Ontology Alignment Using Multiple Contexts

Jeffrey Partyka[1], Neda Alipanah[1], Latifur Khan[1], Bhavani Thuraisingham[1], Shashi Shekhar[2]

Department of Computer Science

University of Texas at Dallas[1]

University of Minnesota[2]

{jlp072000, na061000, lkhan, Bhavani.thuraisingham}@utdallas.edu[1]

shekhar@cs.umn.edu[2]

ABSTRACT

Ontology alignment involves determining the semantic heterogeneity between two or more domain specifications by considering their associated concepts. Our approach considers name, structural and content matching techniques for aligning ontologies. After comparing the ontologies using concept names, we examine the instance data of the compared concepts and perform content matching using value types based on N-grams and Entropy Based Distribution (EBD). Although these approaches are generally sufficient, additional methods may be required. Subsequently, we compare the structural characteristics between concepts using Expectation-Maximization (EM). To illustrate our approach, we conducted experiments using authentic geographic information systems (GIS) data and generate results which clearly demonstrate the utility of the algorithms while emphasizing the contribution of structural matching.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *semantic networks*, *representations (procedural and rule-based)*

General Terms

Algorithms, Measurement, Design, Reliability, Experimentation, Human Factors

Keywords

Ontology, Ontology Alignment, Schema Matching, Geographic Information Systems, Dataset

1. INTRODUCTION

Ontology alignment is the most recent incarnation of the information integration problem. A popular definition of an ontology is that of a "formal, explicit specification of a shared conceptualization", proposed by Gruber. In practice, ontologies for a given domain consist of a series of classes (or concepts) along with their properties, restrictions and instances, many of which are related by various types of relationships. The alignment of ontologies, therefore, entails deriving correspondences between concepts and their associated properties and instances.

2. PROBLEM STATEMENT AND PROPOSAL

Given 2 data sources, S_1 and S_2, each of which is represented by ontologies O_1 and O_2, the goal is to find similar concepts between O_1 and O_2 by examining their names, respective instances and structural properties. Let us assume that O_1 and O_2 are derived from the GIS domain.

The challenge involved in the alignment of these ontologies, assuming that they have already been constructed, is based on the derivation of procedures that will maximize the semantic similarity between any two concepts between the ontologies.

The ontology matching process consists of the matching of names, content and structure between compared concepts. The name match attempts to determine the degree of synonymy between the concept names. The content match determines similarity between the instances of each concept by measuring their mutual information, and it accomplishes this by the extraction of N-grams from the compared columns. The structural match determines similarity by leveraging the EM algorithm and the respective neighborhoods of all concepts to determine the most likely correspondences that occur between the ontologies. The overall similarity between two concepts is an equally weighted normalized sum of the name similarity, content similarity and structural similarity.

3. ONTOLOGY MATCHING ALGORITHM

3.1 Name Similarity

The first part of our approach attempts match concepts between two ontologies by measuring similarities between their names. The process consists of three steps. First, we check to see if an exact match exists between the compared concepts. If so, then a value of 1.0 is assigned to the name matching component of the overall similarity. If not, then we proceed with verifying whether the compared concept names are synonyms. To do this, an external dictionary such as WordNet is used to compute a semantic similarity score of the names between 0 and 1. If the words have any relation whatsoever, the semantic score returned by WordNet will represent the name matching component of the overall similarity. If there is no relation at all between the words, then the name similarity between the concepts is determined via the Jaro-Winkler string similarity metric.

3.2 Content Similarity

Content matching is accomplished by extracting instance values from the compared attributes, subsequently extracting a characteristic set of N-grams from these instances, and finally comparing the respective N-grams for each attribute. An N-gram is simply a substring of length N consisting of contiguous characters. For our experiments, the value of N was set equal to 2. The measure that was used to quantify similarity between compared attributes is known as Entropy Based Distribution (EBD), and it takes the following form:

$$EBD = \frac{H(C|T)}{H(C)}$$

In this equation, C and T are random variables where C indicates the union of the column types C_1 and C_2 involved in the comparison and T indicates the value type (2-gram for an instance value). EBD is a normalized value from 0 to 1, where 0 indicates no similarity between compared attributes, and 1 indicates that the attributes are identical. In our experiments, $C = C_1 \cup C_2$. H(C) represents the entropy of a set of instance values for a particular attribute (or column) while H(C|T) indicates the conditional entropy of a set of instance values for a particular value type.

