Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 394 – 398 Sept. 2006 ISSN 1003—7985

Ontological similarity network reasoning framework

Wen Guihua Jiang Lijun

(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract: To properly compute the ontological similarity, an ontological similarity network-based reasoning framework is proposed. It structurally integrates extension-based approach, intension-based approach, the similarity network-based reasoning to exploit the implicit similarity, and the feedback from the context to validate the similarity measures. A new similarity measure is also presented to construct concept similarity network, which scales the similarity using the relative depth of the least common super-concept between any two concepts. Subsequently, the graph theory, instead of predefined knowledge rules, is applied to perform the similarity network-based reasoning such that the knowledge acquisition can be avoided. The framework has been applied to text categorization and visualization of high dimensional data. Theory analysis and the experimental results validate the proposed framework.

Key words: ontology; similarity network-based reasoning; graph algebra; integration framework

As ontology development becomes increasingly widespread and collaborative, ontology management inevitably meets many challenges. In these challenges, a fundamental problem is measuring the similarity between concepts and ontologies. For example, ontology learning largely depends on the clustering algorithm that groups instances of a concept together based on the similarities that are measured between instances[1]. Ontology mapping needs similarity measure to determine which concept is most similar to each other, so that semantic relationships between them can be defined[2]. Therefore the similarity has been extensively studied. With shared ontology, the similarity between concepts is generally calculated by network distance-based methods and information theory-based methods[3]. If such an ontology is not available, the similarity between concepts can be calculated using a matching process over synonym sets, semantic neighborhoods, and distinguishing features[1]. Simultaneously, a similarity graph is also applied to represent the semantic neighborhood of a concept such that the similarity can be calculated using graph computations instead of formal conceptual reasoning[4]. The similarity between concepts can also be calculated based on the extension of concepts using the probability theory[5]. Based on the conceptual similarity, the ontological similarity can be calculated from the lexical and the conceptual levels[6], particularly from the structured cosine similarity by exploiting the hierarchical domain structure between ontologies[7]. The structured cosine similarity has been applied to text processing domains such as ontology-based document summarization[8].

To the best of our knowledge, many works have studied the similarity from intension and extension independently[9], while the similarity between ontologies has not been considered in the framework. Particularly the similarity network-based reasoning has not been applied to exploit the implicit similarity. This paper proposes an integrated framework that structurally integrates extension and intension-based approaches, the similarity network-based reasoning, and the feedback from the context.

1 Structured Framework for Ontological Similarity

Any two concepts or ontologies can be regarded as similar from different perspectives such as from their intensions and extensions. We present an integrated framework with three layers to calculate similarity: concept layer, ontology layer, and application layer, as shown in Fig. 1.

The concept layer aims to calculate the similarity between two concepts based on conceptual intensions (properties), extensions (instances) and specific domain context. The calculated similarity between concepts can be applied to construct the concept network, on which the final conceptual similarity can be calculated by network-based reasoning. The inferred conceptual similarity can be applied to ontology mapping or

Received 2006-04-25.

Foundation item: The National Natural Science Foundation of China (No. 60003019).

Biography: Wen Guihua (1968—), doctor, associate professor, crghwen@ scut. edu. cn.

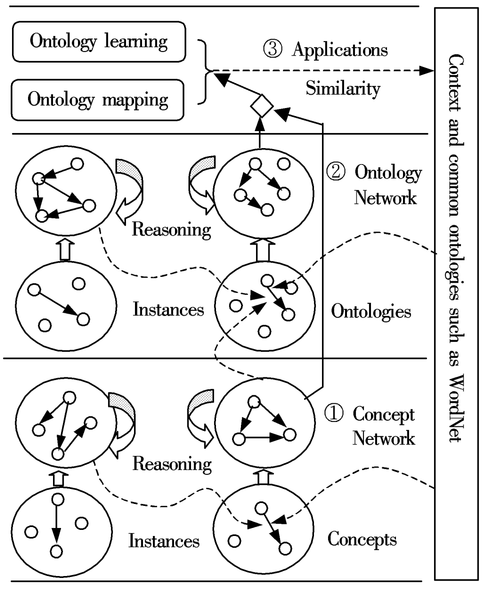


Fig. 1 Ontological similarity network framework

to calculate the similarity between ontologies.

The ontology layer is structurally similar to the first layer. But when computing the direct similarity between ontologies, except involving the results from instance similarity network of ontology and specific domain context, it largely depends on the reasoning results of concept network. Based on the same principle, we also construct the ontology network and apply it to the network-based reasoning. The inferred ontological similarity can be applied to ontology engineering such as ontology learning.

