ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(2), 103–107, January 2015* DOI : 10.17485/ijst/2015/v8i2/57771

Social Information Retrieval Based on Semantic Annotation and Hashing upon the Multiple Ontologies

S.Vigneshwari[1] and M. Aramudhan[2]

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai, India; jayam3@rediffmail.com
[2]Perunthalaivar Kamarajar Institute of Engineering and Technology, Karaikal, Tamilnadu, India

Abstract

Ontology is the best way for representing the useful information. In this paper, we have planned to develop a model which utilizes multiple ontologies. From those ontologies, based on the mutual information among the concepts the taxonomy is constructed, then the relationship among the concepts is calculated. Thereby the useful information is extracted. There is multiple numbers of ontologies available through the web. But there are various issues to be faced while sharing and reusing the existing ontologies. To resolve the ambiguity which exists, when comparing two concepts are semantically similar, but physically different, an approach is proposed here to index and retrieve the documents from two different ontologies. The ontologies used are WordNet and SWETO ontology. The results are compared based on semantic annotation based on RMS and hashing between the cross ontologies using Rabin Karp fingerprinting algorithm. Also the datasets are trained to yield better results.

Keywords: Concept Similarity, Information Extraction, Hashing, Ontological Relationship, Semantic Annotation, Training the Ontology

1. Introduction

The existing systems for web information gathering focuses on the user satisfaction by meeting their requirements. So web page personalization has become a crucial phenomena which can be semantically met using ontologies[3]. Ontology is defined as a formal blueprint of a mutual perception of a particular domain of interest. Ontology should be shared so that it is accepted by a group or community usually ontology merging involves two source ontology to be merged. Manual ontology merging is also tedious, comprehensive and sometimes contains flaws. The main problem in the existing systems is the polysemy and synonym matching. The next problem is polymorphism in identifying semantic similar concepts. Also there is no restriction on the concept count in the ontology.

Hence, a variety of frameworks have been proposed for merging more than one ontology[6]. In this work, we have planned to extend the existing work with effective ontologies and ontology mining algorithm.

2. Review of Related Works

Berendt al.[2] suggested a variety of user friendly web mining techniques. Buitelaar et al.[4] presented methodologies for automatic extraction of text based information. Here the author proposed many methodologies and metrics on ontology learning and evaluation. Such types of metrics are applied in real applications such as bio-informatics, telemedicine, geographic information systems and so on.

*Author for correspondence

Xiaohui Tao et al.[5] proposed a novel information gathering model across the web. Such a model is very much useful for formalizing the ontological user profiles.

Jayasree et al.[8] proposed a cross ontology similarity for medical databases using distance based similarity measures. We propose an enhanced technique by using information.

Jung Ae Kwak[3] proposed various dimensions of similarity like lexical, structural, instance and inference similarity. These similarity based approaches can be organized under property based similarity method.

3. Proposed Method

An effective method is proposed for retrieving the social data from the document repository of SWETO database. A novel cross ontology measure is proposed where two important ontologies like WordNet and SWETO ontology are applied. A successful query refining schema is designed. A comparative study with existing research is done which yielded better precision and recall rates.

3.1 Definition of Ontology

Ontology is defined as a collection of Synsets of concepts which in turn is a collection of Hyperyms/Hyponyms and Holonyms/Metonyms. Usually the relationships between the synsets will be of the type is-a or part-of. Table 1 represents the following relations for the types.

3.1.1 WordNet Ontology

A sample definition of the word "country" on WordNet[11] looks like the following. "people who live in a nation or country". Here the Synset is country. Hypernym is nation. Hyponym is people.

3.1.2 SWETO Ontology

Introduced by LSDIS (Large Scale Distributed Information Systems). There are three versions of SWETO namely SWETO small, SWETO medium and SWETO large[12].

