A User-Friendly Data Mining System

J. Raul Ramirez, Ph.D.
The Ohio State University Center for Mapping
raul@cfm.ohio-state.edu

1. *Introduction*

Image acquisition of the Earth's surface has become a common event. LANDSAT, SPOT, IRS, IKONOS, are some examples of satellite platforms that are continuously collecting images of the surface of the Earth. Also, hundreds of aerial photo missions are flown each year all over the Earth to acquire images of its surface. With all this activity there is no shortage of images of the Earth today.

In general, the use of these images is limited to a small sector of the population: those scientists with a background in geo-science. A major reason for the small number of users is the fact that is not easy to use these images. Usually, they are stored in large computer files, requiring specialized software to display and manipulate them and an understanding of what they represent. Versions of software necessary to display and manipulate these images are becoming more and more common today. But, even if the ordinary person could display these images, it is very difficult to find specific information of interest to the user. There is very little explicit information (the kind of information a computer can use) in these images. Most of the information in these images is implicit. As a consequence, these images are used mostly in scientific projects.

This paper describes our ongoing research in developing a user-friendly approach to retrieve user-selected information from these images with minimum user effort. We are designing a data mining system for Earth images. Data mining (also known as Knowledge Discovery in Databases - KDD) is defined by Frawley et. al. (1991) as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data." Conventionally, it uses machine learning, statistical and visualization techniques to discover and present knowledge in a form that is easily comprehensible to humans.

Our data mining system will allow the general public to retrieve information from Earth images with very little effort and with only minimum information. This is accomplished by developing and implementing the idea of graphic metadata (graphic data about data), by using fuzzy logic as part of the search engine, by using machine learning, and by using the logic used in navigation and orientation of daily tasks. Our data mining system will be Internet based and is tested with Landsat 7 data, but the system is developed in such a way that it could be used with any kind of imagery.

2. *Data Mining and Images of the Earth's Surface*

There are many definitions of data mining. Grossman (1999) indicates, "Data mining is the semi-automatic discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data. That is, data mining attempts to extract knowledge from data." In the Zucker-Kodratoof Data mining Glossary (1998) data mining is defined as:

> An information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotion, and credit risk analysis.

Simply put, data mining is basically a modeling activity. You need to *describe* the data, *build* a predictive model describing a situation you want to investigate based on patterns determined from known results, and *verify* the model. Once these things are done, the model is used to test the data to see what portions of the data satisfy the model. If you find that the model is satisfied, you have discovered something new about your data that is of value to you.

Today, data mining activities are concentrated mainly in tabular data (Elder and Abbot, 1998), (Piatetsky-Shapiro, 1999), (Graettinger, 1999), and in a lesser quantity in text data (Hearst, 1999). We did not find any application of data mining to images but Grossman (1999) mentioned the use of data mining for images as a new application under "multi-media documents".

In this research we follow closely the general concept of data mining. We build models of objects to be found, then we will test the data for those objects, and if we find them, we will extract them. Our specific approach is described below.

3. Development of Our User-Friendly Data Mining System

We are investigating the following topics:

(1) Navigation and directions taxonomy
(2) Fuzzy operators
(3) The model for data mining of images: Graphic Metadata
(4) Database design for graphic metadata
(5) Data mining system design

A brief discussion of topics (1), and (2), and a more in-depth discussion of topics (3) and (4) are presented next. Topic (5) is not discussed in this paper and it is part of our ongoing research.

3.1 Navigation and Directions Taxonomy

We have investigated the terms and facts that are most commonly used for navigation and giving directions. We were interested in learning what are commonly the terms and facts most frequently used in everyday situations to provide geographic directions. For example, if you need to go from Point A to Point B in a city, and you need to ask for directions, how are these directions provided? How is this done if you need to go from City A to City B? Etc. Our findings are similar to those found in the literature. In general, geographic directions are given using geographic landmarks, their names, approximate distances, and basic direction commands: straight forward, turn left, turn right, etc.

