



A Value-Driven System for Autonomous Information Gathering*

JOSHUA GRASS

SHLOMO ZILBERSTEIN

Computer Science Department, University of Massachusetts, Amherst, MA 01003, USA

jgrass@cs.umass.edu

shlomo@cs.umass.edu

Abstract. This paper presents a system for autonomous information gathering in an information rich domain under time and monetary resource restrictions. The system gathers information using an explicit representation of the user's decision model and a database of information sources. Information gathering is performed by repeatedly selecting the query with the highest marginal value. This value is determined by the value of the information with respect to the decision being made, the responsiveness of the information source, and a given resource cost function. Finally, we compare the value-driven approach to several base-line techniques and show that the overhead of the meta-level control is made up for by the increased decision quality.

Keywords: autonomous information gathering, resource-bounded reasoning, planning, model-based reasoning, real-time control

1. Introduction

This paper is concerned with the construction of a system for autonomous information gathering from a large network of distributed information sources such as the World-Wide-Web (WWW). Special attention is paid to the different characteristics of each information source in regards to its role in the decision, the quality of the information it contains, the cost of accessing the site, how quickly the site responds to a query, and how easy it is to extract information from the response and incorporate it into the decision model.

The rapid growth of the WWW over recent years presents a serious challenge for AI researchers to build systems that can exploit this information in order to solve problems and make decisions. The WWW presents a domain in which information is so abundant, so varied and so loosely organized that often times the question is not "Where can I find relevant information?" but "How quickly can I access it? How accurate is it? How does it affect my decision?" Viewing the WWW as a large homogeneous database may be unproductive. It is more beneficial to view it as a varied landscape in which some resources are more useful than others and some resources are more difficult to reach than others. The lack of a standard format for this information must also be considered as well as the computational resources required to translate information from its raw form into a form the decision model can use.

*Support for this work was provided in part by the National Science Foundation under grants IRI-9624992, IRI-9634938, and INT-9612092.

These aspects of the WWW make it a challenging domain that requires effective systems to take into account the characteristics of the environment in which they operate. The ability to judge information sources both by their content and the resources required to access and use that information can be applied to many other complex domains that involve information gathering actions.

To achieve the goal of building an intelligent information gathering planner in the WWW our system addresses several problems: locating useful information sources, requesting the right information, extracting it from the response, determining the accuracy of the information, and integrating the results into the decision making process.

Value-driven information gathering (VDIG) complements current work in search on the Internet, as it focuses on the query selection problem using response time and cost as opposed to the coverage problem. We are not attempting to find the perfect information source, or all of them. Instead, we are attempting to query a subset of available resources until it is no longer advantageous for us to continue because the cost of doing so outweighs the potential benefits. Many systems today do an adequate job of finding information given a query or searching a large number of sources for the best instance, but none consider the problem of selecting a good subset of sources that provide high-quality information. Our approach differs from existing work on information gathering in which the primary goal is to achieve coverage over a set of information sources (Doorenbos et al., 1996), to discover appropriate information sources (Selberg and Etzioni, 1997), or to merge disparate information sources (Ashish and Knoblock, 1997; Ambite and Knoblock, 1997).

VDIG is concerned primarily with using a decision model and a set of information sources in order to gather information and arrive at a decision in a timely manner. Figure 1 shows an overview of the value-driven information gathering system. The user selects a decision problem, specifies the value of resources (time and money) and a set of preferences used to modify a decision model constructed by an expert. Using this information, the system selectively queries a subset of available information sources and reaches a decision within the set of resource constraints specified by the user.

Value-driven information gathering is applicable in domains that have the following characteristics.

1. A decision model is available that allows the system to make decisions with incomplete information. The quality of the decision increases as more information becomes available.
2. The decision model can be used to determine the value of information (Howard, 1966; Howard and Matheson, 1984; Pearl, 1988) for a set of information items used in the decision.
3. There are multiple information sources for each piece of information used in the decision. These sources have different monetary costs and response times.
4. The user operates with limited resources (time and money) and cannot gather all the relevant information before making a decision.
5. Information sources return information items that can be incorporated into the decision model.

This paper shows that in such environments, value-driven information gathering can greatly increase the quality of a decision over other approaches given the same resources.

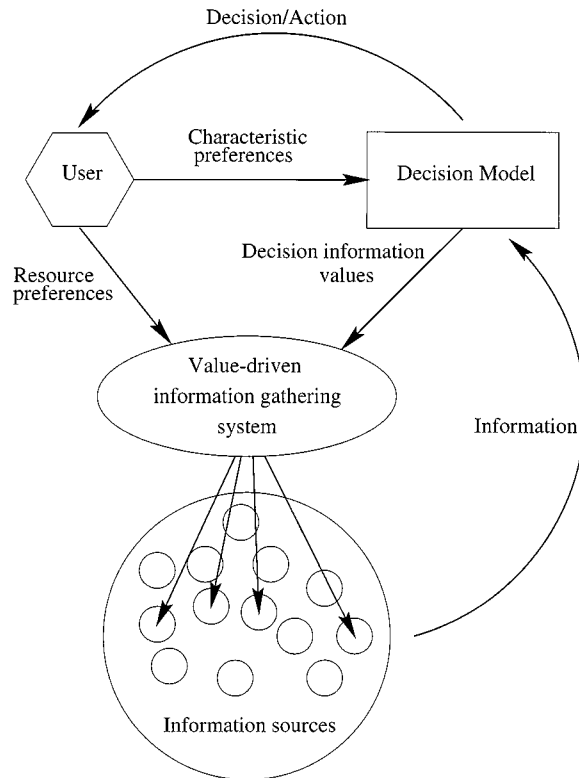


Figure 1. Value-driven information gathering.

We have implemented a prototype system and used it for gathering information on the WWW using influence diagrams to represent decision models. The WWW meets the criteria outlined above and VDIG has worked effectively in both simulations and real-world testing. It is important to note that, although we have built this system for testing in the WWW, value-driven information gathering can also be applied to other domains: Examples include image understanding (Jaynes et al., 1998; Marengoni et al., 1999) and signal processing (Klassner, 1996) where computationally expensive image/sound processing modules are treated as information sources. The WWW offers an excellent domain for value-driven information gathering systems to accomplish a number of tasks: product purchasing, vacation/travel planning, evaluation of a job offer, finding technical support, and other tasks.

The VDIG system architecture integrates techniques from several different fields: decision-making under uncertainty, information extraction, database systems, and resource-bounded planning. The system maintains a dynamic pool of queries and a comprehensive utility function that gives the expected net gain in decision quality as a function of time (given that no more queries are made in the future). As long as new queries have positive marginal values, the query with maximal value is added to the pool. As long as the comprehensive utility function increase in the future, the system continues to gather information.