Table 1. Relation table

Type	Relation
Hypernym/Hyponym	Is-a
Holonym/Meronym	Part-of

3.2 Multiple Ontological Similarity Measures

The ontology mapping should also undergo three phases of similarity measures. They are concept similarity, property similarity, and ontological inference. When compared to the mere character comparison, WordNet similarity approach is more predominant. Each concept in the ontology should undergo a variety of taxonomy and constraints. Figure 1 shows a flow diagram for document indexing based on WordNet and SWETO ontology.

3.2.1 Concept Similarity

There are various methods for finding the concept similarity. They are edge counting based methods, information content methods, feature based methods and hybrid methods, which are the combination of all or some of the combination of the above methods.

3.2.2 Similarity Calculation

The similarity calculation is initial step of the proposed method, which is characterized by calculating the

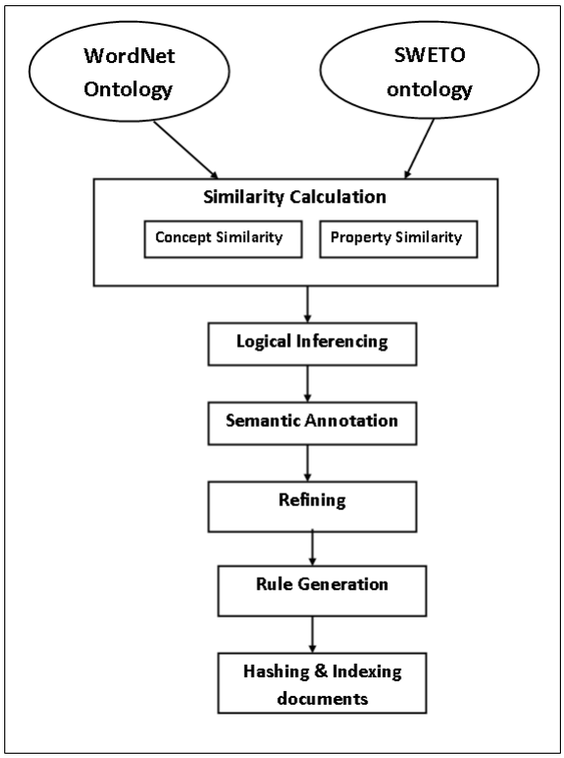

Figure 1. Flow diagram.

similarity between the concepts of different ontologies. In the following equations ci,cj refers to concepts and P() represents the probability function.

$$Mutual\ Similarity = \frac{\log\left(p(c_i)p(c_j)\right)}{p(c_i, c_j)} \quad (1)$$

$$P(C_i) = \frac{W_c}{W} \quad (2)$$

P (c_i,c_j) is the joint probability distribution of common terms incident on the same window, and the P (ci) is the probability of a particular keyword ki appearing in the text window. A text window is a frame of text sequences in a web document. To resolve the ambiguity of the obscure concepts we need a fuzzy membership function which is similar to the mutual similarity[9]. Let μ_i be the fuzzy membership function of the j^{th} concept C and α is a constant and its value is set to 0.5.

$$\mu_i(c_j) \cong Mutual\ Similarity(c_i, c_j) \quad (3)$$

$$= \propto \times P(c_i, c_j)\log_2\left(\frac{P(c_i)p(c_j)}{p(c_i, c_j)}\right) \quad (4)$$

3.2.3 Property Similarity

The important properties to be considered are Object Property (OP) and Data Property (DP).

$$Property\ similarity = \propto \frac{OP_i}{\sum_{i=1}^{m} OP_i} + (1-\propto)\frac{DP_j}{\sum_{j=1}^{n} DP_j} \quad (5)$$

3.2.4 Logical Inference

Logical inference is achieved based on search queries over the web. Let us take for example, "search for the countries except SriLanka". This can be done by the logical Inferencing technique as follows:

$$Q = country(x) \wedge (x, Srilanka) \quad (6)$$

3.2.5 Semantic Annotation and Refining

The ontologies are trained by validating the Root Mean Square (RMS) deviation of the query with that of the document searched.

$$E = \sqrt{\frac{wq_i^2 + wd_i^2}{wq_i + wd_i}}, \forall 1 \le i \le n \quad (7)$$

Here wqi is the query weight and wdi is the document weight. The weight is based on the mutual similarity rank and the relationship rank.

$$Wx = MSM + \frac{Threshold}{Priority\ value} \quad (8)$$

Wx stands for weight of a query or a document. Apriori algorithm[10] is applied for training the documents based on RMS.

