A Value-Driven System for Scheduling Information Gathering

Joshua Grass and Shlomo Zilberstein

Computer Science Department

University of Massachusetts at Amherst

CMPSCI Technical Report 98-9

February 18, 1998

A Value-Driven System for Scheduling Information Gathering

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, scheduling, decision theoretic planning

Abstract

This paper presents a system for scheduling 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 actions (queries) are scheduled myopically by 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 time cost function specified by the user. We describe the benefits of the system and the lessons learned from its development.

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). The rapid growth of information sources over recent years presents a serious challenge to AI to build systems that can exploit this information in order to solve problems and make decisions. To achieve this goal, one must address the problems of locating useful information sources, requesting the right information, extracting it from the response, and integrating the results into the decision making process. In addition, an effective system must operate with limited computational resources taking into account the characteristics of each information source in terms of accessibility, reliability and associated costs.

This paper presents a system for value-driven information gathering (VDIG) that can schedule information gathering actions in order to make a decision within a set of resource constraints. This differs from existing work on information gathering in which the primary goal is to achieve coverage over a set of information sources (Doorenbos 1997), or to discover appropriate information sources (Selberg 1997). VDIG systems are 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 and specifies the value of resources (time and money) and a set of preferences over decision outcomes (e.g. in purchasing a car, the importance of total cost of ownership, number of doors, etc.). The system also uses a decision model constructed by an expert to propagate the influence of various factors on each other and on the overall decision (e.g. determining the total cost of ownership from gas-mileage, resale value, average maintenance cost, etc.). 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 are characterized by the following 5 features:

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 (Pearl 1988) for a set of information items used in the decision.

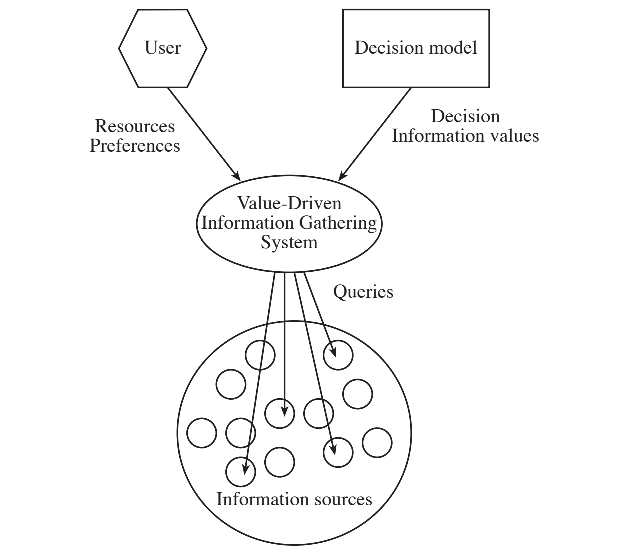


Figure 1 - Value-Driven Information Gathering

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 resource restrictions. We have implemented a prototype system and used it for gathering information on the WWW using influence diagrams to represent decision models. This is just one example of an environment with the above features making VDIG an effective approach. However, it is important to note that value-driven information gathering can also be applied to other problems such as image understanding (Jaynes 1998), treating costly image processing modules as information sources. It can also be applied to signal processing (Klassner 1996), and numerous other domains.

The VDIG system architecture integrates AI techniques from several different fields: decision-making under uncertainty, information extraction, and resource-bounded planning. The system maintains a dynamic pool of queries and a comprehensive utility function that gives the net gain in decision value as a function of time. As long as new queries have positive marginal values, a query with maximal value is added to the pool. As long as further delay in reporting the current best decision has positive value, the system continues to gather information. The VDIG system has an open architecture (shown in Figure 2) 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 on the Internet and how it differs from our value-driven approach. The complete implementation of the system is described in Section 3. Section 4 illustrates the operation of the system with a simple problem instance. Section 5 includes an experimental evaluation of the operation of the system. Finally, Section 6 summarizes the lessons learned from the construction of the system and further work.