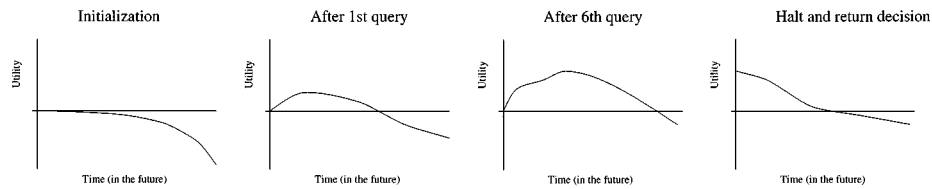


Figure 2. Comprehensive utility function over the course of making a decision.

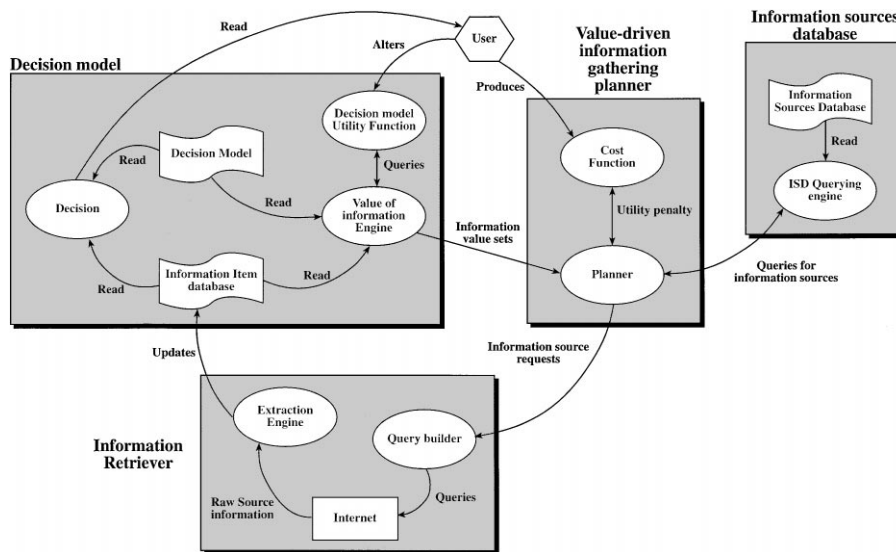


Figure 3. The components of a value-driven information gathering system.

Figure 2 illustrates the evolution of the comprehensive utility function over the life-time of a decision making event. Initially the comprehensive utility function is simply equal to the cost function with monetary resources being spent equal to zero. As queries are added to the query pool the comprehensive utility function changes depending on the value and cost of the information sources. Since information sources do not return information immediately, the effect of adding an information source to the query pool is that the curve of the comprehensive utility function increases at some point in the future. Eventually, the potential information gained by querying information sources will be outweighed by the time and monetary costs of making a query and the system will make a decision using the information currently available. The decision-point occurs when the graph representing expected utility has reached a maximum at the current time.

The VDIG system has an open architecture (shown in figure 3) that employs an object model in which components communicate with each other through a defined vocabulary of messages. This architecture makes it relatively simple to replace individual components and apply the system to other domains.

The rest of this paper describes the components of the system in detail and evaluates its operation. Section 2 describes related work in automated information gathering and how it differs from our value-driven approach. Section 3 formally describes the problem which VDIG attempts to solve. The implementation of the system is described in Section 4. Section 5 illustrates the operation of the system with a simple problem instance. Section 6 includes an experimental evaluation of the operation of the system. And Section 7 summarizes the lessons learned from the construction of the system and further work.

2. Background

Information gathering from the World-Wide-Web is a major challenge and an expanding research area within AI. Recent applications focus on such problems as complex query answering and product selection. In this section, we discuss several integrated approaches to information gathering that share some features and motivation with our approach. We also point out the distinctions between these systems and VDIG.

At the University of Washington, a number of researchers are currently working on building high-level information agents that process information directly from web pages and from low-level agents (e.g. search engines). They view the Internet as an information ecology which at the present time is filled with raw information and information herbivores (Etzioni, 1996). Their goal is to construct information carnivores that can use information generated by low-level agents as well as by directly accessing web pages in order to further process information before it is passed on to the user. Specific projects include Metacrawler (Selberg and Etzioni, 1997), which queries a set of search engines on the web and then processes the results from the search engines in order to return a list of pages that best match a query, and ShopBot (Doorenbos et al., 1996) that queries several software sites to find the best price for a particular product requested. It uses a simple natural language extraction engine to both identify the price and make sure that the product is the latest version and for the correct platform. VDIG differs from these systems in its ability to integrate the results of the search process into a decision-making process and in its ability to reason about resources and a broad variety of information sources and decision models to schedule and monitor the information gathering process.

The Stanford information group is currently working on a system called Infomaster. This system is similar to the high-level information agent work in that it correlates information from a broad variety of information sources in order to return answers to more advanced queries than any one information source could return individually. Infomaster does this by using the Agent Communication Language (A simplified predicate logic). This language specifies how to break down queries and also how to compute a result from the components. An example of an Infomaster query would be, "Find me all available housing within two miles of campus." Infomaster would translate this query into a rule that would query a listing of available housing, and for each free house query a map engine to determine the distance to campus, and then only return houses that were within two miles of campus. Unlike Infomaster, VDIG deals with incomplete information and resource restrictions. For the Infomaster system, each component of the query rule is of equal importance because all are necessary to returning the final result. Infomaster may be able to use alternative

information sources if a source fails, but there is no information about the cost, responsiveness, or value of the information source. Also, there is no uncertainty involved with any of the decisions that the system makes, making it unable to return a partial response with information about its confidence in the result.

The integration of information gathering with problem solving has been studied extensively within the Multi-Agent Systems (MAS) Laboratory at the University of Massachusetts. Researchers at the MAS lab have developed an effective system for information gathering using TAEMS task-structures (Lesser et al., 1997) and the RESUN planner (Carver and Lesser, 1995). These task structures represent possible search strategies to use for collecting information needed for making a decision (e.g., deciding which type of car is a better purchase). These task structures contain not only the method for using the information (for example, finding the car with the minimum purchase price), but also pointers to sites where the information may be located. Also contained in the TAEMS task structure is information about the quality, cost and time expectations of each information source. The design-to-criteria planner can then use the TAEMS structure to build an information gathering plan that not only meets the users cost, duration and quality requirements, but it can also control the variance of these three factors. For example a user could specify that they not only want a high quality result, but they also value low variance on the amount of time the plan will take. Because selecting an appropriate information gathering plan from a TAEMS task structure is computationally intractable, a heuristic search algorithm has been developed that decreases the time requirements, while finding high-quality plans (Wagner et al., 1998). Unlike this work, VDIG does not have an explicit information gathering tree, with branches representing various information gathering tasks and information operations. Instead the VDIG system builds such a plan by evaluating the decision model, the resources available and the user preferences.