3.3 Rabin Karp Document Fingerprinting Algorithm

Let D be the document and ti be the ith term in the document. Let Q be the Query String and pi be the ith pattern in the Query string. The above algorithm explains the indexing methodology for a time efficient search of a document relevant to the search Query string. This algorithm follows hash based indexing of the documents. The algorithm is given below:

(0) Assign n ← size[D]
(1) Assign m ← size[Q]
(2) Assign h ← d^{m-1} mod q
(3) p ← 0
(4) t_o ← 0
(5) for i ← 1 to m
(6) p ← (d_p + P[i]) mod q
(7) t_o ← (d_p + P[i]) mod q
(8) end
(9) Repeat the following for each s varying from 0 to n-m by 1
(10) Check if p == t_s
(11) Check if P[1 . . m] == T [s + 1 . . s + m]
(12) Return the match in the document
(13) Check if s < n – m
(14) t_{s+1} ← (d (t_s – T [s+1] h) + T [s+m+1]) mod q
(15) End Repeat
(16) Retrieve the documents with hash index above the threshold, τ

4. Experimental Evaluation

The dataset comprises of SWETO small dataset (21,134 Kb) and WordNet Browser (Version 2.1). The social cross ontology similarity is implemented in Net Beans IDE 7.4 in Windows 8 Operating system (with RAM configuration of 4GB), as a client server model, where the client sends the search data which is common to both the ontologies and

the server processes the requests by computing semantic similarity measures and suggests the semantic annotations. Based on the semantic annotations, the documents relevant to the word Barcelona are indexed and retrieved. The total number of documents used is 250. The experimental analysis shows that the documents which are semantically annotated are more relevant when compared to executing the same without semantic annotation.

Semantic annotation of rules can be generated using the properties based on the cross ontology similarity measure by mapping WordNet ontology with that of the combined rules of SWETO small and SWETO medium ontologies.

4.1 Analysis Based on SWETO Dataset

In this section, we conduct the performance analysis of the proposed multi ontology based information extraction based on sweto ontology. The sweto ontology is a standard ontology created based on the semantic web technology data. There three version of the SWETO ontology[12], the SWETO Big, SWETO medium and SWETO small. In the analysis process, we use the SWETO small and SWETO medium for performance analysis. Initially, according to the definition of the proposed approach, we conduct a similarity calculation between the two ontologies, namely SWETO small and SWETO medium, based on the mutual information value. The important concepts in both the ontologies are extracted according to their mutual information values.

The Figure 2 represents the combined ontology of the SWETO small and SWETO medium. In the next phase we search for the rules obtained from the proposed approach. The rules are generated with the different relation values assigned over the different concepts. The brief list of the rules generated based on the relationship values defined in the ontologies. Figure 3, show a portion of the generated rules.

According to the proposed approach, we conduct a filtering process to extract the most relevant information based on the user's request. So for the same operation, a threshold value (α) is considered by taking the average of the mutual information values of the concepts. So, by applying the threshold values, a set of most relevant information are extracted from the ontologies.

