Value-Driven Information Gathering*

Joshua Grass and Shlomo Zilberstein
Computer Science Department
University of Massachusetts
Amherst, MA 01003 U.S.A.
jgrass,shlomo@cs.umass.edu

Abstract

We describe a decision-theoretic approach to information gathering from a distributed network of information sources. Our approach uses an explicit representation of the user's decision model in order to plan and execute information gathering actions. The information gathering planner issues requests based on the value of information taking into account the computational resources and monetary costs of information gathering. At any given time, the system assesses the marginal value of dispatching new queries and selects the one with maximal value. When no further improvement of the comprehensive utility function is possible, the system stops gathering information and reports the results. We show that this approach has significant advantages including high performance, interruptibility, and adaptability to dynamic changes in the environment.

1 Introduction

This paper is concerned with the problem of information gathering from a large network of distributed information sources each of which is characterized by a different level of accessibility, reliability and associated costs. The problem is motivated by the rapid growth in on-line information sources such as digital libraries, professional reviews, news agencies, government agencies, as well as human experts providing a variety of services. A continued growth in information services is expected over the coming years. In addition, we anticipate that improved information retrieval (IR) and information extraction (IE) technologies will become available [Callan *et al.*, 1992; Riloff and Lehnert, 1993]. These technologies will allow a system not only to locate but also to extract necessary information from unstructured documents.

The large number of information sources that are currently emerging and their different levels of accessibility, reliability and associated costs present a complex information gathering planning problem that a human decision maker cannot possibly solve. Manual navigation and browsing through all the *relevant* information is not always feasible. The computational tradeoffs offered by a large collection of information sources and the dynamic nature of the environment make the information gathering process a complex planning and monitoring problem.

A fundamental premise of our approach to the problem is that information gathering is an intermediate step in a decision making process. We provide the system with an explicit representation of the user's decision model so that information gathering activity can be organized on the basis of its marginal contribution to the quality of the decision. This work extends the scope of current state-of-the-art information gathering systems by providing an answer to a decision problem rather than collecting the relevant data. Preliminary evaluation shows that when operating under resource constraints (related to cost of information access and limited amount of time), our information gathering strategy leads to significant performance improvements.

Although much work has been done on information gathering [Birmingham *et al.*, 1995; Gio, 1992; Oates *et al.*, 1994] and decision making [Howard and Matheson, 1984; Jensen and Liang, 1994; Horvitz and Peot, 1996] separately, little work has capitalized on the synergy that can develop when these two problems are solved together [Knoblock *et al.*, 1995; Nagendra Prasad *et al.*, 1996; Zilberstein and Lesser, 1996]. Previous work on using information value theory has concentrated on a small set of information sources with little uncertainty regarding the computational resources and costs of gathering information [Pearl, 1988; Jensen and Liang, 1994]. In contrast, the large number of alternative sources and the high degree of uncertainty regarding response time are the focus of the Value-Driven Information Gathering (VDIG) system described in this paper. (The problem of information extraction is not addressed directly in this paper, but it is part of a collaborative project that we are working on.)

VDIG uses the value of information derived from the decision making process to prioritize the search process. It also uses the partial results found during search to reevaluate its decision. The system has four major components shown in Figure 1.

The decision model is a probabilistic model of the user's decision-making task. We represent decision models using influence diagrams and use standard algorithms for belief propagation and for calculating the utility of actions.

*Support for this work was provided by the National Science Foundation under grants IRI-9624992 and IRI-9634938, and by the U.S. Air Force under grant F30602-95-1-0012.

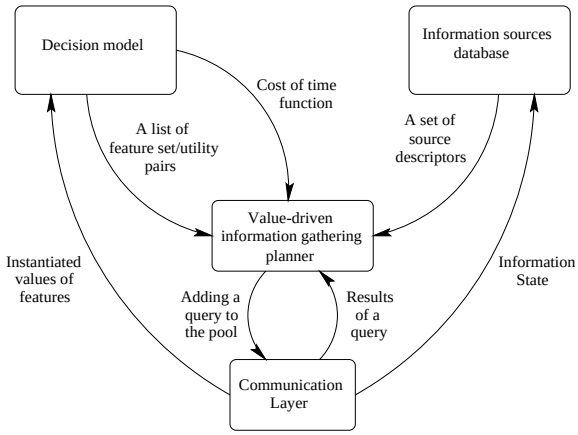

Figure 1: The major components of the Value Driven Information Gathering System.

