CO-119

DATA MINING METADATA TO AUTOMATE THE IDENTIFICATION OF DATASETS COMMONLY FOUND ON A GENERAL REFERENCE MAP

SMITH R.

University of Georgia, CORPUS CHRISTI, UNITED STATES

BACKGROUND AND OBJECTIVES

When cartographers are determining the best symbol choice to represent a spatial phenomenon on a general reference map, they do not necessarily need to see the geometric representation of the feature to make an initial informed design decision on symbolization. Cartographers recognize patterns in the geospatial metadata and datasets due to their experience with similar datasets. For instance, if a cartographer is provided with a metadata document describing a dataset containing a national park boundary, the cartographer would draw on their past experiences and other clues provided in the metadata to consider symbol choices. They do this without having ever seen the geometric representation of the park. The initial symbolization decision can be made with information provided such as theme, purpose, extent, scale and attributes; information which is often included in metadata documents.

It is this idea of recognizing patterns in metadata that prompted the following research question: Do metadata of freely available GIS datasets provide sufficient information for automating the identification of datasets commonly found on general reference maps? This paper will present the initial findings of a research project designed to address this research questions.

The methods and results discussed in this paper examine the results of an initial exploratory journey of describing and mining datasets and their respective metadata for patterns useful for identification. This research serves as a starting point in the conceptualization of how metadata mining can inform the symbolization process of a general reference map. Building on research on data mining, this research examines how an artificial intelligence recognizes spatial datasets and then provides symbol choices, providing a time savings to experienced cartographers or learning opportunity for new mappers.

APPROACH AND METHODS

Data mining falls under the umbrella of knowledge discovery in databases (KDD). KDD has many definitions, however, the two definitions that best describe the purpose of this research are as follows: i) “the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data” (Sowa 1998) and ii) KDD is a secondary data analysis of databases where “secondary” means that the database was never originally designed for data analysis (Usama, Gregory, and Padhraic 1996). This research examines the potential for a relationship between metadata and data mining. Metadata is essentially ancillary data about a data set. It contains a wealth of information in the form of keywords, purpose, attribute field names and so on. The major uses of metadata are to organize, maintain, catalog and aid in data transfer and not data analysis (Hand 1998; Federal Geographic Data Committee 2000). Exploring the relationship between KDD techniques and metadata can possible yield knowledge which will assist in the symbolization decision making process for general reference maps.

The first step in this research project was to prepare the metadata for data mining. This preparation required multiple prerequisite activities. First, a large sample of datasets which are commonly found on general reference maps were downloaded. To narrow the scope and expanse of this research, freely available datasets, covering the United States of America were only considered. The second step in in preparing for data mining was to develop a software program written using Python to produce descriptive statistics which were used to describe the datasets. The third step was to develop an additional software program, again using Python, to extract relevant information (such as keywords and field names) from the datasets and metadata to be used as inputs to the data mining program. The last step was to utilize the information extracted from the datasets and metadata as input into a data mining program which would uncover and classify useful classifying patterns.

Download Datasets

As a first step, freely available vector (point, line and polygon) datasets were downloaded from a variety of freely available sources. Datasets were downloaded if they were considered to be a layer commonly found on a general reference map. These layers included datasets such as roads, political boundaries, urban areas, places, and topography. Once the datasets were downloaded, they were converted into shapefile format (as this was the most widely used data format) to reduce the complexity of crawling through the datasets and metadata programmatically.

Once a dataset was downloaded, it was placed into a directory coinciding with a theme. Four top-level themes where chosen that categorized the downloaded data. These were, “transportation”, “boundaries”, “physical features”, “points of interest”. Each top-level theme was further broken down into sub-themes. For example, the top-theme “points of interest” was broken down into sub-themes such as “airports”, “hospitals”, “monuments”, “museums”, “parks”, “police” and “fire”. This categorizing of datasets was primarily required for training the data mining software to recognize potential patterns.

