Ontology Mapping: An Information Retrieval and Interactive Activation Network Based Approach

Ming Mao

School of Information Sciences
University of Pittsburgh
mingmao@mail.sis.pitt.edu

Abstract. Ontology mapping is to find semantic correspondences between similar elements of different ontologies. It is critical to achieve semantic interoperability in the WWW. This paper proposes a new generic and scalable ontology mapping approach based on propagation theory, information retrieval technique and artificial intelligence model. The approach utilizes both linguistic and structural information, measures the similarity of different elements of ontologies in a vector space model, and deals with constraints using the interactive activation network. The results of pilot study, the PRIOR, are promising and scalable.

Keywords: ontology mapping, profile propagation, information retrieval, interactive activation network, PRIOR

1 Introduction

The World Wide Web (WWW) now is widely used as a universal medium for information exchange. Semantic interoperability among different information systems in the WWW is limited due to information heterogeneity, and the non semantic nature of HTML and URLs. Ontologies have been suggested as a way to solve the problem of information heterogeneity by providing formal and explicit definitions of data. They may also allow for reasoning over related concepts. Given that no universal ontology exists for the WWW, work has focused on finding semantic correspondences between similar elements of different ontologies, i.e., *ontology mapping*. Automatic ontology mapping is important to various practical applications such as the emerging Semantic Web [3], information transformation and data integration [2], query processing across disparate sources [7], and many others [4].

Ontology mapping can be done either by hand or using automated tools. Manual mapping becomes impractical as the size and complexity of ontologies increases. Fully or semi-automated mapping approaches have been examined by several research studies, e.g., analyzing linguistic information of elements in ontologies [15], treating ontologies as structural graphs [12], applying heuristic rules to look for specific mapping patterns [8] and machine learning techniques [1]. More comprehensive surveys of ontology mapping approaches can be found in [9][14].

This paper proposes a new generic and scalable ontology mapping approach, shown in **Fig. 1**. The approach takes advantage of propagation theory, information retrieval technique and artificial intelligence model to solve ontology mapping problem. It utilizes both linguistic and structural information of ontologies, measures the similarity of different elements of ontologies in a vector space model, and integrates interactive activation network to deal with constraints.

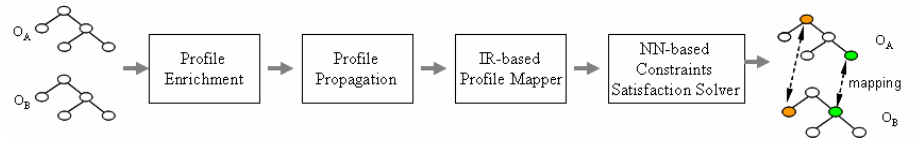

Fig. 1. The architecture of the proposed approach

2 The Proposed Approach

2.1 Profile Enrichment

Similar as the virtual document used in Falcon-AO system [15], the *profile* of a concept is a combination of linguistic information of a concept, i.e. the profile of a concept = its name + label + comment + property restriction + other descriptive information. The Profile Enrichment is a process that uses a profile to represent a concept in the ontology, and thus enrich its information. It is based on the observation that sometimes the information carried in the name of a concept is restricted, but other descriptive information like labels and comments may contain words that better convey the meaning of concepts. A sample profile of a concept "book" looks like *Profile(book)= (book, book, book, monograph, collection, write, text)*. Afterwards the *tf·idf* (term frequency–inverse document frequency) weight will be used to assign larger weight to the terms that have a high frequency in given document and a low frequency in the whole collection of documents. Here each profile is treated as a document and all profiles in two ontologies are treated as the collection of documents.