The database of information sources characterizes each source of information in terms of the type of information it can provide, the associated cost, probabilistic information about response time, and additional information that tells the system how to interact with the source.

The communication layer maintains the pool of queries that have been sent out but have not yet been answered. It launches queries in the correct format and protocol, and it extracts information from the returned data and sends it to the decision model component.

The value-drive information gathering process is responsible for planning and monitoring the overall information gathering activity. It uses information from the other three components to determine what information gathering action to take.

The rest of the paper describes the components of the VDIG system in detail. Section 2 describes the representation of decision models and the environment of information sources. In Section 3, we describe the decision-theoretic information gathering strategy that we implemented. A complete implementation of the system is described in Section 4 along with preliminary results and evaluation. We conclude with a discussion of the benefits of VDIG and further work needed in order to apply the system to large-scale practical problems.

2 Decision Models and Information Sources

This section describes the two primary components that define the information gathering task, namely the user's decision model and the description of the environment of information sources from which information can be gathered.

2.1 Decision Models as Influence Diagrams

We use a standard influence diagram to represent the user's decision model. Diagrams representing many useful tasks can be stored in a library so that users would only need to retrieve the diagram and modify the utility function to reflect their specific preference structure. We prefer this representation because of the availability of efficient algorithms for belief propagation and because it provides sufficient data to assess the value of missing information. However our open system architecture allows for future integration with different representations of decision models.

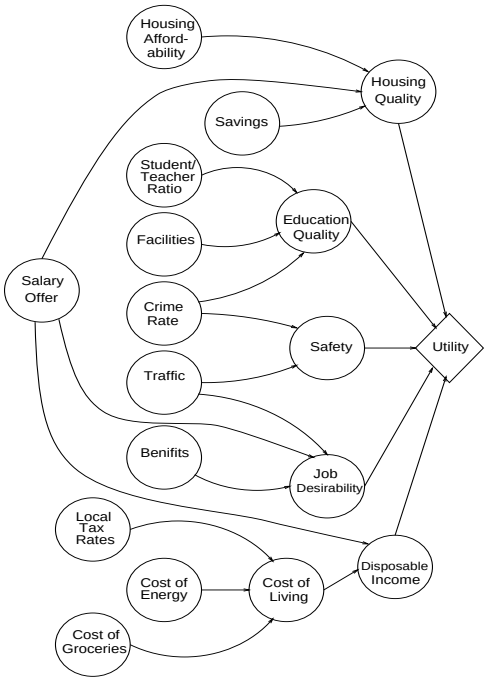

Figure 2: An influence diagram used to decide whether to take a new job that requires relocation.

To illustrate the operation of the VDIG system and to evaluate its behavior, we have constructed a sample decision model, shown in Figure 2, and a corresponding simulated information environment. The decision model is designed to help a person evaluate a new job offer that requires relocation. To make the diagram more readable, the binary decision (accept or reject offer) is not represented explicitly in the graph. The user's utility function is based on five major factors: `housing quality`, `education quality`, `safety`, `job desirability` and `dispensable income`. Each factor, referred to as features in this paper, has a small set of discrete values. For example, `education quality` can be excellent, good, or poor. Each feature may depend on a variety of other features shown in the diagram.

At any given time, standard algorithms for evaluating influence diagrams allow the system to compute the best action and the its associated utility. The value of information (VOI) associated with a single feature is the increase in the expected utility as a result of knowing the precise value of this feature. (The current version of the system is based on information sources that return perfect information. We will relax this assumption in future versions of the system.) It is well known that the

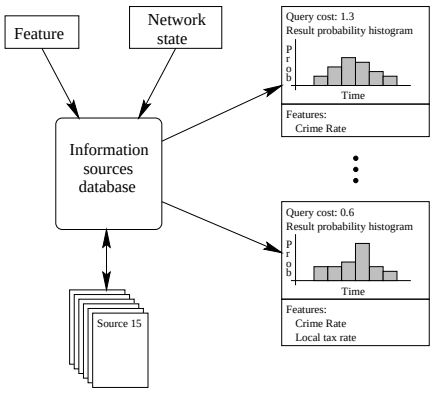

Figure 3: The database of information sources

value of information is, in general, non-additive [Howard, 1966]. In other words, the value of information of a set of features is not necessarily the sum of the value of each feature. Our system is capable of calculating the exact value of information of any set of features, but due to the high computational complexity we limit the size of the set by a small constant. This allow the system to make decisions in a non-myopic way with limited computational effort.