Information gathering has also been studied by the natural language processing community. Recent research efforts have mainly addressed the problem of information extraction (Riloff and Lehnert, 1994; Ashish and Knoblock, 1997; Heflin et al., 1998), allowing systems to obtain the information they need from unrestricted text documents. Progress in this area will help systems like VDIG to expand the range of information sources they can utilize.

To summarize, although much work has been done on the problem of information gathering, little work has capitalized on the synergy that develops when the information gathering problem is solved in conjunction with the decision problem for which the information is needed. VDIG is the first system that uses a decision model and the value of information to plan, monitor, and control an information gathering process using a set of World-Wide-Web information sources.

3. Problem definition

This section defines more formally the terms that are used throughout the rest of the paper. It also includes a concise definition of the problem which value-driven information gathering is attempting to solve.

Definition 1. A *decision model* is an influence diagram (Shachter, 1986).

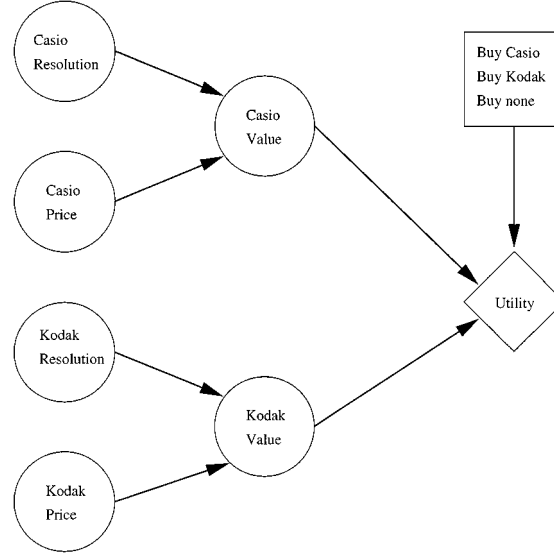


Figure 4. A simplified digital camera decision model.

An influence diagram is an acyclic directed graph composed of three types of nodes: chance nodes, decision nodes, and utility nodes (see figure 4). Chance nodes (ovals) are random variables with conditional probability distributions based on the nodes that connect to them. We represent the i -th chance node as V_i and its domain (a finite set of possible values) as D_{V_i} . Decision nodes (rectangles) represent possible decisions that the system can make. This is the variable that the system controls in optimizing utility. Utility nodes (diamonds) represent the utility for the agent given the state of the world and the decision made by the system.

Definition 2. An *information item* instantiates a leaf chance node in the decision model with a value $v \in D_{V_i}$.

Definition 3. An *information source*, s , is an independent process that has a probability, $P_s(t)$ of returning a set of information items I_s at time t after being queried for a monetary cost of m_s .

Definition 4. A *query*, $q = (s, t)$ is a request sent to an information source, s , at time t .

Definition 5. The *query pool* is a set of all of the queries made by the VDIG system that have not yet returned, $Q = \{(s_i, t_i) \mid \text{a query } i \text{ was sent to site } s_i \text{ at time } t_i\}$.

Definition 6. The *total monetary cost* M is equal to the sum of all the queries in the query pool as well as the cost of all of the queries that have been answered.

Definition 7. The *time-dependent cost function* gives the cost, u , based on the time since the start of querying, t , and the money, m , spent by the system, $C(t, m) \Rightarrow u$.

Problem statement: Given a decision model, a database of information sources and a time-dependent cost function, VDIG attempts to selectively query information sources in order to make a decision that optimizes overall utility while taking time and monetary resources into account.

4. Components of the VDIG system

The components of the VDIG system, shown in figure 3, are grouped into three sub-systems. These subsystems are all controlled by the value-driven information gathering planner: the decision model, the information sources database, and the information retriever. The value-driven information gathering planner is given a set of resource constraints and preferences by the user. The preferences (along with information provided by the expert that constructed the decision model) are used to create the utility function used by the decision model. The resource constraints provided by the user are used to create the cost function used by the value-driven planner. Below is a description of each of the three sub-systems and the VDIG planner.

4.1. Decision model

The decision modeling component of the value-driven information gathering system has three functions:

1. Return the best decision given the information that is currently available
2. Return the value of the information received from a particular information source
3. Incorporate evidence from information sources into the decision model

The decision model used in value-driven information gathering is an influence diagram as described in (Howard and Matheson, 1984). An influence diagram is an intuitive, widely-used technique for representing decision problems under uncertainty. Figure 4 shows an example of an initial influence diagram used by the system for purchasing a digital camera based on the resolution and price of the camera. When an information source returns a result, the raw data is passed to the extraction engine, where it is converted into a set of information items. *Information items* are records that contain the node which is instantiated and the node's value. Information items are generated by the extraction engine when it is passed raw data (an HTML file).

One difficulty in using a decision model is in assigning prior probabilities to the leaf chance nodes in the decision model. One approach is for the value-driven system to maintain statistical information over multiple gathering sessions in order to assign prior probabilities. Another approach is to assume that the "expert" that has built the decision model for use by the system has some knowledge of the domain and can estimate the probabilistic distribution of the leaf nodes. If the expert lacks this domain knowledge the prior probabilities can be

assign an even distribution to represent lack of information about their true distribution in the world. In such an event the decision model is still an extremely powerful technique for determining the value of information as more information becomes available to the system. High quality prior probabilities improve the efficiency of the system but are not mandatory for the value driven information gathering system to operate. In our experimental systems, we assigned prior probabilities based on our experience from finding information sources for the extractor to use.

In cases where the decision model contains a large number of instances to choose from and the system has little or no knowledge of the domain, a value-driven system may be forced to gather information about a large number of instances before it can begin to eliminate instances from further consideration. The breadth of information gathering policy is based entirely on the utility function created by the user and the expert at the beginning of the information gathering process. In many domains making a near optimal decision is acceptable. The decision model contains in it the value which the user places in making a perfect compared to a good decision. If the utility function has a large cost for not making an optimal decision and the cost function provides adequate resources to thoroughly analyze all available choices then the value-driven system will attempt to collect information on all of the instances before focusing on those with the highest likelihood of being a good decision. If, on the other hand, the utility function does not penalize the system for make near-optimal decisions or the cost function is high, the system will sample the set of instances and focus only on the best of that subset. The larger the penalty for making mistakes and the larger the resources allocated for gathering information, the larger the sampling set will be. This flexibility—the ability of the system to adapt its behavior to any given user objectives—is one of the strengths of VDIG.

