Expertise Evaluation using PathFinder Networks
Scaling in Ranking of Satellite Images

Adrian S. Barb
Information Science Department, Penn State Great Valley
Malvern, PA 19355
Email: adrian@psu.edu

ABSTRACT

In this article we propose a methodology to evaluate the level of expertise of image analysts when searching domain-specific images by semantics. We apply our methodology to ranking high-resolution satellite images by semantics. Our methodology applies PathFinder Network Scaling methods to create concept maps for representing associations of semantics to regions of a feature space for each image analyst. The relevance of each node in a concept map is evaluated using a hits authority algorithm. The expertise of each image analyst is then evaluated by comparing to ground truth models using the Kendall tau rank correlation coefficient. Our system allows us to identify areas of expert disagreement by evaluating the relative difference of individual models placed on features as well as recommend areas of that needs to be stressed by novice image analysts.

Keywords: data mining, Pathfinder Network, expertise, content-based image retrieval, ranking, geospatial images.

1 INTRODUCTION

Supervised learning for domain-specific image ranking is a very important research area of machine learning. A critical assumption is that supervised training methods can guaranty successful exploitation of existing information to generate useful patterns that can generalize well when applied to new, not yet evaluated data sets. Part of such assumption is that the training data is covering the semantic domain at a high level of accuracy. Consequently, supervised methods rely heavily on the quality of semantic metadata as assigned by image analysts. Previous research [3] shows that such ranking models are likely to overfit [11] fact that may render the models as unfit to predict new, unlabeled data. A diverse range of researches [5, 7, 10, 21, 22] address the issue of ranking quality by providing several data mining algorithms, to represent the complex domain knowledge found in images. Among such methods, ensembles [17, 19, 24] address overfitting by building multiple models on subsets of the training data and combining the results to produce a more accurate predictive model. However these methods are highly dependent on the quality of manually assigned semantic labels by image analysts.

Attaching semantic information to domain-specific images, such as geospatial, is difficult due to complex visual patterns and to the coexistence of visual patterns related to multiple semantics in each image. For example, Figure 1 shows some examples of geospatial images that contain visual patterns related to multiple semantics of interest. For example image tile in Figure 1(a) shows an example of a barren area which occupy more than 50% of the tile. While analysis would have little issues in assigning this barren as semantic metadata to this image, they would have a more difficult choice when evaluating the existence of urban areas or wetlands which occupy less than 12% of the image tile. Similar patterns are shown in Figure 1(b) and Figure 1(c) where grassland and shrubland may render subjective interpretations by coexisting with urban area and forest respectively. In this context, it is important to recognize and evaluate the degree of expertise of individuals that contribute to a community that evaluates images by semantics.

Several data reduction algorithms [14, 23, 12] can be applied to reduce dimensionality and correlation of data, as well as to maximize data information and increase understandability of the generated models [8]. Pathfinder Network analysis [15, 20] is a methodology that evaluates models’ structure by comparing the similarity of model constituents using co-occurrence measures. It can be applied for elicitation and analysis of knowledge structures in raw data. Each model is considered a network where concepts are represented as nodes connected by their strongest links to proximal/related concepts. The concept of Pathfinder Network analysis can be extended to image analysis where feature regions are concepts that explain a visual pattern. They have the advantage of better representing the localized rather than global characteristics of the model structure [6].

In this paper we explore methods for evaluating expertise levels of image analysts using PathFinder Network Scaling methods in conjunction with additive associative models for ranking domain-specific images by semantics. We have chosen associative ranking due to their white-box approach that allows us to pinpoint to the main areas of disagreement among image analysts. For each image analyst we generate cross-folder associative ranking models [2], which are averaged for better representation of experts’ models, reduced with a PathFinder network...
algorithm, and pruned of irrelevant association. This paper is organized as follows: In Section II we present the methodology used to evaluate the level of expertise, in Section III we perform experimental evaluation, and then in Section IV we conclude the article with future work.

Figure 1. Geospatial image tiles with multiple possible semantic assignments.

2 METHODOLOGY

In this section we present our methodology for evaluating image analysts’ expertise when ranking satellite image tile by semantic. The key feature of our methodology is the construction of additive, associative semantic models by choosing feature spaces that maximize area under the curve (AUC) value on the training set. Features are grouped in categories of interest such as color, object shape, etc. so that each association mined contain only features from same category. Once a combination of features is selected, we model the assignment using sigmoid possibilistic functions. Further, sets of feature spaces are used additively to model the correlation to a semantic. To evaluate which subspace is the most relevant we also apply genetic operations at this level.

