed system. Application of the work is shown in section VI. Paper is concluded in Section VII.

2 *Related Works:*

Similarity measure also known as similarity function is a function which determines the similarity between two objects.In ontology mapping, two types of similarity measures involved are lexical similarity measure and the semantic similarity measure.Lexical similarity measures is the measure of degree to which the word set of two given languages are similar.Semantic similarity or semantic relatedness is defined as a metrics , where the distance between them is based on the likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation. Lexical similarity

Corresponding Author: Sangeetha B., Assistant Professor, Department of IT, PSG College of Technology, Coimbatore – 641004.
E-mail: sangeethabalaji05@gmail.com

measures includes jaccard similarity, jaro wrinkler distance, Levenshtein distance and n gram distance. Semantic similarity measures includes Hirststonge, jiang and cornath, leacock and chodrow, resnik, lesk, lin, path and Wu and palmer (David Sánchez, 2012).

2.1 Distance based similarity:

Distance based similarity is also known as path based similarity measures.

Leacock and chodorow[2] measures the similarity based on the path length.

$$sim(c1,c2) max[-log(length(c1,c2)(2.D))] \quad (1)$$

Hirst and St Onge (1998) determines semantic relationship based on the close meaning of the concept.

Weight= C–path length-KXnumber of changes in direction… (2)

where C and K are constants, $pathlength$ is the length of the path between words and $numberofchanges\ in\ direction$ denotes the number of direction changes while traversing the path between the concepts.

Wu and palmer's (1994) determine the similarity by comparing the number of relationship between the concepts.

$$sim(c1,c2) = 2H/(D1 + D2 + 2H) \ldots\ldots\ldots \quad (3)$$

where $D1$ and $D2$ are the shortest paths from $c1$ and $c2$ to c, and H is the shortest path from c to the root.

*Jaro distance metric(1995)*is also a similarity metric used to check the spelling mistakes. The Jaro similarity metric is defined by,

$$d_j = \begin{cases} 0 & , \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right), & \text{otherwise} \end{cases} \quad (4)$$

Where m is the number of matching characters, t is the number of transposition, s1 and s2 are the length of two string.

Jaro Winkler distance (1999) mainly focuses on finding similarity between words.

The *Jaro Winkler distance* is given by,

$$d_w = d_j + (lp(1 - d_j)) \ldots\ldots\ldots\ldots\ldots \quad (5)$$

Where l is the length of the common prefix and p is the constant scaling factor.

*Jaccard Similarity(*Suphakit Niwattanakul, 2013*)* is defined as the ratio between commonalities of the sets to the total number of properties.

$$J(A,B) = \frac{A \cap B}{A \cup B} \ldots \quad (6)$$

2.2 Information content based:

Information content also known as corpus based approach which measures the information contained in the concept. Information content increases with increase in the frequent occurrence of words.

$$IC(c) = -logp(c) \ldots\ldots \quad (7)$$

Where $p(c)$ is the probability of occurrence of concept c in the corpus.

Resnik (1999) uses a similarity measure based on information content.

$$sim(c1,c2) = \max(-\log P(c)) \quad (8)$$

where c is the set of concepts that subsume both $c1$ and $c2$.

Lin considers both the information content shared by two concepts and the information content of each concept respectively.

$$sim(c1,c2) = \frac{2\,XIC(c)}{(IC(c1)+IC(c2))} \ldots\ldots \quad (9)$$

Jiang and Cornath (1997) measures that the relatedness of two concepts can be scaled by the difference between the information content of the subsuming concept and of the individual concepts.It improves correlation by combining edge counting and information content method.

$$dist(c1,c2) = IC(c1) + IC(c2) - 2\,XIC(c) \quad (10)$$

$$sim(c1,c2) = \frac{1}{dist(c1,c2)} \ldots\ldots\ldots \quad (11)$$

2.3 String based similarity:

It is used to retrieve similar sentence from the source language in the database.It includes three categories such as character based,token based and hybrid based similarity measures. The character based similarity measures includes common word and overlap coefficient.The token based includes taglink token,Euclidean distance, smith waterman,jaro wrinkle, Needleman,Levenshtein distance,Dice similarity and cosine similarity measures.Based on the length of the sentence, the similarity is determined (Church, K.W. and P. Hanks,1990).