4.2. *Information sources database*

The information sources database (ISD) maintains information about each potential information source that the VDIG system may access. For each information source listed in the database, the system maintains:

- The cost of accessing the information source.
- The responsiveness of the information source (The probability of the source returning the information at any point in the future after the query is sent). Figure 5 show a response probability graph used by the system.
- The nodes in the decision model that the information source returns.
- Information about how to extract the information from the information source.

Learning the responsiveness of an information source can be done off-line by querying information sources a number of times. In the case of the WWW, the responsiveness of an information source depends on the server and global factors that affect the “load” on the network. Many information sources have the same response graph because they are physically located on the same server (e.g., pages on a camera purchasing site are considered different information sources, but have similar responsiveness because they are located on

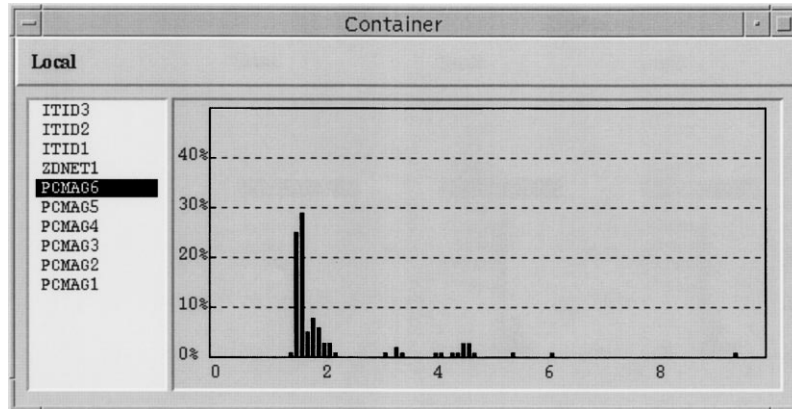


Figure 5. The response probability graph for an information source used in the digital camera purchasing decision. The X axis is measured in seconds.

the same server). There are methods for testing the global “load” on the Internet by testing specific highly used servers (such as name-domain servers) and using that information to modify the response graphs of all of the information sources.

Finally, the ISD Querying engine can also return a list of potential information sources that match the *focus-set* (see below) and have not already been queried. This list is evaluated by the planner using the value of information and the response expectations to select the best source to query.

4.3. Value-driven information gathering planner

The VDIG planner uses information from the user, the decision model, the ISD and its own internal record of the currently active queries in order to decide which, if any, queries to send at any given moment. In order to do this the system must be able to calculate the *value of a query*. This value is determined by evaluating the value of the information, the cost of making the query, and the responsiveness of the information source. In this section we first describe how the value of a query is calculated and then describe the algorithm which the VDIG system uses to decide which (if any) queries to make and when to halt to return a decision.

4.4. Calculating the value of a query

Value-driven information gathering activates queries based on their marginal query value. This value is determined by the value of the information the query will return, the expectation of the information source returning the result, the current information known by the system and the set of queries that have been sent but have not yet returned results. The current state of the information gathering process is characterized by the query pool

$Q = \{(s_1, t_1), \dots, (s_n, t_n)\}$. The system maintains the activation time for each query (t_i) and uses it to dynamically update the response expectation. For each information item, f_i , the system calculates a comprehensive probability over time of the information arriving taking into account all of the relevant queries in the query pool (assuming independence of information sources in terms of their response probability).

The marginal query value, Δ , is determined by comparing the utility of the query pool to the utility of the query pool and the candidate query (q).

$$\Delta(q) = U(Q \cup q) - U(Q) \quad (1)$$

The value of a query pool ($U(Q \cup q)$ or $U(Q)$) can be reduced from a time-dependent function to a single value because the VDIG system controls the time at which the information gathering process will halt and return a decision. The gathering process will halt once the expected utility as a function of time has reached its maximum.

$$U(Q) = \arg \max_t [U(Q | t)] \quad (2)$$

The comprehensive utility of the query pool at any given time t , is the difference between the expected value of the returned information and the cost function $C(t, m)$. Where t is the time and m is the total money spent. 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 information items, $\alpha \subset \{f_1, \dots, f_k\}$, and the corresponding value of information $V(\alpha)$.

$$U(Q | t) = \sum_{\alpha \subset \{f_1, \dots, f_k\}} P_\alpha(t) V(\alpha) - C(t, m) \quad (3)$$

The probability of a particular subset of information items being returned at any time given a specific query pool is calculated by multiplying the probability of finding each information item, f , in α by the probability of not finding each information item not in α .

$$P_\alpha(t) = \prod_{f \in \alpha} P_f(t) \prod_{f \notin \alpha} (1 - P_f(t)) \quad (4)$$

The probability of finding an individual information item f in time t is simply the sum of probabilities for each time step between 1 and t .

$$P_f(t) = \sum_{i=1, \dots, t} H_f(i) \quad (5)$$

Finally, the value of $H_f(i)$ represents the probability of query pool Q (or $Q \cup q$ if we are using the query pool plus the potential new query) returning an individual information item f at time i (all times are measured relative to the current time). The histogram ($H_f(i)$) is calculated by examining all of the information sources that return information items on the feature f . $H_f(i)$ is also dynamically updated as time progresses using Bayes rule to

- | |
|--|
| <ol style="list-style-type: none"> 1) Initialize the utility curve. Repeat 2) Calculate the value of information for each information item in the decision model 3) Select the focus set, FS, of the most valuable items 4) For each information source that returns items \in FS, determine the marginal query value 5) Query the information source with maximal value <p style="text-align: center;">Until the utility curve is non-increasing</p> |
|--|

Figure 6. The VDIG algorithm.

determine the correct probability of the query pool returning a value given the fact that it has not yet returned a value.

Using these equations to evaluate any query pool, and defining the value of a query for q as the difference between the current query pool and the query pool with q , it is possible to rank potential queries and determine which, if any, should be added to the query pool.

4.5. The querying algorithm

Figure 6 shows the algorithm used by the value-driven information gathering system to determine which action to take: add a new query, wait, or halt and return a decision.

Another task of the value-driven information gathering system is to select a “focus” set of features which will be used to calculate the value of information for the decision model. The focus set is composed of the N most valuable individual features in the decision model (The size of N depends on the speed of the computer and the complexity of the decision model being evaluated). The focus set is then used to restrict the largest set of features that can be evaluated together at any one time. The value of a set of features is not equivalent to the sum of the value of information for each individual feature and thus must be considered as sets. The amount of time it would take to exactly calculate the value of information for information sources that return a large number of features is combinatorially infeasible, so a limit must be placed on the number of features that are considered together. The features that are pruned are those with the lowest individual value of information.

