SCIENCE CHINA
Information Sciences

• LETTER •

December 2021, Vol. 64 229101:1–229101:2
https://doi.org/10.1007/s11432-020-3214-2

Matching weak informative ontologies

Peng WANG[1*] & Baowen XU[2*]

[1]*School of Computer Science and Engineering, Southeast University, Nanjing* 211189*, China;*
[2]*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China*

Received 10 November 2020/Accepted 9 March 2021/Published online 3 November 2021

Citation Wang P, Xu B W. Matching weak informative ontologies. Sci China Inf Sci, 2021, 64(12): 229101, https://doi.org/10.1007/s11432-020-3214-2

Dear editor,

Most existing ontology matching methods utilize literal information to discover alignments [1, 2]. However, some literal information in ontologies may be opaque and some ontologies may not have sufficient literal information. These ontologies are named weak informative ontologies (WIOs) and it is challenging for existing methods to match WIOs. On one hand, string-based and linguistic-based matching methods cannot work well for WIOs. On the other hand, some matching methods use external resources to improve their performance, but collecting and processing external resources are still time-consuming.

To address the issue of matching WIOs, we propose a practical matching method which is inspired by our previous work [3, 4] about the semantic subgraph and similarity propagation. Figure 1 depicts an overview of the proposed method, which involves three steps: (1) building the ontology graph from the WIO, (2) extracting semantic subgraphs from the ontology graph, and (3) calculating similarity propagation to obtain similarity matrix and the alignment.

Ontology graph. We first use the hybrid ontology graph to represent the WIO for distinguishing multiple properties between concepts, then explicitly describe the containers and collections in the ontology graph, afterward, enrich the ontology by discovering hidden semantics, furthermore, refine the ontology graph by removing annotation and definition triples. As a result, according to the original source WIO and target WIO, we build two ontology graphs, which can clearly describe the semantic information in ontologies.

Semantic subgraph. For each concept or property in the ontology graph, we extract the corresponding semantic subgraph, which can precisely describe the meaning of the concept or property.

Definition 1. Given an element e in a hybrid ontology graph G_h, its semantic subgraph $G_s(e)$ is composed of top-k (top-$k \in \mathbb{N}$) related triples that describe e. $G_s(e) \subseteq G_h$ and $G_s(e)$ has the following features. (1) The size of $G_s(e)$ is limited; (2) $G_s(e)$ does not emphasize semantic completeness; (3) $G_s(e)$ is unique; (4) $G_s(e)$ prefers triples related to e.

We apply a circuit model to efficiently rank triples and then extract semantic subgraphs. More concretely, in the circuit model, the conductivity C simulates the capability of conveying information, the voltage V indicates the capability of preserving information, and the current I denotes the semantic information flows on edges in the ontology graph. Let $I(u, v)$ denote the current from vertex u to vertex v, $V(u)$ and $V(v)$ be the voltages on u and v, and $C(u, v)$ be the conductivity on the edge between u and v. z is a sink node and each vertex has an edge to z. Then an ontology graph is converted into a circuit, which has the initial conditions: $V(s) = 1, V(z) = 0$. The delivered current $\hat{I}(P)$ in a prefix-path $P = (s = u_1, \ldots, u_i)$ is the volume of electrons that arrives at u_i through P, and it can be calculated by

$$\hat{I}(s = u_1, \ldots, u_i) = \hat{I}(s = u_1, \ldots, u_{i-1}) \frac{I(u_{i-1}, u_i)}{I_{\text{out}}(u_{i-1})}, \quad (1)$$

where $I_{\text{out}}(u)$ is the total current coming from u. Therefore, the captured flow of subgraph G_s is the sum of all the delivered current in the prefix-path in G_s:

$$\text{CF}(G_s) = \sum_{P=(s,\ldots,t)\in G_s} \hat{I}(P). \quad (2)$$

For all subgraphs with k triples, the subgraph with the maximum capture flow is the semantic subgraph. In other words, a semantic subgraph is determined by the captured flow on paths, which contains relevant triples about s.

In an ontology graph, the conductivity for any triple $t = \langle s, p, o\rangle$ can be obtained based on the weights of s, p and o. Since s and o are relevant to other triples, their weights should be divided by the degrees:

$$w(t) = \frac{\frac{w(s)}{\text{degree}(s)} + w(p) + \frac{w(o)}{\text{degree}(o)}}{3}. \quad (3)$$

Matcher based on semantic subgraph. For an element, we organize relevant literal information based on semantic subgraphs as a virtual document [5] and call this virtual document the semantic description document (SDD). Each concept, property, or instance has a basic SDD D_{base}, which consists of the local name, label, and annotation. The SDD of concept C is organized by concept hierarchy, axioms, related properties and instances. The SDD

* Corresponding author (email: pwang@seu.edu.cn, bwxu@nju.edu.cn)

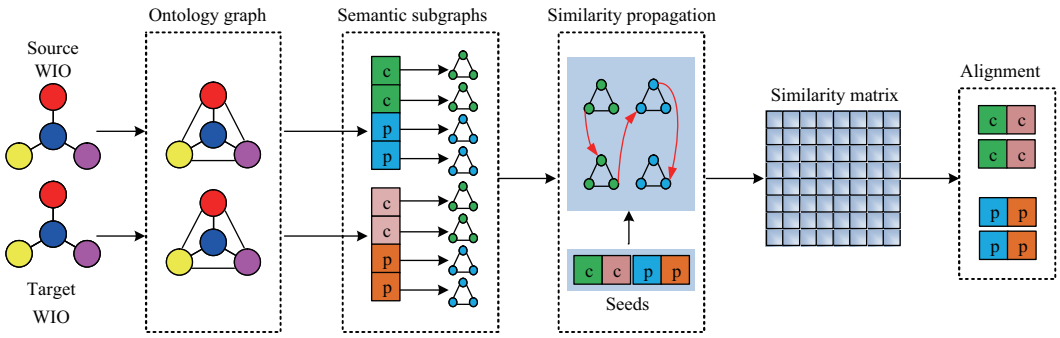

Figure 1 (Color online) Overview of matching weak informative ontologies.

of property P is organized by domain and range statements. The SDD of a blank node b in an ontology is calculated recursively according to triples that contain b. A SDD is a set of vocabularies with weights, namely, SDD $= \{p_1 \times W_1, p_2 \times W_2, \ldots, p_x \times W_x\}$.

