From HTML to usable data
Problems in meaning and credibility in the WWW *
Position paper

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

Introduction

This paper describes issues in information integration that relate to Value-Driven Information Gathering (VDIG)(Grass & Zilberstein 1996; 1997; 1998). Value-driven information gathering is the process of querying multiple information sources for information items which are used to make a decision. VDIG works in a resource-bounded environment where it is not possible to gather all the information needed to make a perfect decision. Instead, VDIG keeps statistics on the response expectation of particular sites and the decision model can operate with partial information. The process is referred to as value-driven, because the algorithm determines the *value of a query* for potential sites and queries the best candidate. The value of a query is determined using the value of information from the decision model, the expectations of a site returning a result at any time in the future, the information the system already knows, and the cost function, which represents the resources the system is allowed to spend in order to make the decision.

In this paper we will focus on one aspect of the value-driven process, taking raw information from sites and converting it into a form usable by our decision model[1]. The decision model we use is an influence diagram which uses information passed to it from an extraction engine to instantiate nodes. At the present time we rely on hand coding extraction algorithms that convert web sites into a list of feature/value tuples. For our prototype system this approach works well, and we have received good results after testing the system in the domain of making a decision about purchasing a digital camera. Numerous other groups are developing much more open-ended extraction engines (Doorenbos, Etzioni, & Weld 1997; Ashish & Knoblock 1997a; 1997b; Konopnicki & Shmueli 1995; Genesereth, Keller, & Mueller 1996). Figure 1 shows the influence diagram use by VDIG to evaluate digital cameras.

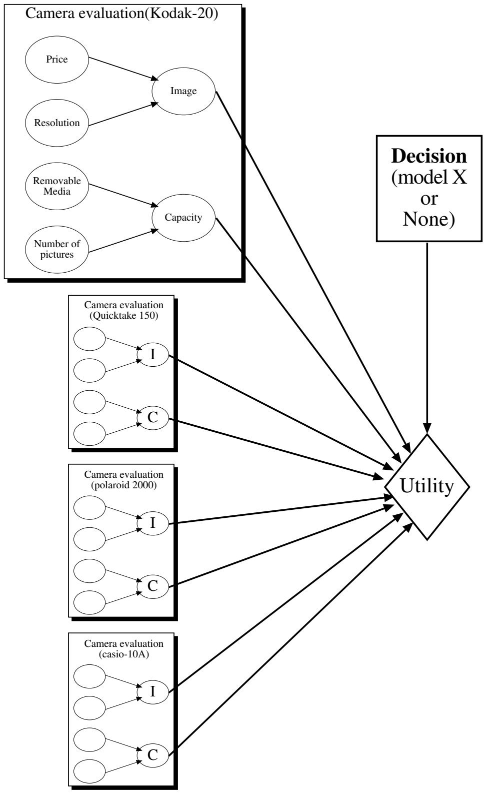
Camera evaluation(Kodak-20)
Price
Resolution
Image
Removable Media
Capacity
Number of pictures
Camera evaluation (Quicktake 150)
I
C
Camera evaluation (polaroid 2000)
I
C
Camera evaluation (casio-10A)
I
C
Decision (model X or None)
Utility

Figure 1: An influence diagram used for making a purchasing decision about a digital camera

One of the challenges facing VDIG at this point, and also facing other systems that integrate and use information from the WWW, is how to apply *consistent meaning* to information extracted from sites and how

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

[1] It should be noted that although we are dealing with unifying information from distinct information sources in the context of value-driven information gathering, these techniques are valid in domains with few information sources.

to assign *credibility* to those sites. Consistent meaning represents the process of translating the information from the site into a universal scale that allows comparison. And credibility represents how much weight to assign to the information in the decision process. For example, two sites that evaluate cars could differ in the rating systems that they use: One site could use a four star method and the other could use a scale from one to ten; or one site could break the rating into several categories; or one site could consistently rate all cars lower then the other. It could also be the case the I might share the views of one site more then the other, and would want it's information to have more influence in the decision. *If the problem of meaning and credibility are not addressed, then research in automated information gathering will never develop powerful techniques for integrating information.*

At this time, most extraction wrapper systems are given a piece of HTML from a site and return a list of facts based only on the document. In this paper we will argue that we can create more accurate systems if we expand wrappers to look not only at the HTML from a source, but to also take the source itself into account[2]. Further, we will argue that the best approach for doing this is to maintain external information about the sites and to expand wrapper systems to return a probability distribution, using a standard scale, of results from a HTML document. The credibility and meaning of information might be maintained internally by the automated information gathering system, or might be maintained in a 3rd party database. This information is not easy to maintain, but the benefits of deeper knowledge about the sources from which the system gathers information will be worthwhile in the increased accuracy of the decisions made using that information.

