Ontology Driven Content Mining for Semantic Queries in Satellite Image Databases

Adrian S. Barb and Chi-Ren Shyu
University of Missouri

Extracting knowledge from image databases is challenging and requires a deep understanding of domain-specific knowledge. This is especially true in the geospatial domain where the huge amounts of generated satellite imagery make knowledge discovery computationally expensive. While existing content-based knowledge representation approaches have resulted in varying degrees of success, they normally rely on human inputs. Moreover, insufficient training data may lead to gaps between computer and human models in describing visual patterns. To tackle these issues, we incorporate other available domain knowledge resources into the discovery process, namely domain ontology, to computationally reduce the semantic gaps and improve the precision of retrieval results.

Ontology is a valuable resource that captures declarative knowledge from the perspective of the domain [Gruber]. A fragment of an ontology for the geospatial domain is depicted in Figure 1. In this figure, each ontological concept is depicted by a rectangle with relationships between concepts depicted by arrows. There are two types of relationships in this figure: “is a” relationships (solid lines) and “part_of” relationships (dotted lines). Typical association rule mining systems try to mine rules for each concept individually and seldom make use of the relationships between concepts in an ontology. This may be sufficient for datasets with balanced distributions of semantics. However, it is well known that association rule mining performs poor in the case of unbalanced distributions [Jovanoski 2001]. Unbalanced class distribution affects classification results in two ways: (1) minority classes are ignored since generated itemsets from these classes rarely pass the support threshold, and (2) classification models tend to be overfitted with respect to the training samples resulting in high error rates for newly discovered cases. The example in Figure 1 shows the number of training images that were labeled for semantics in parentheses. Because less than one percent of the satellite images in our training set contain highway, we can say that highway is an under-represented semantics, and thus it is unlikely that association rule mining will yield any rules for highway because they may not pass the support threshold level.

One of the most common techniques for dealing with underrepresented class data is resampling, which minimizes class unbalance by altering the distribution of training examples. Resampling algorithms include undersampling [Kubat], which discards samples from the majority classes, and oversampling [Chawla], which adds synthetic data to minority classes. Although resampling methods have been proven to increase classification accuracy, they also have several drawbacks. Undersampling may degrade system performance by potentially discarding useful majority-class examples, while oversampling may lead to overfitting and lower time efficiency [Chawla]. These issues become more evident when trying to model visual patterns in satellite image databases due to the fact that several visual patterns (semantics) may coexist in the same image. For example if an image analyst wants to retrieve images that include highway, it is likely that the image also includes...
**grassland, farmland, or forest.** In the geospatial domain, oversampling is preferred over under-sampling due to the high cost of expert-in-the-loop processes.

We developed a new approach to include domain knowledge from an ontology in the mining process. More specifically, we use ontological relationships between geospatial concepts to oversample the under-represented semantics provided by experts for training. This is expected to improve the knowledge discovery process, reduce semantic gaps, as well as increase the accuracy of semantic queries. We do this by applying three steps: First, we embed the image feature space into a higher-dimensional hyperspace that includes information in the ontology \( O = \{ C, R \} \), where \( C \) is the set of concepts defined by \( O \) and \( R \) is the set of relationships between concepts in \( O \). The ontological information is modeled by a feature space in which each relationship \( r \in R \) is represented by a feature \( f_r \). The value of the feature \( f_r \) for an ontology concept \( c \in C \) is given by Equation 1. In this equation, \( w \) is a weight that depends on the characteristics of the dataset, \( \text{depth}(r) \) is the depth of the relationship \( r \) in \( O \), and \( \text{root}(O) \) is the root concept in \( O \). Second, we add synthetic data to an under-represented class by applying an algorithm derived from SMOTE [Chawla] to the newly-created hyperspace. According to this algorithm, synthetic examples are randomly generated on the lines between two neighboring points regardless of their class assignment. After oversampling, the original feature space is used for association rule mining.

\[
\phi(f_r, c) = \begin{cases} 
  \frac{w}{\text{depth}(r)} & \text{if } r \in \text{path}(c, \text{root}(O)) \\
  0 & \text{else}
\end{cases}
\]  

(1)

To evaluate our approach, we apply it to a computer model that maps visual patterns to feature spaces using association rules [Shyu, et al]. In this model, each rule has two parts: the antecedent, which models low-level feature subspaces, and the consequent, which represents a geospatial semantic.

In conclusion, using only data mining techniques for knowledge representation might lead to overfitted models, especially when the training set has an unbalanced class distribution. Our approach uses declarative knowledge from a domain ontology to provide a better description of geospatial visual patterns that can lead to better precision of semantic queries. This project was supported in part by the National Geospatial-Intelligence Agency University Research Initiatives (NURI) under grant number HM1582-04-1-2028.

**Bibliography**


P. Domingos, Toward knowledge-rich data mining. Data Mining and Knowledge Discovery, 15, 1 (2007), 21-28


C.-R. Shyu, A. Barb, and C. Davis. Mining image content associations for visual semantic modeling in geospatial information retrieval and indexing, in Proc. of International Geoscience and Remote Sensing Symposium, No. 5622, July 2005