2.2 The Environment of Information Sources

The VDIG system uses a database to describe all the available information sources and their characteristics[1]. The database maintains information about the type of information that is available, the associated cost, probabilistic information about response time, and additional information needed to interact with the information source and extract the necessary information from the data it returns.

Figure 3 shows the contents of the information sources database. The responsiveness of each information source is described by a histogram showing the probability of information being returned at a given time after the query is sent (using a discrete representation of time), we refer to this as the result probability histogram. The histogram depends on the state of the information network, capturing such parameters as load and time of day. The histogram represents the uncertainty regarding the delay between the time a query is issued and the time the information becomes available, including both communication time and information extraction time. Once a query q is issued, the system retrieves from the database appropriate response histogram, $H_q(f|t)$, representing the probability of feature f becoming available t time steps from the current time.

Each information source has a fixed cost charged when a query is issued (regardless of whether the user waits for the response). We have examined more complex cost models that may be used in future versions of the system. Additional information related to interaction with each information source is not discussed in this paper.

3 Information Gathering Strategies

An information gathering strategy is a method for making the following two decisions:

1. Selecting a new query and sending it to a particular information source.
2. Deciding when to stop the process of information gathering and report the best decision based on all the available information.

The VDIG system uses a decision-theoretic approach to making these decisions in an attempt to maximize the comprehensive utility function. At any given time, the system identifies a small set of k features that are the focus of information gathering activity. Those are the k features with highest *individual* VOI[2]. Once the set of features is determined, the system starts to issue queries and monitor their execution. The following two subsections give a detailed description of how the system selects queries and how it decides to stop gathering information.

3.1 Best query selection

The VDIG system activates queries based on their marginal value with respect to the currently active information gathering process. The current state of the information gathering process is characterized by the existing active queries in the query poll $Q = \{q_1, ..., q_n\}$. The system maintains the activation time for each query q_i and uses it in order to dynamically update the probability histogram of information arrival time. For each feature, f, the system calculates a comprehensive histogram of information arrival time taking into account all the relevant queries in the query pool (assuming independence of information sources in terms of their response time). Let $H_Q(f|t)$ represent probability of the query pool Q returning the value of feature f at time t (all times are measured relative to the current time). Then, the probability of the query pool returning the value of a feature within n time units is:

$$P_Q(f|n) = \sum_{t=1}^{n} H_Q(f|t) \quad (1)$$

The system can also determine the probability that the current query pool will provide the value of a specific set of features s at a specific time t. This value is calculated

[1] We do not refer in this paper to any specific information environment. An information source can be a single WWW site or, more interestingly, a smart search engine. We plan to integrate the VDIG system with a specific search engine currently under development.

[2] The complexity of computing information value over subsets of features forces us to focus on a small set of features that *appear* to be the most valuable ones. The selection of the set is *myopic*, but once queries are activated, their VOI takes into account mutual information of features included in the focus set.

by multiplying the probability of finding each feature in s by the probability of not finding each feature not in s.

$$P_Q(s|t) = \prod_{f \in s} P_Q(f|t) \prod_{f \notin s} (1 - P_Q(f|t)) \quad (2)$$

To compute the comprehensive utility of the current query pool at any given time t, the system computes the expected value of the returned information minus the associated cost. When computing the expected value the system does not know what subset of information will be available at time t. Therefore, it averages over all possible subsets of features, s, and the corresponding value of information $V(s)$.

$$U(Q|t) = \sum_{s \subset \{f^1, ..., f^k\}} P_Q(s|t)V(s) - C(t) - C(Q) \quad (3)$$

Note that $V(s)$ is the comprehensive value of information of the set of features s with respect to the decision model (this value is computed in a non-myopic way, taking into account mutual information). $C(t)$ represents the cost of time and $C(Q)$ represent the monetary cost of all the queries in the query pool. Obviously, the complexity of computing the comprehensive value of the query pool grows exponentially with the size of the feature set under consideration. That is why we limit the system to gather information on a small focus set of up to k features, $\{f^1, ..., f^k\}$.

