Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A TECHNIQUE TO USER PROFILING ONTOLOGY MINING AND RELATIONSHIP RANKING

[1]S.VIGNESHWARI, [2]DR. M. ARAMUDHAN

[1]Research Scholar, Department of Computer Science and Engineering, Sathyabama University, Chennai, Tamilnadu, India
[2]Associate Professor and Head, Department of Information Technology, Perunthalaivar Kamarajar Institute of Science and Technology, Karaikal, Tamilnadu, India

E-mail: [1] jayam3@rediffmail.com , [2]aramudhan1973@yahoo.com

ABSTRACT

Ontologies play a crucial role in emerging fields of web technology like e-commerce, expert systems and so on. User preferences and user interesting topics can be captured through User Profiling Ontologies (UPO). The proposed framework discussed here, is to construct a personalized UPO, based on the interest of a particular user. The constructed UPO can be much more strengthened, by establishing relationships among the concepts. Web Ontology Language (OWL) is a standard ontology language used for knowledge representation on the web. Using the proposed information gathering model, the similarity measures between the concepts are identified. Based on the priority of the similarity measures, strong relationships can be established among the concepts in the ontology. It is obvious that more relevant results can be produced using this model, based on precision, recall and weighted harmonic mean. As well as the time of concept retrieval is also less using this approach of establishing relationships. This can be implemented by utilizing the OWL properties and the results are given.

Keywords: *Ontology, Preprocessing, Crawler, UPO, Precision, OWL, Object Properties, Semantic Similarity, Relationships*

1. INTRODUCTION

Personalization helps to reduce ambiguity and to return results which are interesting or important for a particular user. Context sensitive mining is used to extract the useful or meaningful context and to construct the ontology in that context thereby contributing to the semantic world. User profiling ontology can be constructed based on the frequent user queries. This User profiling ontology is useful for extracting the personalized information of the users. Contexts are defined through ontological user profiles. An ontological way of establishing the relationships and calculating the semantic similarity among the concepts is a better approach when compared to the existing textual taxonomies. Ontology, once constructed needed to be strengthened by assigning relations between the concepts. We need a semantic way of expressing the RDFs without changing their meaning. This can be achieved using the Web Ontology Language (OWL). Some of the important properties of OWL are Object Properties and Data Properties. Using these properties, strong relationships based on the priority values can be established among the concepts.

1.1. Related Work

Tao *et al.* [6] created a model which uses a World Knowledge Base (WKB) and user local repositories in order to capture the user history and information needs of the user. This model is a contribution to the web information gathering system.

Heasoo *et al.* [2], proposed a robust approach to organize user queries into groups dynamically and automatically. In this study, search behavior graphs like query reformation graph, query click graph and query fusion graph are generated, with the help of which, it is experimentally proved that query automation is very much useful for a collaborative search. Dynamic query grouping has also played a significant role in organizing the user search queries, which is also important in the construction of ontologies.

Yanhui and Chong [3] proposed a flexible mechanism to integrate ontologies in a multi ontology database system. In his study, a framework for ontology integration has been suggested, which combines both ontology similarity measures and ontology integration algorithms. The integrated ontology is evaluated and checked for consistency.

Haijun Zhang and Tommy W.S. Chow [5], proposed a multilevel matching method to synthesize, global and local semantics in documents, thereby improving document accuracy. It is an optimized approach even with lengthy documents.

Chin-Ang *et al* [4] , proposed a framework for user preference ontology using fuzzy linguistic terms. The framework is an active multidimensional association mining framework, where data are fetched from heterogeneous resources and finally knowledge interpretation is made. This framework is intended for automated learning purpose.

2. MATERIALS AND METHODS

2.1. User Profiling

There are two types of users like single user or user communities. For a single user only one person will have a particular profile which is personal to him alone. Other users cannot have the same personalized user context as the single user. In user communities a group of users will have a common social network and they can share their contexts. In this study, the single user profiling approach is followed.

