CONCEPTUAL SIMILARITY MEASUREMENT ALGORITHM FOR DOMAIN SPECIFIC ONTOLOGY

Zin Thu Thu Myint[1] and Kay Khaing Win[2]

[1]University of Technology, Yadanarpon Cyber City, Near Pyin Oo Lwin, Myanmar
[2]Department of Advance Science and Technology, Nay Pyi Daw, Myanmar

ABSTRACT

This paper presents the similarity measurement algorithm for domain specific terms collected in the ontology based data integration system. This similarity measurement algorithm can be used in ontology mapping and query service of ontology based data integration system. In this paper, we focus on the web query service to apply this proposed algorithm. Concepts similarity is important for web query service because the words in user input query are not same wholly with the concepts in ontology. So, we need to extract the possible concepts that are match or related to the input words with the help of machine readable dictionary WordNet. Sometimes, we use the generated mapping rules in query generation procedure for some words that cannot be confirmed the similarity of these words by WordNet. We prove the effect of this algorithm with two degree semantic result of web mining by generating the concepts results obtained form the input query.

KEYWORDS

Semantic Similarity, Ontology, Concepts, Triplets & Query Processing

1. INTRODUCTION

Ontology based data integration system can support the virtual web portal to the users' view because users can get the uniform access to the multiple data sources that are located in separated locations. By considering the architecture, it has three main processing phases that are ontology creation, ontology mapping and web query service. In ontology creation, the expert at each local source builds the local ontology by using their domain concepts. So, the domain concepts at each local source may be various and it causes the semantic conflicts when integrating these local ontologies [1, 2]. It is a reason to create the mapping rules that help to solve the semantic conflicts among multiple ontologies. Mapping rules mean to construct the semantic relations such as equivalent, hyponym and homonym among multiple ontologies [3, 4]. These rules are applied not only to assist the integrated ontology to handle the semantic conflicts but also to help the web query service by the matching the words contained in the user input query with the domain concepts in ontology. However, these mapping rules are not sufficient to extract the terms relating the domain concepts from the user input query because user can submit the query without knowing exactly the domain concepts in ontology. The words in the input query may not be wholly equivalent to the terms contained in the mapping rules. For this reason, other similarity measurement methods become to add in term matching process of query service.

DOI : 10.5121/ijitmc.2014.2204

Query service in ontology based data integration system contains the triplets' extraction, query generation and retrieval of knowledge from the ontology. Similarity measurement is mainly suppose the triplet extraction process that can extract the specific triplets from the user input query and can add the necessary information to build the ontology understanding query such as SPARQL [5, 6]. There are many similarity measurement methods to match the keywords in the input query and the domain concepts in the ontology. However, they return the many possible concepts that are similar to the input keywords and so; their precision and recall degrees are low at query processing in ontology based data integration system [7, 8, and 9]. The proposed semantic similarity algorithm can reduce the problem of retrieving unnecessary information from the ontology by finding the most closet concepts in ontology that are similar to the terms in user input query with the help of machine readable dictionary WordNet. This also gives the specific triplets to suppose the query generation procedure. So, the precision and recall degree of this proposed algorithm is very high.

This paper is organised as follows. In Section II, this paper presents the overview of ontology based data integration system. The detail of this proposed similarity measurement algorithm is described in Section III. Section IV will fully explain the experimental results based on the precision and recall rates by generating the SPARQL query and retrieving the require information from ontology. Finally, we conclude the presentation of this paper with some remarks in Section V.

2. ONTOLOGY BASED DATA INTEGRATION SYSTEM

Using ontology in data integration systems is an ideal solution to handle the semantic conflicts between various data sources [10]. There are two trends to use the ontology in data integration system: one use for translating query or their result and the other uses ontology for the generation of global schema [11]. The system presented in this paper uses both of these two trends for data integration and accessing data on integrated ontology. The system architecture is depicted in figure 1.

