SEMANTICS-BASED SIMILARITY DECISIONS FOR ONTOLOGIES

Anne Yun-An Chen, Dennis McLeod
University of Southern California, 941 W. 37th Place, Los Angeles, CA 90089-0781, U.S.A.
Email: yunanche@usc.edu, mcleod@usc.edu

Keywords: Data mining, Ontology, Semantics, Similarity decision

Abstract: Many data representation structures, such as web site categories and domain ontologies, have been established for semantic-based information search and retrieval on the web. These structures consist of concepts and their interrelationships. Approaches to determine the similarity in semantics among concepts in data representation structures have been developed in order to facilitate information retrieval and recommendation processes. Some approaches are only suitable for similarity computations in pure tree structures. Other approaches designed for the Directed Acyclic Graph structures yield high computational complexity for online similarity decisions. In order to provide efficient similarity computations for data representation structures, we propose a geometry-based solution. Similarity computations are based on geometric properties. The similarity model is based on the proposed geometry-based solution, and the online similarity computation is performed in a constant time.

1 INTRODUCTION

Data representation structures have been developed to support online information search and retrieval. Interconnections of concepts define the relationships among the concepts, and hierarchical structures are commonly employed data representation structures. Examples of hierarchical structures are web site directories, subject categories, web page links, and some general or domain ontologies (Phillipi, 2004) (De Lazzari, 2003). There exits ontologies that are represented by Directed Acyclic Graph (DAG) structures. Ontologies contain sufficient information to facilitate information retrieval processes in order to match user expectations of retrieved results (Latifur, 2004).

Traversing DAG structures for similarity decisions is complicated since there may be more than one possible path from one concept node to another. Several possible traversing paths indicate the ambiguity of the similarity decision. Similarity computations in DAG structures have been studied in the filed of knowledge discovery in databases (KDD) (Shekar, 2002).The relatedness of two items is calculated by traversing all possible paths. Traversing all possible paths costs O(|E|), |E| is the number of edges. The maximum number of |E| is $[(n-1)(n-2)/2]$, where n is the number of nodes in the structure. The computational complexity can be expressed as O(n^2) for online computations.

A geometry-based solution is proposed to provide systematic similarity computations, uncomplicated online similarity decisions, and data representation structure configurations. The similarity is determined within the data representation structure, and is decided with the consideration of the direct inheriting relationship quantity. The solution is also suitable for DAG structures with simple adjustments. The proposed solution enables the utilization of current data representation structures. If the similarity computation is required to be performed online to support the information search or recommendation, the data adaptation and the similarity model construction is performed offline. The online similarity computation cost O(c), c is a constant.

2 GEOMETRIC-BASED DATA ADAPTATION

Geometry enables the study of properties of elements that remain invariant under specific transformations. In the proposed geometric-based solution, vertices are represented by points, and edges are represented by vectors in geometric space. The similarity can be decided based on the geometric properties of coordination.

2.1 Data Adaptation

The data adaptation manipulates data representation structures with defined geometric properties in a 3-dimentional space. The assumption here is that the structure for the similarity computations only has one root node. If DAG structures are involved in the data adaptation process, a virtual parent node of the root nodes will be inserted before the data adaptation begins. The proposed algorithm shown below performs the data adaptation. The input is G {V, E}, where V are nodes (vertices) and E are edges. The outputs are the coordinates of points and vectors between points.

```
while(not all edges are traversed)
{
     Get the next available node;
     If (current node is not marked as
     VISITED)
     {
       Accumulate the minimum number of
       edges to reach the root node as
       ID;
       if(the plane of X=ID does not
       exist)
         Create a plane X=ID;
       //x representing the id value,
       //y representing the next
       //available incremental value,
       //and z representing 1 larger
       //number from the maximum z
       //value of the previous plane.
       Assign the x, y, z value of the
       coordination to the current node;
       Define the vector between the
       node and its parent node;
       Mark the current node as VISITED;
       for (each plane X=value, value<x)
       {
         //x= value,
         //y=the y value(s) of the
         //parent(s)
         //z= current z value.
         Assign the   mapping
         coordination x, y, z;
       }
     }
     else
     {
         if(the coordination has been
         adjusted before)
           Replace the adjusted
           coordination to the original
           coordination;
         Calculate the difference
         between the z-axis values of
         two parent nodes of the
         visited node;
         while(From the parent node
         with larger z-axis value Z,
         not reach the node has the
         same z-axis value Z' as the
         other parent node)
         {
             Change the z value of
             mapping coordination for
             each reached node of the
             same y-axis value and its
             descendant to Z'+[(z-
             Z')/(difference + 1)];
             for (each plane X=value,
             value<x value of the new
             parent)
             {
                 //x= value,
                 //y=the y value(s) of
                 //the parent(s)
                 //z= current z value.
                   Assign the
                   mapping
                   coordination x, y,
                   z;
             }
         }
     }
}
```