Background

Information gathering from the World-Wide-Wed 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 1997) that 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 1997) 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 multi-agent systems lab at the University of Massachusetts has been working on an information gathering framework based on their work on TÆMS task-structures (Lesser 1997) and the RESUN planner (Carver 1995). These task structures represent possible search strategies to use for collecting information on-line to make 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 specific areas where that information can be collected. Also contained in the TÆMS task structure is information about the quality, cost and time expectations of each information source. The design-to-criteria planner can then use the TÆMS 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 TÆMS task structure is computationally intractable, a heuristic search algorithm has been developed that decreases the time requirements, while finding a high-quality plan (Wagner 1997). Unlike this work, VDIG does not have an explicit, user defined acceptable decision. Instead, the criteria used to evaluate a decision is determined implicitly by the decision model. In addition, there is a separation between sources of information and the use of the information in making a decision. Separating these two aspects of information gathering allows the system to add new information sources without altering the decision model.

Information gathering has also been studied by the natural language processing community addressing mainly the problem of information extraction (Riloff 1994). This technology allows systems to extract useful information from unrestricted text documents. By integrating this

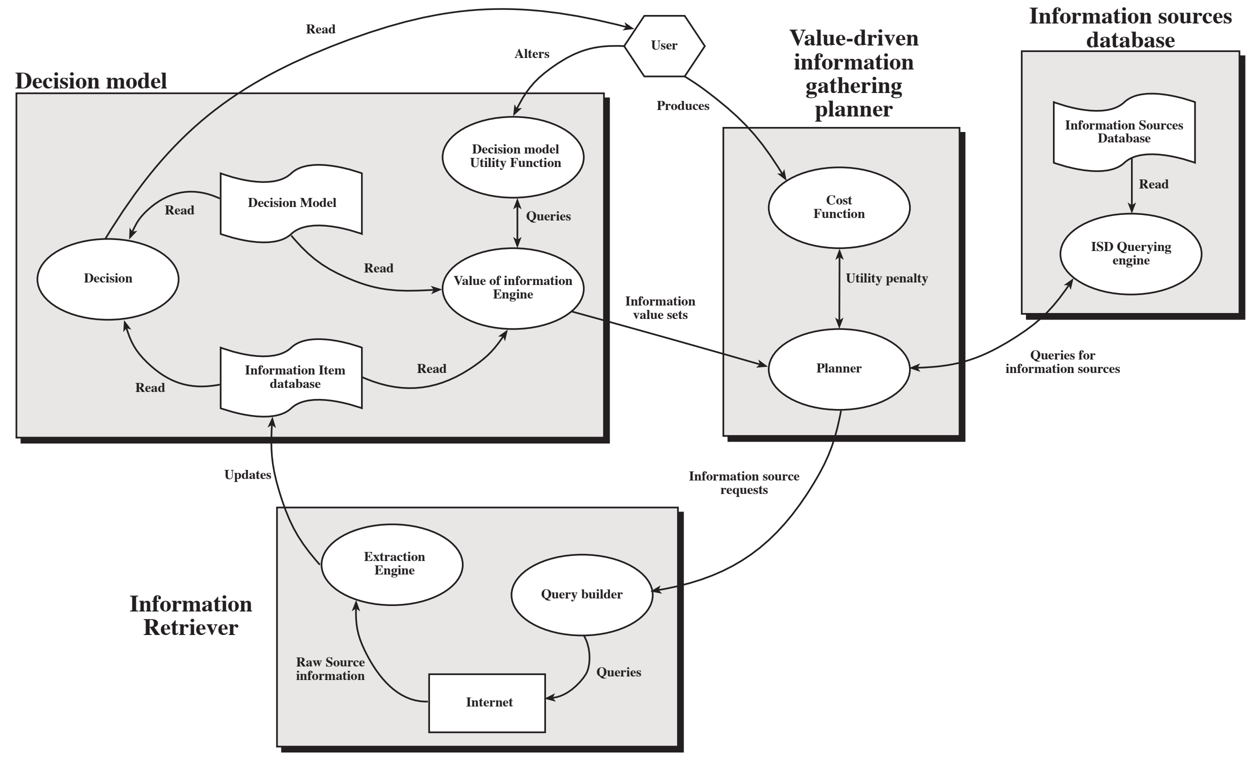


Figure 2 - The components of the Value-Driven Information Gathering System

technology with the VDIG system we allow it to expand the range of information sources it can handle.

