AI Magazine Volume 18 Number 2 (1997) (© AAAI)

SAVVYSEARCH

A Metasearch Engine That Learns Which Search Engines to Query

Adele E. Howe and Daniel Dreilinger

■ Search engines are among the most successful applications on the web today. So many search engines have been created that it is difficult for users to know where they are, how to use them, and what topics they best address. Metasearch engines reduce the user burden by dispatching queries to multiple search engines in parallel. The SAVVYSEARCH metasearch engine is designed to efficiently query other search engines by carefully selecting those search engines likely to return useful results and responding to fluctuating load demands on the web. SAVVYSEARCH learns to identify which search engines are most appropriate for particular queries, reasons about resource demands, and represents an iterative parallel search strategy as a simple plan.

Companies, institutions, and individuals must have a presence on the web; each are vying for the attention of millions of people. Not too surprisingly then, the most successful applications on the web to date are search engines: tools that assist users in finding information on specific topics.

A variety of search engines are available, from general, robot based (for example, ALTAVISTA [www.altavista.digital.com] and WEBCRAWLER [webcrawler.com]) to topic or area specific (for example, FTPSEARCH [ftpsearch.unit.no/ftpsearch] and DEJANEWS [www.dejanews.com]). Each uses different algorithms for collecting, indexing, and searching links; thus, each returns different results for similar queries. Empirical results indicate that no single search engine is likely to return more than 45 percent of the relevant results (Selberg and Etzioni 1995). To find what they desire, users might need to query several search engines; metasearch engines automate this process by simultaneously submitting a single query to multiple search engines.

The simplest metasearch engines are forms that allow the user to indicate which search engines should be contacted (for example, ALL-IN-ONE [www.albany.net/allinone] and METASEARCH [members.gnn.com/infinet/meta.htm]). PROFUSION (Gauch, Wang, and Gomez 1996; www.designlab.ukans.edu/profusion) gives the user the choice of selecting search engines themselves or letting PROFUSION select three of six robot-based search engines using hand-built rules. METACRAWLER (Selberg and Etzioni 1995; metacrawler.cs.washington.edu:8080/home.html) significantly enhances the output by downloading and analyzing the links returned by the search engines to prune out unavailable and irrelevant links.

Metasearch engines reduce the burden on the user. They make available search engines that might have been unknown to the user. They handle the simultaneous submission of queries; some direct the query to appropriate engines, and some postprocess the results as well. They provide a single interface (on the down side, they might not support all the features of the target search engines).

Unfortunately, metasearch can lead to the *tragedy of the commons problem* from economics in which an individual's best interests run counter to society's. Individual users appear to be best served by simultaneously searching every possible search engine on the web for desired information. However, the process might waste web resources: network load and search-engine computation.

We believe that a metasearch system can be a good web citizen (Eichmann 1994) by targeting those search engines likely to return useful results and responding to changing

Copyright © 1997, American Association for Artificial Intelligence. All rights reserved. 0738-4602-1997 / $2.00

load demands on the web. To provide this function, we incorporated simple AI techniques in a metasearch engine. Our metasearch engine learns to identify which search engines are most appropriate for particular queries, reasons about resource demands, and represents an iterative parallel search strategy as a simple plan.

Our Metasearch System: SAVVYSEARCH

SAVVYSEARCH is our metasearch system (Dreilinger and Howe 1997, 1996) available at guaraldi.cs.colostate.edu:2000. It runs on five machines (three Sun SPARCSTATIONS and two IBM RS 6000S) at Colorado State University. The system was first made available in March 1995 and has undergone several revisions since the original design. At present, two versions of the system are available: (1) the one described here and (2) an experimental interface that is mentioned in the section entitled Retrospective and Prospective Views of the SAVVYSEARCH Project.

SAVVYSEARCH is designed to balance two potentially conflicting goals: (1) maximizing the likelihood of returning good links and (2) minimizing computational and web resource consumption.

SAVVYSEARCH is designed to balance two potentially conflicting goals: (1) maximizing the likelihood of returning good links and (2) minimizing computational and web resource consumption. The key to compromise is knowing which search engines to contact for specific queries at particular times. SAVVYSEARCH tracks long-term performance of search engines on specific query terms to determine which are appropriate and monitors recent performance of search engines to determine whether it is even worth trying to contact them.

In this section, we describe SAVVYSEARCH from a user's perspective. We follow a running example, indicating what a user sees and what goes on behind the scenes in processing a search request.

