Genetic programming approach to geospatial data mining

Supat Rattanasuksun
The University of Montana
Permission is granted by the author to reproduce this material in its entirety, provided that this material is used for scholarly purposes and is properly cited in published works and reports.

**Please check "Yes" or "No" and provide signature**

Yes, I grant permission  
No, I do not grant permission

Author's Signature: ____________________________

Date: ________________

Any copying for commercial purposes or financial gain may be undertaken only with the author's explicit consent.
A GENETIC PROGRAMMING APPROACH
TO GEOSPATIAL DATA MINING

by

Supat Rattanasuksun

B.S. Bangkok University, 1995

presented in partial fulfillment of the requirements
for the degree of

Master of Science

The University of Montana

2003

Approved by:

Chairperson

Dean, Graduate School

Date
Although Geographic Information Systems (GIS) can graphically represent spatial and non-spatial information, it is difficult for a user to view attribute tables associated with each spatial object on a map. Therefore these systems are not conducive for analyzing or discovering spatial regularities concerning large and complex databases. While there have been several attempts at providing data mining systems for geospatial databases, these systems have been unable to effectively discover information that is of interest to the user. This project presents a system based on Genetic Programming that exploits user interaction to discover knowledge of interest. This system works by reforming a user specified Spatial Structure Query Language (SSQL) query into a tree representation, then genetically reproducing new members using two standard genetic operators, mutation and crossover. Results demonstrate the efficacy of the system for searching for better populations. Thus this project takes an initial step toward solving the ability to dynamically discover nuggets of interest for a user in geospatial databases.
Acknowledgements

First of all I have to thank you Dr. Mathana Santivat, Asst. Prof. Somchit Likithaworn from Bangkok University and Dr. Lloyd Chesnut from The University of Montana, who gave me a life-time opportunity to come and study here in The University of Montana. Dr. Robert Haussman, the Director of English Language Institute, All of my Teacher in English Language Institute, and Dr. Alden Wright, the Chairperson of Computer Science Department, who have been so supportive since the beginning of my student life in USA.

Dr. Joel Henry and Dr. Don Morton for the priceless knowledge and experiences in every classes. Dr. George McRae from Mathematics Department for supporting me to finish this project. Dr. Alden Wright, again, for wonderful knowledge from every classes, especially, Evaluation Computation that helps me gone through this project. Linda, Kathy, Eric from CS Department. Also my friends in Machine Learning Lab, HaiYing, Navin, Konda, and Challeam. Thank you and Thank you.

Finally and specially, Dr. David W. Optiz, my Advisor, who has been so patient and so supportive. This project could not have finished without him. Thank you so much.
# Table of Contents

List of Illustrations ................................................................. v
List of Tables ........................................................................ vi
List of Equations ..................................................................... vi
1. Introduction .......................................................................... 1
2. Backgrounds and Related Works ........................................... 3
   2.1 Data Mining and GeoMiner ............................................... 3
      2.1.1 Association Analysis and Apriori Algorithm ................ 6
      2.1.2 Spatial Association Rules and GeoMiner ................... 8
   2.2 Genetic Programming ....................................................... 10
   2.3 GIS Concepts and Datasets .............................................. 11
3. The Algorithm ...................................................................... 13
   3.1 The Detail of IDM ............................................................ 15
   3.2 Examples and Results ...................................................... 22
      3.2.1 First experiment ....................................................... 23
      3.2.2 Second Experiment .................................................. 27
4. Discussions and Future Work ................................................ 29
5. Conclusions ........................................................................ 30
6. References ........................................................................... 31
List of Illustrations