3.2 Fuzzy Operators

We have investigated a set of operators to express these terms and facts. Our goal was transforming these terms and facts into computer-based operators able to reflect the way people provide directions. Generally, we have investigated three types of operators: (1) For fully defined terms and facts, for example, “starting at point *X* follow highway *Y* for two miles, then exit at point *Z*;” (2) Incompletely defined terms and facts,

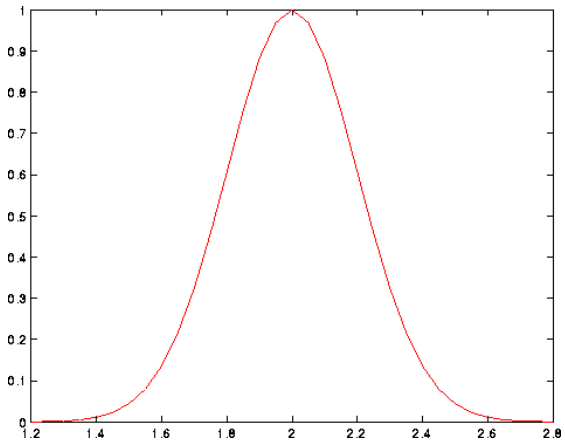

Figure 1. Fuzzy membership function for ‘around 2 miles’

for example, “when near to point *X*, follow a highway for several miles then exit around point *Z*;” and (3) Mixed expressions, for example, “at point *X* take a highway for several miles and exit at *Z*.”

We have used Fuzzy Logic to derive these operators as the theory of Fuzzy Logic provides techniques to handle such imprecise, vague and ill-defined statements. For example, *near point X*, *around point Z* and *around 2 miles*, are fuzzy terms for which equivalent fuzzy operators can be derived. These operators can be characterized by using appropriate fuzzy membership functions. Figure 1 shows a membership function based on a Gaussian error density function that can be used to define the phrase “around 2 miles.” Similarly, other vague phrases can be defined using different fuzzy membership functions. When a decision needs to be made by combining vague phrases using logical operators such as AND, OR, NOT etc., fuzzy aggregation operators can be used to combine them.

3.3 The Model for Data Mining of Images: Graphic Metadata Design

As it was indicated earlier, data mining requires the use of models to find hidden information in a database. We believe that data mining of images also requires such models. Models are built based on hypotheses and facts. For example, if a TV cable provider starts offering Internet cable connection and it wants to increase its current base of Internet cable connection users, the provider could use data mining to select possible new users. In order to do that, the provider needs to build a model describing users of Internet cable connection. By selecting a set of parameters, such as annual income, level of education, credit history, etc. and by comparing and analyzing the current users of Internet cable connection, the provider will build the model. Then, the model will be run on all the provider's TV cable users, and the outcome will be a list of those individuals that may be interested in using the Internet cable connection service. Then, the provider can target them in an ad campaign. We should notice that in the case of tabular information, the model is based on a set of known facts and assumptions. This allows the construction of generic models.

In the case of images, we could think of two different approaches to build the models. In the first approach, we build generic models. For example, a road could be described based on the number of lanes, the width of each lane, the type of geometric alignments, the surrounding geographic features, etc. The drawback of this type of model is that because the restricted explicit information to find roads in images requires an exhaustive search of the whole image, and because of shadows and other circumstances, it may be possible that not all roads of that type are found. Besides, the user may be interested in a very specific road and not in all the roads of this type. In the second approach, models are built for specific terrain features. In this case, one of the parameters to be used is the location of the feature. In general, these two models (generic and specific) complement each other. We believe that from the viewpoint of images, you would need to start with specific models (because the limited explicit information on the images), and later on to use generic models to refine the search of information. This is the approach we are following in this research.

Specific models need to carry explicit information about the content of the images. Vector landmarks and their names could be used to build the type of specific models we could use for image data mining. Vector landmarks are the computer representation (in vector format) of those terrain features that are well known in a region, such as major highways, buildings, sport facilities, etc. Vector landmarks and their names could be connected to images in order to facilitate the location of specific objects on those images. How to geo-define those landmarks and their respectively names, and how to connect those landmarks and their names to the images, in an efficient way, are open questions.

Most users of geo-spatial data are familiar with the word *Metadata* (which is data about data). In this context we will call *Graphic Metadata* the kind of data we believe is needed to help the search of images. Graphic metadata is geo-referenced data about images. As it was indicated above, it is a very rough representation of the ground, with a minimum number of attributes about the images. Graphic metadata is the specific model to be used in data mining of images.