Submitting a Query

To find out about artificial intelligence conferences, we enter the query, select the Integrate Results option, and click on the SAVVYSEARCH! button, as shown in an image of the interface in figure 1. The *search form*, the query interface to SAVVYSEARCH, asks the user to specify a set of keywords (query terms) and options for the search. Users typically enter two query terms.

The options cover the treatment of the terms, the display of results, and the interface language. Query terms can be combined with logical *and* (all query terms must be included in documents) or *or* (any query term should be present) or can be presented as an ordered phrase. Three aspects of the results display can be varied: (1) the number of links returned, (2) the format of the links description, and (3) the timing. By default, 10 links are displayed with the uniform resource locators (URLs) and descriptions when available, and the results of each search engine are listed separately as they arrive. Alternatively, we could change the number of links to 50, return less or more description of the links, and interleave the results of the separate search engines. The interface is also available in 23 different languages.[1]

Processing a Query

When a user submits the query, SAVVYSEARCH must make two decisions: (1) how many search engines to contact simultaneously and (2) what order the search engines should be contacted in. The first decision requires reasoning about the available resources and the second about ranking the search engines.

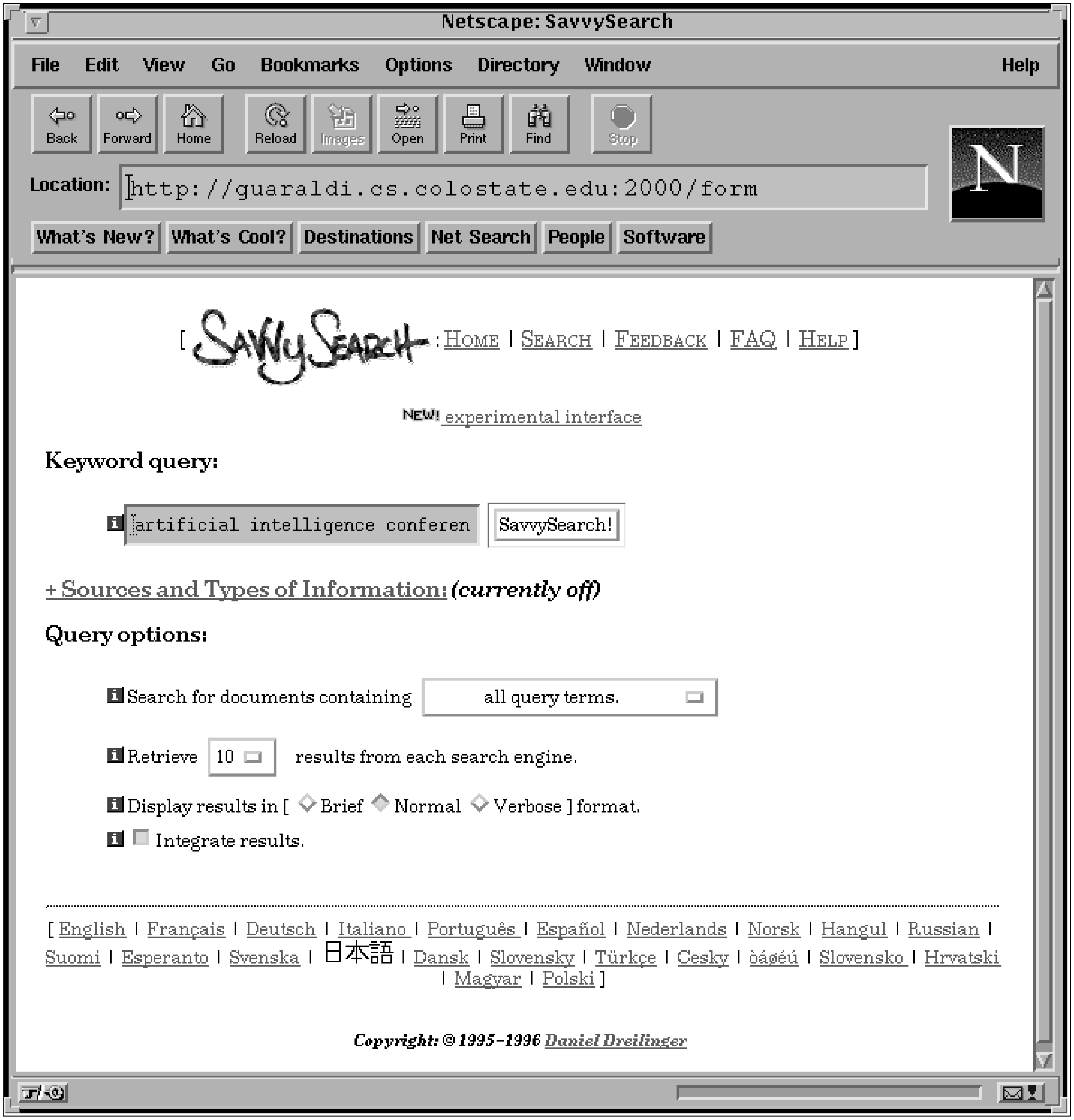
Resource Reasoning Each search engine queried expends network and local computational resources. Thus, modifying concurrency (number of search engines queried in parallel) is the best way to moderate resource consumption. Concurrency is a function of *network load estimates,* which are determined from a lookup table created from observations of the network load at this time of day in the past, and *local central processing unit load,* which is computed using the UNIX uptime command.

