A Bayesian framework for automated dataset retrieval in Geographic Information Systems

Arron Walker Binh Pham Anthony Maeder
Queensland University of Technology
ar.walker@qut.edu.au b.pham@qut.edu.au a.maeder@qut.edu.au

Abstract

Existing Geographic Information Systems (GIS) are intended for expert users and consequently, do not provide any machine intelligence to assist users. This paper presents a Bayesian framework that will incorporate expert knowledge in order to retrieve all relevant datasets given an initial user query. The framework uses a spatial model that combines relational, non-spatial and spatial data. This spatial model allows efficient access of relational linkages for a Bayesian network, and thus improves support for complex and vague queries. The Bayesian network assigns causal probabilities to these relational linkages in order to define expert knowledge of related datasets in the GIS. In addition, the framework will learn which datasets are best suited for particular query input through feedback supplied by the user.

This contribution will increase the performance and efficiency of knowledge extraction from GIS by allowing users to focus on interpreting data, instead of focusing on finding which data is relevant to their analysis. The initial user query can be vague and the framework will still be capable of retrieving relevant datasets via the linkages discovered in the Bayesian network.

1. Introduction

Most experimental and commercial GIS remember dataset configuration through a manual process called "workspace creation". Here users must manually select the datasets[1] of interest and explicitly save them to a workspace. Consequently, a workspace for each user analysis task is required. These workspaces are a static record of the datasets loaded into the GIS, and thus do not dynamically update themselves as new datasets become available.

Setting up workspaces requires expert knowledge of the type of data that would best assist in a particular analysis task. Ideally, the datasets would be ranked by suitability to the particular task. In addition, knowledge of availability and location of the dataset is also required. For example, to analyse possible human population growth in an area, data on historical sites would not be necessary.

Spatial data itself is getting cheaper, in fact some governments and organizations give data away free of charge [1]. In addition, the Open GIS Consortium (OGC) [2] released Web Map Service (WMS) to promote and facilitate the sharing of map datasets across the Internet in 2000. The major commercial GIS products (e.g. MapInfo Professional 7.5 and ERSI's ArcGIS version 8.2) are supporting the WMS specification. In addition, Intergraph provides a free web based WMS viewer called "OGC WMS Viewer" [3]. WMS provides the ability to download maps from WMS servers, thus, allowing different datasets from different WMS servers as well as local data to easily be combined into single map visualization. WMS uses client-server architecture as shown in Figure 1.

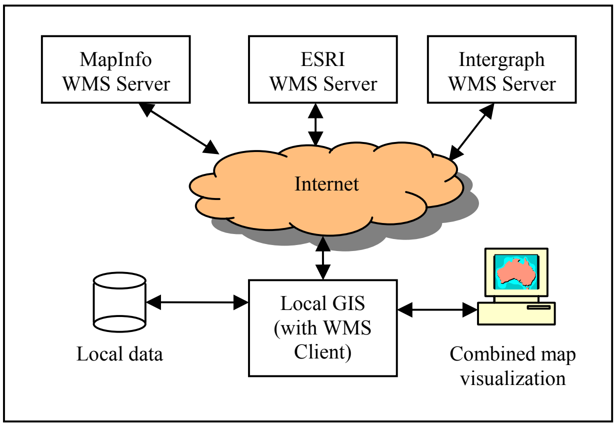


Figure 1 - Web Map Service (WMS)

Not all users possess the expert knowledge to match datasets to analysis task. Furthermore, users usually do not have the time to study the meta data for all available datasets to make these decisions. Technology such as WMS will greatly increase the number of datasets available for analysis. With so many data sources available, the manual process of selecting datasets for particular analysis tasks is not trivial, hence the need for an automatic process. A static workspace requires users to

[1] When a dataset is loaded into a GIS it is often referred to as a theme or layer. Themes will be explained in more detail in section 2.1.

constantly check for new datasets, but a dynamic environment that automatically loads new datasets would ensure that users' decision making is based on the best available data. These factors provide the motivation for a Bayesian framework, which can automatically select and retrieve appropriate datasets.

