Ryan Cooper
SAND2016-0279PE
1/12/2016

Research Board Content (to be displayed on the research board)

Introduction

Template generation via data mining is an analytical process in which large amounts of data are analyzed against a known true positive data set to determine a variance to define or differentiate them from the rest of the data. This project arose from a necessity for optimization of a time-consuming process previously done by analysts. This particular template was being used for searches regarding large overhead images analyzed through a Geospatial-Temporal Semantic Graph (GTSG) format. The information for the GTSG is stored in a SQLite database; therefore, it can be queried using structured query language (SQL). The data mining utility used was the Waikato Environment for Knowledge Analysis (WEKA), an open source data mining utility that allows researchers “easy access to state-of-the-art techniques in machine learning” (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009). Through documentation provided by WEKA, SQLite databases were accessible and manipulable through WEKA, creating an opportunity to query and analyze directly from the GTSG. This project was inspired by and built off of prior research detailed by Brost et al. (2014) and was used to expand upon their research to increase the overall accuracy and efficiency of the template generation process, and uses this basis as a fundamental element on which the project was designed to improve.

Procedure

- The environment consisted of the two main parts: WEKA, a data mining utility, and GeoSearch, the program that outputted the data to be examined.
- A search used previously as detailed by Stracuzzi et al. (2015) involved finding high schools in Anne Arundel County.
    - This example served as a proof of concept for the idea that data mining was a viable method to generate templates for searches, and that it would inherently be more accurate and find more correlations when compared to human analysts.
    - Previously, this search had been used as a test for the search function and had been used to build quality score matches (Stracuzzi, Brost, Phillips, Robinson, Wilson, and Woodbridge, 2015), so it had an analyst's interpretation of the template already in place. This template had been established to the best accuracy they could determine which showed identification accuracy capable of limiting the 1.2 million nodes down to a discrete 67 potential results. Due to the existence of a template already in place, this gave way for possible practical improvements.
- The data collection and template generation process can be broken down into seven discrete steps:
    - First, an established baseline of land cover types would be applied to each sub-search in the process, in the high school search, a football field would be an example of a sub-search, and this is classified as a grass field type.

- Second, this baseline would run GeoSearch on the aforementioned data set, as detailed by Brost et al. (2014), resulting with the StoredGraph that contained various types of land covers and land cover data to be analyzed later – at this point it is important to note that no criterion have been applied to any empirical properties, simply a search by land type.
- Then the user would use Quantum GIS (QGIS), software that allows the visualization of the generated SearchGraph, to analyze and find the true positive subsets for each true positive; meaning that for one high school, the user would have to find all the corresponding sub-features.
- With all of these noted, the user would then input that information into a few different SQL queries, and then run the resultant data through WEKA.
  - This imputed data will have variant values, but the essential goal of WEKA is to define a variance for what can be considered a true positive while minimizing false positives.
  - WEKA is a GUI for various data mining algorithms, including supervised and unsupervised. This allows for the user to access powerful algorithms easily through WEKA's built in SQLite compatibility.
- Once in WEKA, the user could run the data through C4.5, and if it showed inconclusive results, and the ratio of the true positives to false positives is relatively small, then the user could apply a spread subsample and test again to lessen the ratio of the data.
  - C4.5 works by defining what the target variable is, in this case it would be true positive, then builds a decision tree to classify the target variable based on the other information provided.
- They would do this for each empirical property to discover true positive indicators.
- Once all empirical data points were finished, the user would compile these into the original GeoSearch format, and re-run to ensure the accuracy and conclude the results.

Abstract

Data mining plays a key role in search template generation for the analysis of large overhead image sets, particularly that of ontological storage, or geospatial-temporal semantic graph (GTSG). It provides an efficient method for determining the median of accuracy and consistency for template generation, one of which human analysts are required to provide substantial time and effort to create comparable results. The implementation of template generation is mostly autonomous and fairly straightforward when compared to current techniques. These templates are used in feature analysis of height and landform fused data, and allow the easy construction and analysis of any desired query. This process of template

generation has useful implications in a wide variety of fields, and can transform correlations of random data into insightful and useful information.