Exploring and Extracting Metadata Information

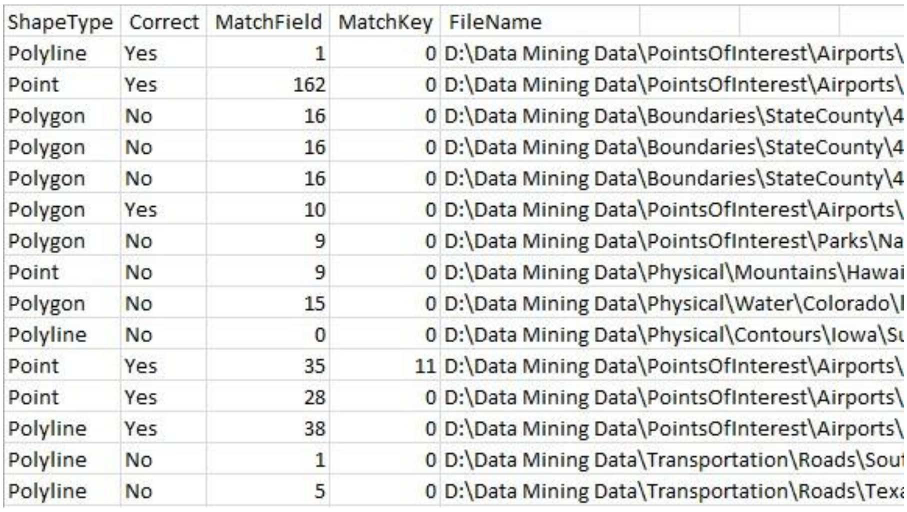
Once the data was downloaded and parsed into themes and sub-themes, descriptive statistics were run on both the datasets and metadata to assess what was available for extraction and analysis. At this point, two additional programs developed in Python were written to derive the statistical descriptions. The first program compiled a count of the number of valid datasets, valid metadata files, metadata file types and terms used to identify keywords in the metadata. The second program extracted all keywords in the metadata and then determined how many keywords were in a plural form. The program then identified and replaced all plural keywords with their singular counterpart to reduce keyword redundancy.

Using the count derived from the first program, an additional program was written to extract and store the keywords, shape type and field names from each dataset and metadata document. This extracted information was used later to compare the individual datasets with the cumulative counts of keywords and field names. At this point, the individual keyword and field name information were aggregated into files containing a count of keywords and field names for each of the themes.

Classification of Extracted Information

To classify the extracted information, using data mining techniques, the software package Rapid Miner (Mierswa et al. 2006) was utilized. Rapid Miner is an open-source software package that provides analysis and data mining algorithms and visualization. Rapid Miner has the capability to perform data modeling and develop Decision Tree and Random Forest data mining algorithms.

To develop the decision tree, training datasets were built using Rapid Miner. To achieve this, keywords, shape types, and field names which had been previously filtered into themes, to be used as input. Multiple training datasets were created, one for each theme. Each training dataset had a target theme which would evaluate to true based on the shape type, keywords, and field names. All other themes would evaluate to false based on this criteria. For each theme, the keywords and field names of individual datasets were compared against the target theme’s keyword and field names list. The result of this comparison was saved to a training dataset along with whether the compared file was a member of the target dataset (for an example, see Figure 1).



ShapeType	Correct	MatchField	MatchKey	FileName
Polyline	Yes	1	0	D:\Data Mining Data\PointsOfInterest\Airports\
Point	Yes	162	0	D:\Data Mining Data\PointsOfInterest\Airports\
Polygon	No	16	0	D:\Data Mining Data\Boundaries\StateCounty\4
Polygon	No	16	0	D:\Data Mining Data\Boundaries\StateCounty\4
Polygon	No	16	0	D:\Data Mining Data\Boundaries\StateCounty\4
Polygon	Yes	10	0	D:\Data Mining Data\PointsOfInterest\Airports\
Polygon	No	9	0	D:\Data Mining Data\PointsOfInterest\Parks\Na
Point	No	9	0	D:\Data Mining Data\Physical\Mountains\Hawai
Polygon	No	15	0	D:\Data Mining Data\Physical\Water\Colorado\I
Polyline	No	0	0	D:\Data Mining Data\Physical\Contours\Iowa\Su
Point	Yes	35	11	D:\Data Mining Data\PointsOfInterest\Airports\
Point	Yes	28	0	D:\Data Mining Data\PointsOfInterest\Airports\
Polyline	Yes	38	0	D:\Data Mining Data\PointsOfInterest\Airports\
Polyline	No	1	0	D:\Data Mining Data\Transportation\Roads\Sout
Polyline	No	5	0	D:\Data Mining Data\Transportation\Roads\Texa

