Geospatial-Temporal Semantic Graph Search Template Generation via Data Mining

Ryan Cooper and Diane Woodbridge

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.

Introduction

Data mining is the discovery process that analyzes data and generates useful information from it (Pandya, 2015). This process of data mining is useful for a variety of applications, but specifically, for template generation to analyze large overhead image sets. This 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 certain "groupings'" composition and properties, whether it is a field or a building, its 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 query is not searching on a pixel-by-pixel basis, rather through a "natural ontological query"; for example, finding a building that is one year old, within 50 meters of a large body of water, and also next to a parking lot. This type of querying can give a variety of benefits when "compared to traditional search strategies", as expressed by Brost et al. (2014). These benefits include: complex image searches, time-sensitive changes, extraneous node elimination, data fusion - in the form of Light Detection and Ranging (LiDAR) data, normalized digital surface model (nDSM) and Geographic Information System (GIS) data, and decomposing images into large areas within the ontology.

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 is due to the

potential optimization available and definite efficiency benefits that will occur as a result. As stated previously by Brost et al. (2014), the analyst would construct the search query through "a primitive ontology based on regions that have been previously identified in the land cover map" and by the "relationships that have been assigned to specific pairs of nodes". This means that analysts have to meticulously plan, analyze, and compute empirical data in order to design a query template; with the automation provided by data mining, only a few key items need to be identified in order to generate a query template, and the whole process of identification and selection can be simplified to an automated process.

Data mining is one of the most important steps in the process of knowledge discovery in databases (KDD). Traditional methods of turning raw data into useful information rely on manual analysis and interpretation (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), which is inefficient and sometimes inaccurate. Data mining provides an alternative to conventional methods that have proven to be efficient and accurate in almost all cases. It works by finding correlations in large amounts of data by "understanding the application domain" (Fayyad et al., 1996), and building a model that can be analyzed against another target dataset - in this case, a true/false positive/negative system. It then analyzes sets of data and determines useful correlations that can be interpreted as useful information. Fayyad et al. (1996) defines that process in a nine step manner: identifying the KDD goal, creating a target dataset, data cleaning and pre-processing, data reduction and projection, matching goals of KDD process to a particular method, exploratory analysis and model selection, searching for patterns of interest through data mining, interpreting mined patterns, and understanding the outputted data. This details the entire process that one would go through in the process of data discovery and serves as a foundation for the effectiveness of this method of data analysis.

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. It incorporated ideas that were generated by Brost et al. (2014), and uses these as a fundamental element on which the project was designed to improve. In the remainder of the paper, the process of developing templates using data mining will be explored, and how it can account for variance. The balance of accuracy and consistency that machine learning analysis can provide will be explored, as well as tested forms of typical data analysis. Finally, the validity of this method will be evaluated, as well as other applications using these methods will be explored.

Materials & Methods

This process of template generation via data mining was inspired by an opportunity to contribute to a pre-existing project, and help improve the efficiency of the overall project. Due to no existing process regarding automated template generation being in place, a process needed to be created. The information for the GTSG is stored in a SQLite database; therefore it can be queried using structured query language (SQL). Also, 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 usable through WEKA, creating an opportunity to query and analyze directly from the GTSG.

With the environment established, work could begin on template generation and experimentation. 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. This search was broken 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. These criteria were based on the land cover region labels that were assigned to the aforementioned "groupings"; these include: buildings, trees, grass/shrub, dirt, water, road, and other paved areas (Brost et al., 2014). These land cover regions represent the node's type or its composition in relativity to the proposed question. It may also be important 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 (Stracuzzi et al., 2015), all containing empirical property information for analysis, which correlates to a land area of just over 600 square miles as seen in *Figure 7*. Previously, this search had been used as a test for the search function and had been used to build quality score matches (Brost, Phillips, Robinson, Stracuzzi, Wilson, & Woodbridge, Accepted), 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 was an identification accuracy of 99.97% (statistic calculated with overall potential amount of high school nodes compared to actual nodes found). Due to the existence of a template already in place, this gave way for possible practical improvements.