4.6. Information retriever

The information retriever dispatches server requests to the Internet and extracts information items from the results returned by information sources. This sub-system acts as a buffer between the value-driven planner and the real world. Information extraction is an extremely difficult problem and we do not suggest here that we have solved it. Instead, what we have worked on building an open framework in which different extraction engines can easily be added. In our current system, the high-level extraction engine executes a set of rules in order to change HTML into a list of information items and values. This approach has been used with several purchasing domains (digital cameras, removable media, restaurants) and

appears to work well. Our main goal in building our own extraction engine was to develop a working prototype and lay the groundwork for incorporating more advanced extraction engines developed by others in the future.

The extraction engine uses pattern matching in order to find the information the system is searching for. For example, when attempting to extract the resolution of a digital camera, the extraction engine looks at the page and tries to find text of the format *number x number*. One difficulty we faced in attempting to extract information from HTML documents is the increased use of tables and other formatting constructs. This has forced us to develop specialized modules to deal with such concepts as table headings, rows and columns. Through the course of adding these specific extraction components we have learned that it is often difficult for an autonomous system to analyze documents which are very easy for the human eye to analyze quickly. Several current research efforts are aimed at constructing powerful tools for information extraction (Doorenbos et al., 1996; Etzioni et al., 1996; Etzioni and Weld, 1994). Such tools would contribute to the scalability of our system.

5. Walk-through

This section illustrates the operation of the value-driven information gathering system as it executes a gathering session (see figure 6). In order to keep this walk-through simple, the system only has to choose whether to purchase a Casio-10A or a Kodak-40 digital camera. The walk-through example uses only price and resolution to make the decision. Figure 4 shows the simplified decision model. In our experimental system the digital camera decision model contained six nodes for each camera being considered (see figure 11). Finally, this walk-through has access to only two information sources. The first information source returns the price of the Casio digital camera. The second information source returns both the price of the Kodak digital camera and the resolution.

The first step for the system is to calculate the value of information for each information source. The value of information for the first source (www.itid.com/Casio.html) is 2.23 and the value of information for the second source (www.itid.com/Kodak3.html) is 2.64. The value of information is then factored by the response probability graph for the information source.

Figure 7 shows the result of evaluating both information sources. The two bottom lines show the response probability graph for the sources (the probability of receiving a response by time t), and the top two lines are the utility gained by querying the sources. At this point, the Kodak information source appears to be the better of the two information sources because of its higher value of information.

The next step in the selection process is to take the cost function into account. Figure 8 shows the expected utility for querying the candidate information sources minus the cost function. It now becomes apparent that the Casio source is better to query because its faster expected response outweighs the less valuable information.

At this point the VDIG-system queries the Casio site, and begins the evaluation process again, taking into account the query already in the query pool when determining the value of information for potential information sources the system might query. In this simplified example, the system has only one other information source to consider, the Kodak source.

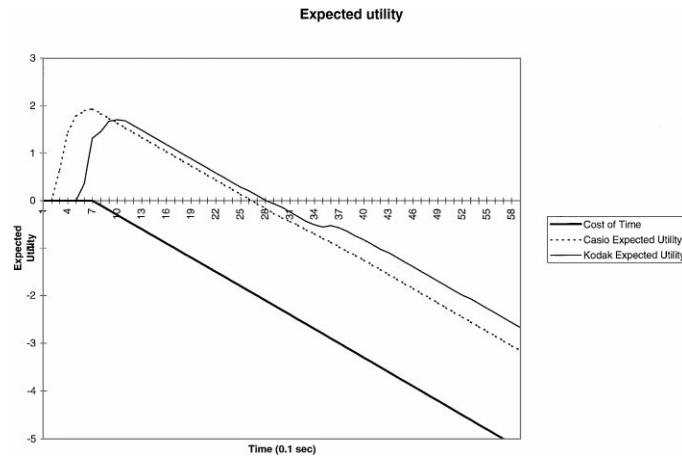


Figure 7. The response probability curve and expected utility (without the cost of time) for querying the two potential information sources.

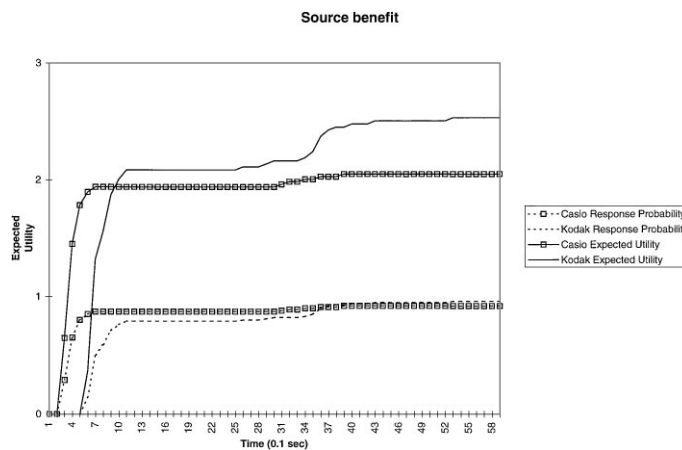


Figure 8. Expected utility for querying the two information source with the cost of time included.

The system evaluates the expected increase in decision quality for querying the Kodak information source and activates the query. Figure 9 shows the expected utility for the information gathering process with both the Casio and Kodak sites added to the query. The Casio site was added at time unit 0 (measured in seconds) and the Kodak site was added at time unit 1.

In this simplified case the system has no more potential information sources to add to the querying pool. The system must now monitor the queries and the expected utility function and determine when to quit the gathering process and return a decision. This can happen once an information source has returned a result or when the expected utility curve becomes negative.

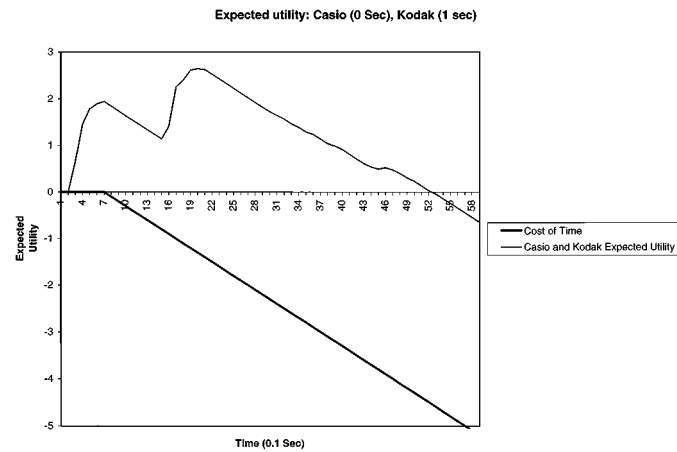


Figure 9. Expected utility for querying the Casio information source at time 0 seconds and the Kodak information source at time 1 second. This includes the cost of time.