Concurrency has a base value of two; as many as two additional units are added for each load estimate for periods of low load. Thus, the maximum concurrency value is six.

Ranking Search Engines SAVVYSEARCH includes both large robot-based search engines and small specialized search engines in its set. The large search engines are likely to return links for any query, but these links might not be as appropriate as links returned by a specialized search engine for a query in its area.

The purpose of ranking is to determine which search engines are most worthwhile to contact for a given query. Search engines are ranked based on learned associations between search engines and query terms (stored in a metaindex) and recent data on search-engine performance.

The metaindex—A compendium of search experience: The *metaindex* maintains associations between individual query terms (simplified by stemming and case stripping) and search engines as effectiveness values. High positive values indicate excellent perfor-



Location: http://guaraldi.cs.colostate.edu:2000/form

Keyword query:

artificial intelligence conferen

+ Sources and Types of Information: *(currently off)*

Query options:

Search for documents containing all query terms.

Retrieve 10 results from each search engine.

Display results in [◇ Brief ◆ Normal ◇ Verbose] format.

□ Integrate results.

[English | Français | Deutsch | Italiano | Português | Español | Nederlands | Norsk | Hangul | Russian | Suomi | Esperanto | Svenska | 日本語 | Dansk | Slovensky | Türkçe | Cesky | òáøéú | Slovensko | Hrvatski | Magyar | Polski]

Copyright: © 1995–1996 Daniel Dreilinger

Figure 1. User Interface to SAVVYSEARCH *for Entering a Query.*

mance of a search engine on queries containing a specific term; high negative values indicate extremely poor performance.

The effectiveness values are derived from two types of observation of the results of users' searches. We used observations (passive measures) because we obtained a low rate of response to requests for user feedback as well as some questionable responses. For each search, we collect two types of information: (1) *no results,* the search engine failed to return links, and (2) *visits,* the number of links that are explored by the user. No results reduces confidence that the search engine is appropriate for the particular query, and Visits indicates that the user found some returned links to be interesting and so increases confidence.

SAVVYSEARCH uses a simple weight-adjustment scheme for learning effectiveness values. No results and Visits are treated as negative and positive reinforcement, respectively, amortized by the number of terms in the query. Thus, if a search engine returned nothing for the example query, the effectiveness values for *artificial*, *intelligence*, and *conferences* would each be reduced by one-third. Although simple, this scheme proved to be effective (see Retrospective and Prospective Views of the SAVVYSEARCH Project for a brief description of our evaluation of the learning).

Tracking recent performance: Search engines occasionally are inaccessible or slow to respond. Their network connections might be at fault, or the engines themselves might be experiencing problems or upgrades. SAVVYSEARCH monitors recent performance by recording the number of links returned (hits) and the response time for the last five queries submitted to each search engine. Given current use of the system, 5 queries corresponds to about 45 seconds for the large, general search engines and about 15 minutes for the infrequently used search engines.

Just as primary search engines are now a resource for metasearch engines, so metasearch engines should become a means of achieving higher-level goals rather than an end in itself.

Calculating rank from experiences: For a given query (q), a search engine's (s) rank ($R_{q,s}$) is its query score reduced by a penalty for recent poor performance on hits (h) and response time (r):

$$R_{q,s} = Q_{q,s} - (P_{s,h} + P_{s,r}) \ .$$

Each term is normalized to a point between 0 and 1.