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 together with the decision problem for which the information is needed.* This allows VDIG to use the value of information derived from the decision making process to prioritize the information gathering process.

Components of the VDIG system

The components of the VDIG system, shown in Figure 2, are grouped into three sub-systems 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 are used to create the utility function used by the decision model and the resource constraints 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. The VDIG planner and the information sources database were built in LISP and the two low level components of the system (the Internet querying engine and the value-of-information tool layer on top of HUGIN) were built in C.

Decision model

The decision model has two functions: to return the best decision given the information items returned by queried information sources, and to return the value of information for a set of information items. The current implementation uses influence diagrams to represent decision models. An influence diagram is an intuitive, widely-used technique for representing decision problems under uncertainty. There are many standard software packages for construction of and reasoning with influence diagrams. Most importantly for our application is the ability to monotonically increase the quality of a decision as evidence becomes available and the ability to determining the value of missing information (Pearl 1988). Figure 9 shows an example of an influence diagram used by the system.

We use a commercial package called HUGIN in order to construct decision models, propagate evidence and evaluate alternative decisions. On top of the HUGIN library, we have built a set of tools for communicating with the value-driven planner, building utility functions, and calculating the value of information for sets of information items. An expert can build a decision model using the HUGIN tools, a user can use the VDIG system to modify the preference structure or utility table, and the VDIG planner can use a set of functions to communicate with the HUGIN engine and determine the value of information for a set of information items.

Information sources database

The information sources database maintains information about each potential information source that the VDIG system may access. The information sources database also contains tools for querying the database to find potential information sources, and to automate the process of maintaining the response expectation histograms.

Each entry in the information sources database contains information about the location of the source (in the WWW this consists of a server address and a file path), how the source is accessed (for sources that return information via a form), how information is extracted from the source (a pointer to various extraction engines), and the response expectation for the source. The response expectation for an information source represents the probability of the source returning a result at any given period of time after the source has been sent a query. Figure 3 shows an information source database entries for the digital camera domain we used to test the system.

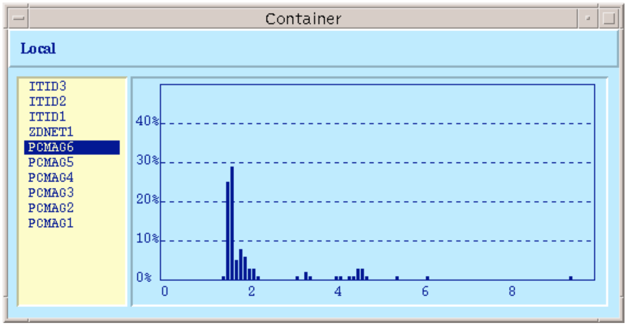


Figure 3 - An IDS entry used in the digital camera domain

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

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. Figure 4 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. In order to restrict the number of information sources examined, and the number of information item subsets that have the value of information calculated, the VDIG system restricts itself to a focus set of the N most valuable information items. This number is set by the user depending on the size of the problem and the speed of the system. It is well known that the value-of-information for a set of features is not additive. In order to correctly determine the value-of-information, subsets of features must be evaluated together. In large decision models it is combinatorically unfeasible to evaluate all subsets of features. In order to still correctly evaluate the model, the planner selects a restricted set of features (the focus set). The planner than calculates the value-of-information for all subsets of features in the focus set.

```
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) Find the value of information for each subset of FS
5) For each information source that returns an item in FS,
   determine the marginal query value.
6) Query the information source with maximal value.
   Until the current utility curve is at the maximum value.
```

Figure 4 - The VDIG algorithm

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 = $\{q_1, ..., q_n\}$. The system maintains the activation time for each query q_i and uses it to dynamically update the response expectation. For each information item, f, 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 expectation).

The marginal query value is determined by comparing the utility of the query pool with the query added to the current query pool.

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

The utility of a query pool is calculated by selecting the time at which the comprehensive utility curve is maximal. We can do this because the VDIG system determines when to stop and return a decision.