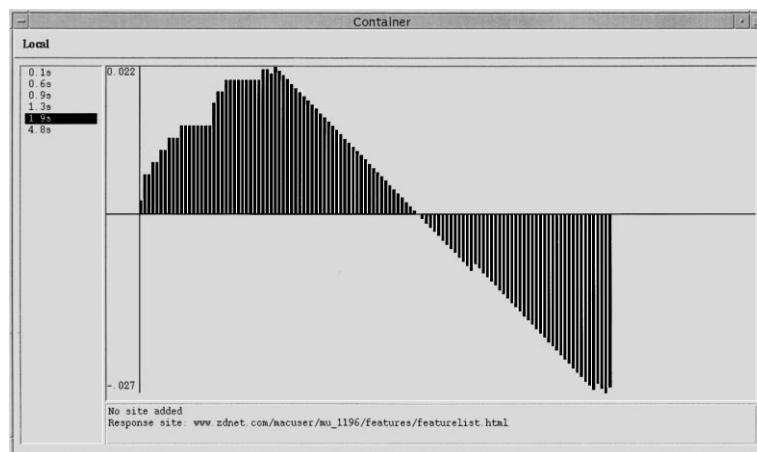


Figure 10. The VDIG system working on a camera decision.

Figure 10 shows the output from our working system, displaying the expected utility at 1.9 seconds into the information gathering process.

6. Evaluation

We have tested the VDIG system in three different real-world domains: purchasing a digital camera, purchasing a removable-media drive, and selecting a restaurant. We focus on the digital camera purchasing decision in this paper. A review of the removable-media purchasing result and the restaurant selection system will be available in (Grass, 2000). In general, the most difficult process in adding new domains is the construction of a decision

model and the construction of new information extractors for the information used by the system. A number of software tools have been developed to facilitate this task. The most recent implementation of the system has been written in Java and will become available as part of the first author's Ph.D. dissertation.

For the digital camera purchasing decision, we have constructed a simple decision model representing the usefulness of the key features of a digital camera. For example, if the camera uses flash storage cards than the system has greater storage capacity. These low-level features are then used to evaluate high-level features of the camera. The relationship between low-level features and high-level features is defined by an expert. This simplifies the specification of a utility function by a novice user (e.g., one may want a camera that is expandable, but is a camera that uses 8MB flash cards more expandable than one that uses floppy disks?). In other domains it may be possible to gather information at multiple levels directly (rather than low level only), giving the system a choice between gathering the high-level information or attempting to extrapolate it from low-level features.

In our experiments, the system had five choices; to recommend one of the four cameras presented to it or to decide not to purchase any of them. For each of the four digital cameras being considered a small decision model was built (see figure 11). We kept the maximum

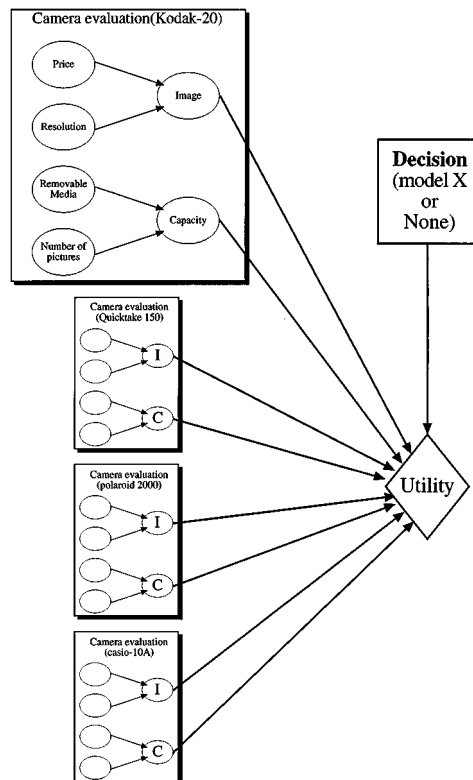


Figure 11. The decision model for the digital camera purchasing decision. Note that the same Bayesian network is replicated four times in the decision model.

	Coverage	VOI only	VDIG	“Ideal” Utility
Utility	0.122	0.213	0.255	0.295
Accuracy	46%	58%	80%	100%

15 seconds of time for free and a utility penalty of 0.1 for every second after that.

	Coverage	VOI only	VDIG	“Ideal” Utility
Utility	0.130	0.213	0.251	0.295
Accuracy	44%	58%	80%	100%

5 seconds of time for free and a utility penalty of 0.1 for every second after that.

Figure 12. Comparison of VDIG to other retrieval methods.

number of digital cameras being considered to four because of some limitations of the Bayesian network package we were using. We have since implemented our own Bayesian network package that has eliminated this problem.

We compared the value-driven system against three base-line systems. The results are presented in figure 12. The first system attempted to collect information for each item in the decision model with equal weight (coverage). The second system used the value of information with no additional knowledge about the information sources in order to decide which sources to query (VOI only). The third system always returned the correct decision instantly (an “Ideal” system). Of course, this third system cannot actually exist, but it is useful as an upper bound on performance. The coverage and VOI only systems continue to query until the cost function is non-zero. We tested all four systems under a variety of resource constraints, utility functions, cost functions and available choices of cameras (in total we had ten cameras, but for each experiment we selected four for the systems to pick from). For each set of resource constraints we ran each of the systems fifty times, each time with a different utility function and set of cameras to pick.

Value-driven information gathering performed well compared to the base-line systems. Compared to the ideal system, the value-driven system had high-quality accuracy in picking the best camera (80%) and an even higher average utility compared to the best choice. This discrepancy comes from the value-driven system not spending resources distinguishing between two high-quality choices. Value-driven systems attempt to maximize average utility more than selecting the optimal answer. This behavior can be changed by altering the utility function to penalize non-optimal decisions, but in practice it is often satisfactory to find a good match instead of spending the extra resources looking for the optimal one.

Compared to the coverage system a value-driven approach nearly doubles both the average utility and accuracy given the same resources. Demonstrating that spending internal resources to increase the quality of the queries made has a large impact on the overall quality of the decision.

Compared to a system that uses only the value of the information, the value-driven approach demonstrates that tracking the difficulty in acquiring information improves both the quality and the accuracy of the decision about 20%. The extra time spent by a value-driven system calculating the expected response is small compared to the amount of time spent determining the value of information (which both systems must do), so this increase in decision quality comes at very little additional computational overhead.

7. Conclusion

We have described a system for value-driven information gathering and have demonstrated its operation over the World-Wide-Web. The benefits of using a value-driven approach have been demonstrated when the system is compared to similar base-line systems that use less sophisticated approaches to organizing information gathering.