2.2. Measures to Evaluate Context Sensitivity

The dimensions of context related approaches depend on logic based or probabilistic descriptions, semantics or meaningfulness of the contexts, static or dynamic contexts, cross ontological mappings, user profiles as a single user or a user community, scalability and preciseness of the contexts, the quality of the contextual representation and so on.

2.3. Uses of Context Sensitivity

Context sensitive mining is used for page ranking as well as for mining large scale repositories. Context sensitivity measures help to improve the precision rate of the documents. Also in order to extract precise information from the documents, context sensitivity is used. Automation of tasks is also achieved using this.

2.4. Textual Model in OWL

The textual model in OWL comprises of two different properties. They are SameClassAs construct and the Object Property construct, as represented in *Figure. 1.*

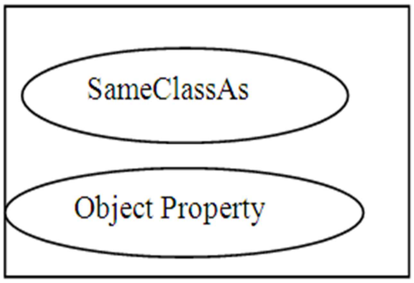

Figure. 1. The OWL Textual Model

The SameClassAs construct is used to represent equal concepts in OWL. The Object Property is used to represent the relationship between the two concepts. The other property, which is the Data type Property, represents the binary relationship between instances of the class and XML. Object Properties, which enhance the relationships among OWL classes, can be listed as [is, has, part_of, union_of, synset...]. Relationships among the concepts are established by considering the class hierarchies, which are developed using the above discussed properties.

2.5. OWL Descriptors

Two types of OWL descriptors are Textual Descriptors and Parametric descriptors. Textual descriptions contain concept names, whereas Parametric descriptors contain (attribute, value) pair.

2.6. Definition

Let c_i be the i^{th} concept and p_i be the i^{th} property in an Ontology O, then the object property can be represented in the parametric descriptor form like Equation (1):

$$\text{Object Property} = OP_O(c_i,p_i) \qquad (1)$$

Both object property and data property play a vital role in establishing semantic relationship among the OWL entities.

2.7. Richness Types

There are various types of richness described in ontological schema. The richness types are given below:

- Relationship richness
- Inheritance richness
- Attribute richness
- Class richness

2.8. The User Profiling Ontology Model

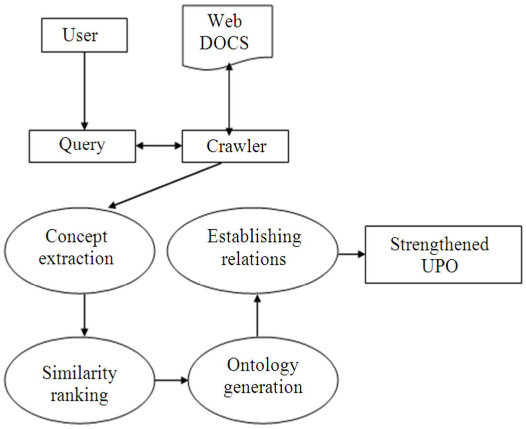

Figure. 2. UPO model

Figure 2 represents the User Profiling Ontology (UPO) model. The data from the World Wide Web (www) is extracted. The data collection from the web, can be done with the help of crawlers as proposed by Dimitrios and Georgios [1]. We can use the, crawler which downloads the web pages visited and encodes the web pages using vector space representation, thereby extracting the most important terms. Then the raw data are cleaned. After preprocessing, ontology is constructed based on the important contexts from the retrieved documents. Each node in the ontology corresponds to a particular concept. Thereby the database is created for the ontology to store the weighted concepts in descending order and thus the user interesting topics are ranked.In this work Wordnet 2.0 is used as a crawler.

2.9. Step for Preprocessing

- Semi-structuring the raw data
- Standardizing the terms
- Local consistency checking
- Global consistency checking
- Document redundancy checking

2.10. Methodologies for Mapping web Pages Used in this Model

- Document clustering using hierarchical agglomerative approach
- Document classification where keyword extraction is performed

The documents can be clustered using k-means approach, where a centroid is calculated and the relevant documents form a cluster. Classification is done to distinguish between the created clusters. PLSA technique (Dimitrios and Georgios, [1]) can be used for removing the redundancy.