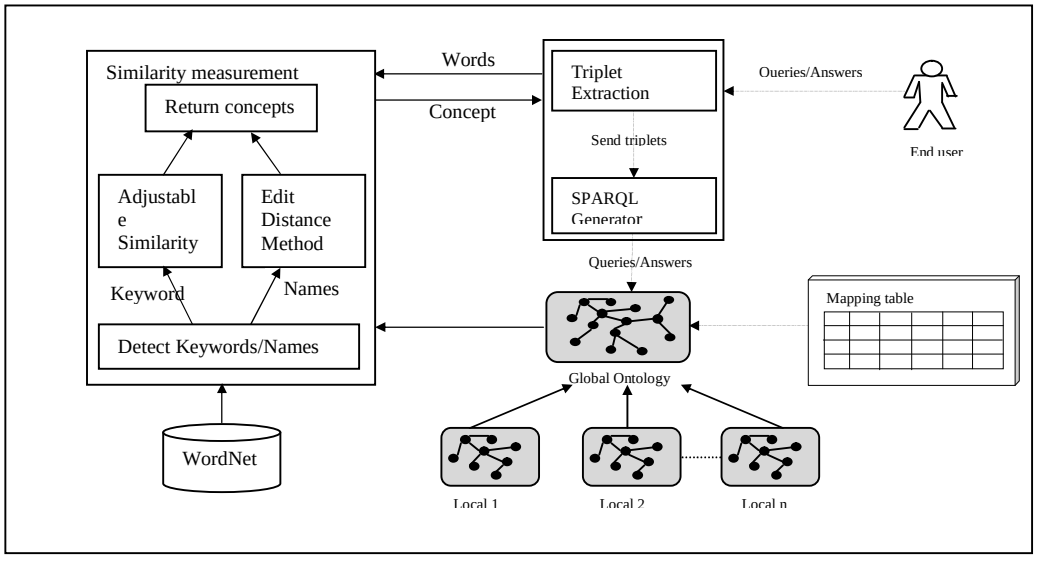

Figure 1. Ontology based data integration system

As architecture, the system follows the framework of Global As a View (GAV) approach [12]. Local ontology is firstly created to represent the relational structure of database at each local source as the semantic model: table names are recognized as the ontology classes, column name in each table are recognized as the data-type properties of each class that are defined for the corresponding table and the between the classes are defined by the object-type properties. Global ontology is built by using the data collected from the existing local ontology [13].

When users make queries and submit them to the system, the global ontology and mapping schema are used to retrieve the information needed from the sources. Mapping rules mean to construct equivalent, homonym and hyponym between the words in the input query and domain ontology concepts [14, 15]. These mapping rules are constructed by referring to the semantic similarity. Moreover, users' input query may not be contained the terms that are wholly equivalent to the concepts in ontology. So, it is needed to match the terms in input query with the concepts in ontology by utilizing the proposed similarity measurement algorithm to suppose the specific concepts in building the ontology understanding query such as SPARQL. This proposed similarity measurement algorithm will be fully explained in the following section. SPARQL query language has a graph-based structure and can be built by combining triple patterns (subjects, predicates and objects). To fulfil the requirements in query generation procedure, triplets extraction process can continually takes the necessary steps [16]. After achieving the triplets from the input sentence, these are used to build the ontology understanding query SPARQL and retrieve the required information from the ontology to reply to the user.

3. SEMANTIC SIMILARITY MEASUREMENT

For accessing the data on ontology, ontology understanding query such as SPARQL is needed. According to the structure of SPARQL, it is needed to extract the specific triplets from the user input query. Here, the terms contained in the user input query are not same exactly with the concepts in ontology. So, similarity measurement algorithm is applied to match the terms extracted from the user input query with the concepts in domain specific ontology. The proposed algorithm is mainly based on the machine readable dictionary WordNet to define the forms of each terms contained in the user input query. According to the senses such as words or names, it uses the different similarity measurement methods to estimate the closest entities. This type definition function in proposed algorithm is shown in the figure 2.

WordNet is the machine readable dictionary and it is widely used for confirming the semantic relationships between overlapping domain concepts of ontology. In this system, WordNet is used to discover the form of words (such as noun, adjective, name, etc.) that are contained as the stream of words in the user input query. And then, the proposed algorithm finds the information of the input words whether it is unknown or it has form. The rules for deciding on each word according to its form are as follows.