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

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 associated costs. These cost are represented by the cost of time ($c(t)$) and the monetary cost of the queries in the query pool ($C(Q)$). 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, s, and the corresponding value of information *V(s)*.

$$U(Q|t) = \sum_{S \subset \{f1,...,fk\}} P_Q(s|t)V(s) - c(t) - C(Q)$$

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

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

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.

$$P_Q(f|t) = \sum_{i=1...t} H_Q(f|i)$$

Finally, the value of $H_Q(f|t)$ represents the probability of query pool *Q* returning an individual information item *f* at time *t* (all times are measured relative to the *current time*). The histogram ($H_Q(f|i)$) is dynamically updated as time progresses using Bayes rule to 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 simple to rank potential queries and determine which, if any, should be added to the query pool. The value-driven information gathering planner uses a message-passing paradigm to simplify the process of creating replacement components. Each component in the system has responds to between three and fifteen messages with are explained in the system documentation.

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 is building an open framework in which different extraction engines can easily be added. In our current system we have created a high-level extraction engine that attempts to execute a set of rules in order to change HTML into a list of information items and values. These rules have been tested only in the digital camera domain although they could be expanded to deal with most purchasing decisions without much difficulty. 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 other groups in the future.

Walk-through

This section will describe the value-driven information gathering system as it goes through one loop in a session (see Figure 4). 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 system will also only use price and resolution to make the decision, Figure 5 shows the simplified decision model.

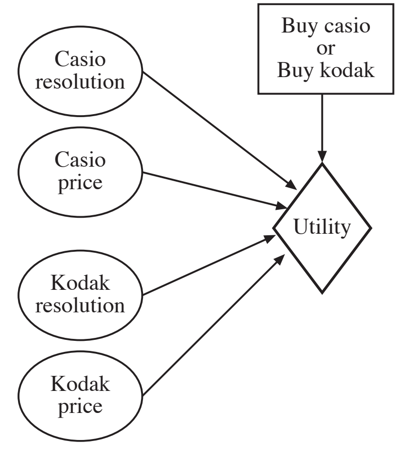


Figure 5 - A simplified digital camera decision model

The first step for the system is to calculate the value of information for each subset of the decision model based on the structure of the decision model and the utility function. This generates fifteen pairs of node subsets corresponding values-of-information. Of those, only two are used in this example: The value of Casio-price(2.23) and the value of Kodak-price and Kodak-resolution(2.64). The system evaluates each potential information source using the list of information values. Figure 6 shows the result of evaluating two information sources: http://www.itid.com/Casio.html and http://www.itid.com/Kodak3.html. The first source returns the price of the Casio-10A and the second source returns the price and resolution of the Kodak-40. The two bottom lines show the response expectation 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 it's higher value-of-information.

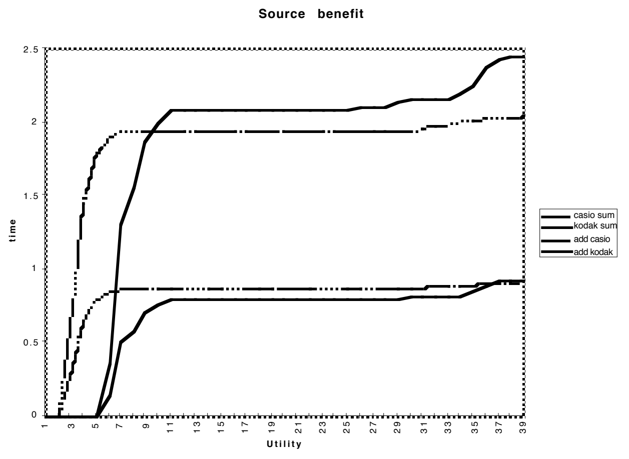


Figure 6 - Source benefit

The next step in the selection process is to take the cost function into account. Figure 7 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 of the short time frame for it to respond to a query.

