A Theory of Efficient Search in GIS

Yongjian Ye
Institute of Mathematical Geology and Remote Sensing Geology
China Geoscience University
Wuhan, Hubei
P.R. China 430074

Abstract

In this paper, we present a theory of efficient search in GIS. By efficient search we mean to efficiently search for a given feature over a given spatial region or an information system, such as searching for a wildlife specie in a forest, searching for an oil field in a region, or searching for certain information in a GIS database. The task of Geo-information search is formulated as an optimization problem - maximizing the tendency of obtaining the information to be searched within a given effort constraint. Our formulation relates the search action (such as dig a well to find whether the region contains oil) with its effects (such as the action really find the oil field). We give a general action selection strategy that can determine which action should be applied when the number of actions that can be applied is restricted. We also give a strategy of how to integrate the result of the search action with our previous belief.

1 Introduction

The research described in this paper conforms to the area of Geo-information search, such as searching for a lost animal in a forest, searching for an oil field, etc. The meaning of "GIS" includes a wide range of activities within the geographic information, ranging from investigations into how people think and reason with geographic information, so that better systems can be designed that are easier to use, to studies of the legal and intellectual property issues raised by widespread sharing of geographic information. The Geo-information search problem is important in many "GIS" activities. For example, a traditional view of GIS is that its purpose lies in building a digital representation of some existing set of geographic data, thus allowing the data to be subject to analysis. For this view, the task of searching for a specific geographic date becomes important if we want to convert the relevant information into digital form and to process numeric information much more rapidly than can a human being. Some GIS practioners incorrectly treat maps as reality, but maps are only representations of reality, and thus database constructed from maps are also representations. The quality of a spatial database depends on how well it represents reality, which depends in turn on numerous factors in the processes used to create the database, such as measurement error, uncertainty in interpretation, and digitizing error. In some situations, it is even impossible to give an exact map - for example, the possible oil field distribution over a region or the distribution of a certain kind of animal over a region. For these situations, the spatial database uncertainty must be described and modeled. Easing spatial error modeling into the user domain will be far more easier if error definition, visualization, and management are based on the same fundamental statistical principles. For example, in the Geo-information search task, if we assume that we are able to divide the $3D$ spatial region into $3D$ grids or the $2D$ spatial region into $2D$ grids, and we encode our beliefs about the possible presentation of the target such as the oil field etc. into these grids as a probability value, than the Geo-information search task will be much more easy. This paper is about how to find a given fact within a spatial region and how to distribute the search effort efficiently during the search process in order to find the given fact based on the GIS map. The main concern of this paper is how to form the GIS map based on the available information and how to design the search strategies. To make the discussions easier, we assume that our goal is to find an oil field within a spatial region and our task is to determine where should we dig wells so as to maximize the probability of finding the oil field when the number of wells that can be dig is a fixed constant. All the strategies should be generated from the GIS map.

2 The Formation of the GIS Map

It is obvious that we are not able to generate a GIS Map for every possible point of the spatial region because of the limited computer memory and other factors. So, we divide the spatial region with identical $3D$ grids. The size of the grids represent the resolution of our tessellation. To be able to focus on specific geographical areas and to handle portions of as well as complete maps, we define a hierarchy of maps with increasing size of the grids. We start with the smallest size of the grids (the highest resolution) and generate a progression of maps with increasingly less detail. This allows certain kinds of selection to be performed at lower levels of resolution with correspondingly less computational expense. Each action selection process should based on a given level of resolution. The resolution of the maps is increased by dividing a given grid of the map into 8 identical small grids in $3D$ situation, or by dividing a given grid of the map into 4 identical small grids.

Each grid g_i of the map is associated with a probability value $0 \leq p(g_i) \leq 1$ that specifies the belief that whether this grid contain oil or not. $p(g_i) = 0$ if we believe that there is exactly no oil within this grid; $p(g_i) = 1$ if we believe that there must be oil within this grid. $p(g_i) = 0.5$ if no information concerning it is available.

It is interesting to ask how can we obtain the initial probability value for each grid. The initial probability has a great influence on the performance of the search system. A good initial distribution can lead to a quicker discovery of the oil field, while misleading knowledge can greatly degrade the performance of the system.

The initial probability can be determined approximately by our initial knowledge

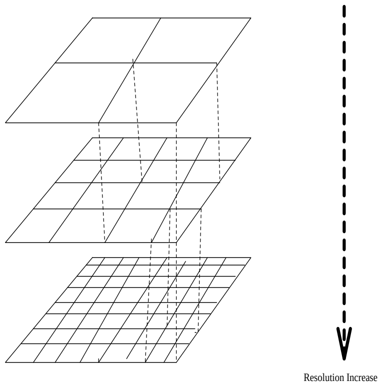

Figure 1: The pyramid structure of the GIS map

of the possibilities of oil within a given grid. Since usually we might have more than one knowledge source, we need to combine the knowledge from all the sources to form the initial probability distribution of the Map.