Because the VDIG system decides when to halt, it determines the utility of the query pool based on the time that maximizes the comprehensive utility defined in Equation 3.

$$U(Q) = arg \max_t U(Q|t) \quad (4)$$

Finally, the system computes the marginal value of a query, q, based on the increase in the expected utility of the query pool.

$$MV(q) = U(Q \cup \{q\}) - U(Q) \quad (5)$$

At any given time, the system considers all possible queries that may contribute information on the features in the current focus set. It selects the query that has the highest marginal contribution, $MV(q)$. If no query has a positive marginal value, the system will not issue any new queries at the current time slice. Later on, the system may issue new queries if some information source fails to return valuable information or if the value associated with some feature is increased based on the actual information acquired.

3.2 Stopping criterion

A decision by the system to stop gathering information is reached when the cost of information gathering becomes greater than the benefits. In other words, when it reaches a *global* maximum of the comprehensive utility function of the query pool. The stopping criterion is:

$$\forall t : U(Q|t) \leq U(Q|0) \quad (6)$$

Since the system re-evaluates the stopping criterion every time slice, it can handle possible changes in the cost model in a dynamic environment or it can be simply interrupted by the user if necessary.

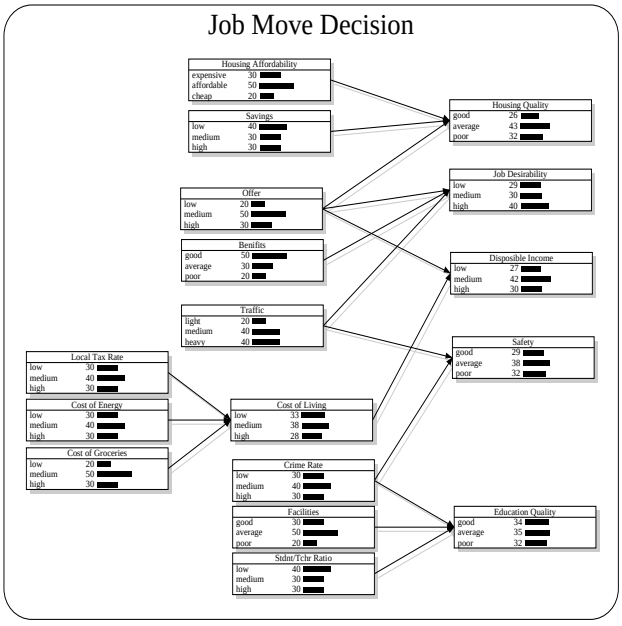

Figure 4: The underlying belief network used by the system to evaluate a new job offer.

4 Experimental Results

We have implemented the VDIG system and tested it using a simulated information environment. The planner, information sources database and communication layer were developed in Common Lisp, and the decision model is implemented using the Hugin belief network library. In part one of this section we describe the implementation of the VDIG system and examine the system gathering information for a moderately sized decision model. Part two compares the VDIG system with three other approaches to solving decision problems.

4.1 Implementation

We tested the VDIG system using the decision model introduced in Section 2.1. The main decision is whether to take a new job that requires relocation. In order to make this type of decision, information has to be collected from many disparate sources. Figure 4 shows the underlying belief network representing this decision model. The probabilistic information attached to each node represent the initial state before information is gathered.