Semantic Model for Image Ranking

Each semantic model is a set of associations between feature regions and a semantic of interest that are optimized to maximize the AUC value. For example, Figure 2 shows a mapping of a semantic into a two-dimensional feature space formed from two object orientation features. When a new image is presented to the system, the features are computed and the relevance of the image to the semantic of interest is computed using the procedure explained below. The extracted features are grouped into natural categories $C$ such as, object orientation, color, or texture. Each association rule maps region of the feature space $\Theta \subset C \subset F$ into the semantic of interest $\varsigma \in S$.

$$\Theta(m_\varphi | C) = \{g(m_\varphi)\} \text{ where } m_\varphi \in C$$ (1)

$$g(m_\varphi) = \frac{w_g}{2} \left[ \tanh \left( \frac{m_\varphi - \lambda_l}{\lambda_l} \right) - \tanh \left( \frac{m_\varphi - \lambda_r}{\lambda_r} \right) \right]$$ (2)

The function $g$ is an asymmetric double sigmoid possibilistic distribution ($L$ - left and $R$ - right) that models the relevance of a measurement $m_\varphi$ to a semantic $\varsigma$. Each half sigmoid is controlled by two parameters: (a) center ($\lambda_l$) and (b) width ($\lambda_r$) while $w_g$ is weight of the relevance retrieved by the $g$ and it is shaped using the relevance assessments provided by image analysts for the semantic of interest. For details of this mapping function, the reader is referred to [5]. The relevance of an image $i$ to an association rule $\Theta$ over a semantic $\varsigma$ is determined by the relevance of the feature values $m_\varphi(i)$ of the image over region of the feature space $\Theta$.

$$T(i|\Theta, \varsigma) = \min \left( g \left( m_\varphi(i) \right) \right) \mid g \left( m_\varphi(i) \right) \in \Theta$$ (3)

Further, for each semantic $\varsigma$ we create a semantic model $M_\varsigma = \{\Theta \rightarrow \varsigma\}$ defined as the set of mappings of subspaces $\Theta$ of $F$ into a semantic space $S$. The overall relevance $T(i|M_\varsigma)$ of an image $i$ with feature measure $m_\varphi(i)$, to a semantic $\varsigma$ is computed by using a weighted average of relevance to each feature subspace $\Theta \in M_\varsigma$:

$$T(i|M_\varsigma) = \sum_{m \in |M_\varsigma|} w_m T(i|\Theta, \varsigma)$$ (4)

Finally, results are sorted in descending order of $T(i|M_\varsigma)$ and AUC is computed as an aggregate measure of
ranking across all the recall levels for each model $M_\zeta$ for $\zeta \in S$ over a feature space $F$.

Figure 2. Example of semantic model for a semantic of interest

Semantic Model Generation

Although associative methods have the advantage in better interpretation, they are difficult to train and optimize due to the fact that the subspace generation is exponential to the number of possible subspaces. This makes brute force associative methods NP-hard [13]. To address this issue, we use a genetic algorithm [2] to reduce the complexity of the problem. Although genetic algorithms are greedy, they have the advantage of avoiding local maximum traps. The key feature of our genetic algorithm is the usage of genetic operations at both feature and subspace levels. At the feature level, we vary the set of features used to mine association rules while at the subspace level we vary the region for the same feature set that will be used in ranking. Using genetic operations we randomly choose and evolve combinations of feature using methods such as crossovers, shrink, constant, or grow mutations. Once a combination of feature is selected, we randomly generate and evolve features subspaces modeled by sigmoid possibilistic functions. Further, sets of feature spaces are used additively to model correlation to a semantic of interest. To evaluate which subspace is the most relevant we also apply genetic operations at this level. For a more in-depth description of the training algorithm, the reader is directed to [2].

PathFinder Network Analysis

Once a semantic model $M_\zeta$ is generated, we can convert it to a concept map that represents the knowledge accumulated in the semantic metadata attached to each training image and derived by mining associations between features $f \in F$ and a semantic $\zeta \in S$. For more accurate results, we construct multiple such concept maps using cross-fold methods which are ensembled into one average model. Such ensemble conceptual maps are better fit for comparing and contrasting semantic models from different image analysts. To achieve a final concept map for a semantic-analyst pair, we apply a PathFinder Network Analysis algorithm to simplify and evaluate each semantic model. The generated conceptual map contains feature categories as nodes and proximal relationships as vertices. This article extends the research in [4, 9] by creating concept map representations of semantic models for each image analyst with the goal of identifying similarities and differences to ground truth semantic information. We generate each concept map using a four steps algorithm: (1) co-occurrence matrices are constructed for each semantic-analyst pair using cross-fold experiments, (2) co-occurrence matrices are aggregated to generate a unique semantic-analyst matrix representation, (3) PathFinder Network Scaling are applied to the resulting network to elicit internal knowledge structures, and (4) irrelevant nodes of the PathFinder network are pruned using a threshold on node strength.