We suggest first use the Dempster-Shafer theory to integrate knowledge derived from a variety of sources. Then, we transform the result of this combination into a reasonable probability distribution.

Dempster-Shafer theory of plausible inference provides a natural and powerful methodology for knowledge representation and combination. The main feature of the Dempster-Shafer's approach is its flexibility in representing knowledge and its ability to combine uncertainty through inferences without requiring prior probabilities. Standard statistical methods do not perform as well in domains where prior probabilities of the necessary exactness are hard to come by.

Let $\Theta = \{a_1, \ldots, a_n\}$ be a set of propositions about the exclusive and exhaustive possibilities in a domain. Θ is called the *frame of discernment*. This Θ corresponds to all the possible grids of different resolutions of the GIS Map. It ought to be possible to partition all of one's beliefs among the different subsets of Θ, assigning to each subset A that portion which is committed to A and to nothing smaller.

If Θ is a frame of discernment, then a function

$$m : 2^{\Theta} \rightarrow [0, 1] \tag{1}$$

is called a basic probability assignment whenever

$$m(\phi) = 0 \tag{2}$$

and

$$\sum_{A \subseteq \Theta} m(A) = 1 \tag{3}$$

The quantity $m(A)$ is called A's *basic belief assignment*, and it is understood to be the measure of the belief that is committed **exactly** to A. Condition (2) reflects the fact that no belief ought to be committed to ϕ, while (3) reflects the convention that one's total belief has measure 1.

To reiterate, the quantity $m(A)$ measures the belief that one commits *exactly* to A, not the total belief that one commits to A. To obtain the measure of the total belief committed to A, one must add to $m(A)$ the quantities $m(B)$ for all proper subsets B of A:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{4}$$

In terms of this image, $m(A)$ measures the total portion of belief mass that is confined to A yet none of which is confined to any proper subset of A. In other words, $m(A)$ measures the belief mass that is confined to A but can move freely to every point of A. This freedom is a way of imaging the noncommittal nature of our belief. i.e.. it represents our ignorance because we cannot further subdivide our belief and restrict the movement.

If the belief functions m_1, m_2 of two knowledge sources KS_1, KS_2 are provided, the information supplied by KS_1, KS_2 can be pooled by computing the "orthogonal sum". The resulting belief function $m = m_1 \oplus m_2$ is

$$m(A) = \frac{\sum_{X \bigcap Y = A} m_1(X) \cdot m_2(Y)}{1 - \sum_{X \bigcap Y = \phi} m_1(X) \cdot m_2(Y)} \tag{5}$$

This operation is commutative and associative.

The search space is tessellated into a series of grids. This tessellation actually forms the frame of discernment $\Theta = \{g_1, g_2, \ldots, g_n\}$. Our purpose is to obtain an initial probability $p(g_i)$ distribution for each grid.

We need to find a rule that allows the construction of a probability distribution to be used by our sensor planning algorithm from the basic belief number of the Dempster-Shafer theory. Let $A \in 2^\theta$ $A = \{c_{i_1}, \ldots, c_{i_k}\}$, where c_{i_j} is those grids that belong to the highest resolution level. The mass $m(A)$ corresponds to that part of the belief that is restricted to A and that, due to lack of further information, cannot be allocated to a proper subset of A. In order to build the needed probability distribution, we distribute $m(A)$ equally among the atoms of A (those grids that belong to the highest resolution level of the GIS Map). Therefore, $\frac{m(A)}{n}$ is given to each c_{i_j}, $1 \leq j \leq k$. This procedure corresponds to the *Insufficient Reason Principle:* if one must build a probability distribution on n elements, given a lack of information, give a probability $\frac{1}{n}$ to each element. This procedure is repeated for each mass m. Thus, we have the following rule to calculate the probability for each grids which is not on the highest resolution level:

$$p(c_i) = \sum_{c_i \in A, A \in 2^{\theta}, M(A) \neq 0} \frac{m(A)}{\mid A \mid} \tag{6}$$

When more than one knowledge sources are available, we should first use the Dempster-Shafer combine law (5) to combine all the available knowledge sources and then transform the resulted probability number into the probability value for all the grids at every resolution level of the GIS map.

3 The Search Action Formulation

A search action means an action of digging a well to check whether a given region contains an oil field or not. It may also refer to other operations aims at determine a certain region contains oil or not. Initially, the data from a search action are preprocessed, screened and interpreted using the probability functions.

Due to various factors, the data and sample obtained from digging a well or other operations provide only indirect information about the location of the oil field. We must combine the original probability for each grid with the new action result to form the new probability for each grid.

A search action is interpreted as making an assertion about the possibility of containing oil for any points of the spatial environment. Suppose A represent a search action, consider a point (x, y, z) that can be sensed by this action. Than the detection function $D(A, (x, y, z))$ gives the probability of detecting oil given the condition that the point (x, y, z) does contain oil and that the search action is A. For each grid g, $D(A, g)$ gives the probability of detecting oil given the condition that g does contain oil and that the search action is A.