The equation for $Q_{q,s}$ is based on a common approach from information retrieval called *term frequency times inverse document frequency* (Witten, Moffat, and Bell 1994). The query score, $Q_{q,s}$, sums the metaindex values for the terms in the query weighted by the ubiquity of the query term and the search engine:

$$Q_{q,s} = \sum_{t \in q} \frac{M_{t,s} * I_t}{\sqrt{T_s}}$$

$M_{t,s}$ is the weight from the metaindex of the term t for search engine s. I_t is the *inverse server frequency* of the term, which estimates the ubiquity of the query term; frequently occurring, common terms provide little information for distinguishing search-engine performance and so are discounted. T_s sums the absolute values of all metaindex values for search engine s; this term estimates the frequency of the overall use of the search engine. Because the most general search engines are likely to be used more frequently and are likely to return something for any query, their weights will tend to grow larger and more quickly than the specialized search engines. T_s mitigates the tendency toward their dominance, allowing more specialized engines to be selected when appropriate.

The penalties ($P_{s,h}$ and $P_{s,r}$) are accrued only if thresholds on the minimum number of hits and the maximum response time are exceeded. The thresholds have been set somewhat arbitrarily to an average of 1 hit and a response time of 15 seconds. Once the thresholds are passed, the penalties increase quadratically to the worst-possible values: zero for hits and a time-out of 45 seconds for response time. Details of these equations are available in Dreilinger (1996).

Dispatching a Query

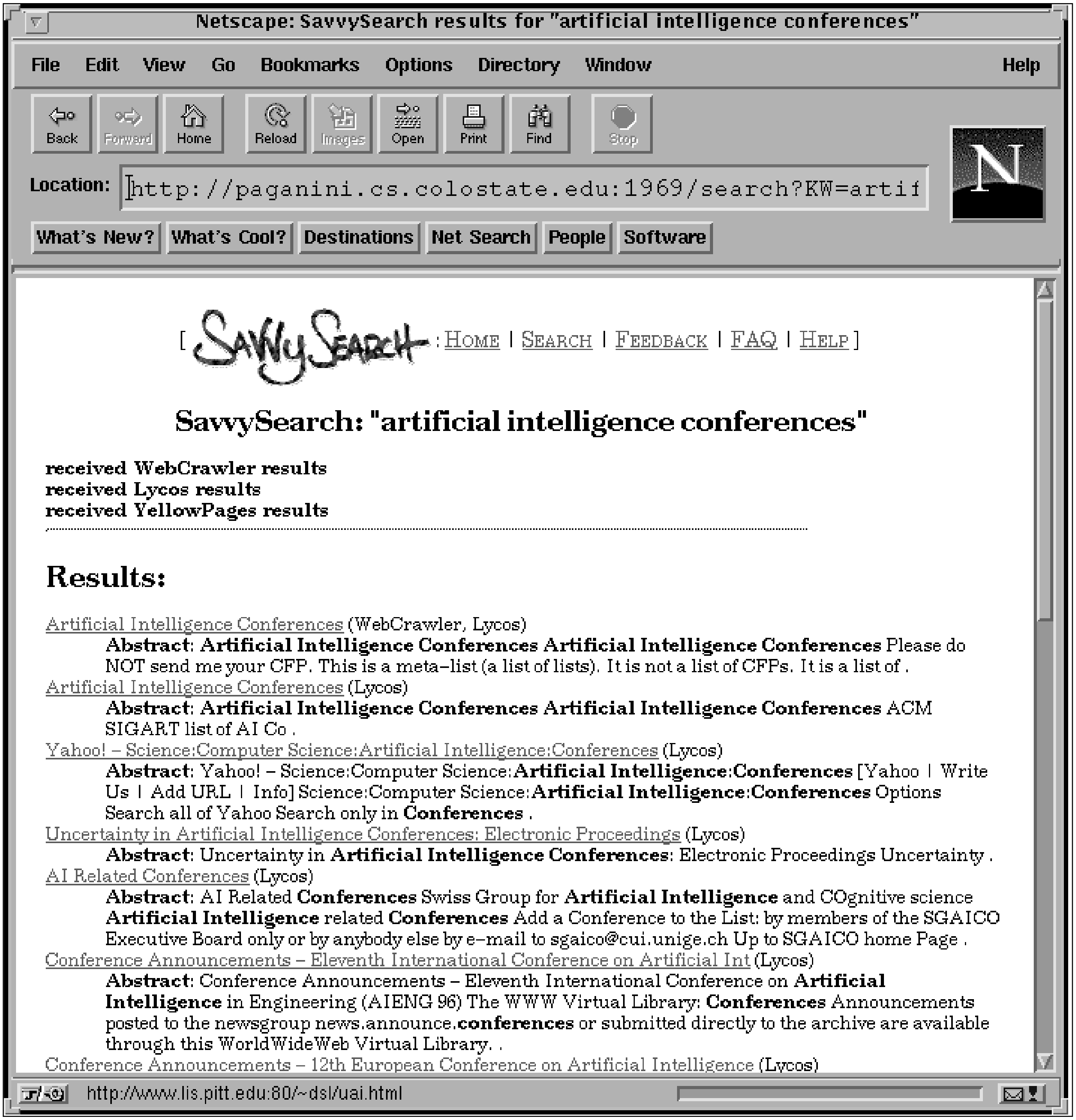
The *degree of parallelism* determines how many search engines to query; the *rank order* determines which search engines should be queried. SAVVYSEARCH dispatches the query in parallel to each of the indicated search engines. For each search engine, a *specialized interface agent* formats the query according to the interface for the search engine and submits it. The interface agent waits a preset amount of time for a response, handles errors that might occur, parses the result into a uniform format, and forwards it to another component for display.