As it was indicated earlier, digital images by themselves carry very little explicit information. The three pieces of explicit information carried by each pixel of a digital image are the row and column of its location in the image, and an attribute value. Usually, the attribute value is a color code. The only way to locate a particular feature or area on an image is to geo-reference all the pixels of the image and search the image

based on coordinate values. This type of approach requires knowledge of geo-spatial science and familiarity with the region in question beyond the general public knowledge.

To find a particular object such as a road without an exhaustive search of the image is impossible, and even an exhaustive search may not always be successful. The fundamental issue here is images do not carry enough explicit information to allow computers to locate complex objects on them. Our goal in this part of the research is to define the minimum set of characteristics for the type of graphic metadata that will allow us to find objects on images. In other words, we are defining the characteristics of the specific model for data mining of images.

If vector datasets exist for the same area as the images, an obvious solution would be to use the vector dataset as the specific model for image data mining. This could be accomplished by geo-referencing both datasets and using the vector information as the base for locating the area to be searched on the images.

We recognize that the kind of vector data and attributes needed to help search images does not necessarily carry the same type of accuracy and completeness as conventional vector data. We believe that from the point of view of the specific model for image data mining, all that it is needed is a very rough vector representation of the landmarks, without any type of symbolization, and the corresponding landmark names.

A basic condition to using graphic metadata is the existence of a vector data set for the region in consideration. In the case of the United States, there is a large digital coverage at scale 1:100,000. There are two versions of this data, one from the U.S. Geological Survey (DLG files), and the other one from the U.S. Bureau of the Census (TIGER files). These data sets can be complemented with information from the Geographic Names Information System (GNIS), developed by the USGS in cooperation with the U.S. Board on Geographic Names (BGN). GNIS contains information about almost 2 million physical and cultural geographic features in the United States. The Federally recognized name of each feature described in the database is identified, and state, county, and geographic coordinates describing a feature's location are also given. The GNIS is our Nation's official repository of domestic geographic names information.

The following information about a selected geographic feature can be obtained from GNIS:

Federally recognized feature name,
Feature type,
Elevation (where available),
Estimated 1994 population of incorporated cities and towns,
State(s) and county(s) in which the feature is located,
Latitude and longitude of the feature location,
List of USGS 7.5-minute x 7.5-minute topographic maps, on which the feature is shown,
Names, other than the federally recognized name, by which the feature may now or have been known.

The GINS dataset was developed from the 1:24,000-scale quadrangle series, but its positional accuracy is inferior.

Our graphic metadata is developed as follows: linear features, such as boundaries, roads, railroads, miscellaneous transportation features, and hydrographic features are taken from the 1:100,000 DLG or TIGER files. Area features such as parks and cemeteries are taken from TIGER files. Point features are taken from the TIGER files and GNIS database. Geographic names are taken from GNIS and the TIGER files. All these datasets are integrated into a consistent database structure. The development of this database structure will be described later. We have collected all the above data sources for a pilot project area in Ohio.

As indicated above, there are two models for the data mining of images: a specific and a generic model. Graphic metadata is the specific model and will help to locate regions on the images where well-known ground objects are represented. If the user is interested in objects that are not well known, then generic

models need to be used. Generic models combined with image processing techniques can be used to find these types of objects.

3.4 Database Design for Graphic Metadata

At The Ohio State University Center for Mapping we have developed a database system for efficient storage, retrieval, and manipulation of geographic data. Ramirez, Fernandez-Falcon, and Schmidley (1993) described this system known, as the Center for Mapping Database (CFMDBF), and Bidoshi (1995) used it for comparing different datasets of the same area.

The Center For Mapping Database Format (CFMDBF) has unique capabilities important for the storage, retrieval, and manipulation of geospatial data. These capabilities include cartographic features with opaque/transparent segments, coordinate systems of n-dimensions, optimization of storage space, expandability, fast feature location tools, and expression of features in a raster-like environment using 3-D Freeman Code (Ramirez, 2001).