The remainder of this paper is structured as follows. Sections 2 and 3 review current work in the areas of relation spatial models and Bayesian networks respectively. Section 4 presents the user requirement analysis. Section 5 proposes the Bayesian framework for automated dataset retrieval and section 6 describes implementation and testing of the prototype system. Finally, section 7 presents the conclusion and plan for future work.

2. Relational Spatial Model

Including relational information into the spatial model will improve support for relating datasets to each other. Current research into relational spatial models has generally limited relational information to adjacency information. The more high-level relationships were excluded from the spatial model and left to the normal relational database to handle. This is done with a linkage to the spatial data by some common attribute fields. This approach has limited the extent to which spatial models can expand their functionality. In addition, the spatial data cannot easily contain knowledge of other objects unless all the relational data is also included. Some researchers believe that including relational information into the spatial model will improve support for complex and vague queries. Some current techniques used to define relationship and spatial data are listed below.

Colour can be used to define relationships between non-connected spatial data [4]. The idea of content-based access to image databases relies on techniques which permit users to abstract images in some space by their visual features. A large number of such models, addressing shape, texture, colour and spatial arrangement of features, have been proposed. Using this method to discover relationships between GIS spatial objects has had limited success due to the size and complexity of the spatial data.

An object's area of influence is another technique for defining spatial relationships [5]. This method allows queries such as "which objects are near another object?" to be easily calculated. Current raster and vector models would need to process this query at a primitive level to discover this type of relationship (i.e. time to process query would be high). These questions are important for a military application for decision support for commanders, but are also useful for urban planning. They help to bring the mental spatial model closer to the system spatial model and hence facilitate successful decision-making. The technique of area of influence has been developed into the Qualitative Spatial System (QSS). This system is claimed to simulate human behaviour of reasoning about spatial and temporal knowledge. Unfortunately, the area of influence is limited by some constant radius value for the object, therefore, this method is unable to provide relational linkages across large distances (e.g. between countries).

A method for manipulating complex features in a 3D-GIS is detailed in [6]. The basis of the model here is that all geographic objects can be broken down into one or more simple features. The paper proposes the use of a multi-list structure to describe any geographic phenomena, therefore, complex features are merely a set of simple features. In this way, relationships are built between spatial objects. However, the relational information does not exist between complex objects. Another possible limitation of this model is the way it performs when queried outside the simple feature structure it is based on (e.g. Can it still be queried from first principles, at the primitive spatial level of points, lines and polygons?).

The relational information can be included as part of an extended spatial Abstract Data Type or as part of a spatial Object Oriented Model. Examples of both these implementation methods have been developed for two dimensional spatial data [7]. This implementation method illustrates current support for a relational spatial model. The main limitation that instigated research here was the ill-designed aggregate/disaggregate functions in the existing relational databases. However, this model was only concerned with improving the relational information for adjoining spatial objects when involved in aggregate/disaggregate functions. It does not address the holistic relationships between spatial objects.

2.1. Spatial Themes

The main principle of data organisation of a GIS is to group the data into themes or spatial data layers. These themes are generally layered one on top of the other in the visualization interface of the GIS (e.g. all customers may be in one theme and all property parcels within another, as shown in Figure 2).

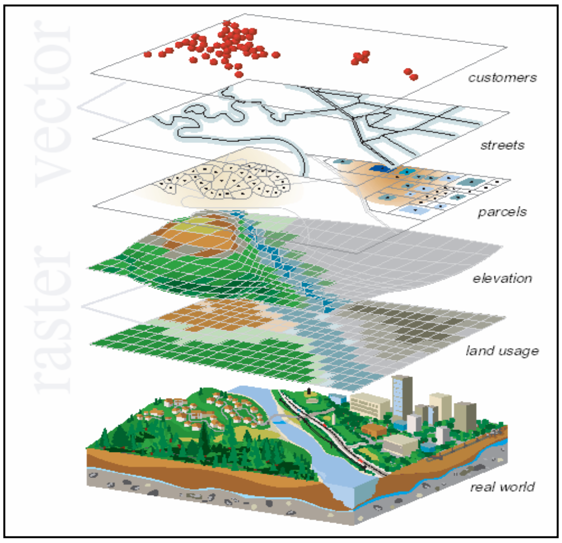


Figure 2 – Spatial Themes [8]

Categorizing data into themes increases the efficiency of data querying. It allows easy addition of new data sets by simple overlay of a new theme layer. However, current methods do not enforce any relational information between themes and rely on primitive operations on spatial data to construct cross theme queries. There is an opportunity to build these relationships as a pre-processing stage as part of the new proposed spatial data model.