Figure 1: Portion of training dataset for airport theme

The extracted information was fed into a decision tree and random forest algorithm to determine which input variables most succinctly and correctly identify the theme of an input dataset. To evaluate the performance of the algorithms, 30% of the training datasets were set aside during the training and then were classified using the calculated decision tree.

The resulting decision trees were incorporated into a final program developed in Python, whose function was assess an input dataset provided by a user and to compare it to each theme’s decision tree to determine which theme the dataset most likely belongs. If the input dataset did not meet the minimum score requirements to be classified into a theme, the theme that the input dataset scored the best in was presented to the user with a message stating that the classification was inconclusive, but the returned theme was the closest match.

RESULTS

In total, 2,099 vector datasets were downloaded from over 65 local, state, federal and private agencies totaling about 18 gigabytes of memory. An attempt was made to download an equal number of datasets in each theme, however, as some themes were not as common (government buildings, monuments and museums) as others (roads, political boundaries and contours) this effort proved unsuccessful.

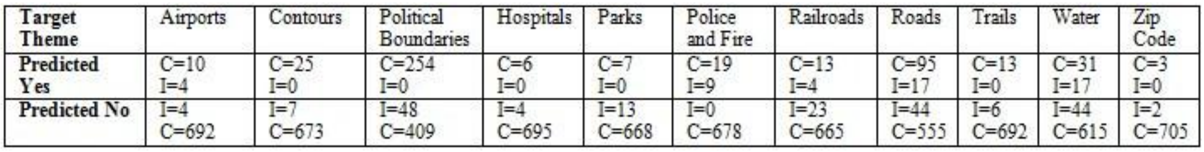
Of the 2,099 datasets downloaded, less than half (1,007) included metadata files. Of the 1,007 metadata files, 214 contained duplicate information saved in different formats (.html, .xml, .txt). Of these different metadata formats, HTML (151) and text (153) files were the most common alternative the prevent XML format (668). Additionally, there were 25 .PDF, 10 .MET and 0 .DOC/X files. While these are not common metadata file formats, they were included in this search to exhaust all options.

Of the 1,007 metadata files, 331 file contained valid keywords in the keyword metadata fields. Files that contained “None” and “Required….” as keywords were ignored. Additionally, if a dataset had multiple metadata files associated with it, the XML file was given preference or text or HTML. This preference was given due to the fact that XML files are highly structured and could consistently be traversed by the programs without issue. If the XML file did not contain any valid keywords however, then the alternate metadata files were similarly searched. Of the 331 metadata files that contained valid keywords, 557 unique keywords were extracted. Of these, 43 keywords were determined to be plurals of other keywords and were added to the singular form’s total. The 2,099 datasets yielded 35,938 field names. After filtering out the feature id (FID) and shape (Shape) fields, which are all shapefiles contain, 3,024 unique field names remained.

After the extraction procedure was executed, a resulting 514 unique keywords and 3,024 unique field names were extracted into individual and aggregate files. Individual files were used to store the keywords and field names in an easy access format. Aggregate files were used to match the keywords and field names which were stored in the individual files and then assigned scores to an input dataset being classified.

Scores assigned were the sum of the total number of instances recorded for each keyword and field name at the aggregate level for the current theme being considered a match for the dataset. As an example, if the aggregate file for the airports theme contained three keywords: “airport_name”, “altitude”, and “type_of_airport” and there were 74, 23, and 44 respective keyword instances recorded, and an individual file contained the keywords “airport_name”, “city”, and “altitude”, then the score assigned to the dataset would be 97 (74 + 23). These scores were compiled into the training datasets and were used by the data mining software.

