A Hybrid Concept Similarity Measure Model for Ontology Environment

Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang

Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology,
GPO Box U1987 Perth, Western Australia 6845, Australia
{hai.dong,farookh.hussain,elizabeth.chang}@cbs.curtin.edu.au

Abstract. In this paper, we present a hybrid concept similarity measure model for the ontology environment. Whilst to date many similar technologies have been developed for semantic networks, few of them can be directly applied to the semantic-rich ontology environment. Before the measure model is adopted, an ontology is required to be converted into a lightweight ontology space, and within it all the ontology concepts need to be transformed into the pseudo-concepts. By means of this model, ontology concept similarities are measured respectively based on the content of pseudo-concepts and the structure of the lightweight ontology space. Afterwards, the two aspects of concept similarity are leveraged as the eventual product. In addition, an experiment is conducted to evaluate the measure model based on a small ontology. Conclusions are drawn and future works are planned in the final section.

Keywords: concept similarity measure, latent semantic indexing, ontology, semantic similarity models.

1 Introduction

Semantic relatedness refers to human judgment about the extent to which a given pair of concepts are related to each other [1]. Studies have shown that most people agree on the relative semantic relatedness of most of pairs of concepts [2, 3]. Therefore, approaches are required in order to measure the extent of semantic relatedness and similarity between concepts. In fact, many such techniques have been developed in the field of information retrieval (IR) [4-6], Natural Language Processing (NLP) [7-10], medicine [11], and bioinformatics [1, 12, 13] etc. In the field of computer science, ontology is defined by Gruber [14] as "*an explicit specification of conceptualization*", which comprises objects (concepts) and relations among objects (concepts). It seems that the existing semantic similarity models can be directly applied to the ontology environment, in order to compute the ontology concept similarities. However, there are two issues observed in the existing semantic similarity models when applying them to the ontology environment, which are described as follows:

- Most of the models are designed for semantic networks. A semantic network is defined as "*a graphic notation for representing knowledge in patterns of*

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 848–857, 2009.

interconnected nodes and arcs" [15]. WordNet is a typical example of a semantic network, in which words or phrases are represented as nodes and are linked by multiple relations. Current semantic similarity models focus on estimating similarities between nodes. Since these nodes normally consist of single words or simple phrases, these models ignore the content of the nodes and deduce similarities based on the relative distance between the nodes or the position of the nodes in the whole semantic network. Nevertheless, in the ontology environment, each ontology concept is defined with semantic-rich contents. For example, in the Web Ontology Language (OWL), each class is defined by its annotation properties, data type properties, object properties, and so on. Hence, it is obvious that the content of ontology concept cannot be ignored when computing concept similarities within the ontology environment.

- Most of the models are designed for definitional networks. A definitional network is a subset of semantic networks, in which the only type of relation is *class/subclass*, or called *is-a* [15]. Therefore, it is easy to visualize that each definitional network is a hierarchy of nodes linked by the *is-a* relations. In contrast, an ontology is more complicated than a definitional network. Although most ontologies also follow a hierarchical structure, the types of relations among concepts are more flexible and customizable. Obviously, the existing semantic similarity models may meet challenges when dealing with the multi-relational ontologies.

In order to address the two issues above, in this paper, we propose a hybrid concept similarity measure model, by considering both the aspect of the concept content-based similarity measure and the aspect of the ontology structure-based similarity measure, with the purpose of measuring concept similarities within the ontology environment.

The remainder of the paper is structured as follows: in Section 2 we present a framework of a lightweight ontology space for dealing with the issue of multi-relations within the ontology environment; in Section 3 we present a hybrid concept similarity measure model based on the framework; in Section 4 we implement an evaluation to the model in terms of a case study; the conclusion and future works are presented in the final section.

2 Converting an Ontology to a Lightweight Ontology Space

Before we present our hybrid concept similarity measure model, in order to solve the problem of multiple relations within the ontology environment, we are required to convert an ontology to a lightweight ontology space in which our proposed model can be applied. Here we define the lightweight ontology space, which comprises two basic definitions as follows:

Definition 1. Pseudo-concept ς

It is well known that in semantic web documenets (SWD), there are various relations available among ontology concepts. Therefore, it is necessary and challenging to take into account these relations when computing the similarity among concepts. To overcome this challenge, we define a pseudo-concept ς for a ontology concept c as a

combination of (c, α, δ, o, γ_o), where in the RDFS/OWL-annotated semantic web documents (SWDs), c is the name (or Uniform Resource Identifier (URI)) of the concept c, α is the annotation property(s) of the concept c, δ is the datatype property(s) of the concept c, o is the object property(s) of the concept c, and γ_o is the name(s) of the object concept(s) to which o relates.