3. Bayesian Networks

Bayesian networks are graphical models for defining probabilistic relationships between a set of variables. These relationships can involve uncertainty, unpredictability or imprecision. The relationships may be learned automatically from data files, created by an expert, or developed by a combination of the two. An advantage of Bayesian networks is that they capture knowledge in a form people can understand intuitively, and which allows a clear visualization of the relationships involved.

Heckerman [9] outlines four advantages of Bayesian networks when compared to the other artificial intelligent methods: 1) Bayesian networks can handle incomplete data sets, whereas neural networks can produce inaccurate predictions in this case, until they have had a chance to learn or adjust to the missing data; 2) Bayesian networks allow users to learn about causal relationships. This is important in data analysis, in order to gain an understanding of the problem domain; 3) Bayesian networks combine domain knowledge and data, hence they merge the ideas of expert systems with statistical analysis; 4) Bayesian networks avoid over fitting of data, as Bayesian networks can handle all available data.

The Bayesian probability of an event x is a person's degree of belief in that event. This is somewhat different from a classical probability of an event x which is a physical property of that event in the world (e.g. probability that a coin will land heads).

Bayesian networks used a directed acyclic graph to represent assertions of conditional independence. The nodes in the graph represent the variables and the directed arcs define the conditional relationships. The advantages of directed graphic models over undirected models are the notion of causality. Causality indicates that if an arc is directed from A to B in the network, then A causes B. Bayes' theorem is used to calculate causal inference about the variables. Bayes' theorem states:

$$p(B_i \mid A) = \frac{p(A \mid B_i)p(B_i)}{p(A)} \quad (\text{ i} = 1, 2, \ldots.., \text{r})$$

Bayes' theorem allows the updating of the probabilities regarding uncertain events when fresh information is received [10]. That is, once you know certain events have occurred then one can recalculate the probability of events occurring.

However, the graphical and probabilistic structure of a Bayesian network represents a single joint probability distribution. This distribution is obtained using the Product (Chain) Rule for Bayesian networks:

$$p(X_1 \ldots. X_n) = \prod_{i=1}^{n} p(X_i \mid pa(X_i))$$

Applying Bayes' decision rule performs classification [11]. For example, assume that there are two hypotheses in the classification domain, Bayes' decision rule states that A should be assigned to the hypothesis for which the posterior probability is a maximum. That is, choose;

B_0: if $P(B_0|A) > P(B_1|A)$
B_1: if $P(B_1|A) > P(B_0|A)$

Where $P(B_0|A)$ and $P(B_1|A)$ can be calculated using Bayes' rule. If the above example was extended to include more than just two hypotheses, then the problem can be viewed as searching through the set of all possible hypotheses with the goal of finding the best hypothesis. The best hypothesis can be defined as the most probable hypothesis given the "evidence" of the Data D in the hypothesis space H. Such a hypothesis is referred to as the maximum aposterior (MAP) hypothesis [12].

$$h_{MAP} \equiv \max_{h \in H} P(h \mid D)$$

From Bayes' rule,

$$h_{MAP} \equiv \max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)}$$

Because P(D) is independent of h, it can be dropped, resulting in

$$h_{MAP} \equiv \max_{h \in H} P(D \mid h) P(h)$$

A Bayesian network will be used to aid non-expert queries. Unlike neural networks, Bayesian networks do not need large sample data sets to train themselves and as a result will be better suited to this application domain. In addition, Bayesian networks allow full visibility of how decisions where made.

With parametric learning (Bayesian learning), we can start using the expert system with an imperfect knowledge base and progressively improve its quality with experience.

Bayesian networks have been used in many different domains [13-17] to provide decision support. For example, medical diagnostic systems based on Bayesian networks compute the best diagnoses given the existence of certain patient symptoms (or evidence) [14]. Of more interest to this research, Bayesian networks have been combined with heuristics with the aim of understanding queries in [16]. Heckerman and Horvitz [18] have developed a Bayesian network for information retrieval (IR), which infers the goals and needs of software users. The approach focuses on the construction of probabilistic knowledge bases for interpreting user queries. The Bayesian network establishes a casual relationship between the query goal and the query terms. The terms are keywords extracted from the text query input and the goals were help topics. The system has become the basis for the Microsoft Office Help program.