- *If the word has the unknown form, assumes these words as name stream and assigns null in its form.*
- *If the word has the form, assigns this form type in its form.*

After defining the form of words, it finds the similarity for these words with the concepts in ontology by applying the *GetSimilarity* similarity measurement method and estimates the similarity of naming streams by applying the *EditDistance* similarity measurement method. This similarity estimation function is shown in figure 3.

```
• Algorithm: WordSimilarity. Estimate the similarity between two words.
    – Input:
    –     Stream of Words
    – Output:
    –     Similarity Measurement

• Function: TypeDefinition (Stream of Words)
    – for (int i = 0; i < words.Length; i++)
    –     wordInfo[i] = Find WordInfo for Words(i) by WordNet
    –       if (wordInfo.partOfSpeech != WordNetlib.PartsOfSpeech.Unknown)
    –         if (wordInfo.text != string.Empty)
    –             words[i] = wordInfo.text;
    –             type[i] = words[i].GetValues(typeof(WordNetlib.PartsOfSpeech));
    –             for (int j = 0; j < type.Length; j++)
    –                 if (wordInfo.senseCounts[j] > 0)
    –                     wordInfos[i] = new MyWordInfo(words[i], type[j])

    –                 else
    –                   Assumes these Words as nameStreams.
    –                   Assign type (wordInfo.word) with null.
    –     wordType = wordInfos U nameStreams
    – return wordType
```

Figure 2. Type definition function in proposed algorithm

As described in above, this proposed algorithm chooses the semantic similarity methods according to the type of words return by the type definition function. So, it can estimate the closet concepts in ontology by adjusting the threshold value of each similarity method respectively. The following section (2.1 and 2.2) will explain the detail of these similarity measurement methods.

```
• Function: SimilarityEstimation(wordType, domainOntology)
    – for each concept of Ontology
    –   If type(wordType.word) is a noun then
    –       wordDistance = wordType.word.GetSimilarity(concept of Ontology), return

    –   else if type(wordType.word) is null then
    –       while (wordType.word is a null)
    –           name[] = wordType.word
    –           wordType.word.next
    –       Build SimilarityMatrix(name[], http://www.owl-ontologies.com/onto.owl#name)
    –       nameDistance = GetEditDistanceMeasurement(SimilarityMatrix), return
```

Figure 3. Similarity estimation function in proposed algorithm

3.1. Edit Distance Similarity Measurement Method

This method is used to estimate how many words are distant between the source and target concepts. It can also estimate the distant degree for the stream of concepts containing space. This method calculates how much distance based on the similarity matrix of two in put strings. Here, it is used to compute the similarity between the words which have no meaning (e.g. the name of the person. We fill each cell of first row in matrix with the word contained in a naming input string and each cell of first column in matrix with the word contained in a naming concept stream.

The remaining cells are filled with the values obtained by applying the rule in equation 3. The following recurrence relations define the edit distance, $d(s_1, s_2)$, of two strings s_1 and s_2 [17].

$$d(\varepsilon, \varepsilon) = 0 \quad // \varepsilon \text{ represents an empty string} \tag{1}$$

$$d(s, \varepsilon) = d(\varepsilon, s) = |s| \quad // |s| \text{ is the length of string } s \tag{2}$$

$$d(s_{1-}+c_1, s_{2-}+c_2) = \min(d(s_{1-}, s_{2-}) + p(c_1, c_2), d(s_{1-}+c_1, s_{2-}) + 1, d(s_{1-}, s_{2-}+c_2) + 1), \tag{3}$$

where c_1 and c_2 are the last characters of s_1 $(= s_{1-}+c_1)$ and s_2 $(= s_{2-}+c_2)$ respectively, and $p(c_1, c_2) = 0$ if $c_1 = c_2$; $p(c_1,c_2) = 1$, otherwise. The threshold value for Edit Distance similarity of two concepts is defined as (*distance* < *name-Length*).

3.2. Get Similarity Measurement Method

This similarity measurement method estimates the similarity for each word that has the form mainly dependent on the adjustable parameter **β**. This addition of adjustable parameter in simple similarity measurement equation can help to overcome the loss of information problem because of the mismatch of one character in input word with the other one character in concept word (e.g. organisation and organization). The similarity between the two concepts is calculated by using the following equation.

$$sim(s_1, s_2) = \sum_{i=1}^{len} \beta_i \, sim_i(s_1, s_2) \tag{4}$$

where $\beta_i (1 \leq i \leq len)$ is an adjustable parameter and len is the number of characters in each word. Moreover, $\beta_i = 1/len$, which reflects the degree contributions to the overall semantic similarity from sim_1 to sim_1. $sim_i(s_1, s_2)$ is respectively semantic similarity of each character contained in a word. The threshold value for semantic similarity of two concepts is defined as 0.8. We set this high threshold to obtain the closet elements from the ontology which are mostly distant one character between the two words.