The data adaptation takes $c \times n^2$ iterations, where c is a constant and n is the total number of nodes. The scale of the possible total number of objects in the queue is $O(n^2)$, the number of edges minus the number of the root nodes. The plane creation in the 3-dimensional space costs O(max (d')), where d' is the minimum number of the edges to reach the root node for the nodes in the structure. For a single-rooted structure, the maximum number of operations for the coordinate adjustment is the total number of all possible edges. Again, the number is bounded by n^2. The final computational complexity for the algorithm is $O(n^2)$.

3 SEMANTIC-BASED SIMILARITY MODEL

The proposed semantic-based similarity model introduces semantic-based operations to provide answers to similarity

decision problems. In this similarity model, three operations are included. The first operation is data adaptation, which is described in detail in the previous section. The second operation, the semantic-based grouping, provides the foundation of efficient online similarity decision processes. The third operation is to decide the semantically similar groups and their priorities. In the following sections, an approach using similarity decisions based on fundamental properties of the geometry embedded in the data representation structures is proposed.

3.1 Semantically Similar

Before the approach of similarity decisions is introduced, definitions of similarity degrees must first be declared. First, a broad definition of the similarity in semantics is declared below.

Definition 1: If two concepts as ending points share the same starting point of the vector, the two concepts are hierarchically similar.

If two concept nodes share the same starting point of their vectors, the two nodes must have at least one common parent node in the domain ontology before the adaptation. Now, a narrow definition of the similarity in semantics is declared.

Definition 2. If two concepts are hierarchically similar and the z-axis values are the same, the two concepts are semantically similar.

Having the same parent node does not imply these two child nodes must have the same generality. The reason is that one node may have more than one parent node, and these parent nodes do not always have the same generality.

3.2 Semantic-based Grouping

Semantic-based grouping is introduced to utilize the results of the geometry-based data adaptation and to facilitate the online recommendation processes. Semantic-based grouping means that groups which contain semantically similar concepts are determined based on the data adaptation results offline. Grouping performed offline decreases the computational complexity of the online similarity decision making. Concepts that are semantically similar are labelled as one group. Priorities are assigned to each group according to the following definition.

Definition 3: If any two concepts share *n* same parent nodes, the priority of the two concepts is n and higher than two concepts that share (*n-1*) parent nodes.

3.3 Online Similarity Decisions

A similarity decision process framework based on the proposed similarity model consists of three gradational approaches, locating semantically similar concept groups, selecting candidates of recommended concepts, and deciding the recommended concept(s). Semantically similar groups containing the concepts being queried are first located. Offline sorting, based the priorities of groups that have one common concept, enables the online group locating to be completed in a constant time of O(C_{top}), where C_{top} is the number of top priority groups needed for deciding the recommended concepts. The value of C_{top} is decided by the information system developers.

After all semantically similar groups are located, concepts in these groups excluding the queried concept are considered as the candidates of the recommended concepts. The candidate(s) with the highest priority will be selected to be recommended. The goal of the selecting candidate approach is to obtain the concept nodes with large similarity degrees. A number C_{limit}, where C_{limit} is a constant, is set to limit the number of selecting. In total, the computational complexity of the online similarity decision is O(*C*), where *C* is a constant.