4. User Requirement Analysis

As stated in the introduction, this paper aims to automatically select datasets that will provide optimal information to assist with a decision making task. This section categorises the types of analysis tasks performed by GIS users and matching these against the types of dataset available. Before progressing further , it should be noted that GIS applications exert an important influence on dataset selection. The analysis of spatial data is predominantly used for planning in following applications: urban, rural, industrial, environmental, tourism, health, historical, military, emergency, navigation, utility and facility management. These applications are self-explanatory and therefore will not be discussed in any more detail other than to establish their relationship to analysis tasks and datasets in section 3.3.

4.1. Types of analysis task

The analysis tasks used in GIS can be categorized into field-based, object-based, and scenario-based.

Field-based tasks include cartographic modelling and digital terrain modelling operations. Cartographic modelling requires the following types of operation: combining map layers to form new ones, reclassifying zones, distance measurement, area calculation. On the other hand, digital terrain modelling requires slope calculation, aspect calculation, interpolation and pattern identification of attribute data over a geographical area.

Object-based tasks include attribute analysis, spatial operations and network modelling. Attribute analysis is performed on the object's attribute values, and generally involve data attribute information about the objects or attribute relationship between objects. Spatial operations use relationships between spatial data objects to return results. Spatial operations include: finding nearest neighbour, equalling, splitting, merging, scaling, rotating, boundary, interior, overlap, cover, and different levels of detail. In addition, spatial operations may include set-oriented tasks (such as equality, membership, subset, disjoint, intersection, and union) and metric calculations (such as distance, proximity, length, area, volume, and perimeter of or between objects). Network modelling include connectivity tests, route analysis, and shortest path calculations.

Scenario-based tasks simulate the effects caused by geo-phenomena. Certain attributes of the geo-phenomena can be changed to visualise different effects. Scenario-based tasks visualise hypothetical and temporal changes in spatial data. Cause and effect analysis is a major focus of scenario-based tasks. Scenario-based tasks are, in fact, the combination of field-based and object-based tasks. For instance, as field-based parameters change, it will cause changes to object-based parameters and vice versa, e.g. how changing flood water level (field-based) would cause elevation changes to boats (object-based).

4.2. Types of dataset

Datasets include spatial, non-spatial and relational information about any natural or man made phenomena, so long as it has some spatial reference in time and space, e.g. roads, rivers, mountains and buildings.

4.3. Mapping analysis tasks to datasets

Bayesian networks will be used to map analysis tasks to datasets. The initial networks will be initially constructed from expert knowledge and available statistical information. The networks will then keep the mapping up-to-date as new information is available through a Bayesian learning process.

For object-based tasks the users are generally interested in the underlying attributes and the spatial position of a particular object or group of objects.

5. Bayesian Framework for automated data retrieval

The framework is illustrated in Figure 3. A typical GIS uses separate spatial and relational database coupled together by a Database manager. The Data layers are selected manual by the users. In the proposed framework, the spatial and relational data are combined into one database, which incorporates improved "spatial awareness". A Bayesian network utilizes this spatial data model in order to automatically select data themes.