The CFMDBF is a vector database with raster characteristics. Partitioning the vector area into grid cells incorporates the raster characteristics. The grid cells are defined as an exact number of cells fitting the vector area of interest. Data manipulation and retrieval becomes easier using a combination of both data structures (raster and vector). Some highlights of the CFMDBF are described next.

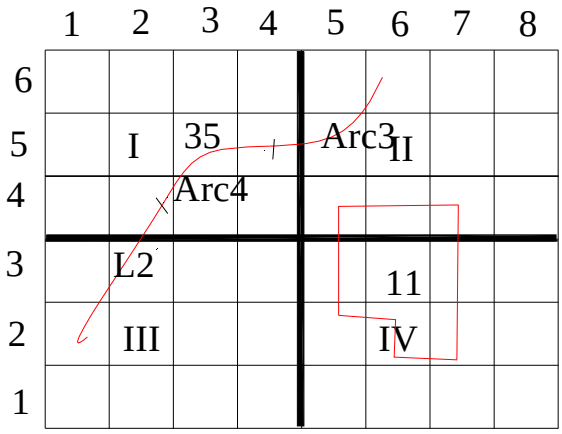

Figure 2. Network of Grids & Cell Covering the Vector Data

3.4.1 Creation of a Network of Grids and Cells Covering Vector Area

To get some specific information of interest by searching all the vector data is inefficient. If the whole vector data area is partitioned into tiles, the extraction of information will be much easier because the scope of query is reduced into one or several tiles.

Therefore, in establishing the CFMDBF it is necessary to create a network of grids and cells covering the vector data area. The method consists in partitioning the area into quadrants called grids. The cell is completely regular and is not finer in the denser regions of data. Then, according to the map condition we group a specific number of grids together into one cell (see Figure 2, with cells I, II, III, and IV).

The relational database derived from such simplicity and regularity is apparent. The location of each grid can be computed from its row and column numbers. Then, every feature will have a grid representation and a vector representation. A cell identification number also is allocated to each cell, and this cell number together with information of all the features that are fully contained or intersect the cell are stored in the database.

With the raster representation of the feature, the scope of the search is not only reduced but also make simple to determine the location and shape of the entire feature. Vector data make every cell and grid meaningful and allow quick query and accurate calculation of the feature information. Therefore, the advantage of this method is this simplicity, resulting in shorter processing time as well as easier software development and maintenance.

3.4.2 Database Structure

CFMDBF is organized in the form of tables. A table is composed of blocks. Blocks are comprised of one or more records. A record is a collection of fields. Fields contain sub-fields. Sub-fields are the minimum unit in the CFMDBF. Records of a given table have fixed size. Tables have two types of records: header and data records. Header is divided into common (with the same structure for all the tables in the database format) and individual (which has unique structure for a given table). Headers carry information that allows the table to be related to its project.

Tables are grouped into four categories: setup, class, data, and description. The detail information about these tables is shown below.

Setup tables: carry information on how the database user wants to organize the data.

Feature Type Table	Stores the type of each feature along with its ID.
Element Type Table	Stores the element type (as point, line) along with the ID the element.
Feature Composition Table	Stores IDs from two previous tables along with a Table ID given by system.
The Data Identification and Setup Tables	Relate Table IDs from the previous table with table names.

Class Tables: carry information to access and manipulate the data.

Feature Table	Stores data using the Feature ID and the Element ID, relating them with the Table ID.
Data Table	Contains information necessary to access the coordinates and attributes of any element.
Cell Table	Carries the unique ID of elements intersecting or contained in a cell.
Cell Key Table	Carries information to access each cell of the grid covering the whole area of the map.

Data Tables: carry the data

Vector Data Table	Carries the geometric coordinate values of any element.
Grid Data Table	Carries the grid representation of the elements that are a combination of grid values and Freeman code.
Attribute Data Table	Contains attribute information for the point, curve, text and line elements in the database.

Description Tables: are related to the characteristics of the elements.

Description Tables contain information such as: Coordinate Axis Description, Reference System, Projection System, Coordinate Units, Primary Color, Line Style, Line Patterning, Level Description, Font Style, Text Justification, and Special Field Codes.

3.4.3 Example of the CFMDBF

The following tables refer to Figure 2.