Data mining is an automated process once the proper information is acquired. The issue, in this case, was retrieving all of the necessary information. In the case of the high school search, a set of true positives was retrieved in order to examine for possible quantitative similarities in order to establish a baseline for the template criteria. For data mining to be as accurate as

possible, large amounts of data are required to make an accurate correlation, in essence, the larger the true positive set, the more variance that will be accounted for. With this set, correlations in the empirical data were determined and the best identifiers were revealed through the process of data mining. To give a step-by-step explanation of how the generation process works (flowchart visualized in *Figure 8)* would look like this: 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 through a GeoSearch, as detailed by Brost et al. (2014), and then a very large database would result with various types of land covers – 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. Once in WEKA, the user could run the data through C4.5, and if it showed inconclusive results, then the user would apply a spread subsample and test again. 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. That is the basis of the process, being able to find relevant correlations through pre-existing data, and turning that raw data into a practical application.

The newly generated criteria that were established through comparing the true positive set with false negatives were essentially the template for searches. To ensure the validity of this

method, the new criteria were applied to the whole dataset, which revealed all matches found in the whole dataset. This was done by the following process: first, a GeoSearch on each sub-search was run with all of the newly generated criteria, so in this case, it was done five times. Then, the newly found nodes were analyzed against the edge distance portion of the template which was established by the true positive set, resulting in all the possible high school results. This method of template generation shows an accuracy percentage of 99.99% (statistic calculated with overall potential amount of high school nodes compared to actual nodes found). This outcome sheds light on the true percentage improvement of this system of data analysis compared to previous methods.

Data mining is only as effective as the algorithm used to analyze the data. In this case, the C4.5 algorithm was applied for a variety of reasons. As a classification algorithm, decision tree algorithms are "easy to understand and cheap to implement" (Chauhan & Chauhan, 2013). The C4.5 algorithm proved to be the most accurate and least intensive classification algorithm that WEKA had to offer, addressed later in the paper. Due to the inherent issues that may arise with varied and random data – noisy data, missing data, and scale of data – C4.5 proved to be a good interpreter and still able to remain statistically significant despite issues with the data (Ruggieri, 2001).

When data is passed through the C4.5 algorithm, sometimes the amount of noise and impurities in the data can cloud the correlations. When this happens, a filter called spread subsampling can be applied to the data. Spread subsampling is a process in which the user specifies the "maximum spread between the rarest and most common class" (Pooja, 2012). This reduces the amount of extraneous and irrelevant data points by simply averaging and scaling down the most common class to the ratio to match the rarest class input by the user. In this case,

the rarest cases would contain around 15 points of data, and the common class could contain upwards of 4,000 points of data, sometimes even more. Essentially, spread subsampling is a method of reducing the noise in the data, maintaining the spread of data at relatively the same levels, while at the same time, reducing the ratio of data (as seen in *Figure 5*). It is important to note that as the data begins to approach a one-to-one ratio between the rarest and common classes, the criteria generated by C4.5 will become less and less accurate, in some cases. However, in other cases, outliers can be removed through this method, allowing more precise criteria to be formed. This is due to the concept of scalability; as the data scales downward, it loses clarity, and data points become arbitrary averages of a conglomeration of former data points. This process proved to be useful as it helped to clear noisy data points and generated overall tighter criteria by eliminating outliers.

Results

The two primary forms of machine learning, supervised and unsupervised, have different methods of accomplishing a similar goal. Supervised learning is a process in which the user supplies a labeled data set, called the training data, and the algorithm uses this to make an informed decision. The other cases, unsupervised, does not use a supplied training set, but rather, works without specifically labeling sets of data, and the algorithm itself is left up to group and distinguish the data (Lindsay & Woodbridge, 2014). Of these two forms of interpretation, there are a variety of machine learning analyses. From supervised interpretation, classification seemed most appropriate. Classification is the form of data analysis that groups data together based on quantitative similarities, often using a tree or other form of flow chart to dictate class. From unsupervised interpretation, clustering provided results that best fit the objective (Awadhesh, 2012). Clustering is an unsupervised learning method that works by assessing the similarities