4. EXPERIMENTAL RESULT

In this section, we show the experimental results of proposed system that estimate the concepts' similarity by applying the proposed semantic similarity algorithm. We build the local ontology with the entities which contained at most twenty classes including main classes and subclasses and twenty individual instances composed by more than 100 data-type properties. And then, we create the global ontology by combining the two sample local ontologies. This data refers to staff profile and history records of an organization. This proposed algorithm is applied to access the data on global ontology. Here, we show the experimental results based on the two degree semantic rates (precision and recall) by generating the concepts from user input query that will be used for retrieving the required information from the domain specific ontology.

Applying the proposed semantic similarity measurement algorithm embedded in the triplets' extraction approach on the sentences listed in below, it takes the process to extract the triplets contained in the input string that are associated with the concept in ontology.

The testing queries are:

Query1: All about John. (*n=3*)
Query2: staff name at the software engineering department. (*n=7*)
Query3: Company's names that are included in advance science and technology department. (*n=11*)
Query4: name, age, NRC, father-name of staff who gets M.C.Sc degree and international paper acceptance. (*n=14*)
Query5: Staff name, degree, position, start-year, department, compensation and bond-year who get English exam marks > 50, Major exam marks >50 and had got Ph.D degree. (*n=27*)

We compare the two results of degree semantic rates (i.e. precision and recall) of the retrieving concepts from the sentences listed in above by applying the proposed algorithm with the retrieving concepts by applying the traditional semantic similarity measurement approach using *cosine* similarity. This comparison is made by submitting the same queries which have the same number of words count for each pair to access the data on the same global ontology. Here, the precision and recall rates can be calculated by using the following equations [18].

$$precision = \frac{extracting\ the\ correct\ number\ of\ information}{extracting\ the\ all\ number\ of\ information} \quad (5)$$

$$recall = \frac{extracting\ the\ correct\ number\ of\ information}{extracting\ all\ correct\ number\ of\ information} \quad (6)$$

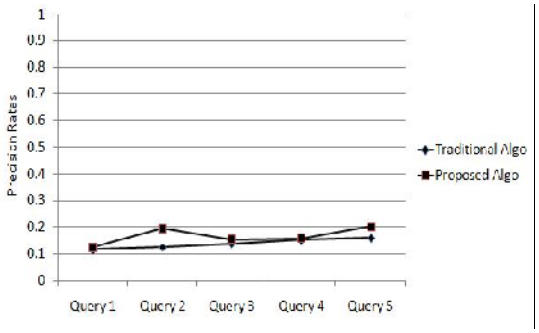

Figure 4. Prediction rates for different queries

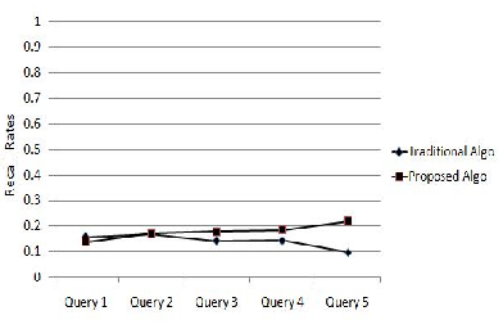

Figure 5. Recall rates for different queries

The illustration of figure 4 and figure 5 represents the obtaining precision and recall rates by testing the five different queries listed in above. By seeing this appraisal, the proposed semantic similarity measurement algorithm can take the preferable results.

The compared results of averaging precision and recall rates based on these five different queries are shown in table 1.

Table 1. Compared results based on two degree semantic

Approaches	Precision	Recall
Traditional Approach	0.695	0.713
Proposed Algorithm	0.836	0.893

By seeing the results in table1, we learn that the proposed semantic measurement algorithm can greatly improve the semantic degree in retrieving the required information from the domain specific ontology. This precision and recall rates are mainly obtained based on the possible outcomes that are associated with the naming concepts because the naming concept results may be variety and here, the semantic degrees of concept words similarity are very high.