Presenting Results

For our example query, the three search engines contacted (WEBCRAWLER, LYCOS, and YELLOW PAGES) returned 30 different links with similar names. The first nine are shown in an image of the result page in figure 2. These results were integrated; SAVVYSEARCH waited until all results had been received and then constructed a single ranked list. Results are integrated by normalizing the scores returned by search engines between 0 and 1.0 and summing them for each link; links for search engines that did not return scores were arbitrarily assigned a score of 0.5. Duplicate links are listed with the names of all the search engines that returned them.

The bottom of each results page displays the search plan, as shown in figure 3. The query was issued during a time of moderately high demand; thus, each step includes only three search engines. Search plans can include from two to six search engines for each step. The current version includes 11 search engines; the number fluctuates as search engines appear, disappear, and merge.

To supplement the results collected so far, the user can execute another step in the


Netscape: SavvySearch results for "artificial intelligence conferences"
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://paganini.cs.colostate.edu:1969/search?KW=artif
What's New? What's Cool? Destinations Net Search People Software
[SavvySearch : Home | Search | Feedback | FAQ | Help]
http://www.lis.pitt.edu:80/~dsl/uai.html

SavvySearch: "artificial intelligence conferences"

received WebCrawler results
received Lycos results
received YellowPages results

Results:

Artificial Intelligence Conferences (WebCrawler, Lycos)
Abstract: **Artificial Intelligence Conferences Artificial Intelligence Conferences** Please do NOT send me your CFP. This is a meta-list (a list of lists). It is not a list of CFPs. It is a list of .

Artificial Intelligence Conferences (Lycos)
Abstract: **Artificial Intelligence Conferences Artificial Intelligence Conferences** ACM SIGART list of AI Co .

Yahoo! – Science:Computer Science:Artificial Intelligence:Conferences (Lycos)
Abstract: Yahoo! – Science:Computer Science:**Artificial Intelligence**:**Conferences** [Yahoo | Write Us | Add URL | Info] Science:Computer Science:**Artificial Intelligence**:**Conferences** Options Search all of Yahoo Search only in **Conferences** .

Uncertainty in Artificial Intelligence Conferences: Electronic Proceedings (Lycos)
Abstract: Uncertainty in **Artificial Intelligence Conferences**: Electronic Proceedings Uncertainty .

AI Related Conferences (Lycos)
Abstract: AI Related **Conferences** Swiss Group for **Artificial Intelligence** and COgnitive science **Artificial Intelligence** related **Conferences** Add a Conference to the List: by members of the SGAICO Executive Board only or by anybody else by e-mail to sgaico@cui.unige.ch Up to SGAICO home Page .

Conference Announcements – Eleventh International Conference on Artificial Int (Lycos)
Abstract: Conference Announcements – Eleventh International Conference on **Artificial Intelligence** in Engineering (AIENG 96) The WWW Virtual Library: **Conferences** Announcements posted to the newsgroup news.announce.**conferences** or submitted directly to the archive are available through this WorldWideWeb Virtual Library. .