Definition 2. Lightweight Ontology Space

Based on the pseudo-concepts, we define a lightweight ontology space as a space of pseudo-concepts, in which the pseudo-concepts are linked only by *is-a* relations. An *is-a* relation is a generalization/specification relationship between an upper generic pseudo-concept and a lower specific pseudo-concept. The lower concept inherits all the properties of the upper concept.

In fact, the *is-a* relation often appears in SWDs, e.g., it can be represented as </rdfs:subClassOf> in RDFS and OWL. As a result, when transforming an ontology to a lightweight ontology space, we need to reserve the *is-a* relations and encapsulate all other properties into the pseudo-concepts.

Subsequently, we use an example in order to illustrate the process that converts an ontology to a lightweight ontology space.

Fig. 1 presents an example of an ontology – a camera ontology, which originates from the cognominal ontology given by the Protégé-OWL tutorial [16]. We can see that there are seven concepts involved in this ontology linked by two different types of relations – *is-a* and *has*. To convert this camera ontology to a lightweight ontology space, we must encapsulate the *has* relations and reserve the *is-a* relations.

According to the two definitions above, the lightweight ontology can be found in Fig. 2, in which each pseudo-concept is represented below:

ς_1= {Purchasable_Item}
ς_2= {Camera, has, Storage_Device}
ς_3= {Storage_Device}
ς_4= {Digital_Camera, has, Memory_Card}

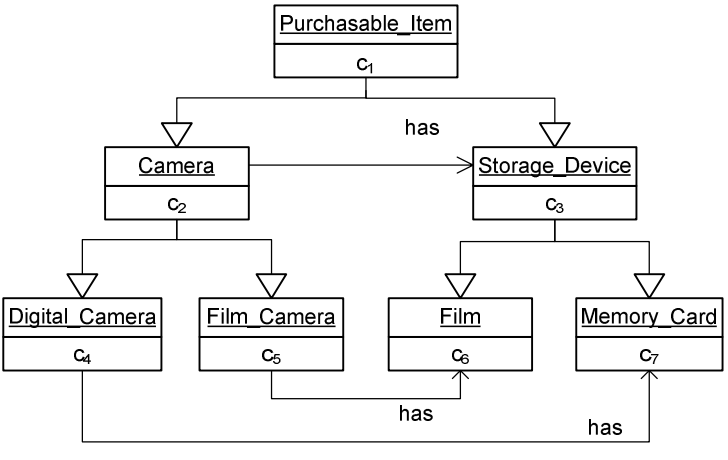

Fig. 1. An example of ontology– a camera ontology

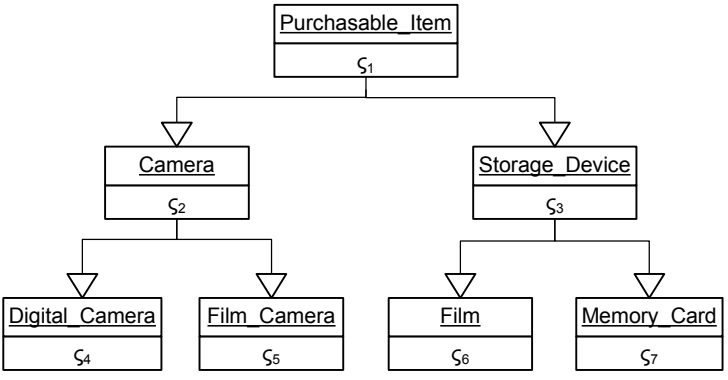

Fig. 2. The lightweight ontology for the camera ontology

ς_5= {Film_Camera, has, Film}
ς_6= {Film}
ς_7= {Memory_Card}

We need to construct a pseudo-concept for an ontology concept because each pseudo-concept comprises almost all the properties that can be used to comprehensively define a concept and thus we can recognize a pseudo-concept as the textual description of a concept, which makes it possible to measure the similarity of two concepts by contrasting the contents of their pseudo-concepts.

3 A Hybrid Concept Similarity Measure Model

It is recognized that many available IR approaches can be utilized to compute concept similarities based on the content of pseudo-concepts. However, we observe an issue here – whereas the similarity of two concepts can be measured by contrasting the content of their pseudo-concepts, this approach is not sufficient to reveal the extent of their similarity. The reason is that an ontology can be represented as a graph in which each concept is a node and relations are arcs among the nodes, and obviously the similarity of two nodes also relates to the structure of the graph and the relative distance between the two nodes [4, 17]. Jiang et al. [17]'s model inspires us to integrate the factor of the pseudo-concept content and the factor of the lightweight ontology structure to compute the extent of similarity between two concepts.