2.11. Procedure for UPO construction

Let d be the document retrieved from the web, ci be the context relevant to the document, wi be the weight term of each document:

```
For every document d
      If $c_i$ of user interest maps with d then
         Sim ($c_i$, d) is calculated where Sim () is the
            Similarity function
         $C_i*w_i$ for each context is calculated ,
thereby
               ontologies are constructed
        A priority queue is constructed to store $c_i*w_i$
for
          ranking the active concept
      End If
End For
```

2.12. Semantic Similarity Algorithm

Steps:

i. Parse OWL ontology
ii. Translate OWL into RDF triples
iii. Assign relationships among the concepts, using the object properties.
iv. Rank the concepts based on the priority of relationships, thereby enhance the user background knowledge.

2.13. Implementation

Term vector is calculated for each document D for all v ∊ D. In a document vector model, mutual information is a non-negative and symmetric one. Let $I(c_g,c_l)$ be the mutual information between the concepts c_g and c_l, where c_g is a concept in the global ontology.Here $P(c_g,c_i)$, is the probability of the co-occurrence of the concepts in both the global ontology and in the local ontology.$P(c_g)$ is the probability of occurrence of the concepts in the global ontology alone and $P(c_l)$ is the probability of occurrence of the concepts in the local ontology alone. and c_l is a concept in the local ontology. $I(c_g,c_l)$ is calculated using the following formula, which is given in Equation 2:

$$I(c_g,c_l) = \sum_{c_l \epsilon L} \sum_{c_g \epsilon G}^{P(c_g,c_l)} \frac{P(c_g,c_l)}{P(c_g),P(c_l)} \quad (2)$$

From different documents indexed under different contexts datasets are created. After preprocessing, the indexed documents are classified into three sets. They are as follows:

- Training set for ontology learning
- Test sets for searching the context in the document collection
- Profile set for analyzing the user profile

Using mutual information mechanism from equation (2), the meaningful concepts are generated and thereby, the relations between the concepts can be discovered.

The concept with higher similarity value will be given higher priority, thereby having the object property 'Is_a'. Similarly for all the concepts, relationships are assigned according to their priority levels of their similarity values. The higher priority value is set as a threshold. Calculation of the priority levels is based on the logical inference reasoning. Figure 3 represents the mapping of similarity value 0.75 with the Is_a relationship

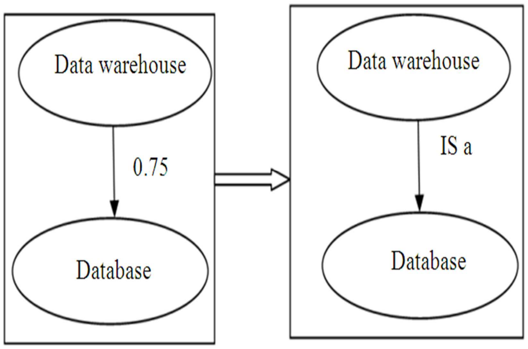

Figure 3. Concept Matching Vs Property matching

Using Wordnet2.0, the Wikipedia pages are fetched. The concepts are extracted and meaningful words are identified using Mutual Information technique. Protégé 4.3, is used for ontology generation. Based on the rank, relationships are established among the concepts.

3. RESULTS

3.1. Logical Inferencing Procedure

Let C_i, C_j be two concepts, T be the threshold and sim (C_i, C_j), be the similarity function, the various levels of relationships will be assigned using the following procedure:

```
Repeat
    If sim($C_i$, $C_j$) >=T, then
    Assign higher relationship between $c_i$ and $c_j$
Else
    Assign next lower level of relationship between $c_i$ and $c_j$
End if
Until all the concepts has been assigned relationships
```

For example we can have relationships like the following:

DB_Model → Has→Network
DB_Model →is a→Database model
Database→ Part of→Database Languages
DBMS→Union of→ Db_systems