The application layer aims to provide feedback to the first two layers. From the viewpoint of determining the similarity, the application context has the critical influence because it is the decisive standard used to make the judgment whether a measure is reasonable or not. It, therefore, has two functionalities: define the explicit similarity measure for specific domain such as gene engineering and to validate the defined measures. This can be performed using the similarity principle from case-based reasoning. That is, if two concepts are used in the same context, then these concepts are similar.

2 Similarity Network-Based Reasoning

2.1 Concept similarity network

Since the similarity between concepts is influenced by their intensions and extensions, it can be integrated by $\mathrm{sim}(x, y) = f(\mathrm{sim}_{\mathrm{extension}}(x, y), \mathrm{sim}_{\mathrm{intension}}(x, y))$, where f can be linear combination function or geometric average. The related parameters can be assigned manually or learned through optimization to a training set.

From conceptual extensions, the similarity between two concepts can be defined by the joint probability distribution:

$$\mathrm{sim}_{\mathrm{extension}}(x, y) = \frac{p(x \cap y)}{p(x \cup y)}$$

where $p(z)$ is the probability that an instance belongs to concept z, which can be calculated by machine learning techniques[5]. From conceptual intensions, similarity between concepts can be studied from the names and attributes of concepts. If the shared ontology is available, the similarity between two concepts can be computed as Eq. (1) using path distance between concepts in the hierarchical structure underlying the shared ontology[9]:

$$\mathrm{sim}_{\mathrm{intension}}(x, y) = \frac{2h_{xy}}{h_x + h_y} \tag{1}$$

where h_x is the height of the concept node x in the hierarchy; h_y is the height of the concept node y in the hierarchy; and h_{xy} is the height of the concept node of greatest depth that is an ancestor of both x and y in the hierarchy. This measure does not consider the contribution of path length to the similarity. A better similarity measure is posited to consider simultaneously the shortest path length l as well as the depth of the subsumer[3]:

$$\mathrm{sim}_{\mathrm{intension}}(x, y) = \mathrm{e}^{-\alpha l}\frac{\mathrm{e}^{\beta h_{xy}} - \mathrm{e}^{-\beta h_{xy}}}{\mathrm{e}^{\beta h_{xy}} + \mathrm{e}^{-\beta h_{xy}}} \tag{2}$$

Because this measure does not scale the similarity using the relative depth of the least common super-concept between these two concepts, we combine the above two measures to define a new measure:

$$\mathrm{sim}_{\mathrm{intension}}(x, y) = \mathrm{e}^{-\alpha l}\frac{\mathrm{e}^{\beta h} - \mathrm{e}^{-\beta h}}{\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}} \tag{3}$$

where $h = 2h_{xy}/(h_x + h_y)$. Now for a set of concepts, the similarity network can be constructed in terms of the above similarity measures.

2.2 Ontology similarity network

Based on the similarity between concepts, the vector-based cosine similarity measure and Euclidean distance can be adapted to define the similarity between ontologies. Suppose that two ontologies x and y can be represented by vectors respectively:

$$x = \sum_{i=1}^{m} a_i x_i, \quad y = \sum_{j=1}^{n} b_j y_j$$

where x_i and y_j are concepts of the ontology and a_i and b_j are their weights. Then we use the structure of the ontology to define the dot product of two ontologies as

$$x \otimes y = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j x_i \otimes y_j = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathrm{sim}(x_i, y_j)$$

From the dot product of the two ontologies, the cosine-similarity and Euclidean distance-based similarity can be respectively defined as

$$\mathrm{sim}(x, y) = \frac{x \otimes y}{\sqrt{x \otimes x}\ \sqrt{y \otimes y}} \quad (4)$$

$$\mathrm{sim}(x, y) = \frac{1}{1 + \sqrt{x \otimes x - 2x \otimes y + y \otimes y}} \quad (5)$$

Similar to concepts, the similarity network between ontologies can be constructed using the above similarity measures.

