

Planning information gathering under uncertainty

Joshua Grass and Shlomo Zilberstein
Computer Science Department
University of Massachusetts at Amherst
CMPSCI Technical Report 97-32
May 21, 1997

Planning Information Gathering Under Uncertainty

Joshua Grass and Shlomo Zilberstein
Computer Science Department
University of Massachusetts
Amherst, MA 01003 U.S.A.
{jgrass,shlomo}@cs.umass.edu
fax: (413)-545-1249

Keywords: information gathering, resource-bounded planning, decision theoretic planning

Abstract

We describe a decision-theoretic approach to information gathering from a distributed network of uncertain information sources. Our approach uses an explicit representation of the user's decision model and a database of responsiveness information about information sources in order to plan and execute information gathering actions. The information gathering planner issues requests based on the value of a query taking into account characteristics of the information sources and the value of information to the decision. At any given time, the system assesses the marginal value of dispatching a new query and selects the one with maximal value. When a set of new queries are dispatched the comprehensive utility function is altered, this function predicts the utility of the system at any point in the future. 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, as well as human experts providing a variety of services. A continued growth in information services is expected over the coming years. In addition, improved information retrieval (IR) and information extraction (IE) technologies are becoming available [2, 12]. These technologies 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 [1, 3, 9] and decision making [6, 7, 4] separately, little work has capitalized on the synergy that can develop when these two problems are solved together [8, 11, 13]. 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 [10, 7]. 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 a query, derived from the decision making process and the responsiveness of potential information sources, to prioritize the search process. We define *responsiveness* as the probability of an information source returning a result at any given point in the future. Responsiveness is represented by a result probability histogram, described latter in this paper. VDIG 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.

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

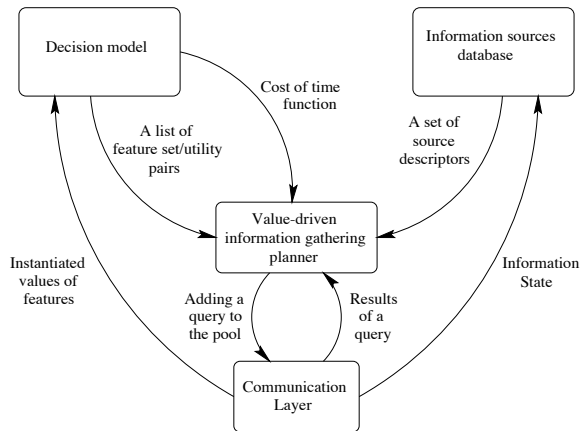


Figure 1: The major components of the Value Driven Information Gathering System.

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 and predict the future state of the system based on the query pool and the responsiveness of each information source.

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 planning 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

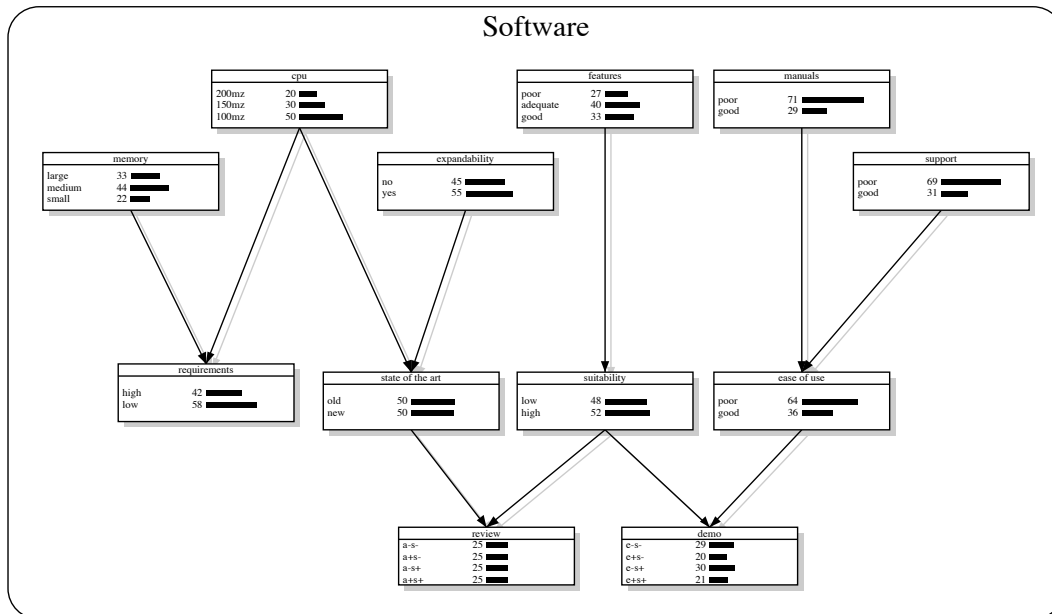


Figure 2: An influence diagram used to decide whether to purchase a piece of software. Note: The decision and utility nodes are not shown.