Feature Point Table

	Road	River	Building
…	…	…	…
F	…	…	121
…	…	…	…
L	35	…	…
…	…	…	…

Feature Table

Record	BP	Feature_ID	Feature_Name	Lo_Grid	Lo_Vector	FP
…	…	…	…	…	…	…
35	0	35	Lane Avenue	96	102	36
…	…	…	…	…	…	…
122	121	122	Federal Building	268	235	0
…	…	…	…	…	…	…

Lo_Grid gives the initial location and length in the Grid table of the features

Lo_Vector gives the initial location and length in the Vector Data Table.

Grid Table

Record	BP	Grid Location	FP
…	…	…	…
96	0	2,1; 3,1; 3,2	97
97	96	4,2; 4,3; 5,3; 5,4;	98
98	97	5,5; 6,5; 6,6	0
…	…	…	…
268	0	4,5; 4,6; 4,7	269
269	268	3,7; 2,7; 2,6; 2,5	270
270	269	3,5	0
…	…	…	…

Vector Data Table

Record	BP	Coordinate Value	FP
…	…	…	…
102	0	xy xy xy xy xy	103
103	102	xy xy xy xy xy xy	104
104	103	xy xy xy	0
…	…	…	…
235	0	xy xy	236
236	235	xy xy xy xy	0
…	…	…	...

Cell Point Table

Record	Cell_ID	Lo_Cell
1	1	1
2	2	61
3	3	63

4	4	100

Lo_Cell gives the initial location in the Cell Table of each cell.

Cell Table

Record	BP	Feature_ID	Lo_Grid	Length1	Lo_Vector	Length2	FP
1	0	35	97	4	103	6	2
…	…	…	…	…	…	…	…
61	0	35	98	3	104	3	62
62	61	71	268	3	235	2	0
63	0	71	269	5	236	4	64
65	64	…	…	…	…	…	…
…	…	…	…	…	…	…	…
100	0	35	96	3	104	3	101
…	…	…	…	…	…	…	…

3.4.4 Summary and Conclusions of CFMDBF

The structure of the CFMDBF takes advantage of the major benefits of the raster and the vector data structures. The manipulation, updating, conflation, extraction, and query of data become easier and faster. And, CFMDBF can be used in many situations according to different requirements. Users can select the tables that are appropriate to their needs or establish new tables to accommodate any special need.

4. Conclusions and Future Work

Based on the idea of graphic metadata, and using fuzzy logic, and an efficient database structure we are implementing user-friendly data mining for images. Graphic metadata adds explicit information to images; fuzzy logic allows searching information in a similar fashion as humans deal with geospatial information and directions; and the CFMDBF provides the searching speed that it is required for real-time queries. All these components are in place together with the data for a pilot project in Ohio. We are currently working on the user interface and visualization system.

References

Bidoshi, K., A Database System as a Tool for Comparing, Updating, and Conflating Spatial Data from DLG and GPSVan, *Unpublished Thesis*, The Ohio State University, 1995.

Elder, J.F., Abbot, D.W. "A Comparison of Leading Data mining Tools," Fourth International Conference on Knowledge Discovery & Data Mining, New York, 1998.

Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J. "Knowledge Discovery in Databases: An Overview," AAAI/MIT Press, 1991.

Graetting, T. "Digging Up $$$ with Data Minining," The Data administrator Newsletter, 1999.

Grossman, R. (Editor) "Data mining Research: Opportunities and Challenges: A report of Three NSF Workshops on Mining Large, Massive, and Distributive Data," University of Illinois, Chicago, 1999.

Hearst, M.A. "Untangling Text Data Mining," Proceedings of ACL'99: the 37th Annual Meeting of the Association of Computer Linguistics, Maryland, 1999.

Two Crows Corporation, "Zucker-Kodratoff Data Mining Glossary", Two Crows Corp., Maryland, 1998.

Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, Two Crows Corp., Maryland, 1999.

Ramirez, "An Extended Representation Model for Geographic Data," Proceedings 20th International Cartographic Conference, Beijing, 2001.

Ramirez, J.R., Fernandez-Falcon, E., Schmidley, R., "Design of The Center for Mapping Database Format," *Center for Mapping Internal Report*, August 1992.