5. CONCLUSIONS

The proposed algorithm presented in this paper can give the grate help in retrieving the require information from the domain specific ontology due to the web resources is massive. This paper also described the overview of the whole proposed system and makes the results comparison with other traditional approach based on the results of two degree semantic. The aim of this paper is to fully present the proposed semantic similarity measurement algorithm and the embedded methods that can help the ontology based data integration system to be more complete and the querying strategy to be improve.

ACKNOWLEDGEMENTS

I am very grateful to Dr. Aung Win and Dr. Soe Soe Khine for fruitful discussions during the preparation of this paper and also specially thank to Rector, Professors and colleagues from Technology University (Yadanarpon Cyber City), Myanmar.

REFERENCES

[1] YANNIS KALFOGLOU and MARCO SCHORLEMMER, "Ontology mapping: the state of the art", UK, 2003.
[2] Jihyn Lee, Jeong-Hoon Park, Myung-Jae Park, Chin-Wan Chung, Jun-Ki Min, “An Intelligent query processing for distributed ontology”, The journal of systems and software (2010), 85-95.
[3] Silvana Castano, Alfio Ferrara, Stefano Montanelli, "Ontology-based Interoperability Services for Semantic Collaboration in Open Networked Systems", Italy, 2006.
[4] Leonid Stoimenov, Aleksandar Stanimirovic, Slobodanka Djordjevic-Kajan, "Semantic Interoperability Using Multiple Ontologies", University of Nis, Serbia and Montenegro, 2006.
[5] Zin Thu Thu Myint and Kay Khaing Win, "Triple Patterns Extraction for Accessing Data on Ontology", International Journal of Future Computer and Communication, Vol. 3, No. 1, February, 2014.
[6] Isabel F. Cruz Huiyong Xiao, "The Role of Ontologies in Data Integration", University of Illinois at Chicago, USA, 2005.
[7] Jiaojie Cai, Yufeng Zhang, Feng Hu and Jianfeng Dong, “Domain Ontology Mapping Based Semantic Web Mining Models Research”, International Conference on E-Business Intelligence, China, 2010.
[8] Fabrice Camous, Stephen Blott, Cathal Gurrin, Gareth J.F. Jones and Alan F. Smeaton, “Structural Term Extraction for Expansion of Template-based Genomic Queries”, Dulbin, 2006.
[9] Jianjiang Lu, Yafei Zhang, Zhuang Mial, Po Zhou, “The Semantic Web Principle and Technology”, Science Press: Beijing, 2007.
[10] Dejing Dou, Paea LePendu, Shiwoong Kim and Peishen Qi, "Integrating Databases into the Semantic Web through an Ontology-based Framework", Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06) 0-7695-2571-7/06 © 2006 IEEE.

[11] F. Hakimpour and A. Geppert, “Relsoving Semantic Heterogenity in Schema Integration: an Ontology Based Approach,” in Proc. of Conference on Formal Ontology in Information Systems, Ogunquit, Maine, USA, October 17-19, 2001.
[12] P. Haase and B. Motik, “A Mapping System for the Integration of OWL-DL Ontologies,” IHIS’05, Bremen, Germany, November 4, 2005.
[13] M. Uschold, Creating, Integrating and Maintaining Local and Global Ontologies, 2002.
[14] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hubener, “Ontology-based data integration – A Survey of Existing Approaches,” in Proc. of IJCIA-01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001.
[15] J. Lu, Y. Zhang, Z. Miao, and P. Zhou, The Semantic Web Principle and Technology, Science Press: Beijing, 2007.
[16] Zin Thu Thu Myint, "Triple Patterns Extraction from Unstructured Sentence Using Domain Specific Ontology", proceeding of 11th International Conference on Computer Application, 2013.
[17] Christopher D. Manning Prabhakar Raghavan Hinrich Schütze, "An Introduction to Informational Retrieval", Cambridge UP, 2009, pp. 58-59.
[18] Till Plumbaum, Tino Stelter and Alexander Korth, “Semanticn Web Usage Mining: Using Semantics to Understand User Intentions”, inG-J,LNCS 5535, Houben, Eds,Herdelberg: German, 2009, pp. 391-396.