In this section, we present a hybrid concept similarity measure model integrating the two factors above. Our proposed hybrid model involves two sub-models. The first sub-model measures the concept similarities based on the content of pseudo-concepts, by means of the Latent Semantic Indexing (LSI) approach [18]. The second sub-model measures the concept similarities based on the structure of lightweight ontology graph, by means of an approach originating from the enhanced topic-based

vector model (eTVSM) [17]. The product of the two sub-models is two concept-concept matrixes. Then we integrate the two matrices to obtain a new concept-concept matrix that indicates the extent of similarity between concepts. To illustrate the working mechanism of the hybrid model, we will compute the concept similarity values for the camera ontology displayed in Fig.1.

3.1 Pseudo-Concept Content-Based Concept Similarity Measure Model

As described earlier, a pseudo-concept can be regarded as a textual description of a concept. In this section, we propose to make use of the LSI model to compute the extent of similarity between each pair of concepts of an ontology based on the extent of their pseudo-concepts.

The main reason for applying the LSI model is to construct a concept-concept matrix for an ontology in which each element is the similarity value between the two corresponding concepts. This matrix is the product of a normalized concept-index term matrix and its transposed matrix. The normalized concept-index term matrix is obtained by the tf-idf model [19] and by normalizing each row to 1. Finally, all pair-wise concept similarity values from the ontology example presented in Section 2 are given in Table 1.

Table 1. Pseudo-concept content based concept similarity values for the camera ontology

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
c_1	1	0	0	0	0	0	0
c_2	0	1	0.5	0.11	0.11	0	0
c_3	0	0.5	1	0	0	0	0
c_4	0	0.11	0	1	0.11	0	0.5
c_5	0	0.11	0	0.11	1	0.5	0
c_6	0	0	0	0	0.5	1	0
c_7	0	0	0	0.5	0	0	1

3.2 Lightweight Ontology Structure-Based Concept Similarity Measure Model

As mentioned earlier, the structure-based approach originates from the topic similarity measure model for the topic map environment. In our model, we employ this method in the environment of the lightweight ontology space. As there is only one type of relation in the lightweight ontology space, the weights of relations can be viewed as equal and the issue of weights of relations can be ignored in the measurement process. The process of computing the extent of the concept similarity can be divided into two processes: 1) determining the pseudo-concept vectors based on a lightweight ontology structure; 2) obtaining a concept similarity matrix by means of the scalar product of the pseudo-concept vectors. The operational vector space dimensionality is specified by the number of pseudo-concepts in a lightweight ontology space. The heuristics behind this process can be found from [17]. Finally, the pair-wise concept similarity values from the camera ontology in Fig. 1 are given in Table 2.

Table 2. Lightweight ontology structure based concept similarity values for the camera ontology

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
c_1	1	0.83	0.83	0.76	0.76	0.76	0.76
c_2	0.83	1	0.4	0.91	0.91	0.36	0.36
c_3	0.83	0.4	1	0.36	0.36	0.91	0.91
c_4	0.76	0.91	0.36	1	0.66	0.33	0.33
c_5	0.76	0.91	0.36	0.66	1	0.33	0.33
c_6	0.76	0.36	0.91	0.33	0.33	1	0.66
c_7	0.76	0.36	0.91	0.33	0.33	0.66	1

3.3 Integrating the Products of the Two Models

Section 3.1 and Section 3.2 present two concept similarity matrixes (*C* and *L*) based on the pseudo-concept content and the lightweight ontology structure respectively. In this section, we leverage both of these matrices in order to yield a new matrix that is able to indicate the similarities among concepts in a more precise manner. We define a matrix *S* in which each element is the weighted arithmetic mean between counterparts in matrices *C* and *L* as shown in Equation (1). The similarity value between two concepts is also the weighted arithmetic mean between the content-based concept similarity value ($sim_c(c_i, c_j)$) and the structure-based concept similarity value ($sim_s(c_i, c_j)$).

$$S_{i,j} = sim(c_i, c_j) = (1-\beta) sim_c(c_i, c_j) + \beta sim_s(c_i, c_j) = (1-\beta) C_{i,j} + \beta L_{i,j} \quad (0 \le \beta \le 1) \tag{1}$$

When β=0.5, $sim(c_i, c_j)$ equals to the arithmetic mean between $sim_c(c_i, c_j)$ and $sim_s(c_i, c_j)$. Returning to the ontology example in Fig. 1, the pair-wise concept similarity values when β=0.5 are shown in Table 3.