The VDIG system is one of the first working prototypes of systems that are capable of not only locating and gathering useful information but also integrating it in real-time with a decision process. This is an exciting new direction for using the vast amount of information available on the WWW. Developing and experimenting with the VDIG system has taught us several important lessons that are summarized below:

- Value-driven information gathering offers important benefits. The significance of these benefits grows when there is a high-level of uncertainty regarding the performance of each information source and as the resources available for making a decision become more restricted. Therefore, we believe that this approach will be essential for the success of agents operating on the WWW.
- Acquiring decision models for many practical tasks is feasible. We found that influence diagrams offer a rich and intuitive representation for a wide range of tasks and that the factors that affect the decision are similar for different users. Relative importance and preference can be modified to reflect the subjective utility of each user. An approximate decision model (one with imprecise prior probabilities) does not degrade performance of the system greatly. Allowing the end user to interact with the system and tune-up the decision model also forces the user to consider what aspects of the decision are most important to them.
- A good model of the responsiveness of information sources can be maintained. This can be achieved by an independent automated process that queries sites and monitors their responsiveness. Discovery of new sites useful for an application is much harder but can be done when the system is not in use. In this case designing a persistent system offers many benefits.
- The cost structure imposed by information sources has a major influence on the behavior of the system. Free information leads to a high degree of parallelism to save time. Costly information leads to a more selective, lengthy process with fewer queries. VDIG adapts its operation to the characteristics of the information environment. In cases where the cost function is at either of these extremes the value-driven information gathering system still does the “right” thing, but often a less complicated approach would also work. One of the experiments run in the removable-media purchasing domain was to compare a

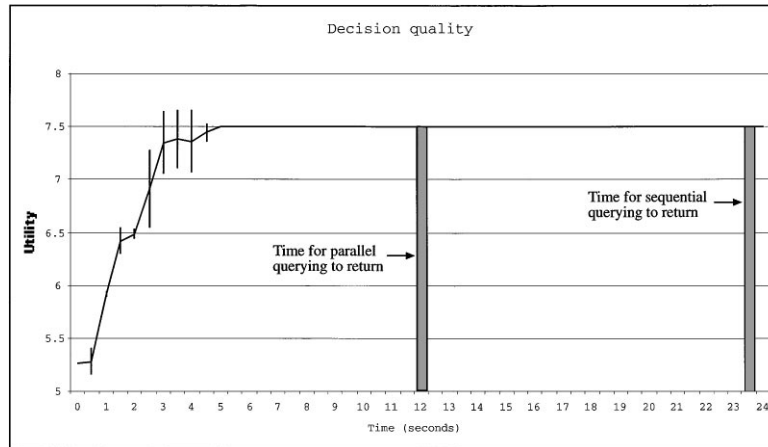


Figure 13. Decision quality of the value-driven information gathering system. The vertical bars represent standard deviation of utility. The graph shows that the system quickly improves the decision quality and reaches a high-quality decision within a fraction of the time required by the two simpler querying approaches (marked by the dark columns).

value-driven approach to these more simplified information gathering strategies without the meta-reasoning overhead from calculating the value of querying an information source. Figure 13 shows how a value-driven approach compares to a parallel approach (where the cost gathering information is very low and all potential sites are queried immediately) and a sequential approach (where the cost of querying an information sources is very high and no two queries are activated simultaneously). The data shows that meta-reasoning about the value of querying an information source provides significant speed advantages. The parallel information gathering approach was slower because the parallel system would query information sources that were not needed and that required a great deal of time to return information. The sequential information gathering approach would continue to query information sources even though their probability of altering the decision was very low.

- The complexity of information extraction varies widely from site to site. It can be as simple as pattern matching and indexing or as complex as natural language processing. We need to limit the information sources we use to those that can be handled by our information extraction module. We anticipate that more advanced extraction techniques will improve the number of information sources that the system can process in the future.
- Information extraction is the computational bottleneck in an environment that provides much information for free. Compared to the meta-reasoning overhead the system spends a majority of its time processing the raw data returned by information sources. The main motivation for limiting the number of queries is the computational cost of processing the results. Future monetary charges for information may change this balance.
- The widespread use of VDIG systems depends on the development of information sources designed to interact with autonomous agents, not just browsers. This will greatly simplify the information extraction problem. To be successful, such interface will have to support

shared ontology to describe objects of common interests. The acceptance of such information tagging languages as XML will facilitate the creation of faster and more reliable extraction engines in the future.

7.1. Further work

7.1.1. Evidential reasoning. We are currently focusing on extending value-driven information gathering to work with evidence instead of direct instantiation. We define *evidence* as pieces of information returned by information sources that may not be accurate and may degrade with time. In order to extend the system to deal with such evidence we are altering the manner in which the decision model is changed as new information items are returned by the extraction engine. Previously, the information items were incorporated directly into the decision model by instantiating a node.

Adding evidence to the influence diagram involves adding new nodes that are causally related to the information node. In figure 4 for example, an evidence node added to the node “Casio resolution” is shown below (figure 14).

Evidence nodes add flexibility to the evaluation of information sources and allow VDIG to represent concepts such as accuracy, bias, and information degradation over time with few variables and little added complexity. We have gone to some effort to limit the amount of information that needs to be maintained about each information source so that the amount of storage is small and the time it takes to learn these values is reduced. We assume that information sources are independent of each other and have a Gaussian distribution based on the *true value* of the node in the decision model (see figure 15). In our example, if the true value of the Casio resolution was 3, the probability distribution of the value returned by the information source would be based on the bias (a horizontal shift) and the accuracy (the

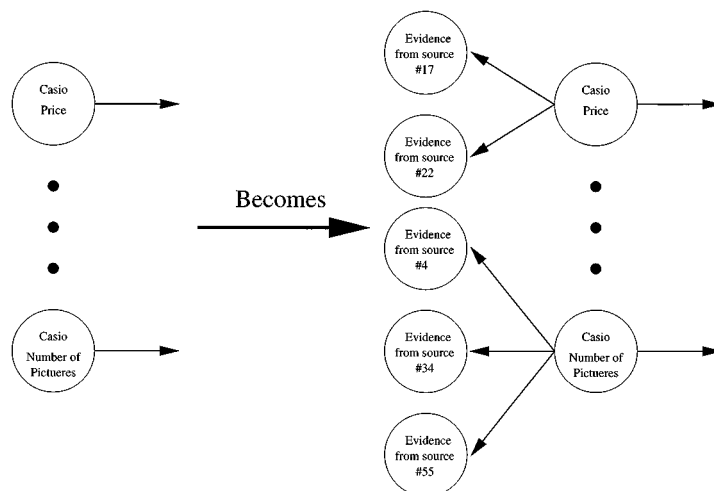


Figure 14. Adding an evidence node to the influence diagram.