Hypothesis

The function of data mining in the case of overhead imagery analysis resides in the advanced search method, and specifically the function of composing templates that can be used on a broad scale, not just for one particular query. This relevance and advantage are due to the potential optimization available and definite efficiency benefits that will occur as a result.

Problem

The process of overhead imagery analysis is described as being a "key technology in commercial and national security" (Brost, McLendon, Parekh, Rintoul, Strip, & Woodbridge, 2014). They detailed a process where they begin by pre-processing large amounts of information through a primitive ontological storage, or geospatial-temporal semantic graph (GTSG). The information held in the GTSG shows relevant ontology through nodes and edges. These nodes show image preprocessing data (O'Neil-Dunne et al., 2013) and reveal things like composition and properties, as in, whether it is a field or a building. This information is classified and stored in the GTSG. The term properties, in this scope, can be defined as empirical data: area, height, perimeter, color, eccentricity, etc., and these properties can be queried to obtain relevant information regarding an analyst's request. The issue resides in this request, as the parameters are often undefined and ambiguous. This project's goal is to create an automated system that combats that inefficiency and is an improvement upon the existing manual method.

Data

The high school search-specific data being analyzed can broke down into six key elements: classroom building, parking lot, football field, tennis court, baseball field, and their relativity to each other. To get a visual representation for how these elements play a role in the determination of search results, see *Figure 1*. The data that each of these elements provided to the template are called the determinant criteria and were based on the land cover region labels that were assigned to the aforementioned image preprocessing data (O'Neil-Dunne et al., 2013); these include buildings, trees, grass/shrub, dirt, water, road, and other paved areas (Brost et al., 2014). These region nodes and distance edges represent the node's type or its composition in relativity to the proposed question. To give some quantitative information about this dataset to provide context, the Anne Arundel County set had generated a GTSG database with over 1.2 million node elements (Brost et al., 2014), all containing empirical property information for analysis, which correlates to a land area of just over 600 square miles.

The data was visualized and cross-referenced with Anne Arundel public school database to determine the results true location for back checking purposes. *Figure 2* gives a school configured in a hub-spoke configuration, giving in to how the relativity factor plays into the determination for criteria.

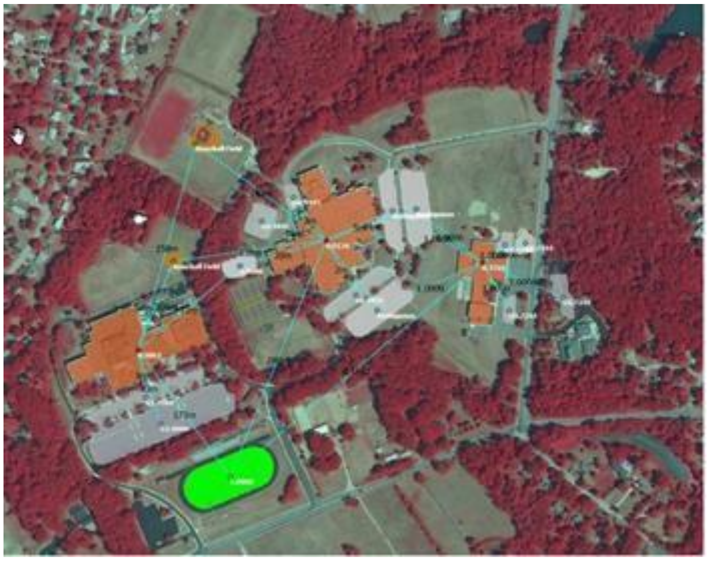

Figure 1. An example of outputs seen in the program Quantum GIS (QGIS), an open-source program that gives visualizations to the results of a specific GeoSearch query (Brost et al. 2014). This image shows a high school, middle school, and an elementary school, all side by side.