Table 3. Combined concept similarity values for the camera ontology (β=0.5)

	c_1	c_2	c_3	c_4	c_4	c_6	c_7
c_1	1	0.42	0.42	0.38	0.38	0.38	0.38
c_2	0.42	1	0.45	0.51	0.51	0.18	0.18
c_3	0.42	0.45	1	0.18	0.18	0.46	0.46
c_4	0.38	0.51	0.18	1	0.39	0.17	0.42
c_4	0.38	0.51	0.18	0.39	1	0.42	0.17
c_6	0.38	0.18	0.46	0.17	0.42	1	0.33
c_7	0.38	0.18	0.46	0.42	0.17	0.33	1

4 Evaluation

In order to evaluate the hybrid concept similarity measure model that we proposed in this paper, based on our ontology example in Fig. 1, we compare the results from the hybrid model with the results from the content (LSI)-based model, in order to decide which group of results is closer to our perception of the objective world.

First of all, let us analyze the results from the content-based model in Table 1. We interpret the results by providing a descending sequence of similar concepts for each

concept in the camera ontology, which are shown in Table 4. By reviewing the interpretation from Table 4, we find the following problems from the content-based concept similarity matrix:

(1) The extent of the similarity between c_1 and c_2 to c_6 should be able to be measured.
(2) Film_Camera and Digital_Camera should be put prior to Storage_Device as the more similar concepts of Camera.
(3) Storage_Device should be similar to Film and Memory_Card other than Camera.
(4) Camera and Film_Camera should be closer to Digital_Camera other than Memory_Card.
(5) The most similar concept of Film should be Storage_Device and Memory_Card other than Film_Camera.
(6) Camera and Digital_Camera should be closer to Film_Camera other than Film.
(7) The most similar concept of Memory_Card should be Storage_Device and Film other than Digital_Camera.

Table 4. Interpretation of the content-based concept similarity values for the camera ontology

Concept No.	Concept	Descending sequence of similar concepts
c_1	Purchasable_Item	
c_2	Camera	Storage_Device>Film_Camera=Digital_Camera
c_3	Storage_Device	Camera
c_4	Digital_Camera	Memory_Card>Camera=Film_Camera
c_5	Film_Camera	Film>Camera=Digital_Camera
c_6	Film	Film_Camera
c_7	Memory_Card	Digital_Camera

Second, we start to analyze the matrix obtained by the hybrid model from Table 3. The interpretation of the model for each concept of the camera ontology can be found from Table 5.

Table 5. Interpretation of the hybrid concept similarity values for the camera ontology

Concept No.	Concept	Descending sequence of similar concepts
c_1	Purchasable_Item	Camera=Storage_Device>Digital_Camera=Digital_Camera= Film=Memory_Card
c_2	Camera	Film_Camera=Digital_Camera>Storage_Device>Purchasable_Item>Film=Memory_Card
c_3	Storage_Device	Film=Memory_Card>Camera>Purchasable_Item>Film_Camera=Digital_Camera
c_4	Digital_Camera	Camera>Memory_Card>Film_Camera>Purchasable_Item>Storage_Device>Film
c_5	Film_Camera	Camera>Film>Digital_Camera>Purchasable_Item>Storage_Device>Memory_Card
c_6	Film	Storage_Device>Film_Camera>Purchasable_Item>Memory_Card>Camera>Digital_Camera
c_7	Memory_Card	Storage_Device>Digital_Camera>Purchasable_Item>Film>Camera>Film_Camera

The defect of this table is that it ranks and returns all available concepts for each concept in the ontology. However, in the real life, users usually only need the most similar concepts for each concept and the less similar concepts may be ignored. Therefore, there is a need for filtering some concepts with weak similarities and we need to set up a filter bar to remove some less similar concepts. There are two usual ways of setting up a filter bar, which are:

- Choosing a threshold value, in other words, the concepts with lower similarities than the threshold value are filtered from the ranked list for each concept.
- Choosing a filter concept, in other words, the concepts ranked behind the filter concept are abandoned from the ranked list for each concept.

In this experiment, we use the second method to filter the less similar concepts. Since the root concept is used to define the domain of the ontology, and all the other concepts are theoretically relevant to this concept, it is unnecessary to include this concept when obtaining concept similarities. Hence, we choose Purchasable_Item as the filter concept, and the filtered results are shown in Table 6.