In the Figure 4, we present the relevant information obtained based on the proposed approach over the ontologies, SWETO small and SWETO medium. The average execution time used by the proposed approach for extracting the relevant information is 1153 milliseconds and the memory utilized is given as 1.18 Megabytes

```
<City rdf:ID="Dili">
<Is_city rdf:resource="#East_Timor"/>
<City rdf:ID="Baucau">
<Is_city rdf:resource="#East_Timor"/>
<City rdf:ID="Bangalore">
<Is_city rdf:resource="#India"/>
<University rdf:ID="Anna_University">
<Is_university_of rdf:resource="#India"/>
<Airport rdf:ID="Chennai_International">
<Is_airport rdf:resource="#India"/>
<Company rdf:ID="Dangote_Cement">
<Is_financial_organization_of rdf:resource="#company"/>
<Airport rdf:ID="Calabar">
<Is_financial_organization_of rdf:resource="#Airports"/>
<University rdf:ID="Abia_State_University">
<Is_university_of rdf:resource="#Nigeria"/>
<University rdf:ID="Ahmadu_Bello_University">
<Is_university_of rdf:resource="#Nigeria"/>
<City rdf:ID="Abuja">
<Is_city rdf:resource="#Nigeria"/>
<City rdf:ID="Enuger">
<Is_city rdf:resource="#Nigeria"/>
<Company rdf:ID="Abarth">
<Is_financial_organization_of rdf:resource="#company"/>
<University rdf:ID="Politecnico_Di_Bari">
<Is_university_of rdf:resource="#Italy"/>
<University rdf:ID="Politecnico_Di_Milana">
<Is_university_of rdf:resource="#Italy"/>
<Airport rdf:ID="Cagliari_Elmas">
<Is_airport rdf:resource="#Italy"/>
<City rdf:ID="Rome">
<Is_city rdf:resource="#Italy"/>
<City rdf:ID="Venice">
<Is_city rdf:resource="#Italy"/>
<Company rdf:ID="Clear_Blue_50000">
```

Figure 2. SWETO ontology.

```
[Pellucidar_Edgar_Rice, Is_financial_organization_of, Books]
[Books, Has_financial_organization_of, Pellucidar_Edgar_Rice]
[The_Black_Star_Passes, Is_financial_organization_of, Books]
[Books, Has_financial_organization_of, The_Black_Star_Passes]
[Triplanetary_E._E._Doc, Is_financial_organization_of, Books]
[Books, Has_financial_organization_of, Triplanetary_E._E._Doc]
[The_Day_of_the_Boomer, Is_financial_organization_of, Books]
[Books, Has_financial_organization_of, The_Day_of_the_Boomer]
[Frankenstein_Mary, Is_financial_organization_of, Books]
[Books, Has_financial_organization_of, Frankenstein_Mary]
[Brigands_of_the_Moon, Is_financial_organization_of, Books]
[Books, Has_financial_organization_of, Brigands_of_the_Moon]
```

Figure 3. A portion of Rules list.

```
[Journels, Has_financial_organization_of, IJAMM]
[Journels, Has_financial_organization_of, IJPA]
[Journels, Has_financial_organization_of, IJSE]
[Journels, Has_financial_organization_of, IJAP]
[Journels, Has_financial_organization_of, JCIB]
[Journels, Has_financial_organization_of, IJWCS]
[Journels, Has_financial_organization_of, IJCIR]
[Journels, Has_financial_organization_of, MMAC]
```

Figure 4. Relevant information.

Figure 5 shows Query Vs Time. Nearly 50 queries were given and the maximum retrieval time is 1500 milliseconds. Figure 6 shows the query versus the number

of relevant documents indexed. The chart shows that more relevant documents were retrieved above the linear separator. Figure 7 shows the comparison chart of hybrid approach with other approaches.

Here in Figure 4 the x-axis represents the number of queries and y-axis indicates the performance measure. The figure shows that the F-Measure of the retrieved documents with both semantic annotation and hashing yields better results when compared to the other two approaches where either semantic annotation or hashing is missing. The formula for calculating the performance measure is given below. Here precision refers to the maximum relevant documents and recall refers to the maximum retrieved documents.

$$F-Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (9)$$

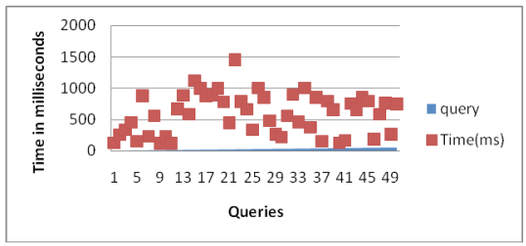

Figure 5. Query Vs Time.