Figure 1: Topological Relationships in GeoMiner ...................................................... 8
Figure 2: The 6 basic geographic relation functions in ArcObjects ......................... 9
Figure 3: Multilayer of GIS ...................................................................................... 11
Figure 4: Map of the “Major Land Uses” Layer ..................................................... 12
Figure 5: Table of the “Major Land Uses” Layer ...................................................... 12
Figure 6: Overview of the Intelligent Data Mining system (IDM System) ............... 13
Figure 7: Detailed Flowchart of IDM ...................................................................... 15
Figure 8: The “Evaluate Fitness” Process ............................................................... 17
Figure 9: Mutation Operations in IDM .................................................................. 19
Figure 10: Crossover Operations in IDM ................................................................. 20
Figure 11: The Main Screen of System
with the Right Top Corner Box for Setting Variables ........................................ 21
Figure 12: Graph shows the growth of Fitness in 1000 generations
with 80% Crossover Rate and 20% Mutation Rate ............................................. 23
Figure 13: Graph shows the growth of Fitness in 1000 generations (1st Experiment)
with each Crossover and Mutation Rate run twice .......................................... 26
Figure 14: The growth of Fitness in 1000 generations (2nd Experiment)
with every Crossover and Mutation rate shown twice ..................................... 27
Figure 15: The Fitness in 1000 generations
with every Crossover and Mutation rates run twice
(Zoom-in version on the first 50 generations from Figure 14) .................... 28
List of Tables

Table 1: List of 10 individuals and fitness values of the initial population .............24
Table 2: List of 10 populations and fitness values of the 1000th generation ...........25

List of Equations

Equation 1: Support (A⇒B) ........................................................................................7
Equation 2: Confidence (A⇒B) ...................................................................................7
Equation 3: Fitness .....................................................................................................17
Equation 4: Individual Fitness ....................................................................................17
Equation 5: Fitness Proportional Reproduction .........................................................18
1. Introduction

Geographic Information Systems (GIS) are systems that provide mapping and query functionality to aid users in analyzing spatial information. A GIS-database management system (GIS-DBMS) can link together both spatial and non-spatial information. It does this by presenting spatial information as three-dimensional object-oriented features on a map, with non-spatial attributes attached to each mapped feature. Despite this advanced functionality, however, it is difficult for a user to view attribute tables associated with each spatial object, and more difficult to answer and discover spatial regularities concerning large and complex databases. The field of Data Mining offers hope to this pressing need.

One system from the Simon Fraser University, called GeoMiner, studied and developed the concepts of spatial data mining [Han, et al. 1997]. This system performs eight different extracting processes for many different kinds of patterns to be mined [Han and Kamber, 2001]; however, this system only implements one-step data mining, without the capability of efficiently exploring and discovering new knowledge, which is the focus of this project.
In order to build an intelligent data mining system, the system should contain the capability of iteratively searching and improving the efficiency of current results. Genetic Programming [Koza, 1992], which simulates biological evolution, has been shown to be effective at searching large spaces. In addition, human interventions along the process are desirable to help guide the system to meet user preferences. This thesis describes and implements such a system.

The organization of the rest of the paper is as follows. The next section gives background information and related previous work. In Section 3, we present the structure of the implemented system, including algorithm details. Section 4 provides results from running the system to demonstrate feasibility. Finally, we discuss the project and future work, before concluding
2. Backgrounds and Related Works

2.1 Data Mining and GeoMiner

Data Mining is an essential step in the process of Knowledge Discovery in Databases (KDD) [Fayyad, et al. 1996]. KDD is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas. It is concerned with the secondary analysis of large databases in order to find previously unsuspected relationships, which are of interest or value to the database owners [Hand 1998].

There are six primary approaches to Data Mining for extracting different kinds of models and patterns of knowledge in databases [Han and Kamber, 2001].

1. **Characterization and Discrimination**

   The pattern for this approach called *Class or Concept Descriptions* explains the characteristic of data or group of data. The data itself has classes and concepts. For example, in a clothing store, classes of the items can be men, women, boys, and girls; while concepts of customers can be “big spenders” and “budget customers.” The *Characterization* approach tries to explain the characteristics of the class or concept, while the *Discrimination* approach compares the target class or concept with the contrasting class or concept defined by user.
2. **Association Analysis**

The pattern for this approach called *Association Rules* explains the associate relationship of the items, which have high occurring probabilities (called *support*) in a group of sets. The rules discover simultaneous occurrences of items, such as A and B, in the same sets frequently enough (called *confidence*) to make a casual statement like “if there is A in a set, then it also likely has B in the same set.” The thresholds for support and frequency, called *Minimum Support* and *Minimum Confidence*, are set by the user.