2.4 Knowledge based similarity:

Lexical similarity measures cannot always identify the semantic similarity of text.To identify the similarity based on semantic, specificity of words should be measured.It is done by using the inverse document frequency.Based on the performance in other language processing applications and high computational efficiency, similarity measures work well in wordnet hierarchy (Lin, D., 1991).

2.5 Structural similarity:

The structural methods are used to evaluate the similarity of entities and relations in two ontologies.It uses semantic flooding where it utilize the graph to compute a structural similarity between the data element.Though two entities are equal,based on the structure its descriptive information might vary and this is one of the disadvantage (Church, K.W. and P. Hanks,1990; Lin, D., 1991).

2.6 Multilayer Perceptron(MLP):

Multilayer perceptron is a neural network algorithm which utilizes the supervised learning technique.It is a acyclic graph with finite sets of

nodes. Each node is considered as neurons which contains non linear activation function .Here, neuron of ith layer work as input for neuron of the succeeding i+1th layer. A nodes which is not having any target connection is known as input neurons. A nodes which have no source connection is called output neuron. Hidden neuron is the one which is neither output neuron nor input neuron.All neurons are arranged as layer with first layer as input and the last layer as output.All neurons are enumerated. A connection exist such that either I is a successor of J as I→J or J is a Predecessor of I as J→I. All connections are having weight .The output and hidden neuron involves bias weight. Neurons use logistic, linear and tanh activation function. Both the logistic and tanh functions are used by output neurons for classification and linear activation for regression.Logistic activation is used by all the hidden layers.Boolean function in Multi layer perceptron works with one hidden layer (Dr Martin Riedmiller, 2010).

2.7 Gradient Boosting algorithm:

Gradient boosting algorithm is an ensemble based machine learning algorithm which is used for classification by reducing it to regression along with loss function. The main issue in the supervised machine learning algorithm is overfitting. Gradient based ensemble learning algorithm solve the problem of overfitting, with the help of shrinkage which is also known as learning rate.The main cause of high rate of misclassification is due to the impact of parameter variation. Gradient vector is calculated by minimizing a smooth and continuous function. It uses binomial and adaboost loss function for classification (Alexei anteing and aloes knoll, 2013).

Algorithm:

- Initialize the model with constant parameter

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

- For a number of iteration M
- ✓ Pseudo residual computation

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

- ✓ Fit a base learner to pseudo residuals
- ✓ Compute multiplier to solve one dimensional optimization

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

- ✓ Model update is done

$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.

- Output execution is done.

3 *Proposed System Model:*

This proposed approach makes the system to train for automated search and also to determine how relatively they are retrieved with closed meaning concepts.In the proposed semantic mapping approach, ontology alignment is done and several similarity measures includes jaccard similarity, jaro winkler similarity, Levenshtein distance, n gram distance, Hirst stonge, jiang cornath, Leacock chodorow, Lesk, Lin, Path, Resnik, and Wu and Palmer are implemented for the mapped ontology .

The similarity matrix is generated in the Cartesian form for the aligned ontology. These matrix is passed on to the Multilayer Perceptron which is a neural network machine learning algorithm used for classification. It consider each concept in the mapped ontology as a node i.e., neuron and the set of input layer is formed for all the concepts.These set of input layer is mapped with the set of output layer. One hidden layer is involved with logistic activation function because the Cartesian matrix uses Boolean function. Output neuron uses tanh functions. Multilayer Perceptron uses backward approach for calculating the weighted bias to minimize the rate of misclassification. But the optimum solution is not obtained from the former algorithm and overfitting still remains. In order to avoid overfitting and also to obtain the optimum solution , gradient boosting algorithm is used. This algorithm uses the base learner to apply the pseudo residual value which is calculated from the gradient vector and it act as a strong learner. Here, it uses base learner as multilayer perceptron to minimize the error.