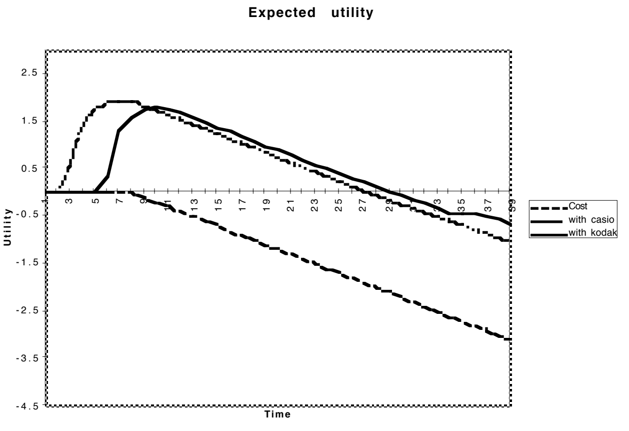
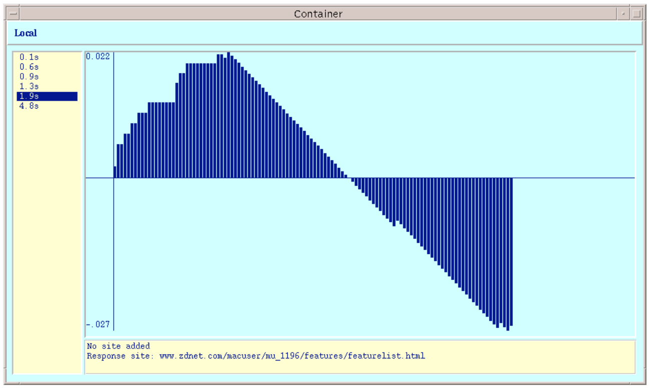


Figure 7 - Expected utility

At this point the VDIG-system would query the Casio site, and begin the evaluation process again, taking into account the query already in the query pool.

Figure 8 shows the output from our s\working system, displaying the expected utility at 1.9 seconds into the information gathering process. At this time, the VDIG planner has decided not to query any sites and has received information from the zdnet site.



No site added
Response site: www.zdnet.com/macuser/mu_1196/features/featurelist.html

Figure 8 - VDIG expected utility curve

Evaluation

We have tested our prototype VDIG system in the domain of making a decision about purchasing a digital camera. 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. The low-level features are then used to evaluate high-level features about the camera. The relationship between low-level features and high-level features is defined by an expert. In this example, this was done in order to make specifying a utility function easier for a novice user (e.g. I want a camera that is expandable, but is a camera that uses 8MB flash cards more expandable then one that uses floppy disks?). In other domains it may be possible to gather some high-level information directly, giving the system two approaches: Attempt to directly retrieve the high-level information, or attempt to extrapolate it from low-level features.

In our experiments, the system had five choices, to choose one of the four camera’s presented to it, or to decide not to purchase any of them. For each camera a small decision model was built (see Figure 9). In each experiment the system had ten information sources available to query.

We compared the value-driven system against three baseline systems: 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, but did not use any information about the information sources in order to decide which sources to query (VOI only). And the third system always returned the correct decision instantly (perfect). Of course, this third system does not actually exist, but it is useful to compare the other systems against. Also the coverage and VOI only systems would continue to query until the cost function was non-zero.