3. **Classification**

The pattern for this approach called *Models or Functions* describes and distinguishes data classes or concepts. This approach uses the concept of inductive learning to construct a model, such as a *decision tree*, to classify the class or concept of new data.

4. **Cluster Analysis**

Unlike the Classification Approach, Cluster Analysis will automatically group, without any human participation, the objects or items based on the principles of *maximizing and minimizing the interclass similarity*. These principles are used for arranging the distribution of the objects or items into groups. This approach can also facilitate *taxonomy formation*, which constructs an organization of observation into a hierarchy of classes.
5. Outlier Analysis

Although there is a cleaning step before the data comes into the data mining process, some data objects do not comply with the general behavior or model. These data are known as Outliers, also called Noise, Exception, Contaminated Data, or Missing Data. Often times this data is considered more interesting than regularly existing data. This approach tries to examine the differences to the main characteristics of these data objects.

6. Evolution Analysis

This approach describes and models the regularity or trends for objects whose behavior changes over time. This broad category includes every approach mentioned above, along with time-series data analysis, sequence, or periodicity pattern matching, and similarity-based data analysis.

The system implemented in this thesis concentrates on Association Analysis using spatial databases
2.1.1 Association Analysis and Apriori Algorithm

As mentioned above, Association Analysis is a famous approach of Data Mining that finds \textit{interesting} association or correlation relationships among a large set of data items [Han and Kamber, 2001] and presents these relationships in a form of $A \Rightarrow B$, where $A$ and $B$ are sets of items. Although Association Analysis is typically used for mining relationships in \textit{Transaction Databases}, it is possible to adapt this analysis with other kinds of databases. Sometimes this analysis is also known as \textit{Market Basket Analysis}.

We can make a casual statement for the rule $A \Rightarrow B$ that,

\begin{quote}
If any transaction contains $A$, it tends to contain $B$.
\end{quote}

The simplicity, certainty, utility, and novelty of patterns are factors used for estimating the \textit{Interestingness Measure} for the rule [Han and Kamber, 2001]. \textit{Support} and \textit{Confidence} are standard measurements used in the Association Analysis Approach. The user defines the \textit{Minimum Support} and \textit{Minimum Confidence} as thresholds for new rules. Rules that exceed these thresholds are considered \textit{Strong Association Rules}. Otherwise, it is considered uninteresting for the user and is dismissed.
Support is measured by the number of transactions that contain both of A and B divided by the total number of transactions.

\[
\text{Support}(A \Rightarrow B) = \frac{\# \text{ of transactions Containing } A \text{ and } B}{\text{total } \# \text{ of transactions}} \quad \text{(Equation 1)}
\]

Confidence is measured by the number of data item sets that contain both A and B divided by the number of data items sets that contain A.

\[
\text{Confidence}(A \Rightarrow B) = \frac{\# \text{ of sets Containing } A \text{ and } B}{\# \text{ of sets Containing } A} \quad \text{(Equation 2)}
\]

In the 20\textsuperscript{th} International Conference on Very Large Databases in 1994, a seminal system called the Apriori Algorithm was introduced to discover the association rules between items in a large database of sales transactions [Agrawal and Srikant, 1994]. This algorithm was designed to address two subproblems for mining association rules: (a) finding all combinations of items, called frequent itemsets, that have transaction support above a minimum support threshold, and (b) using the frequent itemsets to generate desired rules. This popular algorithm provides the foundation for the system presented in this paper.
2.1.2 Spatial Association Rules and GeoMiner

Based on the concept of the Association Analysis, the *Spatial Association Rules* represent the object/predicate relationships containing spatial predicates. With the topology design of the spatial database along with the topological relationships, Spatial Association Rules can reflect the structure of spatial objects with spatial/spatial or non-spatial/spatial relationships [Han and Koperski, 1995]. Samples of Spatial Association Rules are:

\[
\text{Is}(X,T.LUC=2) \rightarrow \text{Closeto}(X,B.FID=47) \quad [10.86\%]
\]

\[
\text{Is}(X,T.LUC=2) \land \text{Closeto}(X,C.FID=48) \rightarrow \text{Closeto}(X,C.FID=30) \quad [85.71\%]
\]