2.3 Path algebra-based reasoning

Once similarity networks for instances, concepts or ontologies are constructed, the path algebra can be applied to calculate their implicit similarity. The similarity network is a directed connected graph $G = (V, E)$, $E \subseteq V \times V$. It has a vertex labeling function $f_v: V \to \Sigma$ and an edge similarity labeling $f_e: E \to [0, 1]$, where Σ can be a set of instances, concepts and ontologies. Similarity network-based reasoning is defined by path algebra on similarity network G. Let $P(v_0, v_m) = \{p \mid p$ is a path from v_0 to v_m in $G\}$, we define the path algebra as a set P with two binary operations $\vee$ and $\cdot$ which have the following properties: ①For all $x, y, z \in P$, the $\vee$ is idempotent, commutative, and associative: $x \vee x = x$, $x \vee y = y \vee x$, $(x \vee y) \vee z = x \vee (y \vee z)$; ②For all $x, y, z \in P$, the $\cdot$ is associative, and distributive over $\vee$: $(x \cdot y) \cdot z = x \cdot (y \cdot z)$, $x \cdot (y \vee z) = (x \cdot y) \vee (x \cdot z)$, $(y \vee z) \cdot x = (y \cdot x) \vee (z \cdot x)$; ③ The set P contains a zero element ϕ and a unit element e such that: $\phi \cdot x = \phi$, $\phi \vee x = x = x \vee \phi$, $e \cdot x = x = x \cdot e$. Obviously, probability reasoning and evidence-based reasoning mechanism can be applied to define the semantics of two path operations. Here two simple reasoning mechanisms are presented.

- Maximum similarity path reasoning can be defined by path algebra ① $x \vee y \equiv \max(x, y)$, for all $x, y \in \mathbf{R}$, $\phi = -\infty$, $e = 0$; ② $x \cdot y \equiv x + y$, for all for all $x, y \in \mathbf{R}$, $\phi = -\infty$, $e = 0$.
- Minimum/maximum similarity path reasoning can be defined by path algebra ① $x \vee y \equiv \max(x, y)$, for all $x, y \in \mathbf{R}$, $\phi = -\infty$, $e = 0$; ② $x \cdot y \equiv \min(x, y)$, for all for all $x, y \in \mathbf{R}$, $\phi = -\infty$, $e = 0$.

3 Experiment and Results

To validate the framework, we apply the framework and similarity network-based reasoning to text categorization and visualization of images. In text categorization, patent documents that consist of 355 samples with six classes are utilized. The reason is that we investigate to learn ontologies from patent documents and then exploit them to perform patent analysis[11]. In experiments, this dataset is randomly split using "ModApte" into two parts: 70% documents for training and the other 30% for testing. This is executed ten times.

We use the conventional TFIDF approach to build the vector for each document[11]. This step does not utilize the stemming algorithm and feature selection. The second step constructs the ontology for each vector using WordNet. The third step establishes the concept similarity networks respectively using similarity measures Eq. (1), Eq. (2), and Eq. (3). It also establishes the ontology similarity network by Eq. (4), where the minimum/maximum similarity path reasoning is applied to calculate the implicit similarity. We test the kNN classifier on the experimental dataset in two cases. The first case is that the similarity measures defined by Eq. (1), Eq. (2), and Eq. (3) are respectively used for kNN, denoted by RHC, LHC, and LXC. The other case is that similarity network-based reasoning is additionally applied, denoted by RHCM, LHCM, and LXCM, respectively. It can be seen from Fig. 2 that similarity network-based reasoning can be applied to improve the accuracy of text categorization with any measure. It is also illustrated that the proposed measure, Eq. (3), is better than any other measures.

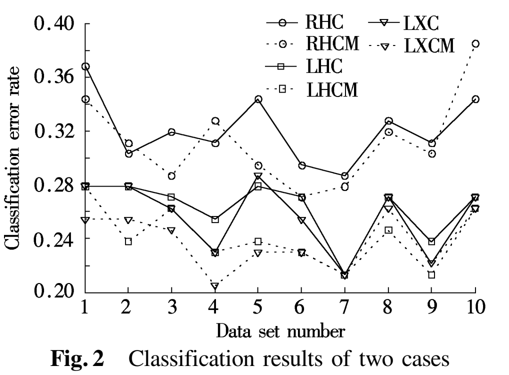
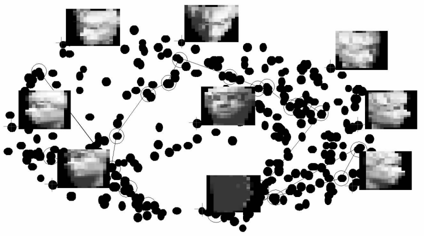


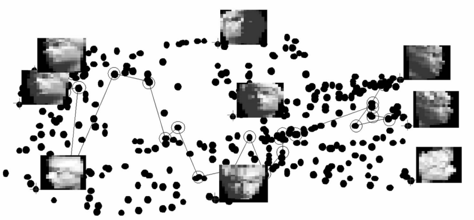
Fig. 2 Classification results of two cases

The other experiment is the visualization of face images[10]. Similar images can be regarded as instances of the same image ontology. When we do not know the structure of the ontology, but we have a lot of its instances that are represented as vectors, we can still apply the proposed framework to reason the similarity between these instances. We choose the dataset that consists of 350 instances (each contains 64×64 pixels) of a human face rendered with different poses and lighting directions. Isomap visualizes this dataset using Euclidean distance to determine the neighborhood[10]. The proposed m2-Isomap utilizes Eq. (5) to build the similarity network and then apply minimum/maximum path-based similarity reasoning to determine the neighborhood for Isomap instead of Euclidean distance.