For example in Figure 4, the following class hierarchy,which has been highlighted in the figure, is considered:

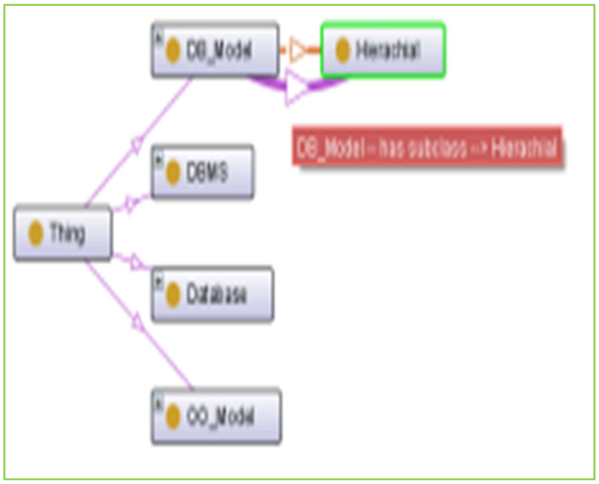

Figure 4. Class hierarchy

DB_Model →has subclass →Hierarchical

Figure 5, depicts the strengthened ontology. For example in Fig 5, the relationship (DB_Modeling is Part_of DBMS), has been highlighted. Thus we can strengthen the existing ontologies by establishing relationships among the classes using object properties based.The dotted lines represent relationships between the concepts.

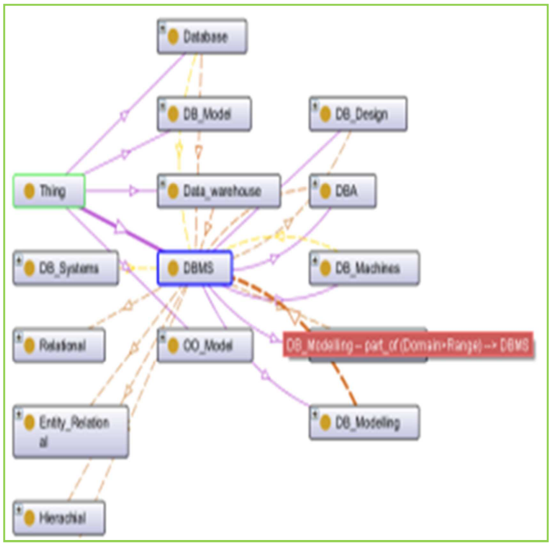

Figure. 5. Strengthened Database Ontology

From the similarity calculation, from equation 2, the concepts having highest similarity values come under Is_a. The priority of relationships is thus calculated, having the priority order as follows. **Table 1**, is the priority table, which represents the various relationships from higher priority to a lower priority.

Table 1. Priority Table

Priority	Relationship
I	Is a
II	Has
III	Part Of
IV	Union Of

4. DISCUSSION

4.1. Evaluating the Strengthened UPO

Documents can be evaluated using TF-IDF where TF is the term frequency, IDF is the inverse document frequency. Precision can be defined as the number of relevant document documents which can be calculated like the one given in Equation 3:

$$\text{Precision}=\frac{|\{relevant\ documents\}\cap\{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (3)$$

Recall deals with the successfully retrieved relevant documents which can be calculated as given in Equation 4:

$$\text{Recall}=\frac{|\{relevant\ documents\}\cap\{retrieved\ documents\}|}{|\{relevant\ \ documents\}|} \quad (4)$$

F-Measure means the weighted harmonic mean which can be calculated from Equation 5:

$$\text{F-Measure}=\frac{2*precision*recall}{precision+recall} \quad (5)$$

Table .2 Evaluation Of Document Retrieval

	Precision	Recall	F-Measure
Without relationships	.44	.52	.45
Using relationships	.88	.78	1.37

Table 2 shows the evaluation metrics of the documents retrieved, on establishing relationships between the concepts, in the ontologies, and without establishing relationships. This was experimented with 100 documents. The precision, recall and F-measure values of many of the concepts are more, for the constructed UPO,on introducing relationships when compared to the one without establishing relationships.The generated UPO can also be evaluated using time based metrics with the help of standard datasets using the proposed technique of similarity measures.