`job desirability`	1.560
`disposable income`	1.405
`safety`	1.180
`housing quality`	1.100

Figure 5: The initial value of information for the four most valuable nodes

When the VDIG system is initiated it first evaluates the value of information for each individual node in the decision model. The value of information calculated at

job desirability	disposable income	safety	housing quality	VOI
x				1.560
	x			1.405
x	x			2.0
		x		1.180
		...		
	x	x	x	2.179
x	x	x	x	2.538

Figure 6: The value of information for each possible set of nodes in the focus set

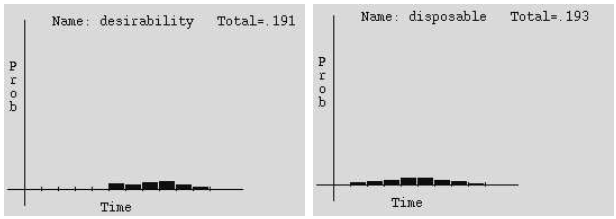

Figure 7: The result probability histograms at time 0.

this stage does not take into account the cost of gathering information. It is calculated using the classical definition of value of information which is the increase in the expected utility of the best action once the information is acquired [Pearl, 1988]. Table 5 shows the four highest scoring nodes in the decision model that together define the current focus set of features. Table 6 shows the value of different sets of features in the focus set. These set values are used in equation 2 to determine the comprehensive utility.

The value of information for each node changes as nodes are instantiated, so when information is returned by the communication layer, the value of information for the remaining uninstantiated nodes must be updated.

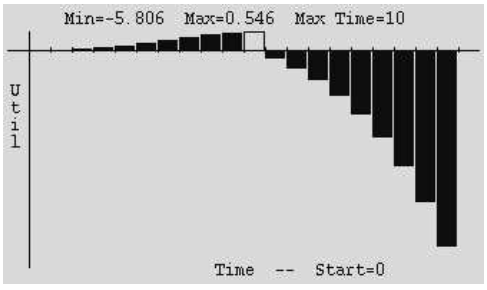

Figure 8: The comprehensive utility at time 0.

Using the value of information for each possible set and the information sources database, the VDIG system can calculate the marginal utility for activating any query. Our implementation picks the two highest scoring queries and activates them. In our example the two queries that were launched during the first time slice return `job desirability` and `disposable income`(see Figure 7). Launching these two queries increased the expected utility by 0.546. The comprehensive utility curve

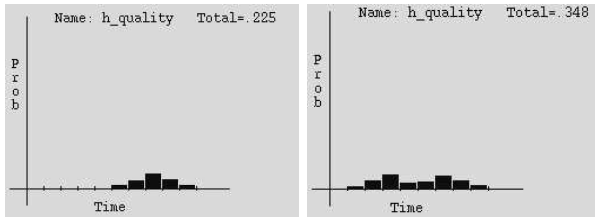

Figure 9: The result probability histogram at time 1(left) and 2(right) for the feature `housing quality`.

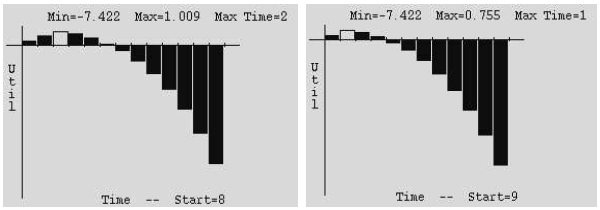

Figure 10: Utility curve at time 8(left) and 9(right).

is shown in Figure 8 with the white bar indicating the maximum utility point.

At time slice 1 the VDIG system launches a query that will return the `housing quality` and `safety`. Both the `job desirability` and the `disposable income` probability histograms are unaffected. At time slice 2 the VDIG system adds two more queries that also return `housing quality` and `safety`. Figure 9 shows how adding another query for `housing quality` changes the result probability histogram.

The VDIG system continues adding queries to the query pool as long as the marginal utility increases. As information starts arriving, the change in the value of additional information can decrease causing the VDIG system to halt. At time slice 9, information on `housing quality` was returned. Figure 10 shows how the comprehensive utility decreased from 1.009 to 0.755. In this example, the VDIG system stops at time 10 after receiving the value of `safety`.

4.2 Comparison

We compared the VDIG system against three other information gathering strategies. The first takes the best action (take the job) given the initial decision model without gathering any information. The second is a naive strategy that does not use the decision model and the responsiveness of the information sources. It gathers information in an attempt to cover all of the relevant features regardless of their corresponding values. The third strategy is an ideal approach that makes the optimal decision in each case (pretending that accurate information is available at no cost). This is the upper bound on the performance of any information gathering technique. The results are summarized in Table 11.