If a search action does not find the oil field, the probability for each grid $p(g)$ of the GIS Map should be updated. Let α be the event that g contains oil, α_{no} be the event that g does not contain oil, and let β be the event that after applying action A, the oil field is detected. Since the above events α, α_{no} are mutually complementary and exclusive, we get the following formula:

$$P(\alpha \mid \neg\beta) = \frac{P(\alpha)P(\neg\beta \mid \alpha)}{P(\alpha_{no})P(\neg\beta \mid \alpha_{no}) + P(\alpha)P(\neg\beta \mid \alpha)} \tag{7}$$

where $P(\alpha \mid \neg\beta)$ is the probability that the grid g contains oil, under the condition that the current action A failed to detect the target. $P(\neg\beta \mid \alpha)$ is the probability that the action A failed to detect the oil given that the g contains oil. $P(\neg\beta \mid \alpha_{no})$ is the probability that the action A failed to detect the target given that g does not contain oil. We have $P(\neg\beta \mid \alpha) = 1 - P(\beta \mid \alpha) = 1 - D(A, g)$ and $P(\neg\beta \mid \alpha_{no}) = 1$.

Therefore, after each A, the previous probability $p(g)$ will be replaced by $P(\alpha \mid \neg\beta)$. Thus, we have the following probability updating rule

$$\mathbf{p}(g) \leftarrow \frac{\mathbf{p}(g)(1 - D(A, g)}{1 - p(g) + p(g)(1 - D(A, g))}, \tag{8}$$

Thus, after each search action, the probability of each grid of the GIS map should be updated according to the above formula.

4 The Selection of the Search Actions

Suppose there are m possible search actions A_1, ..., A_m that can be selected from. But because of various limitations, we can only select N actions out of m (for example, there are n locations that we can dig a well. But because of various limitations, we can only select N out of m to really dig the well and check for the oil field). Then what actions should we select?

For a given action A, we can calculate its tendency of detecting the oil field by the following formula:

$$P(A) = \sum_{\forall g} D(A, g)p(g) \tag{9}$$

Thus, for N actions A_1^*, ..., A_N^*, the tendency of detecting the oil field can be calculated by

$$P(A_1^*, \ldots, A_N^*) = P(A_1^*) + [1 - P(A_1^*)]P(A_2^*) + \ldots + \{\prod_{i=1}^{k-1}[1 - P(A_{N-1}^*)]\}P(A_N^*)$$

The above formula can be directly calculated from the GIS Map. We restrict the calculation to be performed only on the grids of the same resolution (that is, all the grids participated on the calculation should be on a given layer). It is easy to know that the above formula is symmetric, which means that the order of A_1^*, ..., A_{N-1}^* does not matter when calculating $P(A_1^*, \ldots, A_N^*)$ with the above formula.

So, the search strategy is to select N search actions such that $P(A_1^*, \ldots, A_N^*)$ is maximized.

After a set of actions been applied, the probability for each grid should be updated based on formula 8.

5 Conclusion

In this paper, we present a theory of efficient search in GIS. By efficient search we mean to efficiently search for a given feature over a given spatial region or an information system, such as searching for a wildlife specie in a forest, searching for an oil field in a region, or searching for certain information in a GIS database. The feature to be searched has the property that this feature may present in any point of the spatial region. We present a method of obtaining the initial pyramid GIS information map from various sources; a strategy of how to select the best search action set among the possible actions; and a mechanism of incorporating the search action result into the GIS Map.

We have also studied the GIS search problem when the feature to be searched can only present in a given point in the spatial environment and we have proved that this task is NP-Complete [6]. Because of limited space, we have not discussed our theory in this paper.

The results presented in this paper is only a theoretical study of the task. We intend to continue our research along this line and perform some simulation and real experiments to test our theory.

References

[1] Charles R. Ehlschlaeger and Michael F. Goodchild. Dealing with uncertainty in categorical coverage maps: Defining, visualizing, and managing errors. In *Workshop on Geographic Information Systems, at Conference on Information and Knowledge Management*, pages 86–91, Gaithersburg MD, December 1 1994.

[2] P. Fisher. First experiments in viewshed uncertainty: Simulation the fuzzy viewshed. *Photogrammetric Engineering and Remote Sensing*, 58:1345–1352, 1992.

[3] B. O. Koopman. *Search and Screen: general principles with historical applications.* Pergaman Press, Elmsford, N.Y, 1980.

[4] H.J. Miller and M.V. Finco. Spatial search and spatial competition: A probability analysis of basic results from the spatially-restricted theory. *Annals of Regional Science*, 29:312–321.

[5] Lawrence D. Stone. *Theory of Optimal Search.* Metron, Inc., Mclean, Virginia, 1975.

[6] YongJian Ye and Yiming Ye. Searching a point feature using the gis map. In *to submit*, 1997.

[5] [3] [1] [4] [2]