In the first step, the resulting feature subspaces are associated to feature categories as explained in Section 0. Then, the data mining algorithm in Section 0 is applied to cross-fold subsets of the training set. Each association in a model contains only features from one feature category $C$. The set of relevance measures $T(i|\Theta_\zeta)$ is computed for each image $i$ in the training which is ranked in a descending order. The co-occurrence matrix $CM$ is defined over the feature categories as a $|C||C|$ matrix and it is populated by incrementing each cell as follows: (1) for every $T(i|\Theta_\zeta)$, increment $CM[|C|C]|C \rightarrow \Theta$ by $T(i|\Theta_\zeta)/2$ and (2) for every sequence $T_1(i|\Theta_1, \zeta), T_2(i|\Theta_2, \zeta)$ increment $CM[|C|C], CM[|C|C]|C_1 \rightarrow \Theta_1, C_2 \rightarrow \Theta_2$ by a value of $(T_1(i|\Theta_1, \zeta), + T_2(i|\Theta_2, \zeta))/2$. In the second step, we generate a unique cross-fold co-occurrence matrix by summation and normalization of all the fold-specific co-occurrence matrices. In step 3, we apply a PathFinder Network Scaling algorithm to the average network to reduce the irrelevant raw proximal vertices and convert the co-occurrence matrix into a least-weighted path of linked feature subspaces [9, 20]. Finally, in the last step, we compute the HITS authority score [16] of each node and we further simplify the PathFinder network by applying a threshold on authority of nodes at two standard deviations from the average. This network is then used to evaluate the expertise of an image analyst to ranking images by a semantic.
3 Evaluation

For our experiments we used the 2010 WROC satellite imagery data set of Wisconsin [1]. This data set contains 18”, 3-band GeoTIFF image tiles 15,678 x 11,105 pixels which was collected in spring 2010. Each of these images was partitioned into minimal overlapping 1000x1000 tiles. For each tile, a feature extraction algorithm was applied to include color, texture and object features. For each of these features average, quartile, standard deviation, skewness, and kurtosis were calculated resulting in a 292 feature vector for each tile. Each of these features were assigned exclusively to 26 categories based on their meaning. For example, color features were assigned to eight categories for the RGB, HSV, and gray models. Also, object features were assigned exclusively to 14 categories based on area, centroid, bounding box, major and minor axis, eccentricity, orientation, convex area, filled area, Euler number, equivalent diameter, solidity, extent, or perimeter. The rest of categories were for texture and object phase congruency.

Further, we selected a number of 1,124 tiles that were labeled by an expert with one or more labels from the Urban Area (L100), Shrubland (L250), and Other (L260). We will further refer to this expert as groundtruth. In this semantic assignment, 121 tiles were labeled Urban Area (L100), 128 were labeled Shrubland (L250), and 15 were labeled both. Further, 22 tiles were labeled with both Urban Area (L100) and Other (L260) while 8 were labeled with both Urban Area (L100) and Other (L260). Two other image analysts (expert1 and expert2), which are the subject of the experiment, performed the semantic assignment on the same image data set and semantic labels. For example, expert1 ranked 125 images labeled Urban Area (L100), 12 tiles Shrubland (L250), and one tile both. Similarly, expert2 ranked 123 images labeled Urban Area (L100), 118 tiles labeled Shrubland (L250), and nine tiles both. For each of the three image analysts, we trained semantic models on Urban Area (L100) and Shrubland (L250) using four-fold experiments. The information on each fold result was aggregated to construct average semantic models as described in Section 1.