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.

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 whether they should purchase a piece of software. This diagram only shows the belief network component of the influence diagram. The utility node is connected to each of the four high level nodes (**requirements**, **state of the art**, **suitability** and **ease of use**) and the decision node (buy or skip). The user's utility function is based on the value of the four high-level features, and these features may not be directly ascertained. Each features has a small set of discrete values, for example, **memory**, which represents the minimum memory requirements of the software may only have the values of small, medium and large.

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 by incorporating evidence nodes using the Noisy-OR operation in future versions of the system.) It is well known that the value of information is, in general, non-additive [5]. 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

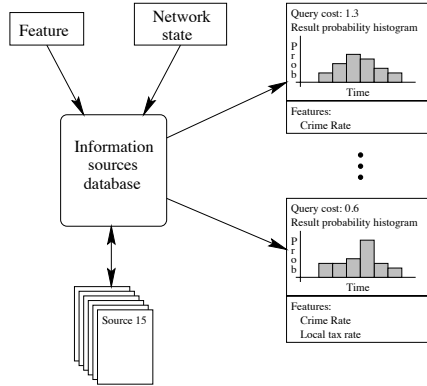


Figure 3: The database of information sources

the exact value of information of any set of features, but due to the high computational complexity we limit the size of this set to a small constant. This allows 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 characteristics¹. The database maintains entries containing 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.

¹We 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.

3 Information Gathering Planning

The information gathering planner is responsible 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 mix of decision-theoretic planning (based on the user's decision model) and planning under uncertainty (based on the database of information sources) 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* VOI². Once the set of features is determined, the system analyzes which information source will increase the comprehensive utility function the most. It then issues these queries and monitors 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 pool $Q = \{q_1, \dots, q_n\}$. 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 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

²The 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.

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, \mathbf{s} , and the corresponding value of information $V(\mathbf{s})$.

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

Note that $V(\mathbf{s})$ is the comprehensive value of information of the set of features \mathbf{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, $\{f^1, \dots, f^k\}$.

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.

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 or not to purchase a piece of software. In order to make this type of decision, information has to be collected from many disparate sources. Figure 2 shows the underlying belief network representing this decision model. The probabilistic information attached to each node represent the initial state before information is gathered.

demo	0.788
cpu	0.466
features	0.390
review	0.327

Figure 4: The initial value of information for the four most valuable nodes

When the VDIG system is first activated, it first evaluates the value of information for each individual node in the decision model. The value of information calculated at this stage does not take into account the cost of gathering information. The value of information is calculated by determining the increase in the expected utility of the best action once the information is acquired [10]. Figure 4 shows the four highest scoring nodes in the decision model that together define the current focus set of features. Figure 5 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.

demo	cpu	features	review	VOI
x				0.788
	x			0.466
x	x			0.884
		x		0.390
		...		
	x	x	x	0.693
x	x	x	x	0.947

Figure 5: The value of information for each possible set of nodes in the focus set

The information sources database contains characteristics for the potential sites that the system might use. In figure 6 we can see the result probability histogram for one information source.

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.

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

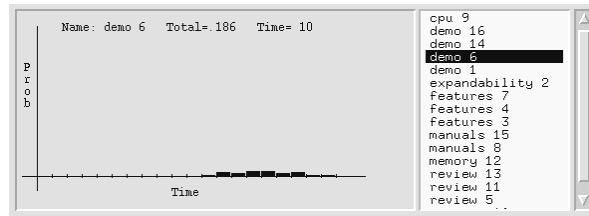


Figure 6: The information sources database

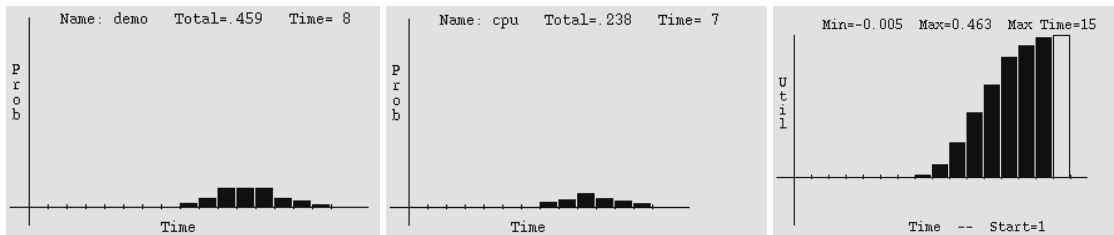


Figure 7: The result probability histograms(left, center) and the comprehensive utility at time step 1.

scoring queries and activates them. In our example the two queries that were launched during the first time slice return `demo` and `cpu`(see Figure 7). Launching these two queries increased the expected utility by 0.463. The comprehensive utility curve is shown in Figure 8(right) with the white bar indicating the maximum utility point.