2.2 Profile Propagation

The Profile Propagation exploits the neighboring information of each concept. That is, the profile of ancestors and descendants are passed to that of the concept itself based on propagation theory [6]. The process of profile propagation can be represented using Equation 1, where *C* and *C'* denote two concepts in the ontologies, *S* denotes the set of all concepts in the ontologies, V_{Cnew} denotes the new profile vector of the concept *C*, $V_{C'}$ denotes the profile vector of the concept *C'*, and *w(C, C')* is a function that assigns different weights to the neighbors of the concept following two principles:

1. The closer two concepts locate, the higher impact they have. 2. The impact from ancestors to descendants is higher than the impact from descendants to ancestors.

$$V_{Cnew} = \sum_{C' \in S} w(C, C') V_{C'} \quad (1)$$

2.3 IR-based Profile Mapper

The insight of the proposed approach is to treat ontology mapping problem as an information retrieval task. That is, if concepts in an ontology are seen as documents in a collection, finding correspondence between similar concepts in ontologies is just like to search the most relevant document in one collection given a document in another collection. Given a query and a set of documents, classical IR methods usually measure the similarity of a query and different documents, and then return the documents with top-ranked similarities as result. In the context of ontology mapping, such IR method can be applied as following: Given two to-be-mapped ontologies, O_A and O_B, all profiles of concepts in O_A are indexed first. Simultaneously queries based on the profile of each concept in O_B are generated. Then searches are executed in O_A using queries generated from O_B one by one. Afterwards the concepts in O_A with top-ranked similarities or above a predefined threshold are returned and stored. Now two ontologies are switched and the whole process is repeated. Finally two result sets are compared and the overlapped ones indicate possible mappings.

Cosine angle between two vectors of the documents is commonly used to measure their similarity. In the context of ontology mapping, the cosine similarity between two concepts C and C' can be measured using Equation 2, where V_C and $V_{C'}$ are two vectors of the profile of concept C and C' respectively, n is the dimension of the profile vectors, V_i^C and $V_i^{C'}$ are i_{th} element in the profile vector of concept C and C' respectively, $|V_C|$ and $|V_{C'}|$ are the lengths of the two vectors respectively. The output of Profile Mapper is a concept-to-concept similarity matrix, where each element represents a similarity between two concepts. Note that such a similarity matrix might be very sparse due to the large size of ontologies and the low overlap between them.

$$Sim_{C,C'} = Sim(V_C, V_{C'}) = \frac{\overrightarrow{V_C} \bullet \overrightarrow{V_{C'}}}{|V_C \| V_{C'}|} = \frac{\sum_{i=1}^{n} (V_i^C * V_i^{C'})}{\sqrt{\sum_{i=1}^{n} (V_i^C)^2} \sqrt{\sum_{i=1}^{n} (V_i^{C'})^2}} \quad (2)$$

2.4 Interactive Activation Network Based Constraints Satisfaction Solver

Constraints satisfaction problem (CSP) [16] arises as an intriguing research problem in ontology mapping due to the characteristics of ontology itself and its representations. The hierarchical relations in RDFS, the axioms in OWL and the rules in SWRL result in different kinds of constraints. For example, "if concept A matches concept B, then the ancestor of A can not match the child of B in the taxonomy" and "two classes match if they have owl:sameAs or owl:equvalentClass relations". To

improve the quality of ontology mapping, it is critical to find the best configuration that can satisfy such constraints as much as possible.

CSPs are typically solved by a form of search, e.g. backtracking, constraint propagation, and local search [16]. The interactive activation network is first proposed to solve CSPs in [13]. The network usually consists of a number of competitive nodes connected to each other. Each node represents a hypothesis. The connection between two nodes represents constraint between their hypotheses. Each connection is associated with a weight. For example, we have two hypotheses, H_A and H_B. If whenever H_A is true, H_B is usually true, then there is a positive connection from node *A* to node *B*. Oppositely if H_A provides evidence against H_B, then there is a negative connection from node *A* to node *B*. The importance of the constraint is proportional to the strength (i.e. *weight*) of the connection representing that constraint. The state of a node is determined locally by the nodes adjacent to it and the weights connecting to it. The state of the network is the collection of states of all nodes. Entirely local computation can lead the network to converge to a global optimal state.

