ANALYSIS OF ONTOLOGY ALIGNMENT ON DIFFERENT DOMAINS: A CROSS ONTOLOGICAL APPROACH

[1]Vigneshwari S, [2]Dr. Aramudhan M

[1]Research Scholar, Sathyabama University, Chennai, Tamilnadu, India
[2]Perunthalaivar Kamarajar Institute of Science and Technology, Karaikal, Tamilnadu, India

Email: [1]jayam3@rediffmail.com, [2]aramudhan1973@yahoo.com

Abstract

Alignment of ontology plays a vital role in mapping ontologies. The ontologies may belong to same domain or different domains. Since there is no generic ontology available for all the domains, a multi ontological method can be very useful. Eventhough various builtin ontologies are available, in order to get precise results on a particular problem we can either reuse the suitable ontology, or some domain specific ontologies can be built.

Key words: ontology, preprocessing, tokenization, chunk, chink, BPN, probabistic weighted measures

I. INTRODUCTION

Ontologies are efficient way of getting a better idea about a particular domain of interest. Various forms of ontologies exist. Certain ontologies are reusable and are available online. Some other ontologies can be developed for a specific purpose and are used to yield precise results in the field of knowledge engineering and automated tasks. If the information between different ontologies of entirely different domain, needed to be gathered, a cross ontology mining method has to be implemented for mining the semantic data. A framework has been discussed here for advanced ontology mining which can be very much useful to the semantic society.

II. RELATED WORK

X. Tao et. al [4] proposed a model for knowledge description and formalization. Ontologies are widely used to represent user profiles in personalized web information gathering. This model learns ontological user profiles from both world knowledge base and user local instance repositories. The ontology model is evaluated by comparing it against benchmark models in web information gathering.

Gerd Stumme et. al [6] determined one of the core challenges for the Semantic Web in the aspect of decentralization. Local structures can be modelled by ontologies. In order to support global communication and knowledge exchange, mechanisms have to be developed for integrating the local systems. We adopt the database approach of autonomous federated database systems and consider architecture for federated ontologies for the Semantic Web as starting point of this work.

III. TECHNIQUES AND DEFINITIONS

A. Ontology Definition

An ontology is a representation of knowledge as a set of concepts and their relationships in a particular domain.

Formal definition of ontology is as follows :

Let C be the class, R the relationship between the classes, A the attributes, and I be the individuals then the ontology O is defined as, *O=(C, R, I, A).*

From the definition of ontology the major components are identified as classes, their relationships, the attributes and individual instances. The components of ontology are given as follows

(i) Individuals which may be instances or otherwise called objects

(ii) Classes which are the collection of concepts or sets

(iii) Atrributes which refer to the properties or the features of objects

(iv) Relationships between the classes and individuals

(v) Events which are the optional components which has the power of changing the attributes and their relationships

B. Cross ontology mapping

Ontology alignment, otherwise called ontology mapping is a process to determine correspondences between concepts . The definition of cross ontology mapping which arises from the definition of ontology is as follows:

Let X be the ontology having the components (Cx, Rx, Ix, Ax) and Y be another ontology having its components as (Cy, Ry, Iy, Ay), then cross ontology mapping referes to the matching of the components of ontology X onto the ontology Y.

The ontologies may be homogenous or heterogenous . Homogenous ontologies represent the same domain where as heterogenous ontologies map onto different domains. The type of query components on the ontologies may be atomic or complex. There is only one –to-one mapping when atomic queries are used and one-to-many mapping relationship will be followed when a complex query is used. The logical axioms of the ontology terms may be similar or different. Finally the type of relationship between the ontologies may be is-a or part-of relationships.

IV. EXPERIMENTAL SETUP

A. Multidimensional Ontology Mapping Model

Based on the probabilistic mapping criteria, local corpora are mapped onto the global corpora and the results are evaluated.The model is given in Fig. 1

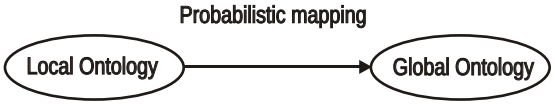

Fig. 1: Multidimensional ontology mapping model

B. Implementation

The cross ontology mapping can be experimented with two monolingual ontologies in order to extract the apt semantic information. Such a method proposed is called cross ontology simulation method.

