A Profile Propagation and Information Retrieval Based Ontology Mapping Approach

Ming Mao, Yefei Peng and Michael Spring
School of Information Sciences
University of Pittsburgh
{mingmao,ypeng, spring}@mail.sis.pitt.edu

Abstract

Ontology has been used widely to help finding relevant information among distributed and heterogeneous sources. Given that no universal ontology exists for the WWW, ontology mapping attracts many researchers' interest in various areas. In this paper we propose a new generic ontology mapping approach based on profile propagation and information retrieval techniques. The features used to establish the profile of a concept include all lexical information (i.e., its name, label, comments, property restriction, etc.). Profile propagation then is used to integrate structural information by adding the profiles of its ancestors, descendants and siblings to the profile of a concept, with different weights. Afterwards profiles are compared in a vector space model and the most relevant one will be returned as mapping results. A search engine is integrated as profile mapper when the size of ontologies is large. Experimental results show that the proposed approach obtained a good performance in OAEI campaign 2006.

Keywords
Ontology Mapping, Profile Propagation, Information Retrieval, PRIOR, Vector Space Model

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. Ontology mapping can be done either by hand or using automated tools. Manual mapping becomes impractical as the size and complexity of ontologies increases. Full or semi-automated mapping approaches have been examined by several research studies. Previous mapping approaches include using linguistic techniques to measure the lexical similarity of concepts in ontologies [14], treating ontologies as structural graphs [11], applying heuristic rules to look for specific mapping patterns [5], and learning to map ontologies through machine learning techniques [1]. For more comprehensive surveys of existing mapping approaches and systems, please see [7][13].

Though these approaches obtain good results in different applications, none of them ever takes the advantage of information retrieval (IR) techniques and few of them utilized the neighboring information of a concept when exploring linguistic characteristics of ontology. In this paper, we propose a new generic ontology mapping approach based on profile propagation and IR techniques. The main procedures of the proposed approach are profile enrichment, profile propagation and IR-based profile mapping. Experiments show the experimental system, the PRIOR [10], obtained a good performance in all tests in OAEI campaign 2006, namely benchmark, web directory, anatomy and food tests.

The paper is structured as follows. §2 depicts the ontology mapping problem. §3 presents the approach. The results of the proposed approach at OAEI campaign 2006 given in §4 are promising and appear to be scalable. Related works are reviewed in §5, and §6 outlines future work.

2. Problem Statement

Ontology is a formal, explicit specification of a shared conceptualization in terms of classes, properties and relations [4]. Figure 1 shows sample ontologies about *Book*. Both the reference ontology (i.e., $Book_R$ on the left) and the test ontology (i.e., $Book_T$ on the right) include three classes (i.e., *Collection*, *Monograph* and *Proceedings* vs. *Compilation*, *Monography* and *ConferenceMinutes*, respectively) with a "*subClassOf*" relationship. Classes can be associated with properties and/or instances. For example, in the reference ontology the "proceedings" has four attributes (i.e., *communications*, *event*, *editor* and *organization*), and the "monograph" has an instance (i.e., "*object-oriented data modeling*" published by the MIT Press at 2000).

Ontology mapping aims to find semantic correspondences between similar elements in two homogeneous ontologies. In this paper, "semantic correspondence" refers to "=" relationship, and the "elements" refers to classes, properties and/or relations. A mapping between similar element e_i and e_j in O_A and O_B respectively can be written as: $m(e_i,e_j)$. In the situation of ontology $Book_R$ and $Book_T$, possible mappings are: $m(Book_R,Book_T)$, $m(Collection,Compilation)$, etc.