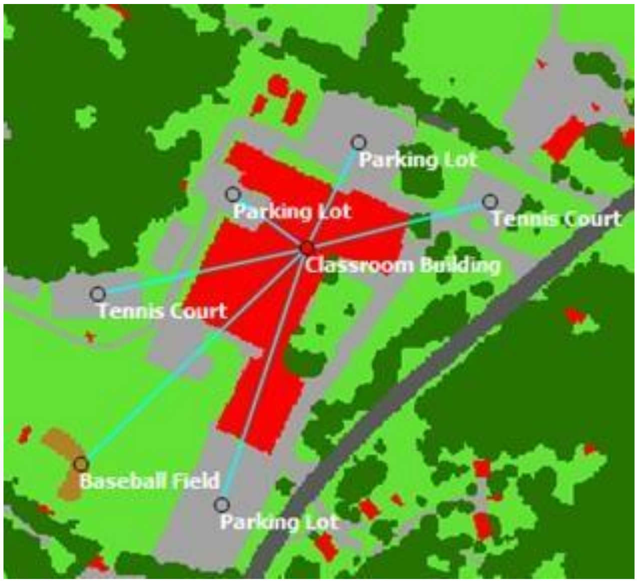

Figure 2. This image gives a visualization of the hub-spoke relationship that these share, and what exactly “distance edge” means. This is a high school building, showing spokes to parking lots, tennis courts, and a baseball field.

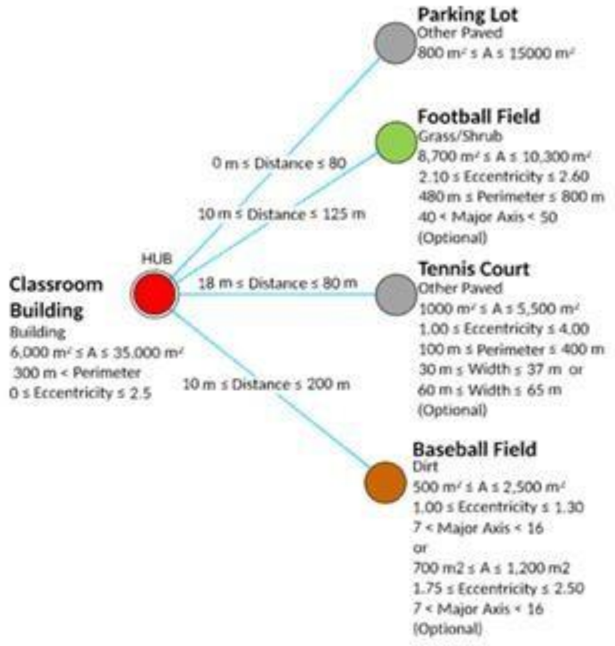

Figure 3. In this image, the final criteria of the template are shown. This gives a visual representation to the limits imposed by the generated criteria, and how specific the correlations can be made to be.

Important note: In this image, the values are rounded for the sake of presentation, and for the discretion of exact values.

Results

Looking into the previous results of the template created by a human analyst, one would see 67 potential "true positive" candidates. It is known, however, that there are only 12 true positives, leading to an 83.08% false positive rate - not so good. Compare this to the procedurally generated template, which resulted in 27 possible high schools. Both sets of data retained the original 12 true positives, the points of information which generated the template, but also constituted a 72.73% reduction in false positives. Even though there was only a difference of 40 results, that itself is an improvement in precision of 59.70%. Due to this process, however, an analyst would now only have to sift through two-fifths of the data that they would have originally had to by using a machine learning based template to find possible queries. This process assumes that if the template were applied to another instance of a GTSG, the pre-generated template held the same ratio of true positives to false positives. An improvement of this magnitude would mean that the criteria were tightened to ensure that it would still be able to account for the variance within the true positive set, but also eliminating extraneous and irrelevant data to the search, essentially providing optimization at no expense.