within the data, and forming groupings to match items of similar characteristics. Tests were run to compare which of these two methods would be a better determinant of template conciseness and overall accuracy. Clustering seemed like a reasonable approach to the task of grouping similar attributes of a high school, as did classification, because it had worked by finding trends in data points, and correlating those to a reasonable conclusion. The classification algorithm used in the testing was the aforementioned C4.5 classification algorithm. The clustering algorithm that was tested was the Expectation-Maximization (EM) clustering algorithm (Bradley, Fayyad, & Reina, 1998). The test was run on the football field sub-search because it would be the best chance that EM clustering would have at clustering the similar features, as all football fields are fairly standard. After the test concluded, it was determined that a total of 70 results were found across all of the six empirical data points: area, eccentricity, major axis, minor axis, orientation, and perimeter. For an example of how this criteria can be seen visually, see *Figure 6.* This is definite on the scope of analysis that occurred and constitutes a refinement of 85.86% meaning that there were 495 possible high school football fields found. It is also important to note that there was a pre-applied eccentricity filter on the football fields so that the algorithm was not attempting to cluster every grass field in all of Anne Arundel County, as it would have originally. Then the classification was examined, and running against the same set of data, it resulted in a possible 27 high school football fields. This seemed a lot more reasonable because it constituted a 94.55% overall reduction in nodes. Though both of these methods showed a 1.0 True Positive (TP) rate, the EM clustering held a higher False Positive (FP) rate, 0.121, when compared to C4.5 classification that had a FP rate of 0.033. Analyzing these statistics, it can be seen that even though clustering works as a method of grouping similarities, in this instance, classification seems to give better results without sacrificing any accuracy.

To give quantitative perspective to the actual reduction of the aforementioned high school search, it is important to examine the analyst's interpretation of the results for context. The analyst established a template that addressed all six of the criteria and resulted in 67 possible high schools, including true and false positives. This is compared 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 accuracy 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 targets. 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 was 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.

To reiterate, the goal of the template generation was to minimize the false 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 (high school, middle school, and 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

that coincidental. Upon further examination, and it was revealed that schools look oddly identical from a purely overhead, quantitative perspective. A high school's hub, the classroom building, and spoke, other features, relationship (which is generated by the search function of the analysis) can be examined to see this exact phenomenon. In *Figure 2,* a high school building can be observed and it can be seen that the building and its features are fairly centralized with no oddities or rare occurrences, just a standard high school grouping. Then *Figure 3* is analyzed, and a very familiar scene can be observed. *Figure 3* looks almost identical to *Figure 2* in the sense that they both have numerous parking lots spread across the campus; they both have baseball fields that are relatively centralized to the classroom building, and neither of them have a football field because certain features were made optional by the template. Because these two buildings share similarities, they are both classified as "high school" under what the template believes. This occurrence does not only happen a few times, but rather 13 times, 86.66% of the false positives. Consequently, the template sees all schools as "high schools" and cannot seem to differentiate them from the true high schools. As Murphy (2012) mentions, extrapolating and interpreting data through methods of data mining can sometimes result in the "wrong result". This is due to the fact that the variance of high school cannot be tightened any further without compromising true positives. If the template were to be designed to only find a few results that were without doubt true positives, then the template would be compromising the accuracy for consistency, but to want to produce results that are both accurate and consistent, a broader variance needs to be accounted for, thus resulting in some false positives. If this is viewed from a different perspective and true positives are classified as schools in general, then it would constitute a 92.59% true positive rate with only two false positives being found. This fact leads to the conclusion that from a one-dimensional, quantitative perspective, a machine learning

software has to either compromise accuracy or consistency in order to achieve a better high school true positive rate. The C4.5 algorithm found a good median and was able to integrate the better accuracy and consistency into one template in order to retain the best possible true positive to false positive ratio as possible.