Next, the training datasets were entered into Rapid Miner and were analyzed using the decision tree and random forest algorithms. The decision trees identified field names as the largest and (in all but one case) only determining factor for classifying input datasets into themes. Keywords were too sparse in the input data sets’ metadata files to be considered a significant and reliable classification measure. The resulting decision trees were interpreted to identify and select the decision tree that yielded the most accurate classification scheme. The selected decision trees yielded a minimum score required to classify a dataset as a member of a theme. Table 1 displays the results of a performance. There are two rows in the table that show the results of a “Yes” and “No” prediction. A prediction of “Yes” means that the dataset is predicted to be a member of the target theme. The results are broken down into how many were correctly predicted (specified as ‘C’) and how many were incorrectly predicted (specified as ‘I’).



Target Theme	Airports	Contours	Political Boundaries	Hospitals	Parks	Police and Fire	Railroads	Roads	Trails	Water	Zip Code
Predicted Yes	C=10 I=4	C=25 I=0	C=254 I=0	C=6 I=0	C=7 I=0	C=19 I=9	C=13 I=4	C=95 I=17	C=13 I=0	C=31 I=17	C=3 I=0
Predicted No	I=4 C=692	I=7 C=673	I=48 C=409	I=4 C=695	I=13 C=668	I=0 C=678	I=23 C=665	I=44 C=555	I=6 C=692	I=44 C=615	I=2 C=705

Table 1: Performance Evaluation

Overall, the performance of the classification was satisfactory for the datasets with a larger training dataset (such as roads and political boundaries). Themes of datasets that did not have enough unique field names were not classified as successfully. The water theme was one collection of datasets that did not have a sufficient number of unique field names to classify correctly. This can most likely be attributed to the fact that the majority of the water datasets had a small number of field names that were vague and commonly used field names such as “name” and “area”. The small number of field names and unique field names provided an insufficient amount of information to classify for the theme on a consistent basis.

CONCLUSIONS

The primary focus of this research project was to determine whether datasets and metadata could be successfully data mined for use in identifying datasets commonly found on general reference maps to assist in determining cartographic symbol choices. The preliminary results of this research show that there is promise in data mining from metadata. However, as seen in this paper, this process is only fruitful when the classification of the datasets parsed into themes is based on field names rather than metadata as a determining variable. The significant lack of completed metadata and lack of shared vocabularies rendered the keywords insignificant in all instances. Even if the research had been re-run using only datasets with completed metadata, vague keywords that were used across many different themes of datasets and inconsistent vocabularies would still hinder their usefulness without the use of additional supporting information.

The surprising result of this research was that field names proved the most useful attribute in determining the theme a dataset. Although the determinations were not overwhelmingly accurate, they did show significant promise to warrant additional research emphasizing field names rather than metadata. One theory as to why field names proved useful was that since the majority of the downloaded datasets were shapefile format in which the attribute table restricted field names to eight characters. With so few characters defining a field name, supplemental information (such as a geographic modifier) were not included in the field name and instead, commonly used acronyms and abbreviations were more widely used.

With these results and lessons learned, future research is being undertaken to determine what other features of datasets and metadata could be data mined to increase the accuracy of feature identification. Additionally, increasing the number of datasets in the weaker theme categories will be targeted to look for increases in accuracy as a result of a more comprehensive training set. Based on the initial findings, the promise of this research path bearing fruit is promising.

WORKS CITED

Gederal Geographic Data Committee. 2000. Content Standard for Digital Geospatial Metadata Workbook.

Hand, D. J. 1998. Data Mining - Reaching Beyond Statistics. Research in Official Statistics 2:5-17.

Mierswa, I. et al. 2006. YALE: Rapid Prototyping for Complex Data Mining Tasks. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06, 935. Philadelphia, PA, USA http://rapid-i.com/content/view/30/214/lang,en/ (last accessed 14 February 2011).

Sowa, J. F. 1998. Knowledge representation: logical, physical, and computational foundations. Boston: PWS Publishing Co.

Usama, M. F., P. Gregory, and S. Padhraic. 1996. From data mining to knowledge discovery: an overview. In Advances in knowledge discovery and data mining, 1-34. American Association for Artificial Intelligence.