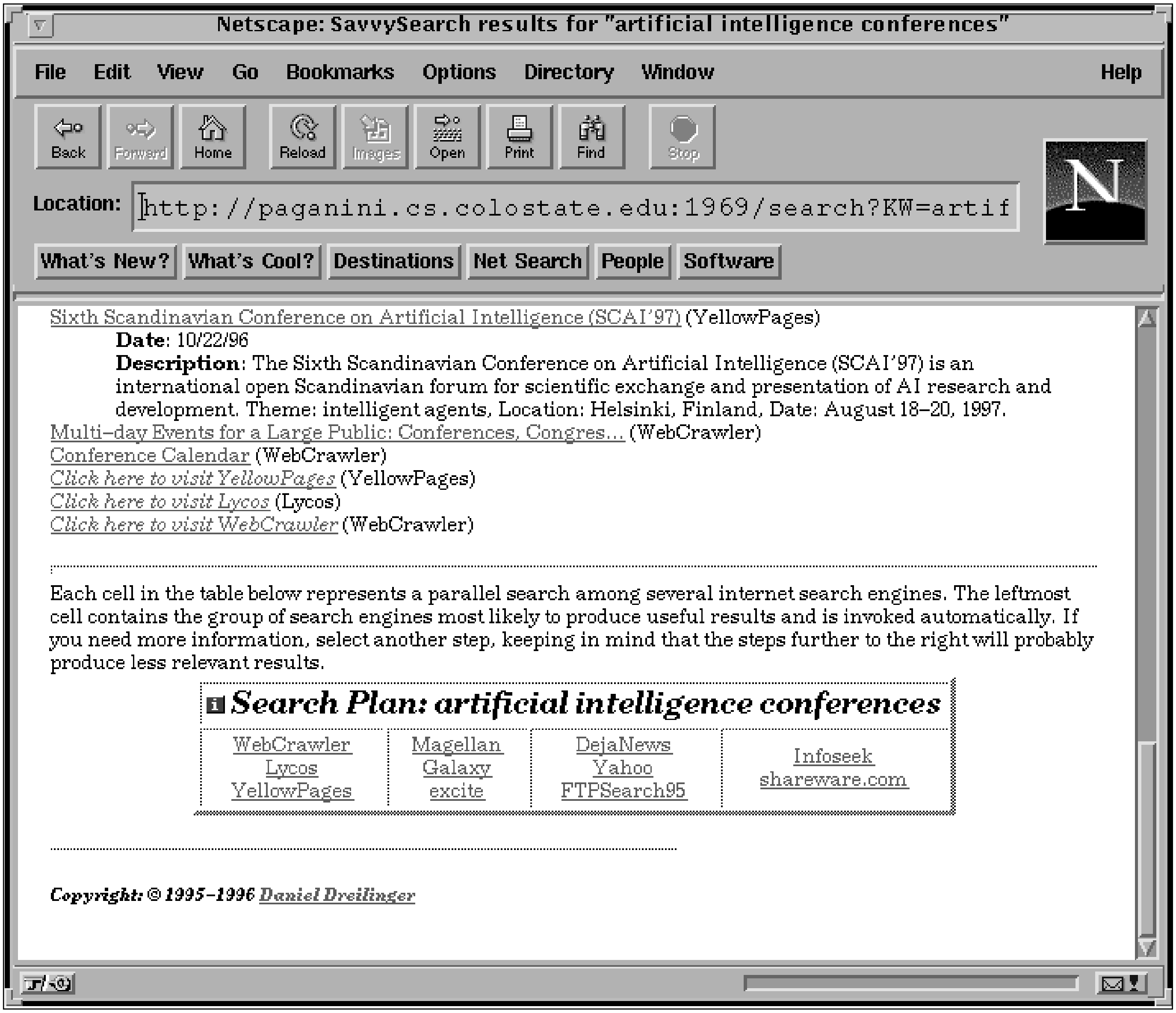
Conference Announcements – 12th European Conference on Artificial Intelligence (Lycos)

Figure 2. SAVVYSEARCH *Display of Results of Artificial Intelligence Conferences Query, Interleaved Display.*

search plan by clicking on the one that looks the most promising. The ordering represents SAVVYSEARCH's best guess about which search engines will return the best results; the user can easily override it by selecting a different step. In one long-term study, we found that users executed 1.42 steps in their search plans on average.

Retrospective and Prospective Views of the SAVVYSEARCH Project Results

The SAVVYSEARCH Project has been quite successful. Currently, SAVVYSEARCH processes over 20,000 queries each day and, based on elec-



http://paganini.cs.colostate.edu:1969/search?KW=artif

Sixth Scandinavian Conference on Artificial Intelligence (SCAI'97) (YellowPages)
Date: 10/22/96
Description: The Sixth Scandinavian Conference on Artificial Intelligence (SCAI'97) is an international open Scandinavian forum for scientific exchange and presentation of AI research and development. Theme: intelligent agents, Location: Helsinki, Finland, Date: August 18–20, 1997.
Multi-day Events for a Large Public: Conferences, Congres... (WebCrawler)
Conference Calendar (WebCrawler)
Click here to visit YellowPages (YellowPages)
Click here to visit Lycos (Lycos)
Click here to visit WebCrawler (WebCrawler)

Each cell in the table below represents a parallel search among several internet search engines. The leftmost cell contains the group of search engines most likely to produce useful results and is invoked automatically. If you need more information, select another step, keeping in mind that the steps further to the right will probably produce less relevant results.