![Figure 3. Aggregated model of semantic representation for the semantic Urban Area (L100) for the image analysts (a) groundtruth, (b) expert1, and (c) expert2.](image)

![Figure 4. Aggregated model of semantic representation for the semantic Shrubland (L250) for the image analysts (a) groundtruth, (b) expert1, and (c) expert2.](image)
Figure 3 shows the resulting PathFinder Networks for Urban Area (L100) semantic models generated by image analysts using the Network Workbench Tool [18]. In this figure the sized of each node in the network in proportional with the authority score generated by the HITS algorithm [16]. For example, in Figure 1(a) node C18, which correspond to features related to object area, has the most relevance to ranking images by this semantic. However, Figure 1(b) shows that expert1 places more emphasis on features related to the category of features related to object bounding box (C11). We can conclude that this expert emphasize more features related to object solidity (C19) than features in the C18 category, which were expected if this image analyst matched the groundtruth expert. Similarly, Figure 1(b) shows the PathFinder network for expert2, which is much more similar to the one of the groundtruth expert. In both these networks features related to object area (c18) were the most authoritative. Overall these networks have less visible differences. We can also see some difference in the number of feature categories used in modeling semantics. For example, the groundtruth image analyst uses 19 feature categories to assess the presence of visual patterns associated with this semantic, expert one uses 18, while expert2 only 15. For example, expert2 does not use color features from the R spectrum (C01) although these features are considered relatively important by the groundtruth expert. Figure 4 shows the generated conceptual maps for ranking images by the Shrubland (L250) semantic. As shown in this figure, there are wider variations in identifying the most relevant feature categories for labeling images by this semantic. Figure 4(a) shows that the groundtruth expert places more importance on color features from the HSV encoding model (C03 and C05) and on object eccentricity (C13). Object phase congruency features (C24) are used predominantly by the expert1 while color saturation features (C05) are used by expert2.

Table 1. Estimation of the level of expertise of image analysts groundtruth, expert1, and expert2 for evaluating images by the Urban Area (L100) and Shrubland (L250) using Kendal tau rank correlation coefficient.

<table>
<thead>
<tr>
<th>L100</th>
<th>groundTruth</th>
<th>expert1</th>
<th>expert2</th>
</tr>
</thead>
<tbody>
<tr>
<td>groundTruth</td>
<td>1</td>
<td>0.450735</td>
<td>0.733309</td>
</tr>
<tr>
<td>expert1</td>
<td>0.450735</td>
<td>1</td>
<td>0.438968</td>
</tr>
<tr>
<td>expert2</td>
<td>0.733309</td>
<td>0.438968</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>L250</th>
<th>groundTruth</th>
<th>expert1</th>
<th>expert2</th>
</tr>
</thead>
<tbody>
<tr>
<td>groundTruth</td>
<td>1</td>
<td>0.416096</td>
<td>0.528562</td>
</tr>
<tr>
<td>expert1</td>
<td>0.416096</td>
<td>1</td>
<td>0.261851</td>
</tr>
<tr>
<td>expert2</td>
<td>0.528562</td>
<td>0.261851</td>
<td>1</td>
</tr>
</tbody>
</table>

To evaluate the expertise level of the image analysts (expert1 and expert2) we compute the Kendal tau ranking correlation between the pairs HITS authority scores. Results for this experiment are shown in Table 1. For ranking images by Urban Area (L100), the correlation between groundtruth and expert1 is 0.45 while the correlation between groundtruth and expert2 is 0.73. Although the two values show positive correlation of the two experts to the groundtruth, we observe that expert2 has a higher degree of expertise than expert1 when evaluating for Urban Area (L100). However, when ranking images by Shrubland (L250), the correlation between groundtruth and the two experts is rather similar at 0.42 for expert1 and 0.52 for expert2. This shows that expertise needs to be evaluated on a semantic assignment basis and different experts can discern different visual patterns with different accuracies. The evaluation of expertise brings also the advantages of pointing out expert training needs. For example, upon reviewing expertise levels of individual image analysts, training specialists can further evaluate the structure of each PathFinder network and focus the training procedure on an area of interest.

4 Conclusions and Future Work

In this article, we have developed an approach to evaluate the level of expertise applied to image ranking by semantic of high resolution satellite images. We apply PathFinder network scaling methods to contrast and compare feature categories used by different image analysts when they assign semantic labels to domain-specific images. Our model has the advantage of being applicable in practice since it can suggest areas of high difference among experts’ perception of semantics, which can be used for training as well as for reaching agreement in organizations that perform such tasks. Our future work includes a more comprehensive evaluation on different domain-specific data and semantic sets as well as a more in-depth evaluation on accuracy of the resulting assessments. We also plan to evaluate methods of assessing the degree of truthfulness of ratings by ground truth image analysts.

Bibliography