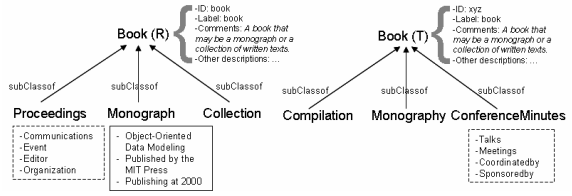

Figure 1. Two sample ontologies about *Book*

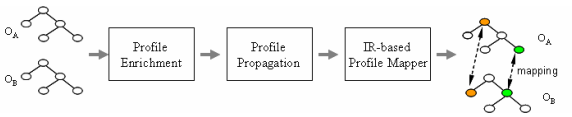

Figure 2. The architecture of proposed approach

3. The PRIOR Approach

The PRIOR approach is based on the insight that ontology mapping is also an information retrieval task. That is, if concepts in an ontology are treated 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. Figure 2 depicts the architecture of the proposed approach.

3.1. Profile Enrichment

First we introduce the term "profile". The profile of a concept in an ontology (i.e. class, attribute and instance) is a combination of linguistic description of a concept, i.e. the profile of a concept = its ID + label + comment + other descriptive information. For example, in *Book* example indicated in Figure 1, $Profile(Book_R)$ = (*book*, *book*, *book*, *monograph*, *collection*, *write*, *text*) and $Profile(Book_T)$ = (*xyz*, *book*, *book*, *monograph*, *collection*, *write*, *text*), after tokenizing (e.g., removing stop words, stemming and separating words via underscore etc.) but keeping all duplicates.

The utilization of *profile* is based on the observation that sometimes the information carried in ID is restricted, especially in the cases where the name is identified as meaningless symbols. Meanwhile other descriptive information such as labels and comments may contain words that better convey the meaning of the concept. Profile Enrichment thus aims to enrich the information of a concept by generating its profile.

Having profiles for each concept, 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. The *tf·idf* weight is defined as Equation 1, where, n_i is the number of occurrences of the considered term, $\sum n_k$ is the number of occurrences of all terms, N is the total number of documents in the collection, and n is the number of documents where the term t_i appears at least once (i.e., $n_i \neq 0$). In our case, each *profile* is treated as a *document* and N equals the total number of profiles in two ontologies. Profile Enrichment outputs a set of vectors. Each vector represents a concept using its *tf·idf* weights.

$$w = tf \bullet idf = \frac{n_i}{\sum_k n_k} \times \log_2 \frac{N}{n} \qquad (1)$$

3.2. Profile Propagation

The Profile Propagation exploits the neighboring information of each concept. That is, the profile of ancestors, descendants and siblings will be passed to that of the concept itself. The propagation is based on the observation a super class in an ontology reflects the *context* of its subclasses and a subclass is a specific *content* of its super class. The profile can be propagated in different levels. In Figure 3 the propagation level is 2 with weights of $w_{itself \to itself}$ = 1,

$w_{parent \to itself}$ = 1, $w_{grandparent \to itself}$ = .5, $w_{children \to itself}$ = .5, $w_{grandchildren \to itself}$ = .25, $w_{sibling \to itself}$ = .125, which are used in the experiments described in §4. Two principles are followed to assign weights: 1. The closer two concepts locate, the higher impact they have. 2. The ancestors have more impact to their descendants than the impact from the descendants to the ancestors.

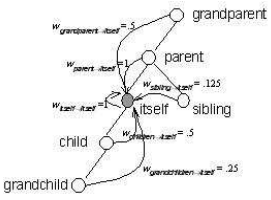

Figure 3. Profile propagation

The process of Profile Propagation can be represented as Equation 2, where N and N' represent two concepts in the ontologies, S represents the set of all concepts in the ontologies, V_{Nnew} represents the new profile vector of the concept N, $V_{N'}$ represents the original profile vector of the concept N', and $w(N, N')$ is the function that assigns different weights to the neighbors of the concept. Profile Propagation outputs a new set of vectors. Each vector is updated from its original vector using Equation 2.

$$V_{Nnew} = \sum_{N' \in S} w(N, N')V_{N'} \qquad (2)$$

3.3. IR-based Profile Mapping

Given a query and a set of documents, classical IR methods usually measure the similarity between a query and different documents, and then return the documents having top-ranked similarities as result. In the context of ontology mapping, if we treat each profile in one ontology as a query, all profiles in the other ontology as a collection of documents, finding the most similar element in two ontologies just as to search the most relevant document from the collection using the query.