Proposed architecture of cross ontology simulation approach is given in Fig. 2

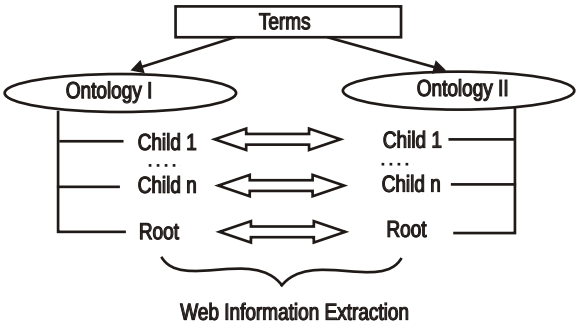

Fig. 2: Cross Ontology Mining Architecture

C. Proposed Algorithm for multidimensional ontology alignment

Steps:

1. Global ontology based on web documents is constructed by pre-processing the stemming words/stop words

2. Local ontology is also constructed in the same way as that of global one. But in the local ontology the user preferred terms are clustered and classified based on the user interests

3. Ontology mapping on both the heterogenous ontologies are done using cross ontology method, a type of multidimensional ontology mapping method

3.1. The documents are classified into positive and negative documents [4] and the weights of the terms are calculated

Let d be the document , tr_i be the terms chosen, fr_i is the frequency of those terms

Definition: $d=\{(tr_1,fr_1), (tr_2,fr_2)...(tr_n,fr_n)\}$ where fr_i is the term tr_i's frequency. In a semantic space the document frequency is calculated as $X(d)=\{(tr_1,wt_1), (tr_2,wt_2)....(tr_n,wt_n)\}$, where $wt_i(1..n)$ is the weight distribution of terms tr_i.

$$wt_i = \frac{fr_i}{\sum_{j=1}^{n} fr_j} \quad ...(1)$$

3.2. A probabilistic function on the selected terms ST was derived as $p_X(tr)$ where

$$p_X(tr) = \Sigma_{d \in Doc^+ (tr, wt) \in X(d)} \text{ support}(d) * wt \quad ...(2)$$

where Doc^+ is a positive document. After calculating the weight of the individual terms, and based on their support in the global database, the document containing these terms is evaluated

3.3. The weight of the document d is calculated as

$$weight(d) = \Sigma_{tr \in ST} pr_X(tr) * ST(tr, d) \quad ...(3)$$

$$\text{where } ST(tr, d) = \begin{cases} 1 & \text{if } tr \in ST \\ 0 & \text{else} \end{cases}$$

4. Based on the weight of the documents, the local ontology is mapped onto the global ontology

The major advantage of this approach is, it has a wide topic coverage when compared to other text based approaches

D. Experimental data setup

More than 100 web documents from en.wikipedia.org were captured . The captured documents were converted into HTML and then preprocessed. The Natural Language Toolkit (NLTK) was used. The database used for both local and global ontologies was SQL RDBMS.

E. Preprocessing the raw text

The Google search results were preprocessed into tokens as a part of preprocessing the global data.

```
>>>url=http://en.wikipedia.org/wiki
>>>raw=nltk.clear_html(html)
>>>html=urlopen(url).read()
>>>tokens=nltk.word_tokenize(raw)
>>>tokens
```

Similarly the cached web pages were taken and based on the precision and recall values of the cached web documents, the local database was constructed after preprocessing. Here the above technique is used, but the url points to the cached web pages.

F. Text Classification

After preprocessing, training and prediction were done as a part of text classification

(i) Training:

Based on the input documents, the features were extracted and sent to some machine learning algorithm. We can use the BackPropagation Network(BPN) algorithm which is effective in pattern recognition.

(ii) Prediction:

Based on the machine learning algorithms results, a classifier model was constructed where a development set and a test set were constructed and trained. This approach is a supervised learning approach.

G. Extracting Information from the text

The raw text contains a set of strings. This is segmented based on the semantic specificity into a set of semantic specific strings which are then tokenized. Then it will be undergoing part of speech tagging and the positive tags are identified. The essential entities are detected and a semantic tree is formed. From the semantic tree, ontologies are constructed and the relations between the entities are identified and the tuples are generated.

The two important techniques in parts of speech tagging are

(i) *Chunking*: Identifying the verb phrase and noun phrase

(ii) *Chinking*: Excluding the unwanted tokens from the chunks

V. RESULTS AND DISCUSSIONS

A. Analysing the semantic structure of the ontologies constructed

The following steps are considered:

Step 1: Linguistic data is taken

Step 2: Ambiguities are removed

Step 3: Parse tree is generated

Step 4: Dependencies, that is dependent terms and their relations are identified

Precision is calculated as a ratio of number of relavant – retrieved documents : no. of retrieved documents

Recall is calculated as a ratio of number of relavant – retrieved documents : no. of relavant documents

F- Measure which is the weighted harmonic mean which is calculated as

$$F = \frac{2.precision.recall}{precision + recall} \quad ...(4)$$

The performance mapping on the ontologies is also done using the following formula.

$$Perf\_increase\% = \frac{1}{n} \times \sum_{i=1}^{N} \frac{ontology\ result - target\ result}{target\ result} \times 100\% \quad ...(5)$$

B. Performance metrices of experimental results

Details of the keywords searched and their precision, recall, f-measure are discussed in Table 1

Table 1. Performance Metrics Table

Keyword	Semantic Mapping	Data Mining	XML parsing	Image Processing	Database systems
Precision	.66	.44	.78	.88	.63
Recall	.57	.52	.55	.78	.77
F-Measure	1.22	.96	1.33	.82	.69

Graphical representation of the above data is given in Fig. 3 The precision, recall and F-Measure values based ontological mapping are then compared with textual mining methods for the same set of keywords.