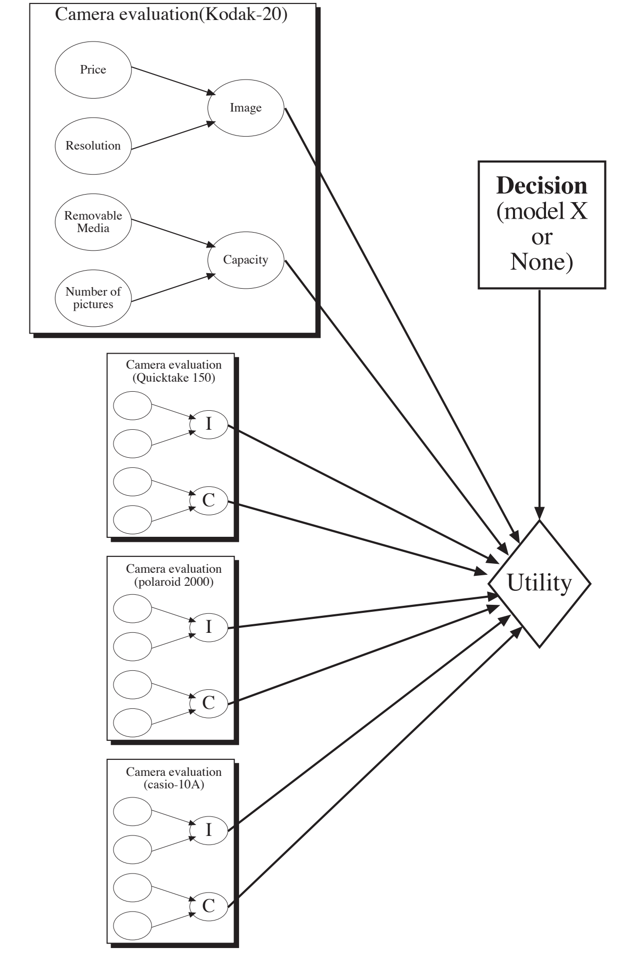

Camera evaluation(Kodak-20)
Price
Image
Resolution
Removable Media
Capacity
Number of pictures
Camera evaluation (Quicktake 150)
I
C
Camera evaluation (polaroid 2000)
I
C
Camera evaluation (casio-10A)
I
C
Decision (model X or None)
Utility

Figure 9 - Decision model for purchasing a digital camera

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 below we ran each of the system 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.

	Coverage	VOI only	VDIG	Perfect
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	Perfect
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.				

	Coverage	VOI only	VDIG	Perfect
Utility	0.137	0.213	0.230	0.295
Accuracy	50%	58%	66%	100%
3 seconds of time for free and a utility penalty of 0.2 for every second after that.				

Table 1 - Results of executing the VDIG system on the WWW in the digital camera domain

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 systems that do not use a value-driven approach with the same resource restrictions.

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 with a decision process. This is an attractive and exciting new direction for using the vast amount of information available on the WWW. Developing and experimenting with the VDIG system taught us several important lessons that are summarized below.

Lessons learned

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. We strongly believe that this approach will be essential for the success of agents operating on the WWW.

Acquiring decision models is relatively easy. We found that influence diagrams represent 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.

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.

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 much less queries. VDIG adapts its operation to the characteristics of the information environment.

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.

Information extraction is the computational bottleneck in an environment that provides much information for free. The main motivation for limiting the number of queries is the computational cost of processing the results. Future 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.

Further work

Much work is needed in order to generalize the components of the VDIG system. In the future we would like expand the system to deal with inaccurate information sources, take the computational resources used by the extraction engine into account in planning, and manipulate the decision model as the system is gathering information. The World-Wide-Web provides an exciting and growing environment for automated information gathering systems to operate in. Even with the increase in band-width and computational resources that are occurring, systems need to be developed that manage time, band-width, monetary and computational resources in order for users to effectively use the WWW as a decision making tool. Fortunately, artificial intelligence has been working on this problem in different forms for several years. Much like the recent success with expert systems, we believe that the WWW offers an opportunity for the field of artificial intelligence to prove its effectiveness again.

Acknowledgments

We wish to thank Guoqing Sun for developing much of the information extraction code for the system. Support for this work was provided by the National Science Foundation under grants IRI-9624992 and IRI-9634938.

References

Carver, N., and Lesser, V. 1995. The DRESUN testbed for research in FA/C distributed situation assessment: Extensions to the model of external evidence. In Proceedings of the international Conference on Multiagent Systems. San Francisco, California.

Doorenbos, R., Etzioni, O., and Weld, D. 1997. A Scaleable comparison-shopping agent for the world-wide web. In 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. In Proceedings of the Thirteenth National Conference on Artificial Intelligence. Portland, Oregon.

Jaynes, C., Marengoni, M., Hanson, A., and Riseman, E. 1998. 3d Model acquisition using a bayesian controller. In Proceedings of the International Symposium on Engineering of Intelligent Systems. Tenerife, Spain.

Klassner, F. 1996. Data Reprocessing in Signal Understanding Systems. PhD thesis, Dept. of Computer Science, 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, 97-34, Dept. of Computer Science, University of Massachusetts at Amherst.

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

Riloff, E. and Lehnert, W.G. 1994. Information Extraction as a Basis for High-Precision Text Classification. *ACM Transactions on Information Systems*, 296-333.

Selberg, E., and Etzioni, O. 1997. The MetaCrawler Architecture for Resource Aggregation on the Web. *IEEE Expert* 12(1):8-14.

Wagner, T., Garvey, A., and Lesser, V. 1998. Criteria-Directed Task Scheduling. *International Journal of Approximate Reasoning*. Forthcoming.