The rest of this paper will discuss some of the approaches we are investigating for solving the problems of meaning and credibility. Section 2 deals with the problem of defining a common meaning. Section 3 addresses the problem of determining the credibility of an information source. Section 4 concludes the paper and discusses what these two problems have in common. Appendix A contains a brief overview of VDIG.

Meaning

One key problem in extraction is understanding the true meaning of the information contained in a web page. Even in a domain as restrictive as product reviews (which is what we have been focusing on recently), there is no dictionary that unifies meaning from publisher to publisher and from time period to time period. A product may be rated as 4 stars in one publication and excellent in another, how do these ratings compare? Or one publication may break ratings up into numerous sub-components, while another publication gives one overall rating. Or products might be rated consistently higher in one publication than another. Also, product reviews are often not updated, and because of this a highly rated feature of a product may only be average by todays standards. For example, a digital camera that had excellent memory capacity three months ago might now only be rated average. We address these three problems of meaning differently.

- Translation records – In some ways translation records are very similar to wrappers. Translation records would most likely have to be constructed by hand for each publisher (e.g. MacWorld has a consistent set of words it uses for product ratings). The translation record could be built by the publisher or a third party. In this case (unlike bias or accuracy) there is no incentive for the publisher to lie. The record would translate their rating system into a universal system. For example, if a universal system used ratings of 0 through 100 to rate a product, a translation record might convert "good" to 80, "fair" to 50 and "poor" to 10.
- Bias records – Bias records would keep track of consistent errors by the system. The bias record could be built by maintaining a history of reviews compared to a set of known reviews, or a 3rd party might maintain a record of bias for a number of sites. The system can make strong or weak assumptions about bias, such as assuming that publications maintain the same level of bias through all reviews or that authors maintain the same level. Other non-trivial bias translations could also be used (e.g. a reviewer that never rates a product higher then an 8 or lower then a 2 on scale of one to ten). To continue our example from above, a bias record might add 17 to the value of any review by a specific author.
- Degradation records – A degradation record would have to be maintained for a specific class of product and it's characteristics, as opposed to the source that it has come from. This record would be built by looking at reviews of the same product at different times by different publications (publications rarely re-review products). Of course, this would be further complicated by having to take bias into account also. A degradation record could be built that keeps track of the relative rating of specific characteristics over time. For example, the rating of the resolution of a digital camera might be decreased by 5 every month after the review is posted.

Maintaining a translation, bias and degradation record for publications and products would allow systems to more accurately extract information from a source. Currently, VDIG and many other information gathering techniques treat all information as if it were

[2] there is also one "internal" factor that may add uncertainty to the facts extracted from a document; uncertainty in the extraction algorithm itself. Although this is an interesting problem to look at, we will deal exclusively with uncertainty from "outside sources".

current and as if it came from one unbiased source. Obviously, this assumption is inaccurate. Different sites can mean very different things, even when the text is similar.

Credibility

VDIG currently assumes that information returned by a source is completely accurate. This allows the VDIG system to fully instantiate a node, recalculate the influence diagram and determine the value of information for the remaining nodes when it receives data from a site. This assumption will have to be relaxed if VDIG is to be effective in real-world situations. There are cases were information is conflicting (different opinions in reviews), information is correct but competing (prices in catalogs), information is old, and instances where information is simply incorrect.

Fortunately, influence diagrams are easily expanded to deal with probabilistic or uncertain evidence. At each feature node we can add evidence nodes and instantiate them as we receive information. Figure 2 shows how this can be accomplished.

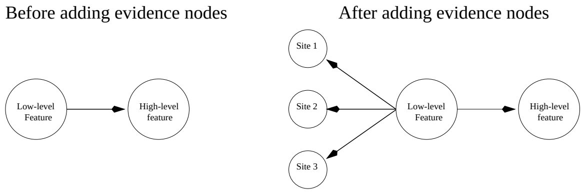

Figure 2: Adding evidence to an influence diagram node

The difficulty is that each of these evidence nodes must have a probability table that relates what they are likely to report based on the true value of the node.

Determining the credibility of a site is not easy. We have been working on four approaches for determining the credibility from external information about the site.

	True value		
Pr(e|true value)	poor	fair	good
poor	0.9	0.1	0.03
fair	0.07	0.8	0.07
good	0.03	0.1	0.9

	True value		
Pr(e|true value)	poor	fair	good
poor	0.5	0.25	0.2
fair	0.3	0.5	0.3
good	0.2	0.25	0.5

Figure 3: The probability table of a highly-credible information source (top) and a low-credibility information source (bottom)

- Relative age of the information as an indicator of credibility. In the domain of product selection, more recent information sources are generally better indicators of the current price.
- Testing against known information. We can test the information source against a set of know information and use that as a indication of credibility.
- Explicit marking. Third parties could set up servers that contained information about the credibility of information from a particular publication. A sort of ”consumer reports” of information from the WWW.
- Connectedness. We could use connectedness to sources with known credibility ratings as a method for theorizing about the credibility of a source. Often, web sites reference each other, sites that are highly referenced or reference by ”high-quality” sites are generally more likely to contain high-quality information. This would be hard to quantify, but the information about how often one site references another or how often a site is referenced in general is easily attained. We can also make assumptions about the quality of articles from one publication. Generally a publication has certain editorial standards that it maintains. Once we have determined the credibility of a publication, it is likely that other articles from the publication willhave the same credibility.