Spatial relations can be constructed as a topology of relationships. Spatial Association Rules can represent topological relationships, such as spatial orientation or ordering, or contain some distance information. GeoMiner [Han and Koperski, 1995] is a system that allows a user to mine for the Spatial Association Rules in any level of the hierarchy.
Our system focuses on the lowest level of the topology and uses the basic geographic relationship functions found in the standard library of ArcObjects [Zeiler 2001] to provide the spatial relations or predicates for the system. It creates one spatial relation or predicate, specifically to represent the distance information called \(*\)CLOSETO.*

---

**Figure 2: The 6 basic geographic relation functions in ArcObjects**
2.2 Genetic Programming

The Genetic Programming is a common and popular evolution technique for searching large search spaces. Using the Darwinian principle of survival and reproduction of the fittest along with a genetic recombination operation appropriate for mating computer programming [Koza 1992], Genetic Programming can generate new generation of the population and keep improving the descendents.

With the unfixed size and shape of a tree representation, Genetic Programming is ideal for evolving spatial logic queries as desired by our system. With Genetic Programming, each individual in the population is measured in terms of how well it performs in a particular problem environment. In our system, fitness measures how well a spatial query can retrieve spatial information.
2.3 GIS Concepts and Datasets

The Geographic Information System (GIS) is a Spatial Database Management System (SDBMS) using the object-oriented and relational database design to represent the groups of features as the layers stacking on top of each other. Each layer has one set of data, called Dataset or Map, collecting the information of the features together (Shown in Figure 3).

![Figure 3: Multilayer of GIS](image)

There are two basic types of map information found in GIS databases. The first is spatial information, which describes the location and shape of geographic features and their spatial relationships to other features. The second is descriptive information about the feature, which is also called the Tabular Data [ESRI 1995]. Figure 4 is an example of a Dataset or Map of “Major Land Uses in Montana” [Montana State Library 2003], and Figure 5 shows the spatial information and descriptive information of the Map (also called Feature) in Figure 4.
Figure 4: Map of the "Major Land Uses" Layer

<table>
<thead>
<tr>
<th>Shape</th>
<th>Area</th>
<th>Position</th>
<th>Lu33</th>
<th>Lu33 of Lu3</th>
<th>Symbol</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polygon</td>
<td>66759136</td>
<td>575908875</td>
<td>2</td>
<td>152647</td>
<td>10 Forest and woodland mostly ungrazed</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>73411622</td>
<td>144022641</td>
<td>3</td>
<td>35</td>
<td>10 Woodland and forest with some cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>171867750</td>
<td>122459625</td>
<td>4</td>
<td>47</td>
<td>10 Forest and woodland mostly ungrazed</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>24697282</td>
<td>73533211</td>
<td>5</td>
<td>46</td>
<td>7 Subhumid grassland and semiarid grazing</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>85330647</td>
<td>880276012</td>
<td>6</td>
<td>33</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>53774003</td>
<td>163079641</td>
<td>7</td>
<td>39</td>
<td>3 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>246977750</td>
<td>510392000</td>
<td>8</td>
<td>56</td>
<td>10 Forest and woodland mostly grazed</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>66867632</td>
<td>111121500</td>
<td>9</td>
<td>66</td>
<td>10 Forest and woodland mostly grazed</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>200687536</td>
<td>324580290</td>
<td>10</td>
<td>42678</td>
<td>7 Subhumid grassland and semiarid grazing</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>107382759</td>
<td>1113838000</td>
<td>11</td>
<td>42709</td>
<td>7 Subhumid grassland and semiarid grazing</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>73163546</td>
<td>148600196</td>
<td>12</td>
<td>71</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>344745612</td>
<td>114154203</td>
<td>13</td>
<td>88</td>
<td>7 Subhumid grassland and semiarid grazing</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>133974960</td>
<td>484632067</td>
<td>14</td>
<td>111</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>223373306</td>
<td>352079699</td>
<td>15</td>
<td>42682</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>220619536</td>
<td>67483914</td>
<td>16</td>
<td>30</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>126526707</td>
<td>146148969</td>
<td>17</td>
<td>138</td>
<td>7 Subhumid grassland and semiarid grazing</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>266148016</td>
<td>87078867</td>
<td>18</td>
<td>35</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>540857660</td>
<td>170476079</td>
<td>19</td>
<td>27</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>639872160</td>
<td>169432391</td>
<td>20</td>
<td>142</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>508174952</td>
<td>898502375</td>
<td>21</td>
<td>42691</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>164253456</td>
<td>22494125</td>
<td>22</td>
<td>22</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>165498347</td>
<td>219621547</td>
<td>23</td>
<td>19</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>712230208</td>
<td>138108703</td>
<td>24</td>
<td>21</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>67482352</td>
<td>128040883</td>
<td>25</td>
<td>28</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>866952800</td>
<td>43890543</td>
<td>26</td>
<td>41</td>
<td>7 Subhumid grassland and semiarid grazing</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>263412296</td>
<td>149359895</td>
<td>27</td>
<td>47</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>875924960</td>
<td>57695504</td>
<td>28</td>
<td>9</td>
<td>0 Open water</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>212735684</td>
<td>76490117</td>
<td>29</td>
<td>36</td>
<td>3 Mostly cropland</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>181276288</td>
<td>249126500</td>
<td>30</td>
<td>67</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>267374048</td>
<td>77037600</td>
<td>31</td>
<td>43</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>679361360</td>
<td>53696637</td>
<td>32</td>
<td>31</td>
<td>0 Open water</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>196473104</td>
<td>71325578</td>
<td>33</td>
<td>20</td>
<td>3 Irrigated land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>346036416</td>
<td>197936725</td>
<td>34</td>
<td>52</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
<tr>
<td>Polygon</td>
<td>199722064</td>
<td>62436629</td>
<td>35</td>
<td>43</td>
<td>9 Cropland with grazing land</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Table of the "Major Land Uses" Layer
3. The Algorithm