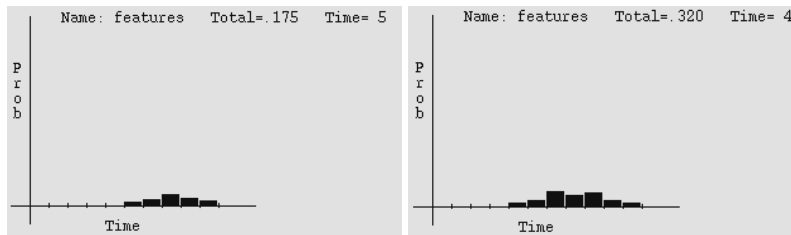


Figure 8: The result probability histogram at time 3(left) and 4(right) for the feature `features`.

At time slice 3 the VDIG system launches a query that will return the `features` node. Both the `demo` and the `cpu` probability histograms are unaffected. At time slice 4 the VDIG system adds another query that also returns `features`. Figure 8 shows how adding another query for `features` 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 12, information on `review` was returned. Figure 9 shows how the expected future comprehensive utility decreased from 0.393 to 0.232. In this example, the VDIG system stops at time

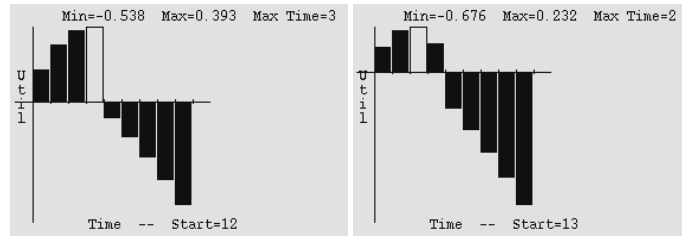


Figure 9: Utility curve at time 12(left) and 13(right).

15 after receiving the value of cpu.

4.2 Comparison

We compared the VDIG system against three other information gathering strategies. The first takes the best action (buy the software) 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 figure 10.

	Always Buy	Coverage	VDIG System	Best case
Utility	1.120	2.090	3.055	4.040
Correct	60%	62%	76%	100%

Figure 10: A comparison of different information gathering strategies

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

We have described a value-driven information gathering planning system which uses information from both the decision side and the acquisition side of an information gathering problem in order to plan 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 take advantage of the growing uncertainty in the responsiveness of information sources to 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 in an uncertain world.

References

- [1] W. P. Birmingham, E. H. Durfee, T. Mullen, and M. P. Wellman. The distributed agent architecture of the university of michigan digital library. In *AAAI Spring Symposium on Information Gathering in Heterogeneous, Distributed Environments*, pages 19–24, Stanford, CA, 1995.
- [2] J. Callan, W. B. Croft, and S. Harding. The inquiry retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.
- [3] W. Gio. The roles of artificial intelligence in information systems. *Journal of Intelligent Information Systems*, 11(1):35–56, 1992.
- [4] E. Horvitz and M. Peot. Flexible strategies for computing the value of information in diagnostic systems. In *Proceedings of the AAAI Fall Symposium on Flexible Computation in Intelligence Systems*, pages 89–95, Cambridge, Massachusetts, 1996.
- [5] R. A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.
- [6] R. A. Howard and J. E. Matheson. Influence diagrams. *Principles and applications of decision analysis*, 2, 1984.
- [7] F. Jensen and J. Liang. A system for value of information in bayesian networks. In *Proceedings of the 1994 Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 178–183, 1994.
- [8] Craig A. Knoblock and Alon Y Levy. 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, CA, March 1995. Stanford University.
- [9] T. Oates, M. Nagendra Prasad, and V. Lesser. Cooperative information gathering: A distributed problem-solving approach. Technical Report 66, University of Massachusetts, 1994.

- [10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, Los Altos, California, 1988.
- [11] 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.
- [12] 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.
- [13] S. Zilberstein and V. Lesser. Intelligent information gathering using decision models. Technical Report 35, Computer Science Department, University of Massachusetts, 1996.