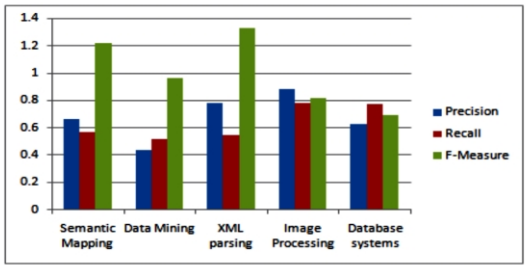

Fig. 3. Measuring precision, recall and F-Measure for different keywords

C. Comparing ontological mapping with web text mapping methods

Table 2. Comparison between Ontological mapping and Textual Mapping

Keyword	Semantic Mapping	Data Mining	XML parsing	Image Processing	Database systems
Ontology map results	90%	80%	75	77	80%
Textual Information mapping results	80%	72.10%	66.7	69.4	72.12%

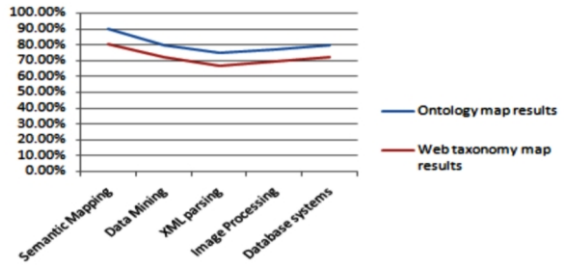

Fig 4. Comparison Chart for Ontology mapping Vs Textual information mapping

The comparison of ontological mapping with web text mapping is given in Table 2 and the comparative chart is given in Fig. 4

VI. CONCLUSION AND FUTURE WORK

The major advantage of this approach is it has a wide topic coverage when compared to other text based approaches. The focus is on HTML documents. The same approach can be extended to XML documents and the analysis is yet to be evaluated. More fast algorithms can be implemented and analysed as a future development of this work.

ACKNOWLEDGEMENT

I am grateful to my research supervisor, Dr. M. Aramudhan for his collaboration and support during preliminary investigations on this work. I would like to thank the reviewers for the efficient prepration of this paper.

REFERENCES

[1] Steven Bird, Ewan Klein,Edward Coper, Natural Language processing with Python, ORelly, 2009

[2] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, When is 'nearest neighbor' meaningful, In Proc. of ICDT-1999, Jerusalem, Israel,1999, pages 217–235, 1999

[3] McClean, S.I., B.W. Scotney, and K. Greer, A Scalable Approach to Integrating Heterogeneous Aggregate Views of Distributed Databases. IEEE Trans. on Knowledge and Data Engineering, 2003. 15(1): p. 232-235.

[4] Xiaohui Tao, Yuefeng Li, and Ning Zhong, Senior Member, A Personalized Ontology Model for Web Information Gathering, IEEE transactions on knowledge and data engineering, Vol. 23, No. 4, pp. 496-511, 2011.

[5] C.D. Manning and H. Schuetze. Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Massachusetts, 1999.

[6] Gerd Stumme, Alexander Maedche, Ontology Merging for Federated Ontologies on the Semantic Web, Institute for Applied Computer Science and Formal Description Methods (AIFB)University of Karlsruhe, Vol. 3, No. 12, pp. 1-9, 2005.

[7] Chin-Ang Wu, Wen-Yang Lin, and Chuan-Chun Wu, An Active Multidimensional Association Mining Framework with User Preference Ontology, International Journal of Fuzzy Systems, Vol. 12, No. 2, pp. 125-135, 2010.

[8] Teena Skaria, Prof.T.Kalaikumaran, Dr.S.Karthik, A Cluster Based Multidimensional Ontology Mining For Personalized Search, International journals of electronics and computer science engineering, Vol.1, No.3, pp.1390-1396, 2006.

[9] Raymond Y.K. Lau, Dawei Song, Yuefeng Li, Terence C.H. Cheung and Jin-Xing Hao, Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning, IEEE transactions on knowledge and data engineering, Vol. 21, No. 6, June 2009.

[10] Paul Buitelaar, Philipp Cimiano and Bernardo Magnini, Ontology Learning from Text: An Overview, DFKI, Language Technology Lab AIFB, University of Karlsruhe, Vol. 3, pp. 1-10, 2003

S. Vigneshwari (Srinivasan Vigenshwari) is currently a research scholar of Sathyabama University in the department of Computer Science and Engineering.She obtained her batchelors degree in Computer Science and Engineering from Madurai Kamaraj University in the year 2002. Then she obtained her Masters degree in Computer Science and Engineering from Sathyabama University in the year 2007.Her research area is web mining and ontologies.