There are many ways to evaluate the similarity between two documents in a collection. The common method is to measure the cosine angle between the two vectors of the documents. In our case, the cosine similarity between two profiles of concepts N and N' can be measured as Equation 3, where V_N and $V_{N'}$ are two vectors representing the profile of concept N and N' respectively, n is the dimension of the profile vectors, V_i^N and $V_i^{N'}$ are i_{th} element in the profile vector of concept N and N' respectively, $|V_N|$ and $|V_{N'}|$ are the lengths of the two vectors respectively.

$$PSim_{N,N'} = Sim(V_N, V_{N'}) = \frac{\overrightarrow{V_N} \cdot \overrightarrow{V_{N'}}}{|V_N||V_{N'}|} = \frac{\sum_{i=1}^{n}(V_i^N * V_i^{N'})}{\sqrt{\sum (V_i^N)^2}\sqrt{\sum (V_i^{N'})^2}} \qquad (3)$$

As a complement to profile similarities, name similarities are calculated using Equation 4, where $EditDist(N, N')$ is the Levenshtein distance between the name (i.e. ID) of concepts N and N', L_N and $L_{N'}$ are the string length of the name of N and N' respectively.

$$NSim_{N,N'} = 1 - \frac{EditDist(N, N')}{Max(L_N, L_{N'})} \qquad (4)$$

Furthermore, the profile similarity and name similarity are integrated to obtain the final similarity between concepts N and N' using Equation 5, where $H_\alpha(x)$ is defined by Equation 6, $\sum w_i=1$, $PSim_{N,N'}$ and $NSim_{N,N'}$ are profile/name similarity of concepts N and N' respectively. In the experiments described in §4, we tentatively set α=.5, w_1=.4 and w_2=.6.

$$FSim_{N,N'} = w_1 PSim_{N,N'} + w_2 H_\alpha(NSim_{N,N'}) \qquad (5)$$

$$H_\alpha(x) = \begin{cases} x & (x \geq \alpha) \\ 0 & (x < \alpha) \end{cases} \qquad (6)$$

The output of Profile Mapping 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. Finally Hungarian algorithm [7] will be used to pick up mapping results from the similarity matrix.

4. Experimental Results and Discussions

This section presents the experimental results of the PRIOR [10] system on benchmark, web directory, food and anatomy test in OAEI campaign 2006, which is a yearly contest on ontology matching organized by Ontology Alignment Evaluation Initiative (OAEI) since 2004. Full results of OAEI campaign can be found at [2]. The evaluation of all tests are precision, recall and f-measure as defined in Equation 7, 8, 9 except anatomy track that is evaluated by a cross-validation:

$$\text{Precision} \quad p = \frac{\#correct\_found\_mappings}{\#all\_found\_mappings} \qquad (7)$$

Recall $r = \frac{\#correct\_found\_mappings}{\#all\_possible\_mappings}$ (8)

F-measure $f = \frac{2 \times p \times r}{p + r}$ (9)

4.1. Benchmark

Benchmark tests include 1 reference ontology O_R, dedicated to the very narrow domain of bibliography, and 51 test ontologies, O_T, but discarding a number of information from the reference ontology in order to evaluate how algorithms behave when this information is lacking. More specifically, benchmark tests can be divided into 5 groups as shown in Table 1. The result of PRIOR on benchmark tests is shown in Figure 4. Figure 5 is the comparison among PRIOR, Falcon-AO [6][14] and RiMOM [9], two dominant systems in the campaign benchmark tests (reviewed in §5).

4.2. Web Directory

The Web directories tests consist of 4640 elementary tests for aligning web sites directories (e.g., Google and Yahoo!). In web directory tests the same approach and parameter sets as benchmark tests are used to obtain alignments. Since all tests are blind, the performance of PRIOR system on Web directories tests is evaluated by OAEI as: *precision*=.337, *recall*= .244, *f-measure*=.283. The comparison among the campaign's participants is shown in Figure 6.