4.2. Using Benchmark Data Sets

In Table 3, Semantic Web Technology Evaluation Ontology (SWETO), the benchmark data set for ontology evaluation is used.

Table 3. Comparison Table Using SWETO Data Set

SWETO small	Time (ms)	Memory usage (bytes)
Without relationships	10920	891048
Using relationships	5350	1311704
SWETO Medium	**Time (ms)**	**Memory usage (bytes)**
Without relationships	12830	982246
Using relationships	6450	1521708

In this study, SWETO is used to compare ontologies using relationships and ontologies which do not use any relationships among the concepts.SWETO ontology is an open source ontology which can be freely downloaded from the following website. (http://archive.knoesis.org/library/ontologies/sweto/) SWETO small and SWETO medium are considered for evaluation.

From the Time Chart given in Figure. 6, it is depicted that the concepts can be quickly retrieved on establishing relationships among them. For both SWETO small and SWETO medium ontologies, the retrieval time of the concepts is drastically reduced on establishing the relationships when compared to the execution time without introducing relationships on the same ontologies.

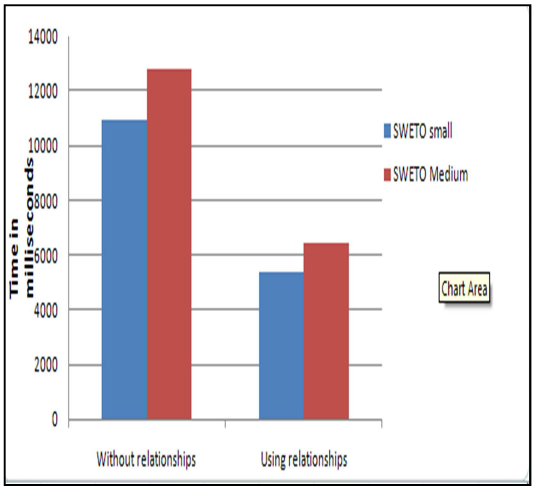

Figure 6. Time Chart

5. CONCLUSION AND FUTURE WORK

Automatic discovery of taxonomies from the user profiles can be done using ontologies. Usually metadata in the ontologies is captured and stored in repositories. Particular user profile can be monitored over time and results can be ranked based on user interests. Based on the similarity between the concepts, relationships between individual classes have been assigned. By applying inference rules we can mine the constructed ontology. Various reasoners like FACT++ and Pellet are available as open source. Using the reasoners we can cross check the semantics of the relations among the concepts. The relations have been applied on a single ontology, that is the UPO, which can also be applied to multiple ontologies and form a cross ontology relationship as well as the documents can be indexed based on similarity measures of the ontologies, which is to be considered as a future development of this study.

REFRENCES:

[1] Dimitrios Pierrakos and Georgios Paliouras. “Personalizing Web Directories with the Aid of Web Usage Data”. *IEEE Transactions on Knowledge and Data Engineering*, vol 22(9):pp 1331-1344, 2010

[2] Heasoo Hwang, Hady W. Lauw.Lise Getoor, Alexandros Ntoulas, “*Organizing User Search Histories*”,IEEE Transactions on Knowledge and Data Engineering, Vol. 24. No.5 May 2012, pp. 912-925

[3] Yanhui Lv ,Chong Xie,"A Framework for Ontology Integration and Evaluation",*Intelligent Networks and Intelligent Systems (ICINIS), 2010* 3rd International Conference on Nov. 2010 Page(s): 521 - 524

[4] 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(2): pp 125-135

[5] Haijun Zhang, Tommy W.S. Chow,” A multi-level matching method with hybrid similarity for document retrieval.” *Expert Systems with Applications* vol 39: pp 2710- 2719.

[6] Tao, X., Y. Li and N. Zhong,. “A personalized ontology model for web information gathering.” *IEEE Transactions on. Knowledge and Data Engineering.*, vol 23: pp 496-511.