The idea of a tradeoff between accuracy and consistency is one that has an intrinsic role in the selection of the value of predictors within the C4.5 algorithm. Because the C4.5 algorithm is a classification decision tree algorithm, it uses a variable selection method of choosing which of the empirical properties would be indicative of a correlation. The main issue is how C4.5 identifies which variables or attributes in a dataset are the best for classification because "irrelevant and redundant variables often degrade the performance of classification algorithms", in terms of speed, consistency, and prediction accuracy (Martinez & Fuentes, 2005). As aforementioned, these datasets are loaded with noisy data, irrelevant data, and overall varied data. This can cause large issues for a data mining set, and sometimes outliers can impede the accuracy of a template. As an example of this idea, for the football field search (*Figure 4)* and it is notable that the edge distance is up to 125 meters. Some may say that this must be an error, as football fields often reside further away from school buildings, and often not even near the campus, and they would be correct in saying so. This is one example of where consistency needed to be valued over accuracy, and outliers from the true positive set were excluded in the formation of the template in order to make the template better suited for variance, and overall more adaptive for other scenarios.

Illustrations

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.

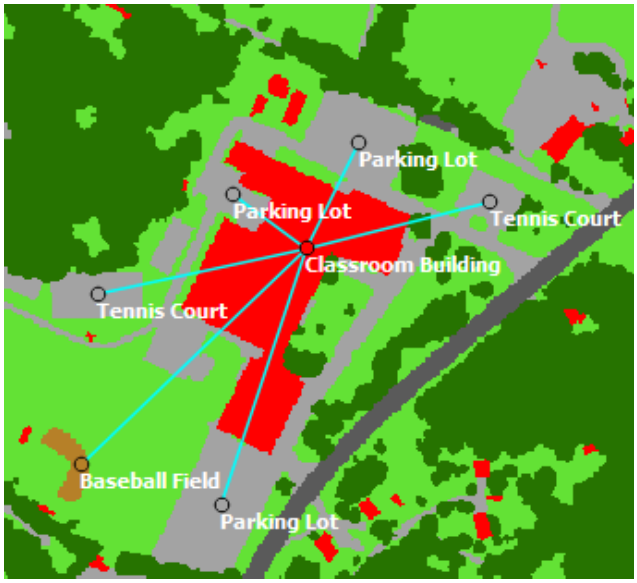

Figure 2. This image gives a visualization of the hub-spoke relationship that these share, and what exactly "distance edge" means. This is a high school building, showing spokes to parking lots, tennis courts, and a baseball field.

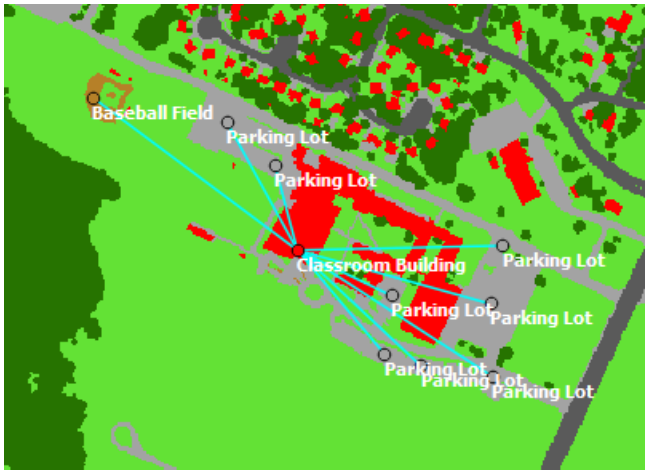

Figure 3. Like Figure 2, this image shows a middle school, with spokes to parking lots, and a baseball field. This image was generated by O'Neil-Dunne et al. (2013) and GeoGraphy (Brost et al. 2014) and has the land cover filter active.