Figure 6: Overview of the Intelligent Data Mining system (IDM System)

Figure 6 gives an overview of the IDM system. There are two main components of IDM, the Processes Under Supervision of the User (PUSU) and the Processes Without Supervision of the User (PWoSU). The PUSU works as a “User-Interface” between the user and the system allowing the user to monitor and guide the system. Meanwhile the PWoSU works as a self-investigator, working without any intervention from the user to search for the best set of results near the initial user specifications.

The first loop of Figure 6, where the user interacts with the PUSU, is a simple loop where the user initiates and keeps track of the investigation. The user guides the system by entering an initial Spatial Structured Query Language (SSQL) for the GIS Spatial Query and setting all the variables the system needs (such as Minimum Support, Minimum Confidence, and Crossover Rate). After initiation, the system starts to run the self-investigation loop.
The second loop in Figure 6 is a self-investigation, which runs without interruption from the user. The system uses Genetic Programming as its search mechanism. The details of the genetic operators and fitness function are discussed in the next section. This search mechanism interacts with the GIS-Database and the final results are brought back to the PUSU. At this point the user can decide whether to continue or stop the search.
3.1 The Detail of IDM

Figure 7: Detailed Flowchart of IDM.
Figure 7 gives the details of the IDM system. The user interacts with the PUSU by initially setting the initial SSQL and all variables. Later the user gives feedback on the results, and terminates the system. The PWoSU genetic programming system uses a steady-state genetic algorithm whereby it deletes the worst individual from the population pool as new individuals are created.

The genetic programming system uses the user-provided query to seed the initial population. This provides the advantage of finding interesting facts that are of interest to the user. The function "Generate All Population" generates this population by randomly picking a branch inside the initial individual tree and randomly generates a new branch using standard genetic programming. This makes a new individual may has some similar part with the initial individual.

The function "Evaluation Fitness" measures the efficacy of each individual by performing a query posting via GeoMiner to the GIS-database. For each query, GeoMiner returns the strong association rules back to the system. Then one fitness value is calculated from these association rules, and returned back to the system (see Figure 8).
Equation 3 shows the equation for calculating the fitness for each association rule, while Equation 4 gives the fitness value for one individual.

\[
\text{Fitness} = \text{Support} + \text{Confidence} \quad \text{(Equation 3)}
\]

\[
\text{Individual Fitness} = \sum_{\text{Every Rule}} \text{Fitness} \quad \text{(Equation 4)}
\]