1 ***Search Plan: artificial intelligence conferences***			
WebCrawler Lycos YellowPages	Magellan Galaxy excite	DejaNews Yahoo FTPSearch95	Infoseek shareware.com

Copyright: © 1995–1996 Daniel Dreilinger

Figure 3. Search Plan for Artificial Intelligence Conferences Query, Which Allows Users to Continue Search if Current Results Are Inadequate.

tronic mail, has attracted a large well-satisfied user base.

SAVVYSEARCH, as described here, is the result of almost two years of work and a series of studies (Dreilinger and Howe 1997). The current design of the resource reasoning and learning algorithm resulted, in part, from the results of these studies (Dreilinger 1996). We studied the effects of the learning by starting from a minimal metaindex (compiled from 2 days worth of data) and allowing it to accumulate experience over a 28-day period. We found that our simple learning algorithm improved performance on both Visits and No results overall (including all search engines and query terms): Visits averaged .36 in the first 5 days and .42 in the last 5 days; No results averaged .142 in the first 5 days and .135 in the last. Thus, this learning algorithm leads to better selection of good search engines than the pruning of poor ones.

In a follow-up study, we examined how much knowledge was needed to significantly improve performance using learning. We found that as few as 10 uses of a query term results in a halving of No Results from .18 to .09. Although Visits increases more slowly with learning, an increase from .34 to .45 is obtained with just 100 examples. By the end of the study, the most used term (5000 uses) had Visits of .76 and No results of .08. Given the level of use of most search engines (millions of queries a day), metasearch engines should have little difficulty in collecting enough data to significantly improve performance through learning.

As web use and users becomes more sophisticated, metasearch will need to be able to be personalized and embedded in other systems. Metasearch currently takes a "one-size-fits-all" approach in which the knowledge underlying query processing is shared by all users. A new experimental version of SAVVYSEARCH (available on the main web page) takes a first step toward personalization by dividing searches into eight categories; each category translates to a set of rules for creating a search plan for this type of search. Ideally, users themselves could create and store their own stereotypical search plans or a system could infer them.

Just as primary search engines are now a resource for metasearch engines, so metasearch engines should become a means of achieving higher-level goals rather than an end in itself. Intelligent-agent technology promises to alleviate the tedium and frustration of mundane tasks and navigate vast information spaces. Metasearch should be an information-gathering tool for helping human users and their intelligent agents find what they need on the web.

The web will continue to grow. Thus, search tools will continue to be critical for managing the information deluge. To keep pace with the expansion, the next generation must include far more sophisticated AI techniques than the current but retain some of the benefits of the current systems: be easy to use, require little feedback from the users, and be mindful of shared resources.

Acknowledgments

The experiments and equipment were supported in part by Advanced Research Projects Agency–Air Force Office of Scientific Research contract F30602-93-C-0100, National Science Foundation (NSF) Research Initiation Award IRI-930857, and NSF Career Award IRI-9624058. We gratefully acknowledge the assistance of Lawrence Livermore Laboratory and the Computer Science Department at Colorado State University for providing machines to run SAVVYSEARCH and Wayne Trzyna and his staff for patiently keeping the machines running.

Note

[1] We thank the users who translated the interface for us.

References

Dreilinger, D. 1996. Description and Evaluation of a Meta-Search Agent. Master's thesis, Computer Science Dept., Colorado State University.

Dreilinger, D., and Howe, A. 1997. Experiences with Selecting Search Engines Using Meta-Search. *ACM Transactions on Information Systems*. Forthcoming.

Dreilinger, D., and Howe, A. 1996. An Information-Gathering Agent for Querying Web Search Engines, Technical Report, TR 96-11, Computer Science Department, Colorado State University.

Eichmann, D. 1994. Ethical Web Agents. In Electronic Proceedings of the Second World Wide Web Conference '94: MOSAIC and the Web, 17–20 October, Chicago, Illinois.