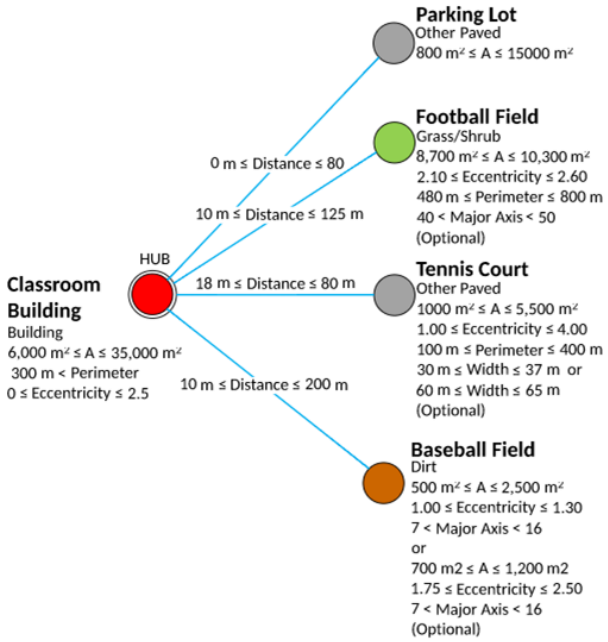

Figure 4. In this image, the final criteria of the template are shown. This gives a visual representation to the limits imposed by the generated criteria, and how specific the correlations can be made to be.

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

Figure 5. This is an example of a 10:1 spread subsampling, showing the reduction in data on the dataset.

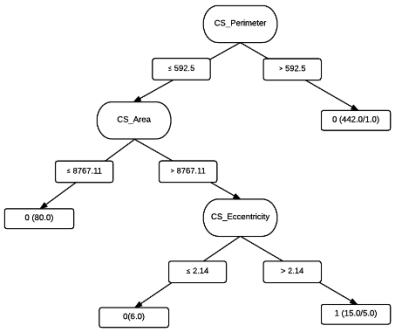

Figure 6. Example C4.5 decision tree outputs.

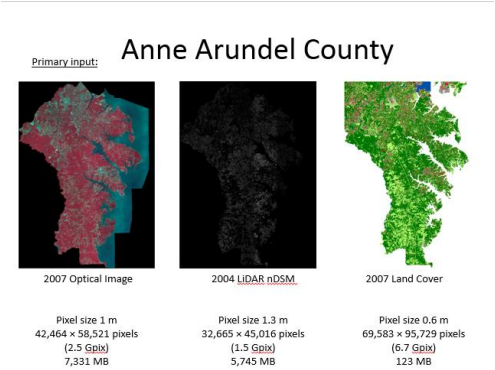

Figure 7. Shows the size of images used in analysis. Ranging from 1.5 gigapixels, to 5.7 gigapixels in clarity.

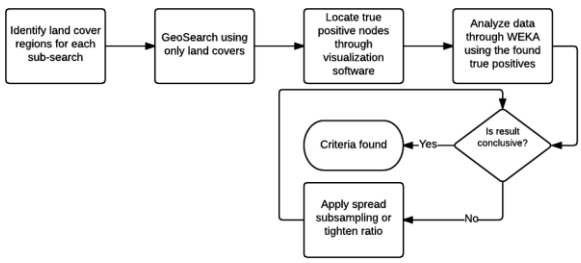

Figure 8. This image shows the process of template generation, and serves as a visual representation of the steps taken in the Materials & Methods section.

Discussion

Template generation is a pre-existing process in place for the analysis of large overhead imagery sets for the purpose of feature analysis. The addition of data mining is a simple addition and a straightforward implementation that would increase the efficiency of the overall search function with few side effects or detracting factors. The weakness of this method, simply put, is that it is not a human analyst. A template is unable to make a decision based on non-quantitative ideas. Data mining and data interpretation share a one-sided perspective, only able to use the information provided to them, and have limited ability to estimate data or provide information

based on inference. This one downside is outweighed by the fact that it is an adaptive learning machine and does what may take a human analyst days or even weeks, in a matter of minutes with almost perfect accuracy. Another limitation faced by machine learning techniques is the "Uncertainty in AI" (Fayyad et al., 1996). This expounds on the thought of one-sided perspectives, as Fayyad et al. breaks it down into three key elements: "issues for managing uncertainty, proper inference mechanisms in the presence of uncertainty, and the reasoning about causality". All three of these ideas reflect the mentality that AI or data interpreting algorithms have a limited perspective on the data provided, and can only infer data in the presence of uncertainty to a certain extent, at which point the error will likely rise. These theories are applicable to this research because the interpretation algorithm that classifies the data analyzes and infers, but will not always see the full picture; therefore, some degree of uncertainty will always be present (Brost et al. Accepted).