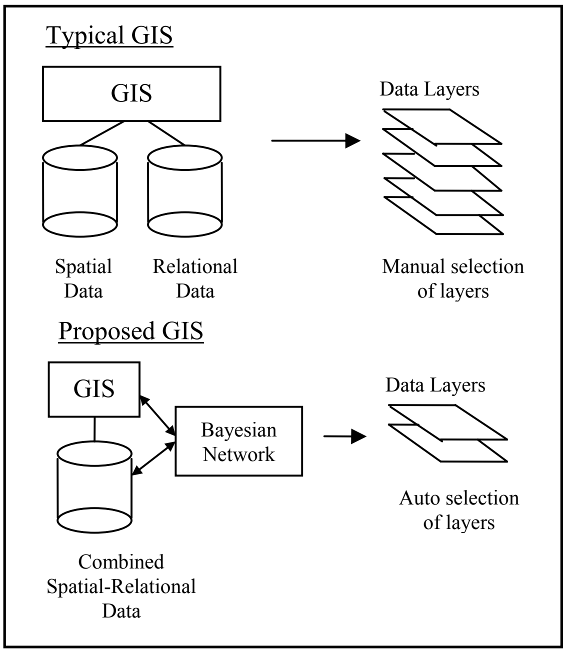


Figure 3 - Bayesian Framework

The Bayesian network will keep a cached file of what datasets exist on the Internet. This list of WMS servers is updated manually by the system administrator. The cached file will contain meta data about the dataset and the location of the dataset. The Bayesian network will search meta data attributes of both local and WMS data in the cached files for datasets that match the posterior nodes selected by the Bayesian network to be conceptually linked to the query input. Once a number of matches is found according to matching criteria, the data is downloaded and sent to the GIS for display. These datasets could be selected to be downloaded according to matching priority, but it is also possible to download more datasets, but only made visible the top matching 5 or 10 datasets.

Regional issues are resolved by looking at the datum setting of the GIS. This limits the search criteria to the region of interest to the user. For example, if a UTM 56 datum id used, the region of interest would be set to South East Asia.

5.1. Feedback to Bayesian network

In order not to continually harass the user for feedback, it was decided to opt for an implicit style of feedback to allow the Bayesian network to learn. The learning process for a Bayesian network is dynamic update of causal probabilities. The causal probability will be updated through considering any post loading of dataset layers, that is, if the users adds additional datasets to the map. If a dataset layer was displayed because of a high causal probability, and if consequently this layer is discarded by the user as unimportant (through the process of removing the dataset layer from the map), then the probability will be decreased in the Bayesian network. This causal probability update would have the effect of making it less likely that the dataset in question would appear in future queries of that type. Similarly, if a dataset layer were manually loaded into the map then the causal probabilities relating that dataset to this query would be increased, thereby making it more likely for that dataset to be automatically loaded with future queries of that type.

5.2. Prototype

A prototype has been implemented using Microsoft .Net's C# programming language and ESRI's MapObjects 2.2 ActiveX component. The resulting prototype is a windows application that automatically loads datasets given an initial user query. The prototype allows visualization of the Bayesian network being utilized in the query process. The local datasets used have been supplied by the school of Built Environment and Design at QUT. The datasets are best suited to an urban planning or rural planning discipline. The WMS servers accessed were CCRS Spatial Data Warehouse, Demis World Map Server and Intergraph World Map [19].

An example program output for the user query input of "Land Parcels" is shown in Figure 4. The Bayesian network returns the following datasets: DCDB, Contours, Waterway and Roads.

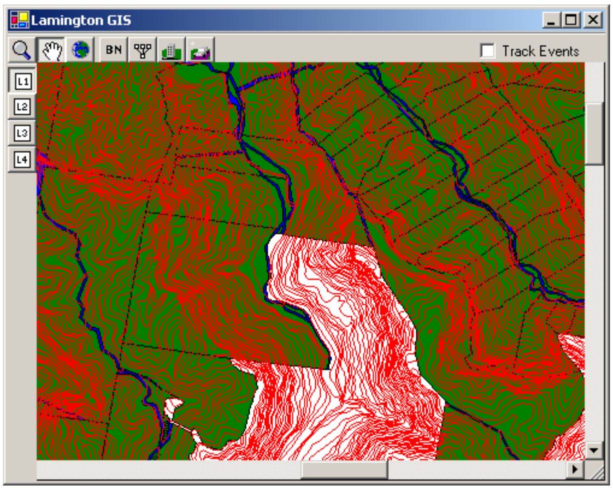
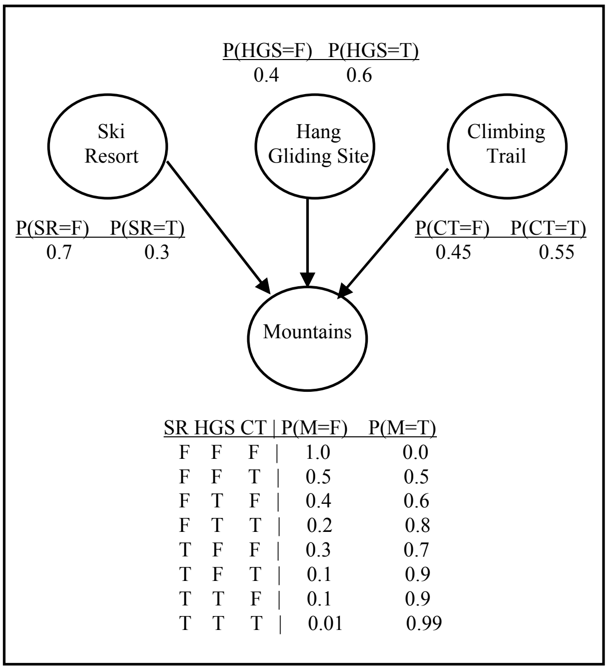