The result analysis and performances of the system is done by using the benchmarked dataset which is provided by OAEI abbreviated as Ontology Alignment Evaluation Initiative.

4 *Implementation:*

The semantic ontology mapping is implemented by the following modules,

- ✓ Creation of ontology using Protégé
- ✓ Implementation of similarity measures
- ✓ Determination of similarity matrix
- ✓ Classification using Multilayer Perceptron
- ✓ Optimize the classification
- ✓ Performance evaluation

4.1 Creation of ontology using protégé:

A bibliography is a list of all the sources which is used for reference in the process of research work.Two different Ontologies are created for the bibliography domain with different set of concepts. These two ontologies are created using the tool called Protégé which generates the owl files.Each ontology contain 67 classes and 39 object These ontology include concept such as title, journal, author, conference, etc.,Concepts in ontologies are arranged in hierarchical and non-hierarchical manner (Dianshuang Wu, 2011).Accuracy is tested and evaluated with OAEI dataset.

4.2 Mapping two ontologies:

Ontology Mapping also known as ontology alignment which is used to integrate two different

ontology in the same domain. Alignment is done to find the problem of semantic mapping in the semantic web and semantic integration solves the problem of interoperability.Two different university ontologies are integrated in the proposed method (http://www.mkbergman.com/ontology mapping-and- alignment-tools/).

4.3 Determination of similarity matrix:

Concepts are extracted from the ontologies and stored in separate files. Then the concepts in the Cartesian form are given to both the lexical and semantic similarity measures to integrated with wordnet for generating similarity matrix (http://wordnet.princeton.edu).

Based on the semantic similarity and difference of the concept, the similarity matrix are generated for the mapped ontology.

4.4 Classification using Multilayer Perceptron:

The similarity matrix are provided as an input to the Multilayer pereptron which is a neural network algorithm is slight correlated with the output (John, A., 2014).

4.5 Optimize the Classification:

The Performance metric includes Mean absolute error, root relative error, root absolute error and root relative squared error is improved when the output of the neural network is provided as input to the optimizer (Jerome, H. friedman, 2001).

4.6 Performance Evaluation:

The Performance is evaluated to determine the range of similarity between the concepts. The Evaluation is done with the OAEI benchmark dataset in order to remove the bias of the system (http://oaei.ontologymatching.org/tests).

5 *Performance Evaluation:*

The performance of the system is evaluated using Precision and Recall.

6 *Application Of The Work:*

- Applied in data integration for interoperability
- Used in the Field of Information retrieval for the automated search when the relevant resources are scattered.

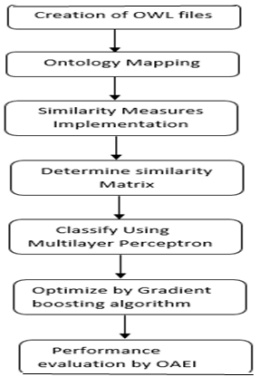

Fig. 1: Methodology.

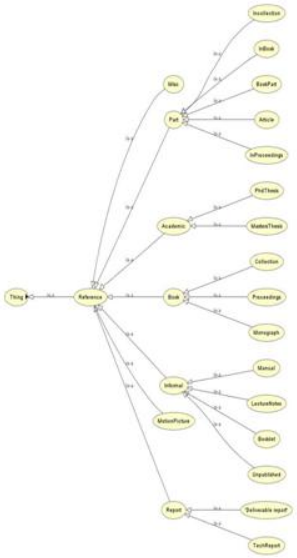

Fig. 2: Base classes of the Ontology1.