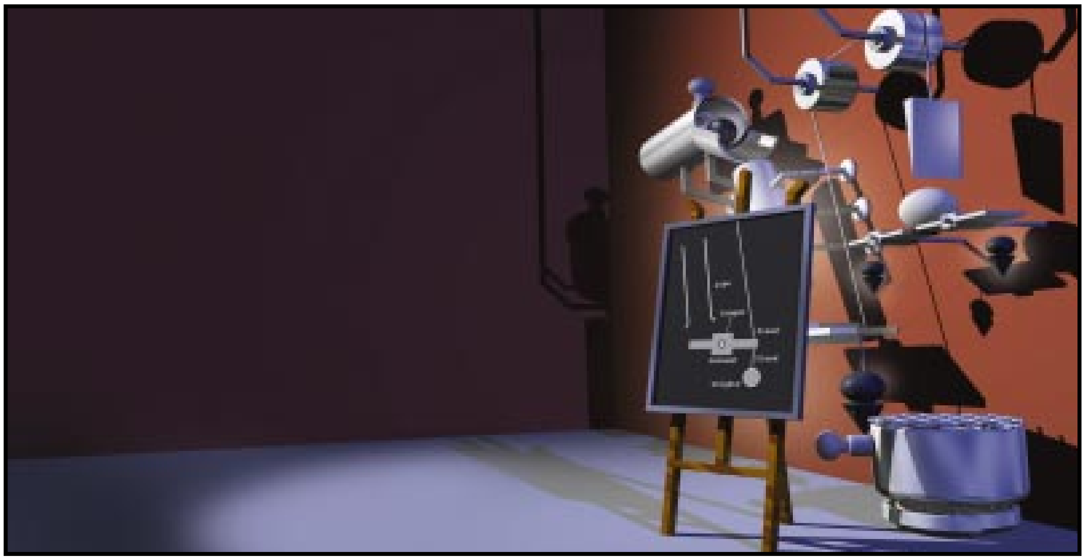
Gauch, S.; Wang, G.; and Gomez, M. 1996. PROFUSION: Intelligent Fusion from Multiple, Different Search Engines. *Journal of Universal Computer Science* 2(9).

Selberg, E., and Etzioni, O. 1995. Multi-Service Search and Comparison Using the METACRAWLER. Presented at the Fourth International World Wide Web Conference, 11 to 14 December, Boston, Massachusetts.

Witten, I. H.; Moffat, A.; and Bell, T. C. 1994. *Managing Gigabytes: Compressing and Indexing Documents and Images*. New York: Van Nostrand Reinhold.

Adele E. Howe is an assistant professor in the Computer Science Department at Colorado State University. Her research areas include planning, agent architectures, information gathering on the World Wide Web, and evaluation and modeling of intelligent systems. She received her B.S.E. in computer science from the University of Pennsylvania and her M.S. and Ph.D. in computer science from the University of Massachusetts. She can be reached at howe@cs.colostate.edu or www.cs.colostate.edu/~howe.

Daniel Dreilinger received a B.A. in mathematics from the University of California at San Diego in 1992. In 1996, he received an M.S. in computer science from Colorado State University, where he created SAVVYSEARCH. Currently, he is a research assistant at the MIT Media Lab working on a second M.S. Research interests include information filtering and retrieval, intelligent agents, and electronic commerce. He can be reached at daniel@media.mit.edu or www.media.mit.edu/~daniel/.



Diagrammatic Reasoning

Cognitive & Computational Perspectives

Edited by Janice Glasgow, N. Hari Narayanan, and B. Chandrasekaran

Foreword by Herbert Simon

> "Understanding diagrammatic thinking will be of special importance to those who design human-computer interfaces, where the diagrams presented on computer screens must find their way to the Mind's Eye.… In a society that is preoccupied with 'Information Superhighways,' a deep understanding of diagrammatic reasoning will be essential to keep the traffic moving." *– Herbert Simon*

Diagrammatic reasoning—the understanding of concepts and ideas by the use of diagrams and imagery, as opposed to linguistic or algebraic representations—not only allows us to gain insight into the way we think but is a potential base for constructing representations of diagrammatic information that can be stored and processed by computers.

Diagrammatic Reasoning brings together nearly two dozen recent investigations into the cognitive, the logical, and particularly the computational characteristics of diagrammatic representations and the reasoning that can be done with them. Following a foreword by Herbert Simon (coauthor of one of the most influential papers on reasoning with diagrams, which is included here) and an introduction by the editors, chapters provide an overview of the recent history of the subject, survey and extend the underlying theory of diagrammatic representation, and provide numerous examples of diagrammatic reasoning (human and mechanical) that illustrate both its powers and its limitations. Each of the book's four sections begins with an introduction by an eminent researcher who provides an interesting personal perspective while he or she places the work in proper context.

ISBN 0-262-57112-9 800 pp., index. $50.00 softcover

The AAAI Press • Distributed by The MIT Press
Massachusetts Institute of Technology, Cambridge, Massachusetts 02142
To order, call toll free: (800) 356-0343 or (617) 625-8569. MasterCard and VISA accepted.