Figure 4 - Map Visualization of Prototype System

Visualization of Bayesian networks allows system administrators or even users to analyse why certain datasets were loaded. A simple example containing three a priori and one posterior is illustrated in Figure 5. Each of the a priori are random variables and their probabilities are given in the figure. Also shown in the figure is the conditional probability table for the posterior, given its parent (aposteriori) nodes.



P(HGS=F)	P(HGS=T)
0.4	0.6

P(SR=F)	P(SR=T)
0.7	0.3

P(CT=F)	P(CT=T)
0.45	0.55

SR	HGS	CT	P(M=F)	P(M=T)
F	F	F	1.0	0.0
F	F	T	0.5	0.5
F	T	F	0.4	0.6
F	T	T	0.2	0.8
T	F	F	0.3	0.7
T	F	T	0.1	0.9
T	T	F	0.1	0.9
T	T	T	0.01	0.99

Figure 5 - Simple Bayesian Network

An example of this Bayesian network as visualized in the prototype system is shown below in Figure 6.

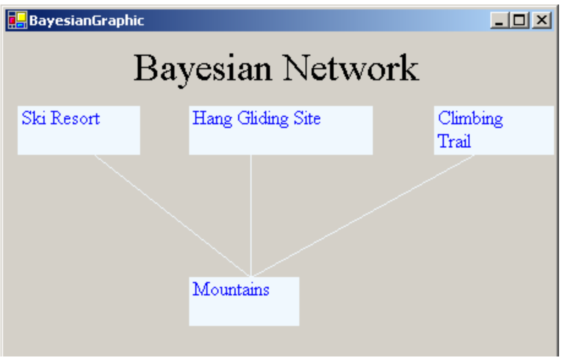


Figure 6 - Visualization of Bayesian Network

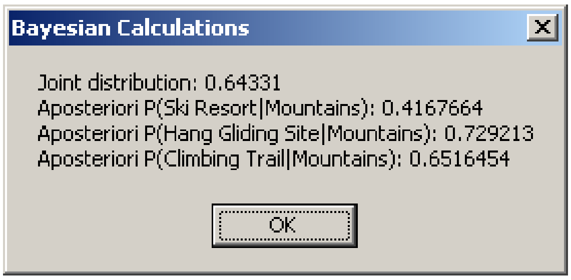
The joint distribution is obtained using the Product (Chain) Rule for Bayesian networks from section 2. Therefore;

$$P(SR,HGS,CT,M) = P(SR)P(HGS)P(CT)P(M|SR,HGS,CT)$$

To calculate the maximum posterior use the equation from section 2, this gives;

$$h_{MAP} = MAX\{ P(SR=T|M=T), P(HGS=T|M=T), P(CT=T|M=T) \}$$

The results as calculated by the prototype system are shown below in Figure 7.



Bayesian Calculations

Joint distribution: 0.64331
Aposteriori P(Ski Resort|Mountains): 0.4167664
Aposteriori P(Hang Gliding Site|Mountains): 0.729213
Aposteriori P(Climbing Trail|Mountains): 0.6516454

Figure 7 - Posterior Probability Calculations

The working prototype is interfacing with the WMS Servers, however, the WMS only provides a pictorial representation of maps. This will limit interaction between GIS and WMS data because it will be harder to query the underlying features in the map. The WMS specification provides limited querying through the “GetFeatureInfo” command. This command only returns information about a number of nearby features to the x,y position selected on the pictorial map display. The user of the WMS map can not be certain that there is even attribute data available on the features they are selecting on the map. If no such data is available, the system will perform an automatic search for new WMS servers.