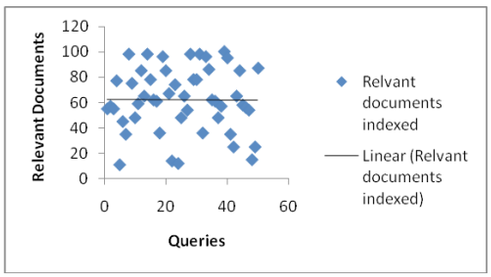

Figure 6. Query versus relevant documents indexed.

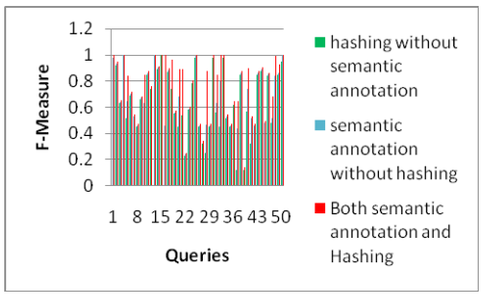

Figure 7. Comparison of various approaches.

5. Conclusion

The data sets are generated with the help of both SWETO and WordNet. The cross ontology mapping was performed with hashing alone and without semantic annotation, with semantic annotation alone without hashing and with both semantic annotation and with hashing. The test was conducted utilizing 50 queries and the hybrid approach which involves both semantic annotation and with hashing yielded better performance when compared to the other two approaches.

6. References

1. Stumme G, Hotho A, Berendt B. Semantic web mining a state of the art and future directions. Journal of Web Semantics. 2006; 124–43.
2. Berendt B, Hotho A, Mladenic D. A Roadmap for Web Mining From Web to Semantic Web. Web Mining: From Web to Semantic Web. Springer; 2004.
3. Kwak J-A, Yong H-S. An Approach to Ontology-Based Semantic Integration for PLM Object. IEEE International Workshop on Semantic Computing and Applications, 2008, IWSCA '08; IEEE; 2008. p. 91–26.
4. Buitelaar P, Cimiano P, Magnini B. Ontology Learning from Text: An Overview. University of Karlsruhe; 2003. p. 1–10.
5. Tao X, Li Y, Zhong N. A Personalized Ontology Model for Web Information Gathering. IEEE Trans Knowl Data Eng. 2011; 23(4):496–511.
6. Stumme G, Maedche A. Ontology Merging for Federated Ontologies on the Semantic Web. University of Karlsruhe; 2005. p. 1–9.
7. Wu C-A, Lin W-Y, Wu C-C. An Active Multidimensional Association Mining Framework with User Preference Ontology. Int J Fuzzy Syst. 2010; 12(2):125–35.
8. Jayasri D, Manimegalai D. An efficient cross ontology based similarity measure for bio-document retrieval system. J Theor Appl Inform Tech. 2013; 54(2):245–54.
9. Lau RYK, Song D, Li Y, Cheung TCH, Hao J-X. Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning. IEEE Trans Knowl Data Eng. 2009 Jun; 21(6):800–13.
10. de Campos LM, Fernandez-Luna JM, Huete JF. Query Expansion in Information Retrieval Systems using a Bayesian Network-Based Thesaurus. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI1998); 1998. p. 53–60.
11. Wordnet. 2014 [cited on 2014 March]. Availaible from: https://wordnet.princeton.edu/wordnet/
12. SWETO Dataset. 2014 [cited on 2014 March]. http://archive.knoesis.org/library/ontologies/sweto