The table highlights the importance of a value driven approach to information gathering. The naive coverage approach, given the same constraints, does not score sig-

	Always Move	Coverage	VDIG System	Best case
Utility	0.900	1.150	2.066	3.090
Correct	59%	60%	86%	100%

Figure 11: A comparison of different information gathering strategies

nificantly higher then the base-line approach of making a decision without any knowledge. This is due largely to the fact that sources in this information environment had a significant uncertainty regarding the timing and success of returning information. The VDIG system could compensate for this by querying multiple information sources when the chance of any one returning a result was low, and not wasting resources trying to gather pieces of information that were of little use to the overall decision.

5 Conclusion

In this paper we have described a value-driven information gathering system which uses information from both the decision side and the acquisition side of an information gathering problem in order to implement an effective information gathering strategy. With the growth in the number of distributed information systems and the increasing strain placed on these systems by users, a value driven approach to information gathering can gather more beneficial information under the same time and resource constraints than other approaches. Although further evaluation is needed to fully understand the benefits of our system, it is clear that value driven information gathering works best with problems in which the information environment offers multiple sources for the same information with different cost and response characteristics. The benefits of the system grow when there is a high degree of uncertainty regarding the performance of each information source.

Further work on the VDIG system is aimed at integration with existing search engines and information extraction techniques, handling inaccurate and conflicting information from different information sources, and enhancing the interaction with the information sources and the cost of information model. Our overall goal in this research is to create a system that takes advantage of a deep understanding of its external and internal resources in order to gather information in the most effective way.

References

[Birmingham *et al.*, 1995] W. P. Birmingham, E. H. Durfee, T. Mullen, and M. P. Wellman. The distributed agent architecture of the University of Michigan digital library. *AAAI Spring Symposium on Information Gathering in Heterogeneous, Distributed Environments*, Stanford, California, 1995.

[Callan *et al.*, 1992] J. Callan, W. B. Croft, and S. Harding. The inquery retrieval system. *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pp. 78–83, 1992.

[Gio, 1992] W. Gio. The roles of artificial intelligence in information systems. *Journal of Intelligent Information Systems*, 11(1):35–56, 1992.

[Horvitz and Peot, 1996] E. Horvitz and M. Peot. Flexible strategies for computing the value of information in diagnostic systems. *Proceedings of the AAAI Fall Symposium on Flexible Computation in Intelligence Systems*, pp. 89–95, Cambridge, Massachusetts, 1996.

[Howard, 1966] R. A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, SSC-2(1):22–26, 1966.

[Howard and Matheson, 1984] R. A. Howard and J. E. Matheson. Influence diagrams. *Principles and applications of decision analysis*, 2, 1984.

[Knoblock *et al.*, 1995] Knoblock, Craig A., and Levy, Alon Y. Exploiting Run-Time Information for Efficient Processing of Queries. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, Stanford University, Stanford, CA, March, 1995.

[Jensen and Liang, 1994] F. Jensen and J. Liang. A system for value of information in bayesian networks. *Proceedings of the 1994 Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 178–183, 1994.

[Nagendra Prasad *et al.*, 1996] M. V. Nagendra Prasad, V. Lesser, and S. Lander. Reasoning and retrieval in distributed case bases. *Journal of Visual Communication and Image Representation*, Special Issue on Digital Libraries, 1996. To appear.

[Oates *et al.*, 1994] T. Oates, M. Nagendra Prasad, and V. Lesser. Cooperative information gathering: A distributed problem-solving approach. Technical Report 66, University of Massachusetts, 1994.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, Los Altos, California, 1988.

[Riloff and Lehnert, 1993] E. Riloff and W. Lehnert. Automated dictionary construction for information extraction from text. In *Proceedings of the ninth IEEE Conference on Artificial Intelligence for Applications*, pages 93–99, 1993.

[Zilberstein, 1996] S. Zilberstein. The use of anytime algorithms in intelligent systems. *AI Magazine*, 17(3):73–83, 1996.

[Zilberstein and Lesser, 1996] S. Zilberstein and V. Lesser. Intelligent information gathering using decision models. Technical Report 96-35, Computer Science Department, University of Massachusetts, 1996.