Let Doc $= \{\mathrm{SDD}_1, \mathrm{SDD}_2, \ldots, \mathrm{SDD}_N\}$, and each SDD contains n items $t_1, t_2, \ldots, t_n$. Thus each document SDD_i can be described as an n-dimension vector $\boldsymbol{D}_i = (d_{i1}, d_{i2}, \ldots, d_{in})$, where d_{ij} is the weight of j-th item. The similarity between two virtual documents is the cosine value of vectors. Therefore, the similarity between $\boldsymbol{D}_i$ and $\boldsymbol{D}_j$ is

$$\mathrm{Sim}(\boldsymbol{D}_i, \boldsymbol{D}_j) = \frac{\sum_{k=1}^{n} d_{ik} \times d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}^2 \times \sum_{k=1}^{n} d_{jk}^2}}. \tag{4}$$

Similarity propagation. Considering the characteristics of ontologies, we design a new similarity propagation model with strong constraint condition (SC-condition). Given two ontology graphs and initial similarity seeds, we first construct a pairwise connectivity graph, then get an induced propagation graph, and finally obtain the new similarities by fixpoint value calculation. This similarity propagation model not only avoids the performance drawbacks but also can handle the property alignment in ontology matching. The similarity propagation between ontology graphs can be computed iteratively until the final similarity matrix is converged.

Definition 2. Given two triples $t_i = \langle s_i, p_i, o_i \rangle$ and $t_j = \langle s_j, p_j, o_j \rangle$, let S_s, S_p and S_o denote the corresponding similarities of (s_i, s_j), (p_i, p_j) and (o_i, o_j), respectively. Similarities can be propagated only if t_i and t_j satisfy the following three conditions. (1) In S_s, S_p and S_o, at least two similarities must be larger than threshold θ; (2) If t_i includes ontology language primitives, the corresponding positions of t_j must be same; (3) t_i or t_j has at most one ontology language primitive.

For each element pair (x, y), which would be subject pair, predicate pair or object pair, its new similarity in the $(i+1)$th propagation contains four parts: (1) the similarity in ith propagation; (2) the propagation similarity when (x, y) is object pair; (3) the propagation similarity when (x, y) is subject pair; (4) the propagation similarity when (x, y) is predicate pair.

The propagation model employs the updating mechanism, credible seeds, penalty, termination condition, and propagation scale strategies in order to ensure a balance between matching efficiency and quality. In particular, the initial credible seeds in propagation are provided by the matcher based on semantic subgraphs, i.e., the matcher calculates the similarities between the semantic description documents, which are constructed from semantic subgraphs. Lastly, after the similarity propagation, we obtain the similarity matrix and then extract the alignment from it.

Results. The proposed method is evaluated on the open OAEI benchmark datasets: benchmark2008 and benchmark2009, which have 110 matching tasks, among them, 78 tasks are WIOs. According to our experimental results, we observe the following facts. (1) The proposed similarity propagation method improves the quality of matching results, especially for the WIOs; (2) For the WIOs, the similarity propagation model can increase the recall of results greatly; (3) For informative ontologies, our method can also produce good results. It means that our method is a general matching method. Moreover, compared with more than 20 ontology matching systems, our method achieves an average of 0.90 precision and 0.76 recall on WIO matching tasks and average 0.95 precision and 0.84 recall on general matching tasks, which are the state-of-the-art performances on the OAEI benchmark datasets.

Conclusion. We proposed a method for matching weak informative ontologies. The success of our method is attributed to two techniques: semantic subgraphs and a new similarity propagation model. Semantic subgraphs can not only precisely describe ontology elements with limited triples, but also can be used to calculate the similarity by constructing the semantic description virtual documents. The similarity propagation model is based on the strong constrained condition, and it is reasonable for handling ontology matching.

Acknowledgements The work was supported by National Key R&D Program of China (Grant No. 2018YFD1100302), National Natural Science Foundation of China (Grant No. 61832009), and 13th Five-Year All-Army Common Information System Equipment Pre-Research Project (Grant Nos. 31514020501, 31514020503).

References

1 Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges. IEEE Trans Knowl Data Eng, 2013, 25: 158–176
2 Ochieng P, Kyanda S. Large-scale ontology matching. ACM Comput Surv, 2018, 51: 1–35
3 Wang P, Xu B. An effective similarity propagation model for matching ontologies without sufficient or regular linguistic information. In: Proceedings of the 4th Asian Semantic Web Conference (ASWC2009), Shanghai, 2009
4 Wang P, Xu B, Zhou Y. Extracting semantic subgraphs to capture the real meanings of ontology elements. J Tinshhua Sci Technol, 2010, 15: 724–733
5 Qu Y, Hu W, Cheng G. Constructing virtual documents for ontology matching. In: Proceedings of the 15th International Conference on World Wide Web (WWW2006), Edinburgh, 2006