However, the uncertainty in this specific instance is merely an account of variance, and thus, compensation in the accuracy of the prediction to account for the consistency. To show the validity of the results, some of the statistics of the results will be analyzed. When looking at the sub-search mean absolute error (MAE), which measures the average magnitude of the errors in a set of predictions, it may be noted that none of them were more than 0.13 in value, which, from a statistical perspective, shows low error in the correlations. The range of MAE is as follows: 0.13 being the most errored with the classroom building sub-search, and 0.05 being the least errored with the football field sub-search, while all of the others were found to be intermediary errors. This trend is due to the variance in each of the subsets, where there are fewer variations in the shape and form of football fields, opposed to classroom buildings, which have a wide variety of sizes, shapes, and form, so it can be difficult to pinpoint a criterion that solely points to true

positives. These statistical analyses reveal that the template was, in fact, able to classify a pattern within the sub-searches, leading to a more accurate and consistent search template to be formed.

When this research is compared to existing documentation in the field, not much emphasis has been recorded for the purpose of making inquiry optimizations, specifically in the field of overhead image analysis. This is a very specific problem, one that has a limited targeted audience. That does not mean that it does not have applications elsewhere, however. Imagine a store that sells food, and they log the customer name and all the purchases that customer makes. They could take this information and create a template to run against their customers as a predictor to see which ones are more likely to purchase another product. This is a common understanding of data mining, using existing information as a predictor, and this is essentially what a template's purpose is, to create a baseline that has serves as an accurate and consistent predictor of who or what will perform the desired outcome. As Fayyad et al. (1996) discuss, data mining and knowledge discovery are at the forefront of discovery, and there are a wide variety of applications and approaches that exist. Information stored in databases have little purpose and meaning, but some of it may be important; it just "has not yet been discovered or articulated yet" (Witten & Frank, 2005). The conclusions drawn from this method are in agreement with the current understanding of data mining and data analysis, the research presented here shows that data mining has a role in almost any medium as it provides consistent and accurate results.

As previously explained, the process to generate templates is somewhat lengthy and user-intensive. To align with the goal of this project, a program was written in C++ using WEKA data mining library to essentially remove almost all user interaction, ensuring that the queries were automatically generated and executed, and that there was little room for human error. Actually, the only human interaction that would take place would be the entering of the true positive

information and some other necessary information into an input file. Then, the user would just have to run the program and multiple criteria lists would be generated from the empirical data analysis, and those would be analyzed, parsed, and outputted for the user. It took what was a seemingly complex user task, and made it into one that was fully autonomous, and resulted in the same template as its human counterpart did using the same operations.

Conclusions and Future Work

The purpose of this process was not simply for one template to be formed, but as a foundation for further research and development to be performed on the topic of template generation. 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 importance of certain sub-searches over others, and allow some sub-features to be optional. In addition, 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.

The issue of finding a way to implement weighted sub-search values into a search will most likely be accomplished through a new algorithm. An algorithm called C5.0 or the "improved C4.5" was an improvement upon the existing C4.5, as R. Pandya and J. Pandya (2015) stated, "C5.0 gives more accurate and efficient result". Not only this, but it has a few

other improvements to C4.5. C5.0 is faster than C4.5, uses less memory, creates more condensed decision trees with improved accuracy, adds support for boosting, allows class weighting, and provides advanced winnowing (Johnson & Kuhn, 2013). Overall the improvements provided by C5.0 seem like a well-fitted algorithm for this purpose; however, the only issue with this is that it is not currently supported by WEKA. What this means for the impact on the project is that the existing methods and queries would have to be re-written for another open-source data mining suite that does support C5.0, of which there are few. Obviously a transition of this scale would require a complete overhaul, which is reasonable considering the added benefits and functionality that would become available with this change. One of the notable options offering C5.0 as an available algorithm is the programming language "R", this language would be an acceptable port, and probably would result in overall faster computations, as well as more user functionality. Obviously, this would be the next step moving forward, taking the current project and improving the functionality as a whole to better conform to any given situation. To conclude, data mining can play a role in improving almost any task that requires a prediction, or has some form of data. It was able to form a template from seemingly random data, and that template proved to lead to accurate and interesting results.

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 Intelligence 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.