4.3. Anatomy

The anatomy task is to find alignment between classes in two medical ontologies, FMA ontology and OpenGALEN ontology. Due to the large size of both ontologies (72559 classes in FMA vs. 9564 classes in OpenGALEN), handling huge similarity matrix in a PC is intractable. In the case, Indri[1], an open source search engine, is integrated as an IR-based profile mapper. In other words, given two ontologies, O_A and O_B, first we index all profiles in O_A as documents. Simultaneously we generate queries based on profile in O_B. Then we do search in O_A using queries generated from O_B by calculating the similarity between queries and documents. Afterwards those concepts in O_A with top-ranked similarity or above a predefined threshold are stored. Now two ontologies are switched and the whole process is repeated. Finally the overlapped results in two processes indicate possible mappings.

The evaluation of the mapping results for anatomy ontologies is problematic. Many standard ways of doing an evaluation cannot be applied in this case. First, there is no gold standard mapping exist for these particular ontologies. Secondly, creating gold standard mappings is not feasible due to the size and complexity

Table 1. The description of OAEI benchmark tests

Tests	Description
#101-104	O_R and O_T have exactly the same or totally different names
#201–210	O_R and O_T have the same structure but different linguistics in some level
#221–247	O_R and O_T have the same linguistics but different structure
#248–266	Both structure and linguistics are different between O_R and O_T
#301-304	O_T are real world cases, which we have more interest in

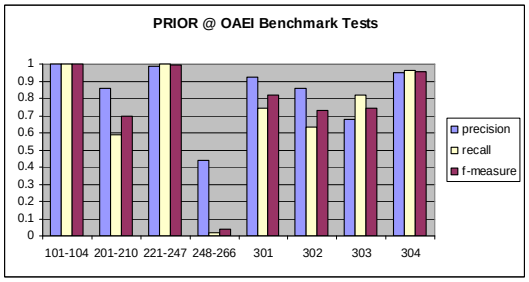

Figure 4. The result of PRIOR at benchmark tests

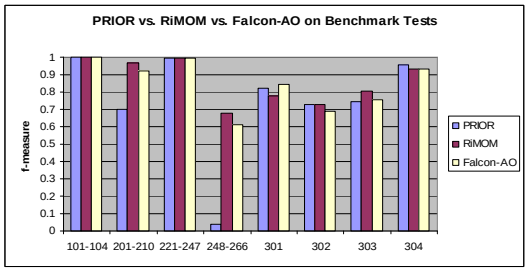

Figure 5. The f-measure of PRIOR vs. RiMOM vs. Falcon-AO at benchmark tests

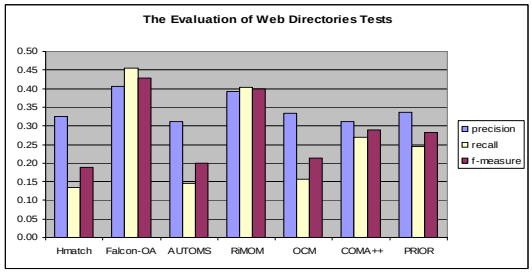

Figure 6. The evaluation of Web directories tests

[1] http://www.lemurproject.org/indri/

of ontologies. Therefore, both OAEI [2] and Zhang [17] adopt a cross-validation (Figure 7) for the evaluation purpose. They both claim that there is a significant overlap of the predicted mapping candidates among evaluated systems. That is, among 2583 predictions of PRIOR, 1455 are the same as other systems and 574 are predicted by PRIOR only.

4.4. Food

The food test requires creating alignment between the SKOS version of the United Nations Food and Agriculture Organization (FAO) AGROVOC thesaurus and the United States National Agricultural Library (NAL) Agricultural thesaurus. In this test, Indri-based profile mapper is used to find mappings due to the same reason as anatomy test that the size of task ontologies is very large (16000 terms, 28179 concepts in FAO thesaurus vs. 41000 terms, 41594 concepts in NAL thesaurus). The evaluation of food test provided by OAEI is shown in Figure 8, where PRIOR has precision=.71, recall=.55, and f-measure=.62.

