A Resource-Bounded Interpretation-Centric Approach to Information Gathering [*]

Victor Lesser Bryan Horling Frank Klassner Anita Raja
Thomas Wagner Shelley XQ. Zhang

Computer Science Department
University of Massachusetts
Amherst, MA 01003
lesser@cs.umass.edu

Introduction

The vast amount of information available today on the World Wide Web (WWW) has great potential to improve the quality of decisions and the productivity of consumers. However, the WWW's large number of information sources and their different levels of accessibility, reliability and associated costs present human decision makers with a complex information gathering planning problem that is too difficult to solve without high-level filtering of information. In many cases, manual browsing through even a limited portion of the *relevant* information obtainable through advancing information retrieval (IR) and information extraction (IE) technologies (Larkey & Croft 1996; Lehnert & Sundheim 1991) is no longer effective. The time/quality/cost tradeoffs offered by the collection of information sources and the dynamic nature of the environment lead us to conclude that the user cannot (and should not) serve as the detailed controller of the information gathering (IG) process. Our solution to this problem is to integrate different AI technologies, namely scheduling, planning, text processing, and interpretation problem solving, into a single information gathering agent, BIG (resource-Bounded Information Gathering), that can take the role of the human information gatherer.

Our approach to the IG problem is based on two observations. The first observation is that a significant portion of human IG is itself an intermediate step in a much larger *decision-making process*. For example, a person preparing to buy a car may search the Web for data to assist in the decision process, e.g., find out what car models are available, crash test results, dealer invoice prices, reviews and reliability statistics. In this information search process, the human gatherer first *plans* to gather information and reasons, perhaps at a superficial level, about the time/quality/cost trade-offs of different possible gathering actions before actually gathering information. For example, the gatherer may know that Microsoft CarPoint site has detailed and varied information on the models but that it is slow, relative to the Kelley Blue Book site, which has less varied information. Accordingly, a gatherer pressed for time may choose to browse the Kelley site over CarPoint, whereas a gatherer with unconstrained resources may choose to browse-and-wait for information from the slower CarPoint site. Human gatherers also typically use information learned during the search to refine and recast the search process; perhaps while looking for data on the new Honda Accord a human gatherer would come across a positive review of the Toyota Camry and would then broaden the search to include the Camry. Thus the human-centric process is both top-down and bottom-up, structured, but also opportunistic. The final result of this semi-structured search process is a decision or a suggestion of which product to purchase, accompanied by the extracted information and raw supporting documents.

The second observation that shapes our solution is that WWW-based IG is an instance of the *interpretation problem*. Interpretation is the process of constructing high-level models (e.g. product descriptions) from low-level data (e.g. raw documents) using feature-extraction methods that can produce evidence that is incomplete (e.g. requested documents are unavailable or product prices are not found) or inconsistent (e.g. different documents provide different prices for the same product). Coming from disparate sources of information of varying quality, these pieces of uncertain evidence must be carefully combined in a well-defined manner to provide support for the interpretation models under consideration.

In recasting IG as an interpretation problem, we face a search problem characterized by a generally combinatorially explosive state space. In the IG task, as in other interpretation problems, it is impossible to perform an exhaustive search to gather information on a particular subject, or even in many cases to determine the total number of instances (e.g. particular word processing programs) of the general subject (e.g. word processing) that is being investigated. Consequently, any solution to this IG problem needs to support reasoning about tradeoffs among resource constraints (e.g. the decision must be made in 1 hour), the quality of the selected item, and the quality of the decision process (e.g. comprehensiveness of search, effectiveness of IE methods usable within specified time limits). Because of the need to conserve time, it is important for an interpretation-based IG system to be able to save and exploit information about pertinent objects learned from earlier forays into the WWW. Additionally, we argue that an IG solution needs to support *constructive problem solving*, in which potential answers (e.g. models of products) to a user's query are incrementally built up from features extracted from raw documents and compared for consistency or suitability against other partially-completed answers.

In connection with this incremental model-building process, an interpretation-based IG problem solution must also support sophisticated scheduling to achieve *interleaved* data-driven and expectation-driven processing. Processing for interpretation must be driven by expectations of what is reasonable, but, expectations in turn must be influenced by what is found in the data. For example, during a search to find information on word processors for Windows95, with the goal of recommending some package to purchase, an

[*] This material is based upon work supported by the Department of Commerce, the Library of Congress, and the National Science Foundation under Grant No. EEC-9209623, and by the National Science Foundation under Grant No.s IRI-9523419 and IRI-9634938, the Department of the Navy and Office of the Chief of Naval Research, under Grant No. N00014-95-1-1198. The content of the information does not necessarily reflect the position or the policy of the Government or the National Science Foundation and no official endorsement should be inferred.

agent finding Excel in a review article that also contains Word 5.0 might conclude based on IE-derived expectations that Excel is a competitor word processor.