In Fig. 3, all input image instances are projected onto two-dimensional space where samples of the images are superimposed. The direction from left to right represents the left-right poses of the faces, and the direction from upside to downside represents the up-down or down-up poses of the faces. The corresponding points of the successive images from left to right in the middle are marked by circles and linked by lines. The nine critical face samples are marked by a plus at the left-bottom corner of each image indicating the point representing the image. It can be observed from Fig. 3 that Isomap can hardly reveal the different face poses. The middle left-right line is heavily curved, and the arrangement of the nine face samples is tanglesome. On the other hand, m2-Isomap renders the middle left-right line better, and the nine face samples are mapped to the approximately correct positions corresponding to the face poses. These results validate the similarity network-based reasoning.



(a)



(b)

Fig. 3 Visualization results of face images in two cases. (a) Isomap, neighborhood size $k = 5$; (b) m2-Isomap, neighborhood size $k = 5$ and extended size $m = 1$

4 Conclusion

This paper proposes a hierarchical framework for calculating ontological similarity, which integrates the extension-based approaches, intension-based approaches, and similarity network-based reasoning to exploit implicit similarity. The preliminary experiments have validated this framework. In future work, we will exploit this framework to map ontology by finding similar concepts among different ontologies. We will also attempt to exploit the probability reasoning and evidence-based reasoning to investigate the semantics of path operations.

References

[1] Egenhofer M J, Rodriguez M A. Determining semantic similarity among entity classes from different ontologies [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2003, **15**(2): 442 – 456.

[2] Ehrig M, Sure Y. Ontology mapping—an integrated approach[A]. In: *Proc of the First European Semantic Web Symposium*[C]. Heraklion, Greece, 2004. 76 – 91.

[3] Bandar Z A, Mclean D, Li Y. An approach for measuring semantic similarity between words using multiple information sources[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2003, **15**(4): 871 – 882.

[4] Andreasen H, Knappe T, Bulskov R. Similarity graphs[A]. In: *Proc of the 14th International Symposium on Methodologies for Intelligent Systems* [C]. Maebashi, Japan, 2003. 668 – 672.

[5] Doan A, Madhaven J, Dhamankar R, et al. Learning to match ontologies on the semantic web [J]. *The International Journal on Very Large Data Bases*, 2003, **12**(4): 303 – 319.

[6] Maedche A, Staab S. Measuring similarity between ontologies [A]. In: *Proc of the 13th International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*[C]. Siguenza, Spain, 2002. 251 – 263.

[7] Molina H G, Widom J, Ganesan P. Exploiting hierarchical domain structure to compute similarity [J]. *ACM Transactions on Information Systems*, 2003, **21**(1): 64 – 93.

[8] Sun J, Yuan S T. Ontology-based structured cosine similarity in document summarization: With applications to mobile audio-based knowledge management [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, **35**(5): 1028 – 1040.

[9] Hefke M, Ehrig M, Haase P et al. Similarity for ontology-a comprehensive framework [A]. In: *Workshop Enterprise Modeling and Ontology: Ingredients for Interoperability* [C]. Vienna, Austria, 2004. 27 – 34.

[10] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction [J]. *Science*, 2000, **290**(5500): 2319 – 2323.

[11] Wen Guihua. *Rotating dynamics for computational creativity* [M]. Beijing: National Defence Industry Press, 2005. (in Chinese)

本体的相似性网络推理框架

文贵华　　江丽君

(华南理工大学计算机科学与工程学院,广州 510641)

摘要:为了更恰当地计算本体相似性,提出了一种本体的相似性网络推理的集成框架. 该框架集成了基于外延的方法,基于内涵的方法,计算间接相似的相似性网络推理,和检验相似性测度有效性的环境反馈. 同时,提出了一种用于构造概念相似性网络的新测度,相似性网络上的推理则采用图论实现而不是预定义知识规则,这样可免去知识获取的困难. 框架已经应用于文本分类和高维数据的可视化,理论分析和实验验证了相似性网络推理框架的有效性.

关键词:本体;相似网络推理;图代数;集成框架

中图分类号:TP393.1