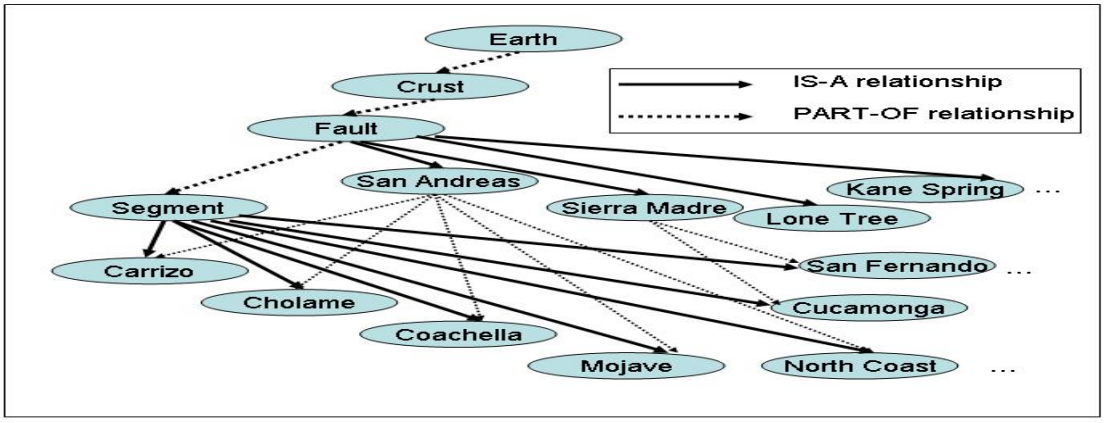

Figure 1: Partial Earthquake Science Domain Ontology

4 APPLICATIONS ON ONTOLOGY

4.1 Earthquake Domain Ontology

In order to access the tremendous amount of heterogeneous geoscience data, a semantic metadata management system and wrappers for web services are required. The domain ontology is the core of the metadata management system (Chen, 2003) and is illustrated in Figure 1. In the following sections, we demonstrate an example of the similarity model application.

4.1.1 Application of the Similarity Model

The proposed approach to determine similarity decisions consists of three elements, the geometric-based data adaptation, the geometric-based similarity definition, and the semantic-based similarity model. The results of the approach are listed in Table 1.

Table 1: Results of grouping and priority defining.

Group of Concepts	Priority
Earth	1
Crust	1
Fault	1
Segment, San Andreas, Sierra Madre, Lone Tree, Kane Spring	1
Carrizo, Cholame, Coachella, Mojave, North Coast, Cucamonga, San Fernando	1
Carrizo, Cholame, Coachella, Mojave, North Coast	2
Cucamonga, San Fernando	2

4.1.2 Recommendation Processes

The recommended concepts associated with queried concepts and decided by the proposed similarity model are demonstrated in the following case.

Case 1: If the queried concept is Cucamonga, the recommended concept is San Fernando. It is because the group these concepts belong to has a higher priority. The assigned priority is 2, and the value of C_{top} here is set to 1.The other group containing the concept Cucamonga only has the priority of 1.

5 CONCLUSION

A semantic-based similarity model is proposed to solve similarity decision problems in data representation structures. The goal of the model development is to perform similarity computations in spontaneous and unambiguous similarity decisions. The data adaptation process is developed to utilize geometric properties. Based on the results of the data adaptation process, the similarity degree is decided by geometric properties. The semantic-based similarity decision model consists of offline computations and online operations. The offline computations include semantically similar grouping for concept nodes and priority computations for semantically similar groups. Performing grouping and computing priorities offline enables the reduction in the computational complexity of online similarity decisions. Online similarity decisions are completed by a sequence of three approaches: locating semantically similar concept groups, selecting candidates of recommended concepts, and deciding the recommended concept(s). The proposed similarity model serves as a good foundation for recommendation processes due to the combination of uncomplicated approaches and results in constant-timed computations.

REFERENCES

Chen, A. Y., Chung, S., Gao, S., McLeod, D., Donnellan, A., Parker, J., Fox, G., Pierce, M., Gould, M., Grant, L., & Rundle, J. (2003). Interoperability and semantics for heterogeneous earthquake science data. Published paper presented to *Semantic Web Technologies for Searching and Retrieving Scientific Data Workshop, Sanibel Island, Florida*.

De Lazzari, C., Guerrieri, E., Pisanelli, D.M., & Murray, A. (2003). A domain ontology for mechanical circulatory support systems **.** *Computers in Cardiology (IEEE Cat. No.03CH37504)*. IEEE Press. xxvii+829, 417-19.

Khan, L., McLeod, D., & Hovy, E.H. (2004). Retrieval effectiveness of an ontology-based model for information selection. *The VLDB Journal.* 13(1), 71-85.

Philippi, S., & Kohler, J. (2004). Using XML technology for the ontology-based semantic integration of life science databases . *IEEE Transactions on Information Technology in Biomedicine*. (IEEE)8, no. 2, 154-60.

Shekar, B., Natarajan, R. (2002). A fuzzy-graph-based approach to the determination of `interestingness' of association rules. *Lecture Notes in Artificial Intelligence Vol.2569*. Berlin, Germany : Springer-Verlag. xiii+648, 377-88.

Acknowledgement This work was supported by NASA's Computational Technologies Project. Portions of this work were carried out by the Jet Propulsion Laboratory, California Institute of Technology under contract with NASA.