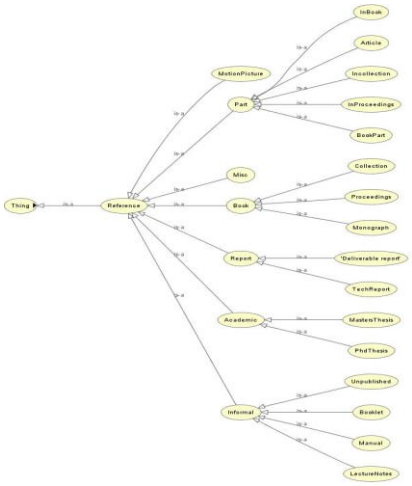

Fig. 3: Base classes of the Ontology2.

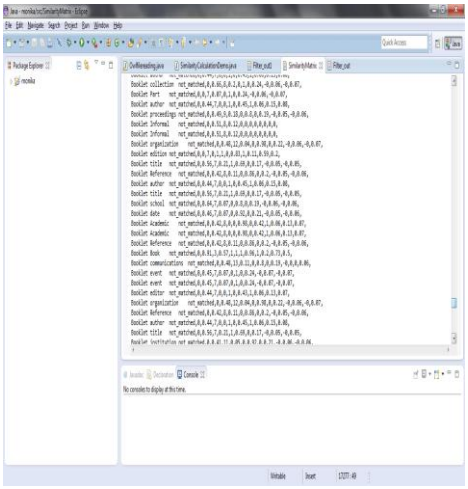

Fig. 4: Similarity Matrix.

Table 1: Result of Multilayer Perceptron for various OAEI dataset.

Dataset No	MAE(Mean Absolute Error)	RMSE(Root Mean Squared Error)	RAE(Relative Absolute Error)
101 and 203	0.0232	0.0721	29.326
101 and 205	0.0249	0.0823	33.269
101 and 224	0.0175	0.0348	26.525
101 and 225	0.0193	0.0547	27.689

Performance Comparision chart for Multilayer Perceptron:

Fig. 5: Mean Absolute error**.**

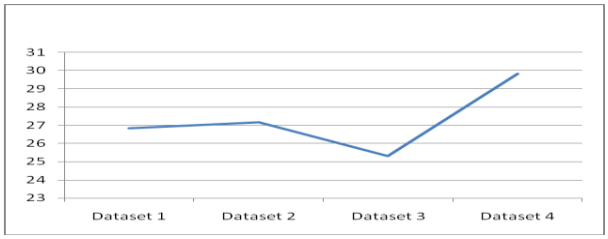

Fig. 6: Relative absolute error.

Table 2: Result of gradient boosting algorithm for various OAEI dataset.

No	MAE	RMSE	RAE
101 and 203	0.0208	0.0521	26.8271
101 and 205	0.0203	0.0771	27.1528
101 and 224	0.0210	0.030	25.3086
101 and 225	0.0213	0.0210	26.8217

Comparision chart for Mean absolute error in Gradient boosting algorithm:

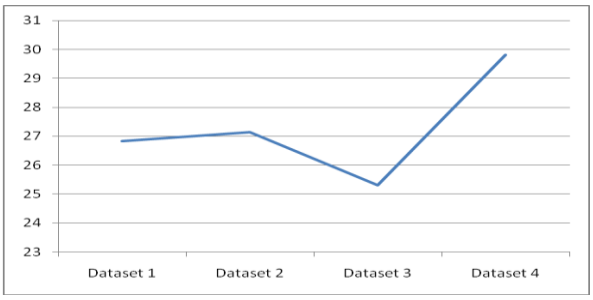

Fig. 7: Mean absolute error.