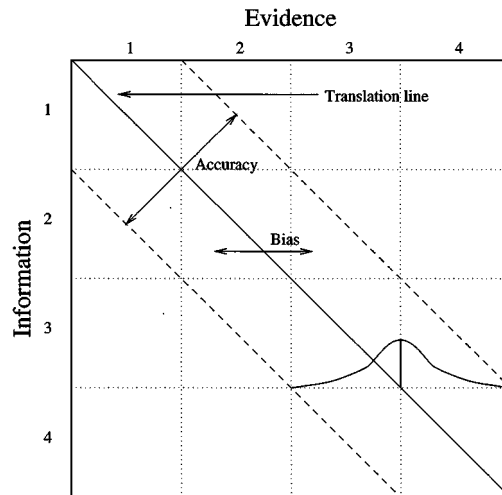


Figure 15. Adding an evidence node to the influence diagram.

inverse of standard deviation). By building this probability table for the information source, and instantiating it with the value the information source returned, the decision model can infer the probability distribution of the true value for “Casio resolution”.

As more evidence nodes are added to the decision model (see figure 14), the VDIG system has a more accurate estimate of the true value of the decision node. To test the increased value of information of adding an evidence node, we add the uninstantiated evidence node to the network and then use the standard value of information algorithm to determine the increased decision quality if the system queries the information node. Because information sources can return more than one information item, this process is complicated by the fact that we have to add several nodes to the decision model to evaluate the multiple pieces of information that the information source will provide. The *value of information for a set* is the increase in decision quality if a set of information sources is returned at the same time.

One advantage to using the accuracy and bias values to generate a probability distribution table for the evidence node is that it can be extended to degrade these values based on the age of the information. For example, if the system is making a decision about purchasing a computer the quality rating of the computers performance might have a negative bias the older the review is. So a system that was reviewed as being very fast six months ago, might now have its true value biased down to having only average speed. In time-critical situations, we can decrease the accuracy of evidence nodes as the querying process progresses, allowing the system to make decisions about re-querying information sources. Another advantage to using only an accuracy and bias measure to build these probability tables is that both of these values can be learned with a relatively small amount of data.

7.1.2. Internal resource allocation. Extending model-based reasoning to control the cost of gathering information has opened a large number of research problems. At this point we

have been studying the notion of independent processes that return information. In fact, translating this information into a form that is usable for instantiating the decision model is a non-trivial task. But this process differs from collecting data in that we have some control over how to prioritize the translation process. There are several benefits to expanding the system in this manner. If two information sources return raw data at the same time, the system should focus on extracting the information that will improve the decision the most. Also, in some domains, there will be different extraction techniques that require different computational resources and return data of varying quality. This leads us to building on our previous work on anytime algorithms, but this time as components of a larger value-driven system.

7.2. Summary

Value-driven information gathering provides an extension of standard decision theory that takes into account information sources that have a cost to access and a probabilistic chance of returning information once they have been queried. This extension allows decision theoretic techniques to be applied effectively to controlling systems that gather and use information retrieved from the World-Wide-Web. As the amount of available information located on the web continues to increase, one of the more challenging problems for agents that use the web will be to selectively decide which information sources to use out of a set that could never be queried completely. Value-driven information gathering provides an architecture to accomplish this task and a metric for evaluating how to spend time and resources in these types of domains. We continue to improve the flexibility and descriptive power of the value-driven approach making it suitable for a wide range of Internet software systems.

References

- Ambite, J.L. and Knoblock, C.A. (1997). Planning by rewriting: Efficiently generating high-quality plans. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Providence, RI.
- Ashish, N. and Knoblock, C. (1997). Semi-automatic wrapper generation for internet information sources. *Second IFCS Conference on Cooperative Information Systems (CoopIS)*, Charleston, South Carolina.
- Carver, N. and Lesser, V. (1995). The dresun testbed for research in fa/c distributed situation assessment. *Proceedings of the International Conference on Multiagent Systems*, San Francisco, California.
- Doorenbos, R.B., Etzioni, O., and Weld, D.S. (1996). A scalable comparison-shopping agent for the world-wide web. *Proceedings of the Agents 97 Conference*, Marina del Rey, California.
- Etzioni, O. (1996). Moving up the information food chain: Deploying softbots on the world wide web. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, Oregon.
- Etzioni, O., Hanks, S., Jiang, T., Karp, R., Madani, O., and Waarts, O. (1996). Efficient information gathering on the internet. *Proceedings of the Symposium on Foundations of Computer Science*, Burlington, Vermont.
- Etzioni, O. and Weld, D. (1994). A Softbot-Based Interface to the Internet, *Communications of the ACM*, 37(7), 72–76.
- Grass, J. (1999). *Value-Driven Information Gathering*. Ph.D. thesis, University of Massachusetts, Amherst.
- Heflin, J., Hendler, J., and Luke, S. (1998). Reading between the lines: Using shoe to discover implicit knowledge from the web. *AAAI 1998 Workshop On AI and Information Integration*, Madison, WI.
- Howard, R.A. (1966). Information Value Theory, *IEEE Transactions on Systems Science and Cybernetics*, 2(1), 22–26.

- Howard, R.A. and Matheson, J.E. (1984). Influence Diagrams, *Principles and Applications of Decision Analysis*, 2.
- Jaynes, C., Marengoni, M., Hanson, A., and Riseman, E. (1998). 3d model acquisition using a bayesian controlle. *Proceedings of the International Symposium on Engineering of Intelligent Systems*, Tenerife, Spain.
- Klassner, F. (1996). *Data Reprocessing in Signal Understanding Systems*. Ph.D. thesis, University of Massachusetts at Amherst.
- Lesser, V., Horling, B., Klassner, F., Raja, A., and Wagner, T. (1997). Information Gathering as a Resource Bounded Interpretation Task. Technical Report, University of Massachusetts, Amherst.
- Marengoni, M., Hanson, A., Zilberstein, S., and Riseman, E. (1999). Control in a 3d reconstruction system using selective perception. *Proceedings of the 7th IEEE International Conference on Computer Vision*, Kerkyra, Greece.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, Los Altos, California.
- Riloff, E. and Lehnert, W.G. (1994). Information Extraction as a Basis for High-Precision Text Classification, *ACM Transactions on Information Systems*, 12(3), 296–333.
- Selberg, E. and Etzioni, O. (1997). The Metacrawler Architecture for Resource Aggregation on the Web, *IEEE Expert*, 12(1), 8–14.
- Shachter, R.D. (1986). Evaluating Influence Diagrams, *Operation Research*, 34(6), 871–882.
- Wagner, T., Garvey, A., and Lesser, V. (1998). Criteria-Directed Task Scheduling, *International Journal of Approximate Reasoning* (Special Issue on Scheduling), 19, 91–118.