A fitness value is calculated for each individual.
There are two genetic operations for this system. The first is mutation (See Figure 9). Mutation is done by selecting one individual, altering the selected individual per standard genetic programming, and sending the survivor to the next generation as an offspring. The second is crossover (See Figure 10), which is processed by randomly selecting two individuals and swapping branches per standard genetic programming. Individuals selected for the genetic operations are selected via fitness proportional selection [Michalewicz and Fogel 2000]. Equation 5 shows the fitness proportional reproduction before the selection.

\[
\text{Fitness Proportional Reproduction} = \frac{\text{Individual Fitness}}{\sum \text{Individual Fitness}} \quad (\text{Equation 5})
\]
Randomly Selecting An Empty Root

Randomly Selecting A Non-Terminal

Randomly Selecting A Terminal

Figure 9: Mutation Operations in IDM
The offspring from Mutation or Crossover is placed into the current population. The population is then maintained by deleting the worst individual from the pool. At this point the process is continually repeated until a constant amount of individuals are created via the genetic operators. Once the genetic programming is terminated, the current population is presented to the user via the PUSU.
Intelligent Data Mining System

Initial Query

DISCOVER SPATIAL ASSOCIATION RULES
FROM ds27 A, d10 B, h490 C
IN RELEVANCE TO L23 T
WHERE (T.LUC=1)

Environment Variables

Iteration (Total Times): 1
Crossover Rate: 80
Mutation Rate: 20
Min. Support (%): 10
Min. Confidence (%): 20
Overall Fitness: 26.14

Query Population From 1 to 10

<table>
<thead>
<tr>
<th>No.1</th>
<th>No.2</th>
<th>No.3</th>
<th>No.4</th>
<th>No.5</th>
<th>No.6</th>
<th>No.7</th>
<th>No.8</th>
<th>No.9</th>
<th>No.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DISCOVER SPATIAL ASSOCIATION RULES
FROM ds27 A, d10 B, h490 C
IN RELEVANCE TO L23 T
WHERE (T.LUC=1)

Evaluation Fitnesses

Total Support: 0.41
Total Confidence: 3.6
Total Fitness: 4.01

Association Rules & Rewards

[S: 10.26%] [C: 80%] is_a(X,T.LUC=1) & closeto(X,B.FID=57) -> closeto(X,A.FID=9)
[S: 10.26%] [C: 100%] is_a(X,T.LUC=1) & closeto(X,B.FID=14) -> closeto(X,A.FID=17)
[S: 10.26%] [C: 80%] is_a(X,T.LUC=1) & closeto(X,C.FID=35) -> closeto(X,B.FID=57)
[S: 10.26%] [C: 100%] is_a(X,T.LUC=1) & closeto(X,C.FID=149) -> closeto(X,C.FID=148)

Figure 11: The Main Screen of System
with the Right Top Corner Box for Setting Variables
3.2 Examples and Results

To illustrate the system, assume we start with a simple question such as:

“What do we know for sure about the Cropland in Montana?”

And

“What do we know for sure about the Forest and Woodland mostly grazed in Montana?”

These questions can be translated to be a SSQL as following:

```
DISCOVER Association Rules
FROM WasteWater A, HazardousWaste B, Lake C
IN RELEVENCE TO LandUses T
WHERE (T.LUC=1)
```

And

```
DISCOVER Association Rules
FROM WasteWater A, HazardousWaste B, Lake C
IN RELEVENCE TO LandUses T
WHERE (T.LUC=6)
```
3.2.1 First experiment

For the first experiment, we set the Minimum Support and Minimum Confidence at 10% and 20% respectively, the Crossover Rate at 80%, which means the Mutation Rate is set to 20% automatically. After entering the first SSQL as an initial individual, the system generates other 9 new “look-alike” queries automatically and evaluates the fitness as shown in Table 1.

The system then searches for new higher fit individuals and replaces the initial population with these individuals. This makes the total fitness of the population continually increase to an asymptote. For this experiment we ran our system for 1000 generations. The resulting total Support, Confidence, and Total Fitness are shown in Figure 12. Table 2 gives the final survivors from the 1000th generation.