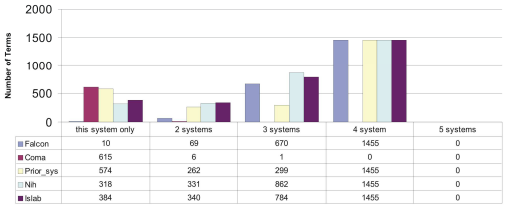

Figure 7. The cross-validation of anatomy test[2]

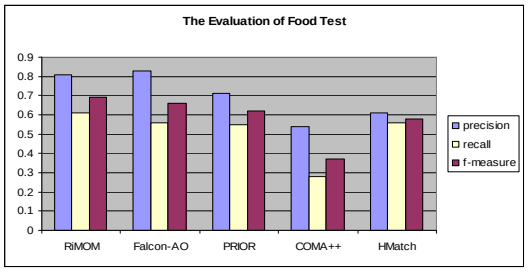

Figure 8. The evaluation of food test

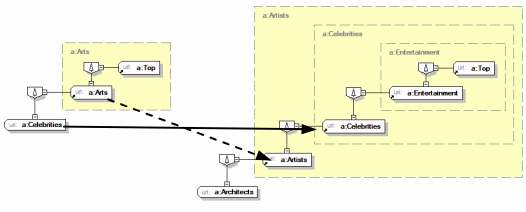

Figure 9. OAEI web directory test case #1

[2] http://oaei.ontologymatching.org/2006/results/anatomy/

4.5. Discussion

The result OAEI campaign 2006 [2] have shown the PRIOR approach based on profile propagation and information retrieval techniques is competitive to other systems in all tests, namely benchmark, web directory, anatomy and food tests. Though PRIOR has a poor performance on pure graphic matching tasks (i.e., benchmark tests #248-266, where both the linguistic and structural characteristics are changed heavily), its performance on real cases (i.e., #301-304) is equal to or slightly better than Falcon-AO and RiMOM. This might be because both Falcon-AO and RiMOM have a separate and independent procedure to measure structural similarities. In the contrast, PRIOR only makes use of structural information by propagating and thus its results rely on linguistic information more.

Another problem the PRIOR faces is many false positive results exist in some tests. For example, in web directory test case #1 (see Figure 9), PRIOR found 2 matches, i.e., “celebrities” to “celebrities” with a similarity score of 0.91, and “arts” to “artists” with a similarity score of 0.42. Since we have higher confidence that “celebrities” to “celebrities” is a true positive match (solid line), the match between “arts” and “artist” is false positive (dashed line), meaning “arts” is unlikely to match “artists” because “arts” is the parent of “celebrities” in the source ontology on the left and “artists” is the child of “celebrities” in the target ontology on the right. This problem is mostly caused by that the PRIOR is lack of validating preliminary results from a global structural view.

5. Related Work

Different approaches have been explored to solve ontology mapping problem. Some comprehensive surveys of ontology mapping and mostly related work, schema matching, can be found in [7][13][15]. Here only two dominant systems Falcon-AO [6] [14] and RiMOM [9] at OAEI campaign 2006 are reviewed.

Falcon-AO [6][14] is a domain-independent, generic ontology matching system. It combines three elementary matchers: V-Doc, I-Sub, and GMO, and one ontology partitioner, PBM. V-Doc constructs a virtual document for each URIref, and then treats the document as bags of words and compares them in a vector space model to evaluate similarity. I-Sub compares the similarity of strings by considering their similarity along with their differences. GMO explores structural similarity based on a bipartite graph. PBM partitions large ontologies into small clusters, and then matches between and within clusters. The *profile* used

in the PRIOR is very similar as the *virtual document* constructed in Falcon-AO. The difference is the *virtual document* only exploits neighboring information based on RDF model. Meanwhile the *profile* does not have any limitation to information type. Any information including instance can be integrated to the *profile* of a concept. Another difference is Falcon-AO design a PBM for large-scale ontologies specifically; while PRIOR take advantage of IR techniques in this case, and thus PRIOR is much more efficient. For example, PRIOR found 2583 mapping pairs in anatomy test within 9 minutes, but Falcon-AO took over 5.5h to complete their process.