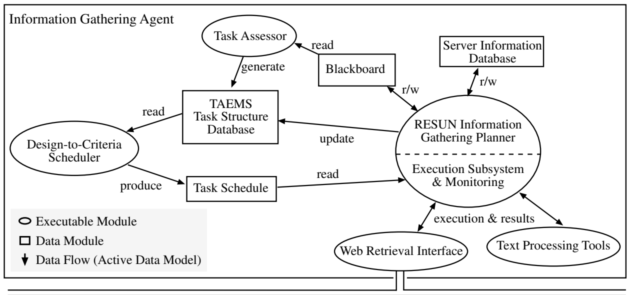
Information Gathering Agent
Task Assessor
read
generate
Blackboard
Server Information Database
r/w
r/w
TAEMS Task Structure Database
read
Design-to-Criteria Scheduler
update
RESUN Information Gathering Planner
Execution Subsystem & Monitoring
produce
Task Schedule
read
execution & results
Web Retrieval Interface
Text Processing Tools
Executable Module
Data Module
Data Flow (Active Data Model)

Figure 1: The BIG Agent Architecture

The BIG Agent Architecture

The overall BIG agent architecture is shown in Figure 1. The agent is comprised of several sophisticated components that are complex problem problem-solvers and research subjects in their own rights. The most important components are:

Task Assessor The task assessor is responsible for formulating an initial information gathering plan and then for revising the plan as new information is learned that has significant ramifications for the plan currently being executed. The task assessor is not the execution component nor is it the planner that actually determines the details of how to go about achieving information gathering goals; the task assessor is a component dedicated to managing the high-level view of the information gathering process and balancing the end-to-end top-down approach of the agent scheduler and the opportunistic bottom-up RESUN planner.

Modeling Framework The TÆMS (Decker & Lesser 1993) task modeling language is used to hierarchically model the information gathering process and enumerate alternative ways to accomplish the high-level gathering goals. The task structures probabilistically describe the quality, cost, and duration characteristics of each primitive action and specify both the existence and degree of any interactions between tasks and primitive methods. The TÆMS models serve as the medium of exchange for the components in BIG.

Design-to-Criteria Scheduler
Design-to-Criteria (Wagner, Garvey, & Lesser 1997; 1998) is a domain independent real-time, flexible computation (Horvitz, Cooper, & Heckerman 1989; Dean & Boddy 1988) approach to task scheduling. The Design-to-Criteria task scheduler reasons about quality, cost, duration and uncertainty trade-offs of different courses of action and constructs custom satisficing schedules for achieving the high-level goal(s).

RESUN Planner The RESUN (Carver & Lesser 1995) blackboard based planner/problem solver directs information gathering activities. The planner receives an initial action schedule from the scheduler and then handles information gathering and processing activities. The strength of the RESUN planner is that it identifies, tracks, and plans to resolve sources-of-uncertainty (SOUs) associated with blackboard objects, which in this case correspond to gathered information and hypothesis about the information.

Information Extractors The ability to process retrieved documents and extract structured data is essential both to refine search activities and to provide evidence to support BIG's decision making. BIG uses several information extraction techniques to process unstructured, semi-structured, and structured information ranging from a full-blown information extraction system, CRYSTAL (Fisher *et al.* 1996), to pattern matchers and table parsers/extractors.

The integration of these components in BIG, and the view of the IG problem as an interpretation task, has given BIG some very strong abilities. First there is the issue of information fusion. BIG does not just retrieve documents. Instead BIG retrieves information, extracts data from the information, and then combines the extracted data with data extracted from other documents to build a more complete model of the product at hand. RESUN's evidential framework enables BIG to reason about the sources of uncertainty associated with particular aspects of product object and to even work to find corroborating or negating evidence to resolve the SOUs. BIG also learns from previous problem solving episodes and reasons about resource tradeoffs to determine the best (satisficing) course of action for the current problem solving context.

For more information on the BIG project, please refer to our AAAI-98 paper (Lesser *et al.* 1998).

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 First International Conference on Multiagent Systems*.

Dean, T., and Boddy, M. 1988. An analysis of time-dependent planning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, 49–54.

Decker, K. S., and Lesser, V. R. 1993. Quantitative modeling of complex environments. *International Journal of Intelligent Systems in Accounting, Finance, and Management* 2(4):215–234.

Fisher, D.; Soderland, S.; McCarthy, J.; Feng, F.; and Lehnert, W. 1996. Description of the UMass Systems as Used for MUC-6. In *Proceedings of the 6th Message Understanding Conference*.

Horvitz, E.; Cooper, G.; and Heckerman, D. 1989. Reflection and action under scarce resources: Theoretical principles and empirical study. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*.

Larkey, L., and Croft, W. B. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR '96)*, 289–297.

Lehnert, W., and Sundheim, B. 1991. A performance evaluation of text analysis technologies. *AI Magazine* 12(3):81–94.

Lesser, V.; Horling, B.; Klassner, F.; Raja, A.; Wagner, T.; and Zhang, S. X. 1998. BIG: A resource-bounded information gathering agent. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*. To appear. See also UMass CS Technical Reports 98-03 and 97-34.

Wagner, T.; Garvey, A.; and Lesser, V. 1997. Complex Goal Criteria and Its Application in Design-to-Criteria Scheduling. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 294–301.

Wagner, T.; Garvey, A.; and Lesser, V. 1998. Criteria-Directed Heuristic Task Scheduling. *International Journal of Approximate Reasoning, Special Issue on Scheduling*. To appear. Also available as UMASS CS TR-97-59.