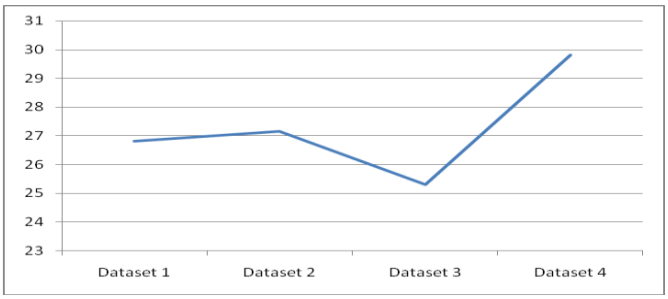

Fig. 8: Relative absolute error.

Table 3: Performance of the system for OAEI dataset.

Algorithm	Precision(%)	Recall(%)
Multilayer Perceptron	66.74	75.01
Gradient Boosting	72.85	79.07

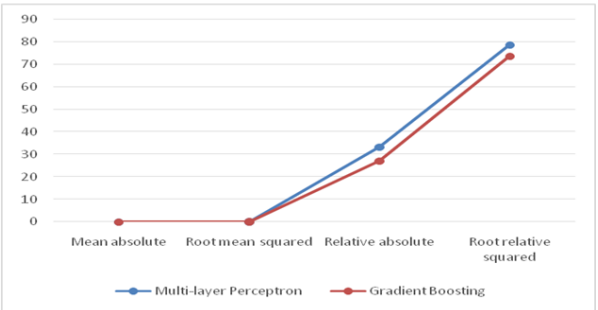

Fig. 9: Performance comparision chart for OAEI dataset.

7 *Conclusion & Future Work:*

The Proposed method trains the system with machine learning algorithm for automated search and also to check the relavant retrieval . The future work and the scope is to improve the precision and recall of the proposed work.

REFERENCES

David Sánchez , Montserrat Batet, David Isern and Aida Valls, 2012. Ontology-based semantic similarity: A new feature-based approach. Information Technology journal, 39(9):7718-7728 .

Leacock and M. Chodorow, 1998. Combining local context and WordNet similarity for word sense identification. An Electronic Lexical Database, MIT Press, 49(2): 265-283 .

Hirst, G. and D. St-Onge,1998. Lexical chains as representations of context for the detection and correction of malapropisms. MIT Press, 305: 305-332.

Wu, Z. and M. Palmer, 1994. Verb semantics and lexical selection. Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp: 133-138 .

Jaro, M.A.,1995. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 14(5-7): 491-498.

Winkler, W.E., 1999. The state of record linkage and current research problems. Statistics of Income Division. Internal Revenue Service Publication.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu, 2013. Using of jaccard coefficient for keywords similarity. International MultiConference of Engineers and Computer Scientists, 1-6.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. AI Access Foundation and Morgan Kaufmann Publishers, 11: 95-130.

Jiang and D. Conrath, 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, arXiv preprint cmp-lg/9709081997).

Church, K.W. and P. Hanks,1990. Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16(21): 22-29.

Lin, D., 1991. Using syntactic dependency as a local context to resolve word sense ambiguity. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, 64–71.

Dr Martin Riedmiller, 2010. Machine Learning: Multilayer Perceptron. Albert-Ludwig's-university, 9(11): 700-712.

Alexei anteing and aloes knoll, 2013. Gradient boosting machines, a tutorial. Department of informatics, technical university munich, Frontiers in neuroacrobatic, 7.

http://wiki.opensemanticframework.org/OntologyTool

Dianshuang Wu, Jie Lu and Guangquan Zhang, 2011. Similarity measure models and algorithms for hierarchical cases.An International Journal, 38(12): 15049-15056.

http://www.mkbergman.com/ontology mapping-and- alignment-tools/

http://wordnet.princeton.edu/.

John, A., 2014. Learning in multilayer perceptrons-backpropagation. University of bullinaria,2014.

Jerome, H. friedman, 2001. Greedy function approximation:A gradient boosting machine. IMS 1999 Reitz lecture, Annals of statistics, 1189-1232.

http://oaei.ontologymatching.org/tests.