Only simple Bayesian networks have been loaded in the system at this stage. Further investigation of performance is required for larger Bayesian networks.

6. Conclusion and Future Work

The framework for Bayesian networks meets the user requirements set out in Section 4 and has certain advantages over previous manual approaches. It has a simple interface that allows continuous improvements by updating the network with new information as they become available. Such information may also come from experts who provide relevant feedbacks.

A prototype application has been developed using simple Bayesian networks in order to evaluate the feasibility of the proposed framework. We are currently investigating more complex Bayesian networks to provide further capabilities to deal with more complex analysis tasks.

7. Acknowledgement

The Cooperative Research Centre for Satellite Systems (CRCSS) is established and supported under the Australian Government's Cooperative Research Centres Program.

8. References

[1] Queensland Government, "Using spatial information for a sustainable SEQ," presented at 2002 SEQ Spatial Information Expo, Brisbane, 2002.

[2] OpenGIS, "Open GIS Consortium," http://www.opengis.org/, Accessed on: 30 - June 2003.

[3] Intergraph, "OGC WMS Viewer," http://www.wmsviewer.com/, Accessed on: 4 - July - 2003.

[4] S. Berretti, A. Del Bimbo, and E. Vicario, "Weighting spatial arrangement of colors in content based image retrieval," presented at Multimedia Computing and Systems, 1999. IEEE International Conference on, Dept. of Syst. & Inf., Florence Univ., Italy, 1999.

[5] D. Kettani and J. Roy, "A qualitative spatial model for information fusion and situation analysis," presented at Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on, Defense Research Establishment Valcartier, 2000.

[6] K.-H. Kim, S.-K. Choe, J.-H. Lee, and Y.-K. Yang, "Efficient method for manipulating complex features in 3D-GIS," presented at Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International, ETRI-CSTL, 2001.

[7] A. Voisard and B. David, "A database perspective on geospatial data modeling," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 14, pp. 226-243, 2002.

[8] ESRI, "ArcView," Environmental Systems Research Institute, http://www.esri.com/, Accessed on: 14-Jan- 2003.

[9] D. Heckerman, *A Tutorial on Learning with Bayesian Networks*. Redmond, WA: Microsoft Corporation, 1996.

[10] T. Leonard and J. S. J. Hsu, *Bayesian Methods - An Analysis for Statisticians and interdisciplinary Researchers*. Cambridge, United Kingdom: Cambridge University Press, 1999.

[11] A. A. Skabar, "Inductive Learning Techniques for Mineral Potential Mapping," Ph.D. Thesis, *School of Electrical and Electronic Systems Engineering*. Brisbane: Queensland University of Technology, 2000, pp. 226.

[12] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[13] A. Jameson, B. Gro[ss]mann-Hutter, L. March, R. Rummer, T. Bohnenberger, and F. Wittig, "When actions have consequences: empirically based decision making for intelligent user interfaces," *Knowledge-Based Systems*, vol. 14, pp. 75-92, 2001.

[14] P. Haddawy, J. Jacobson, and C. E. Kahn Jr., "BANTER: a Bayesian network tutoring shell," *Artificial Intelligence in Medicine*, vol. 10, pp. 177-200, 1997.

[15] N. Liem and P. Haddawy, "Answering queries from context-sensitive probabilistic knowledge bases," *Theoretical Computer Science*, vol. 171, pp. 147-177, 1997.

[16] H. M. Meng, W. Lam, and K. F. Low, "A Bayesian approach for understanding information-seeking queries VO - 4," presented at IEEE International Conference on Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings., 1999.

[17] M.-L. Shyu and S.-C. Chen, "A Bayesian network-based expert query system for a distributed database system VO - 3," presented at IEEE International Conference on Systems, Man, and Cybernetics, 2000.

[18] D. Heckerman and E. Horvitz, "Inferring Informational Goals from Free-Text Queries: A Bayesian Approach," presented at Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, 1998.

[19] Intergraph, "Open Geospatial Network," http://www.geospatial.net/servers/default.asp, Accessed on: 4-July- 2003.