RiMOM [9] is a domain-independent, generic ontology matching system too. To find optimal mappings from source ontology to target ontology, RiMOM integrates multiple strategies: edit-distance based strategy, statistical learning based strategy, and three similarity propagation based strategies. Both RiMOM and PRIOR do propagation based on propagation theory [3]. The difference is RiMOM propagates the similarity of two entities to entity pairs with some kinds of relationship (e.g. superClassOf, siblingClassOf, range, domain etc.) with them. Meanwhile PRIOR propagates original information of a concept instead of their similarity to its ascendant, descendant or siblings, and then compares their similarity based on propagated profiles. Though RiMOM obtains a good performance in benchmark tests, the approach fails in dealing with large-scale ontology mapping problem such as anatomy test, in which PRIOR is efficient.

6. Conclusion and Future Work

In this paper, we propose a new generic ontology mapping approach based on profile propagation and information retrieval techniques. The experimental results of the PRIOR at OAEI campaign 2006 are promising and appear to be scalable.

Future work includes: 1. Using interactive activation model [12][16] to find a state that satisfies domain constraints as much as possible to solve the problem of false positive. 2. Integrating external resources (e.g. WordNet) to distinguish synonym words. 3. Applying supervised machine learning methods to automatically optimize parameters when training data sets are available.

7. References

[1] Doan, A., P. Domingos, et al. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. SIGMOD Conference. 2001.

[2] Euzenat, J et al. Results of the Ontology Alignment Evaluation Initiative 2006. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA. 2006.

[3] Felzenszwalb, P. F. and Huttenlocher, D. P. Efficient belief propagation for early vision. International Journal of Computer Vision, Vol. 70, No. 1. 2006.

[4] Fensel, D. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Springer-Verlag New York, Inc. 2003.

[5] Hovy, E. 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. 1998.

[6] Hu, W., Cheng, G. et al. The result of Falcon-AO in the OAEI 2006 campaign. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA. 2006.

[7] Kalfoglou, Y. and M. Schorlemmer (2003). "Ontology mapping: the state of the art." The Knowledge Engineering Review **18**(1): 1-31.

[8] Kuhn, H. W. "The Hungarian method for the assignment problem", *Naval Research Logistics Quarterly*, **2**:83–87. 1955.

[9] Li, Y., Li, J., et. al. Result of Ontology Alignment with RiMOM at OAEI'06. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA. 2006.

[10] Mao, M. and Peng, Y. PRIOR System: Results for OAEI 2006. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA. 2006.

[11] Melnik, S., H. Garcia-Molina, et al. Similarity flooding: a versatile graph matching algorithm and its application to schema matching. Proc. 18th International Conference on Data Engineering (ICDE). 2002.

[12] McClelland, J. L. and Rumelhart, D. E. Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises. The MIT Press. 1988.

[13] Noy, N. "Semantic Integration: A Survey Of Ontology-Based Approaches." SIGMOD Record **33**(4): 65-70. 2004.

[14] Qu, Y., Hu, W., and Cheng, G. Constructing virtual documents for ontology matching. In Proceedings of the 15th International Conference on World Wide Web. 2006.

[15] Rahm, E. and P. Bernstein. A survey of approaches to automatic schema matching. The VLDB Journal 10(4): 334-350. 2001.

[16] Tsang, E. Foundations of Constraint Satisfaction: Academic Press.1993.

[17] Zhang, S. and Bodenreider O. (2007). Lessons learned from cross-validating alignments between large anatomical ontologies. In Proceedings of 12th World Congress on Medical Informatics. Brisbane, Australia.