Table 6. Interpretation of the filtered hybrid concept similarity values for the camera ontology

Concept No.	Concept	Descending sequence of similar concepts
c_2	Camera	Film_Camera=Digital_Camera>Storage_Device
c_3	Storage_Device	Film=Memory_Card>Camera
c_4	Digital_Camera	Camera>Memory_Card>Film_Camera
c_5	Film_Camera	Camera>Film>Digital_Camera
c_6	Film	Storage_Device>Film_Camera
c_7	Memory_Card	Storage_Device>Digital_Camera

Compared with Table 4, we find that these results correct almost all the problems found from Table 4, which can basically agree with our perception of the extent of similarity between concepts within the camera ontology. Therefore, by means of this experiment, we preliminarily prove that the hybrid concept similarity model is more precise than the content-based model, which provides a better set of answers to human's perceptions of the relations between objects in the world.

5 Conclusion and Future Work

In this paper, by means of analyzing the existing semantic similarity models, we observe two issues when applying them to the ontology environment. The first is that these models ignore the content of ontology concepts; the second is that these models can be used only for the semantic networks in which nodes are only linked with *is-a* relations. In order to solve the two issues, we design a concept similarity measure model for the ontology environment. This model considers both the content and structure of the ontology. In order to realize this model, first of all, we need to convert an ontology to a lightweight ontology space, which preserves the *is-a* relations and combines other relations as the content of the ontology concepts. For the content-oriented concept similarity measure, we employ the LSI model and generate a

concept-concept similarity matrix. For the structure-oriented concept similarity measure, we employ the topic similarity measure approach, in order to create another matrix. Eventually, we combine the two matrices and produce a new concept-concept similarity matrix. In order to evaluate the performance of this model, in terms of a camera ontology, we compare the product of the model with the product of the LSI model. The comparison indicates that the former has better performance than the latter.

Additionally, we intend to implement our concept similarity model in the large-scale knowledge base. Moreover, we intend to compare the performance of our method against contemporary approaches in the literature. At the time of writing this paper, we have conducted the above research comparison work and have documented in [20].

References

1. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40, 288–299 (2006)
2. Miller, G., Charles., W.: Contextual correlates of semantic similarity. Language and Cognitive Processes 6, 1–28 (1991)
3. Rubenstein, H., Goodenough, J.B.: Contextual Correlates of Synonymy. Communications of the ACM 8, 627–633 (1965)
4. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics 19, 17–30 (1989)
5. Srihari, R.K., Zhang, Z.F., Rao, A.B.: Intelligent indexing and semantic retrieval of multimodal documents. Information Retrieval 2, 245–275 (2000)
6. Sussna, M.: Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. In: The Second International Conference on Information and Knowledge Management (CIKM 1993), pp. 67–74. ACM, Washington (1993)
7. Li, Y., Bandar, Z.A., McLean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering 15, 871–882 (2003)
8. Lin, D.: Automatic retrieval and clustering of similar words. In: the 17th COLING, pp. 768–774. ACM, Austin (1998)
9. Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
10. Rosenfield, R.: A maximum entropy approach to adaptive statistical modelling. Computer Speech and Language 10, 187–228 (1996)
11. Steichen, O., Bozec, C.D., Thieu, M., Zapletal, E., Jaulent, M.C.: Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus. Computers in Biology and Medicine 36, 768–788 (2006)
12. Othman, R.M., Deris, S., Illias, R.M.: A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. Journal of Biomedical Informatics 41, 65–81 (2008)
13. Sevilla, J.L., Segura, V.c., Podhorski, A., Guruceaga, E., Mato, J.M., Martínez-Cruz, L.A., Corrales, F.J., Rubio, A.: Correlation between Gene Expression and GO Semantic Similarity. IEEE/ACM Transaction on Computational Biology and Bioinformatics 2, 330–338 (2005)

14. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition 5, 199–220 (1995)
15. Sowa, J.F.: Semantic Networks. In: Shapiro, S.C. (ed.) Encyclopedia of Artificial Intelligence. Wiley, Chichester (1992)
16. Costello, R.L.: OWL ontologies (2009)
17. Kuropka, D.: Modelle zur repräsentation natürlichsprachlicher dokumente. ontologie-basiertes information-filtering und -retrieval mit relationalen datenbanken. In: Becker, J., Grob, H.L., Klein, S., Kuchen, H., Müller-Funk, U., Vossen, G. (eds.) Advances in Information Systems and Management Science. Logos Verlag Berlin, Berlin (2004)
18. Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., Lochbaum, K.E.: Information retrieval using a singular decomposition model of latent semantic structure. In: 11th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 465–480 (1988)
19. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, Harlow (1999)
20. Dong, H., Hussain, F.K., Chang, E.: A concept similarity measure model for enhancing the dependability of semantic service matchmaking in the service ecosystem. IEEE Transactions on Service Computing (submitted, 2009)