By having an indication of the credibility of a site, an automated information gathering system could be adapted to more accurately determine which site will be best to query. Also, by using this information, the decisions that an automated information gathering system will make at the end of the querying process will be more accurate.

Conclusions

Meaning and credibility in the information integration problem present a serious challenge to the field. The set of techniques that we have presented here are methods for mapping the response of an information source to a more accurate representation. These techniques would take a fact extracted from a document (e.g. The food at a restaurant was fair) and translate this information into a probabilistic representation on a universal scale. Figure 4 shows the goal we have in mind for the end result of an improved information extraction system taking the source into account. We want the system to use external information about the site and internal information from the data returned to build a more accurate representation of the extracted information than just a set of facts.

Taking into account adjustments for meaning and credibility in the process of extracting useful data from sites on the web will improve the process of planning information gathering actions and improve the end results returned by the system. This is because applying these methods to VDIG will allow us to relax some of

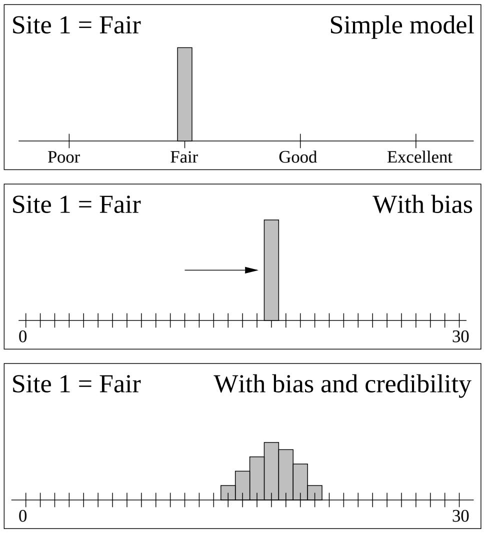

Figure 4: Developing a more complex representation of the information extracted from a WWW source

the assumptions made in using information returned by sites on the WWW. It will allow the system to better represent the information it receives and use it more appropriately in the context of decision making. Overall, reasoning about credibility and meaning should improve the decisions VDIG makes given the same resource restrictions.

References

Ashish, N., and Knoblock, C. 1997a. Semi-automatic wrapper generation for internet information sources. In *Second IFCIS Conference on Cooperative Information Systems (CoopIS)*.

Ashish, N., and Knoblock, C. 1997b. Wrapper generation for semi-structured internet sources. In *ACM SIGMOD Workshop on Management of Semistructured Data.*

Doorenbos, R. B.; Etzioni, O.; and Weld, D. S. 1997. A scalable comparison-shopping agent for the worldwide web. In *Proceedings of the Agents '97 Conference.*

Genesereth, M. R.; Keller, M. A.; and Mueller, G. C. 1996. Stanford information network. Technical report, Stanford.

Grass, J., and Zilberstein, S. 1996. Value directed information gathering. In *AAAI Fall symposium on Flexible Computation in Intelligent Systems.*

Grass, J., and Zilberstein, S. 1997. Planning information gathering under uncertainty. Technical Report 32, University of Massachusetts at Amherst. http://anytime.cs.umass.edu/~jgrass/ps/97-32.ps.

Grass, J., and Zilberstein, S. 1998. A value-driven system for scheduling information gathering. Technical Report 9, University of Massachusetts at Amherst. http://anytime.cs.umass.edu/~jgrass/ps/98-9.ps.

Konopnicki, D., and Shmueli, O. 1995. W3qs: A query system for the world wide web. In *Proceedings of the 21st International Conference on Very Large Databases.*

Appendix A - Overview of VDIG

Value-driven information gathering (VDIG) makes decisions based on information gathered from the WWW under resource restrictions. Although the assumptions needed for VDIG allow them to be employed in a variety of environments, we have focused on the WWW because of it's breadth of cheap information and it's interesting properties for planning (the ability to launch concurrent queries, and the probabilistic response time).

VDIG assumes that for any decision we are making using the WWW, that there are a large number of redundant information sources (sites) that the system may query in order to instantiate nodes in the influence diagram. Each potential information sources has different characteristics about how likely it is to respond at any point in time after it is sent a query and which nodes it can instantiate.