In the context of ontology mapping, a node in an interactive activation network represents a hypothesis that concept C_{1i} in ontology O_1 can be mapped to concept C_{2j} in ontology O_2. The initial activation of the node is the similarity of (C_{1i}, C_{2j}). The activation of the node can be updated using the following simple rule, where a_i denotes the activation of *node i*, written as n_i, net_i denotes the net input of the node.

$$a_i(t+1) = \begin{cases} a_i(t) + net_i(1 - a_i(t)), & net_i > 0 \\ a_i(t) + net_i a_i(t), & net_i < 0 \end{cases} \quad (3)$$

The net_i comes from three sources, i.e. its neighbors, its bias, and its external inputs, as defined in Equation 4, where w_{ij} denotes the connection weight between n_i and n_j, a_j denotes the activation of node n_j, $bias_i$ denotes the bias of n_i, the *istr* and *estr* are constants that allow the relative contributions of the input from internal sources and external sources to be readily manipulated. Note that the connection matrix is not symmetric and the nodes may not connect to themselves, i.e., $w_{ij} \neq w_{ji}$, $w_{ii}=0$.

$$net_i = istr \times (\sum_j w_{ij} a_j + bias_i) + estr \times (input_i) \quad (4)$$

Furthermore, the connections between nodes in the network represent constraints between hypotheses. For example, the constraint that "only 1-to-1 mapping is allowed" results in a negative connection between nodes (C_{1i}, C_{2j}) and (C_{1i}, C_{2k}), where $k \neq j$. Moreover, "two concepts match if all their children match", results in a positive connection between nodes (C_{1i}, C_{2j}) and (C_{1k}, C_{2t}), where C_{1k} and C_{2t} are the children of C_{1i} and C_{2j} respectively. Finally, the complexity of the connections may be very large because of complex constraints.

3 Pilot Study

The proposed approach has been partially implemented in the PRIOR [10][11], an ontology mapping tool based on propagation theory and information retrieval

techniques. The results from OAEI ontology matching campaign 2006[1] show the PRIOR is promising and competitive to all other approaches in different tracks, namely benchmark, web directory, food, and anatomy [5].

4 Future Work

The implementation of the interactive activation network to satisfy constraints in ontology mapping is our major future work. Other work includes integrating auxiliary information such as WordNet to distinguish synonyms.

References

1. Doan, A., J. Madhaven, et al. (2003). "Learning to Match Ontologies on the Semantic Web." VLDB Journal **12**(4): 303-319.
2. Dou, D., D. McDermott, et al. (2005). "Ontology Translation on the Semantic Web." Journal on Data Semantics (JoDS) **II**: 35-57.
3. Ehrig, M. (2006). Ontology Alignment: Bridging the Semantic Gap (Semantic Web and Beyond). ISBN-038732805X. Springer. 2006.
4. Euzenat, J., Bach, T., et al. (2004). State of the art on ontology alignment, Knowledge web NoE.
5. Euzenat, J et al. (2006). Results of the Ontology Alignment Evaluation Initiative 2006. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA.
6. Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. International Journal of Computer Vision, Vol. 70, No. 1.
7. Gasevic, D. and M. Hatala (2005). "Ontology mappings to improve learning resource search." British Journal of Educational Technology.
8. Hovy, E. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), Granada, Spain.
9. Kalfoglou, Y. and M. Schorlemmer (2003). "Ontology mapping: the state of the art." The Knowledge Engineering Review 18(1): 1-31.
10. Mao, M. and Peng, Y. (2006). PRIOR System: Results for OAEI 2006. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA.
11. Mao, M., Peng, Y. and Spring, M. (2007) A Profile Propagation and Information Retrieval Based Ontology Mapping Approach, In Proceedings of SKG 2007.
12. Melnik, S., H. Garcia-Molina, et al. (2002). Similarity flooding: a versatile graph matching algorithm and its application to schema matching. Proc. 18th International Conference on Data Engineering (ICDE).
13. McClelland, J. L. and Rumelhart, D. E. (1988). Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises. The MIT Press.
14. Noy, N. (2004). "Semantic Integration: A Survey of Ontology-Based Approaches." SIGMOD Record **33**(4): 65-70.
15. Qu, Y., Hu, W., and Cheng, G. (2006). Constructing virtual documents for ontology matching. In Proceedings of the 15th International Conference on World Wide Web.
16. Tsang, E. (1993). Foundations of Constraint Satisfaction: Academic Press.

[1] http://oaei.ontologymatching.org/2006/results/