Conclusion

To reiterate, the goal of the template generation, was to minimize the false negative and positive results, while at the same time, retain and discover true positives in the set. Because the foundation for the template was built of off user-inputted true positives, the template built a variance in the 12 data points provided and determined 15 other buildings that met the criteria. Of the results, 13 of the 15 false positives were in some form or another, a type of (private school, high school, middle school, or elementary schools). Of the other non-school results, they fit very well within the bounds of the template and were just coincidental due to the structure of building, being large, and various parking lots scattered around the main building. But re-analyzing the ratio of schools found, it can be seen that there are 25 schools found to the total 27 results, which is more than coincidental. Upon further examination, and it was revealed that many schools look oddly identical from a purely overhead, quantitative perspective.

Future Directions

Even though this system proved to be successful in accomplishing the task at hand, a wide variety of improvements still need to be addressed. For example, a system in which certain sub-searches could hold higher value over others, a weighting system ideally, needs to be implemented into the process to allow the user to specify the importance of certain sub-searches over others, and allow some sub-features to be optional.

The described process was tested on a relatively small scale, controlled environment – small scale referring to one county as opposed to a country or even a continent. There becomes an evident issue in solving this because the computational power required to process that amount of data would be immense; that is not to say that this is impossible, rather a challenging feat. However, the results on this scale prove that this method is a reasonable substitution for the current implementation of template generation and proves to be more accurate, more consistent, and more efficient.

Even though data mining proved to be adequate in accomplishing the task at hand, this method of interpretation may be ineffective when compared to other machine learning methods such as neural networks. This specific type of machine learning has proven to be more effective in interpreting unknown data and determining a learned value from that data, in this case, the intended search query.

Acknowledgements

I would like to gratefully thank Diane Woodbridge for her continued support throughout my internship at Sandia National Laboratories and beyond. She has provided me with the necessary prerequisite information needed to conduct this project and has given me invaluable guidance along the way. This material is based in part upon work performed during an internship at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin

Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Awadhesh, I. (2012). Classification and clustering analysis using WEKA. Vinod Gupta School of Management. http://www.slideshare.net/ishanawadhesh/classification-and-clustering-analysis-using-weka

Bradley, P., Fayyad, U., & Reina, C. (1998). Scaling EM (expectation-maximization) clustering to large databases. Proceedings of the Fourth International Conference of Knowledge Discovery and Data Mining. AAAI Press.

Brost, R., McLendon, W., Parekh, O., Rintoul, M., Strip, D., & Woodbridge, D. (2014). A computational framework for ontologically storing and analyzing very large overhead image sets. Proceedings of the third ACM SIGSPATIAL international workshop on analytics for big geospatial data, 2014.

Chauhan, H., & Chauhan, A. (2013, October 10). Implementation of decision tree algorithm c4.5. International Journal of Computer Applications, 10(3).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. SIGKDD Explorations, 11(1). Retrieved September 11, 2015.

Johnson, K., & Kuhn, M. (2013). Applied predictive modeling. Springer; New York.

Lindsay, S. & Woodbridge, D., (2014). Spacecraft state-of-health (SOH) analysis via data mining. SpaceOps Conferences.

Martinez, J., & Fuentes, O. (2005). Using c4.5 as variable selection criterion in classification tasks. Artificial Iintelligence and Soft Computing.

Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.

O'Neil-Dunne, J. P., MacFaden, S. W., Royar, A. R., & Pelletier, K. C. (2013). An object-based system for LiDAR data fusion and feature extraction. Geocarto International, 28(3), 227-242.

Pandya, R., & Pandya, J. (2015). C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications, 117(16).

Pooja, S. (2012). A comparative study of instance reduction techniques. International Journal of Engineering Sciences 3(3), 7-13.

Ruggieri, S. (2001). Efficient C4.5. 14(2), 438-444.

Stracuzzi, D. J., Brost, R. C., Phillips, C. A., Robinson, D. G., Wilson, A. G., & Woodbridge, D. M. K. (2015). Computing quality scores and uncertainty for approximate pattern matching in geospatial semantic graphs. Statistical Analysis and Data Mining: The ASA Data Science Journal, 8(5-6), 340-352.

Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann:; Burlington.