VDIG evaluates the set of potential queries and gives each one a value. The *value of a query* is based on the value of the information that can be retrieved by making the query (based on decision theory) and the cost of making a query (base on resource-bounded reasoning). Figure 5 shows the four main components of the system:

- **The decision model** – The decision model has two functions: to return the best decision given the information items that have been returned by the queried information sources, and to return the value of information for a set of information items. This is done by using an influence diagram to represent the decision model. Influence diagrams are a widely-used technique for representing decisions under uncertainty and for determining the value of information.
- **The information sources database** – The information sources database has an entry for every potential information source that contains a list of the information it contains, the cost of accessing the source, the wrapper needed to extract the information and the response expectation for the information source. The response expectation record allows the system to determine the probability of a information source responding at any given time after the query has been sent. Figure 6 shows a response expectation record for an information source used in the digital camera domain.
- **The value-driven planner** – The value-driven planner uses the decision model, the information sources database and a cost function to determine

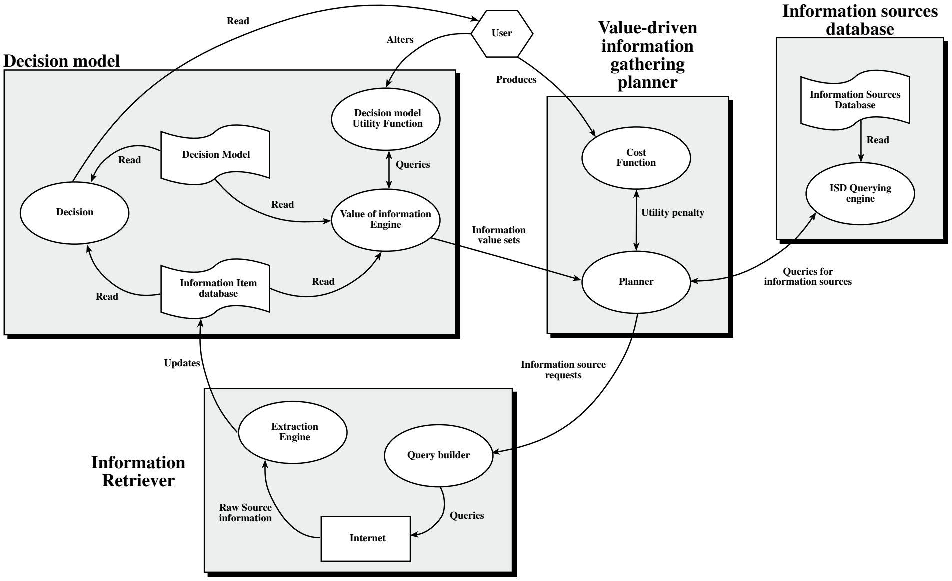

Figure 5: The components of VDIG

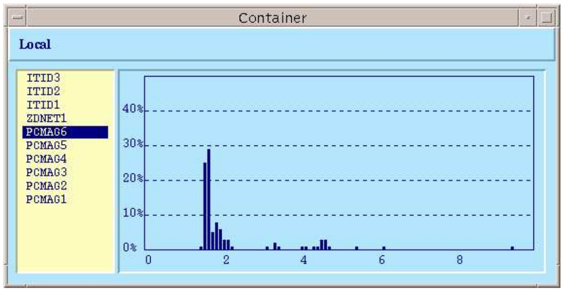

Figure 6: The expected response of an information source

the value for each potential query the system can make. This is done by analyzing the maximum of the expected utility of the system before and after a query is added. The difference is the *value of a query*. Figure 7 shows the expected utility curve along with the cost function and the expected utility of the query pool.

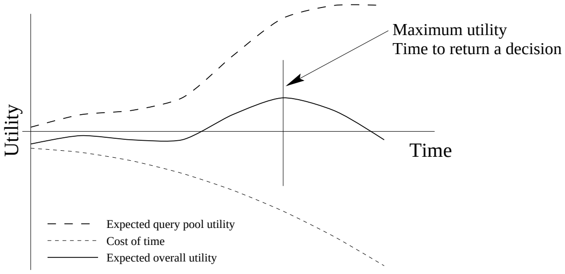

Figure 7: The expected utility curve

- **The communication layer** – The communication layer formats queries for the WWW, sends them out, monitors the pool of active queries that have not yet returned, and extracts the information from queries when they return. When a query returns, the communication layer sends the results of a query to the decision model and the response time to the information sources database. It also sends the state of the active queries and the probability of them returning to the value-driven planner.

The value-driven planner uses the decision model, the information sources database and the communication layer to repeatedly evaluate potential queries, activate the best one(if any), and determine if the expected utility function improves in the future. When the system has reached the maximum point in the expect utility curve, it halts and returns a decision.