Figure 12: Graph shows the growth of Fitness in 1000 generations with 80% Crossover Rate and 20% Mutation Rate
<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1)</td>
<td>4.01</td>
</tr>
<tr>
<td>2.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND ((B.TONS&lt;26000) OR (B.TONS&gt;=23261))</td>
<td>4.01</td>
</tr>
<tr>
<td>3.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND (B.TONS&lt;3127)</td>
<td>4.01</td>
</tr>
<tr>
<td>4.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1)</td>
<td>4.01</td>
</tr>
<tr>
<td>5.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND (NOT (C.LABEL=' '))</td>
<td>4.01</td>
</tr>
<tr>
<td>6.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND ((C.ACREs&lt;=27779) OR (C.ACREs&lt;=23768))</td>
<td>4.01</td>
</tr>
<tr>
<td>7.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND (B.TONS&lt;28963)</td>
<td>4.01</td>
</tr>
<tr>
<td>8.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND (B.TONS&lt;=5550)</td>
<td>4.01</td>
</tr>
<tr>
<td>9.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND (B.TONS&lt;1109)</td>
<td>4.01</td>
</tr>
<tr>
<td>10.</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN REL. TO LandUses T WHERE (T.LUC=1) AND ((C.ACREs&lt;30741) OR (C.LABEL='N'))</td>
<td>4.01</td>
</tr>
</tbody>
</table>

Total Fitness for the first Generation = 40.10

Table 1: List of 10 individuals and fitness values of the initial population.
<table>
<thead>
<tr>
<th>No</th>
<th>Query</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>2</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;4762) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>3</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>4</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>5</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>6</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>7</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>8</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (T.LUC&gt;4) AND (B.TONS&gt;18763) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
<tr>
<td>9</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;18763) AND (C.LABEL=' ') AND (C.ACRE&lt;=25767)</td>
<td>68.21</td>
</tr>
<tr>
<td>10</td>
<td>DISCOVER Association Rules FROM WasteWater A, HazardousWaste B, Lake C IN RELEVENCE TO LandUses T WHERE ((T.LUC=6) AND (T.LUC&gt;4)) AND (B.TONS&gt;4762) AND (C.LABEL=' ')</td>
<td>68.21</td>
</tr>
</tbody>
</table>

| Total Fitness for the 1000th Generation = | 682.14 |

Table 2: List of 10 populations and fitness values of the 1000th generation
Figure 13 gives the results of running the system with the different Crossover and Mutation Rates.

Figure 13: Graph shows the growth of Fitness in 1000 generations (1st Experiment) with each Crossover and Mutation Rate run twice.
3.2.2 Second Experiment

We repeated the experiment for the second SSQL with different Crossover and Mutation rates. See Figures 14 and 15 for results.

Figure 14: The growth of Fitness in 1000 generations (2nd Experiment) with every Crossover and Mutation rate shown twice.
Figure 15: The Fitness in 1000 generations with every Crossover and Mutation rates run twice. (Zoom-in version on the first 50 generations from Figure 14)
4. Discussions and Future Work

Genetic Programming is the process of discovering a computer program that produces some desired output for particular inputs. This process can be reformulated as a search for a highly fit individual computer programs in the space of possible computer programs [Koza 1992]. The previous section demonstrated how our system uses Genetic Programming to solve the process of discovering facts in GIS-databases. However, Genetic Programming can be slow, which is detrimental in a highly user-interactive system.

Artificial Neural Networks (ANNs) [Mitchell 1997] can provide a good solution to the problem. The system can use it to learn which field or range of data is of interest to the user, and allow user can train an ANN to score utility of the individuals the system has found. This score can then be used to guide the system to some specific area of the search space where highly fit individuals are likely to reside. ANNs can also help system to decide amongst individuals with similar or exact fitness values.

Another area of improvement is to perform a complete search for the topological design of the spatial database in order to allow the user to mine knowledge in any level of the GIS database.
5. Conclusions

The concept of structuring an intelligent data mining system that contains the capability of iteratively searching and improving the efficiency of the current results is possible. This paper presents such a system designed especially for GIS data. The system uses Genetic Programming to continually improve a user provided Spatial Structure Query Language query for GIS-Database. The system is able to provide interesting facts found in this database that are of interest to the user.
6. References