3.3 Structural Similarity

In many situations, name and content matching are insufficient for reducing semantic heterogeneity during ontology alignment. As a result, our approach also attempts to match concepts by considering their surrounding structural characteristics. Specifically, we leverage the Expectation-Maximization algorithm to generate a mathematical model which indicates the most likely set of correspondences between concepts of O_1 and concepts of O_2. We compare all neighbors of a concept C_1 from O_1 and compare against all neighbors of a concept C_2 from O_2 to yield the structural similarity between C_1 and C_2. In adopting this algorithm, we decided to treat the concepts of each ontology as observable values while designating the set of correspondences between concepts in O_1 and O_2 as hidden values. Next, we decided that our mathematical model should be a mixture model represented by a similarity matrix SM consisting of $|O_1|$ rows and $|O_2|$ columns, where each individual entry represents an individual component of the mixture. Each entry indicates with a particular confidence value between 0 and 1 (for practical purposes, a probability value) whether or not a correspondence exists between a concept from O_1 and a concept from O_2. If a correspondence is indicated, then the entry has a value of 1, otherwise, the value is 0.

4. EXPERIMENTS

4.1 Datasets

Because data from several different areas of the United States were employed in our experiments, we effectively created a multi-jurisdictional GIS environment. GIS data assigned to concepts for O_1 is disjoint with the data assigned to the concepts for O_2. The number of instances is as low as 24 (Ferry) and as high as 91059 (Junction and Intersection). Meanwhile, the number of attributes is as low as 3 (Ferry) and as high as 26 (Enclosed Traffic Area), and the geographic scope ranges from a particular city (ie. Dallas) to an entire state (Virginia).

4.2 Results

Table 1 below shows the results of concept matching between O_1 and O_2 using name similarity, content similarity, and structural similarity via EM.

Table 1. Name + Content + Structure Similarity between concepts of O_1 and O_2

Ontology O_1 \ Ontology O_2	Intersection	Road	Ferry	Address Area	Enclosed Traffic Area
Road	.12	.78	.13	0.00	.06
Ferry	.10	.19	.76	.06	.11
Junction	.38	.01	.04	.04	.04
Traffic Circle	.40	.48	0.00	.11	.06
Resident-ial Area	.13	.21	.06	.60	.27
Traffic Area	.14	.25	.09	.12	.44

All of the correct correspondences between concepts of O_1 and O_2 are identified by a wide margin. Name similarity makes its strongest contribution to the accuracy of the algorithm regarding obvious correspondences such as Road-Road and Ferry-Ferry while failing to match correspondences such as Residential Area-Address Area and Junction-Intersection whose names are not similar. On the other hand, content similarity solves many of these problems by matching common N-grams existing among the instances of these concepts. While many of the correspondences are identified by name and content similarity, some, such as Traffic Circle-Intersection, remain unidentified, and others, such as Residential Area-Address Area are identified only weakly. To alleviate these problems, structure level matching via EM was applied. After doing this, correspondences that should be strong between concepts such as Residential Area-Address Area are associated with proportionally higher scores. Even in the situation where there does not exist a single correspondence that is significantly stronger than another, the composite algorithm captures the semantics appropriately. This occurs for the correspondences between Traffic Circle-Intersection and Traffic Circle-Road. Since a Traffic Circle is both a Road and an Intersection, the fact that the correspondence values are similar verifies the accuracy of our approach.

5. CONCLUSION

In this paper, we have outlined an algorithm that aligns two separate ontologies from the GIS domain using name similarity, content similarity and structural similarity. We focused on the structural similarity algorithm, which exploits EM to help determine the set of correspondences between concepts of two different ontologies. In regards to future efforts, we will expand our structure-level matching techniques to more accurately and thoroughly examine concept similarity. We will also analyze some of the more traditional techniques, such as sibling relationship similarity, and analyze its effects.