### Planning-Based Information Agents

#### 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington
3935 University Way NE, Box 355754
Seattle, WA 98195

#### 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dept. of the Navy, Office of Naval Research
Seattle Regional Office
1107 NE 45th Street, #350
Seattle, WA 98105-4631

#### 11. SPONSOR/MONITORING AGENCY REPORT NUMBER
ONR

#### 12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release, distribution is unlimited.

#### 15. SUBJECT TERMS

#### 16. SECURITY CLASSIFICATION OF:
<table>
<thead>
<tr>
<th>a. REPORT</th>
<th>b. ABSTRACT</th>
<th>c. THIS PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UU</td>
<td>U</td>
</tr>
</tbody>
</table>

#### 17. LIMITATION OF ABSTRACT

#### 18. NUMBER OF PAGES

#### 19a. NAME OF RESPONSIBLE PERSON
Carol Zuiches

#### 19b. TELEPHONE NUMBER (Include area code)
206-543-4043

---

**REPORT DOCUMENTATION PAGE**

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**
Final Report for "Planning-Based Information Agents"
Grant number N00014-98-1-0147
Daniel S. Weld
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195
weld@cs.washington.edu

SUMMARY
Networked information systems are making so much data available that people can’t find it themselves. Software agent technology promises to amplify human decision making capabilities by gathering information from disparate sources in parallel and integrating it in real time. However, in order to make today’s prototype systems realize their potential, several bottlenecks must be overcome. First, information-gathering agents need robust and efficient execution so they can process large data sets, cope with network failure and site outage. Secondly, in order to scale to the level of thousands of information sources, agents need algorithms for locating information sources; automatically creating wrappers for those sources, processing XML based representations of those sources, and routing queries to the appropriate sources.

PROGRESS
We’ve formulated the problem of wrapper induction, proved some theoretical PAC bounds on the performance of such systems, devised a number of learning algorithms that solve the problems for different classes of sources, implemented the algorithms, and performed empirical tests on the implementations. Many others have extended our seminal results.

We’ve built a prototype system that automatically identifies, classifies, wraps, and query routes to over ten thousand specialized information sources. Key ideas include two novel methods for query routing: intelligent probing of CGI scripts to determine their expertise and using the Yahoo categorization of specialized information sources as a kind of semantic networks.

We’ve built the MULDER system which takes natural language questions, parses them, composes a set of Internet search engine queries of differing specificity using novel paraphrasing technology, sends the queries to an engine such as Google, downloads likely pages returned by Google, parses regions of the resulting pages, extracts candidate answers to the original questions, and votes to determine which are the most likely correct answer(s). MULDER outperforms commercial systems such as Google and AskJeeves. Ablations studies show the benefit derived by each of our techniques.

We’ve developed extensions to the proposed W3C standard XML query language allowing for updates to XML documents. We have implemented a dozen different update methods on a variety of relational encodings of XML data, and performed experiments to
determine which methods work best. We've implemented a highly optimized execution system for data integration. The resulting system, Tukwila, can handle four orders of magnitude more data than its predecessor system Razor. Key ideas include adaptivity at all levels of the architecture, interleaved planning and execution, and a novel double-pipeliined join algorithm which greatly reduces latency when combining data from sources connected via low or medium-speed networks.

We've extended Tukwila to natively handle semi structured, XML information. Our algorithms leverage existing database technology yet incorporates novel query processing operators (such as XScan). Detailed empirical experiments show that our methods vastly outperform previous methods.

We've extended planning technology to handle interleaved query planning and execution as well as traditional AI planning in the context of uncertainty.

We've built two new planning systems. TGP is a temporal planner that uses Graphplan-like mutual exclusion reasoning to achieve impressive performance. LPSAT compiles resource planning problems into a combined linear-programming/propositional satisfiability representation, which is then solved using a novel combination of incremental simplex and Davis-Putnam systematic SAT algorithms.

Finally, we've implemented the Tiramisu web site management system. Tiramisu separates the design of a web site from its implementation, allowing the use of multiple implementation tools while supporting a high-level declarative model of the site.

**ACCOMPLISHMENTS**

Design, implementation and test of next-generation, scalable, fully autonomous wrapper creation system.

Implementation of prototype web resource detector and query routing system.

Implementation and testing of Tukwila adaptive execution system for information integration.

Design, implementation and experimentation on MULDER, the first fully automated question-answering system for the WWW.

Design of XML update language, comparative implementation of update methods and experimental evaluation.

Experiments showing utility of double pipelined join, interleaved planning and execution, and other Tukwila features.

Design, implementation and experimentation on conformant graphplan.

Design, implementation and experimentation on contingent graphplan.
Design, implementation and experimentation on factored expansion graphplan.

Design, implementation and experimentation on TGP temporal planner.

Design, implementation and experimentation on LPSAT resource planner.

Design and implementation of Tiramisu web-site management system.

TRANSITIONS
My primary collaborators are Professor Oren Etzioni and Professor Alon Halevy, both at the University of Washington, and Dr. David Smith at NASA Ames Research Center. Our work on wrapper induction has been adopted and extended by Professor Nick Kushmerick (former student) now at Dublin City University, Ireland, by Dr. Steve Minton and Dr. Craig Knoblock at ISI, and by the group of Professor Tom Mitchell at CMU. Nimble technology (a startup company which I co-founded with Professor Halevy) has licensed the Tukwila data integration system. NASA is interested in fielding our planning work. The W3C standards body is considering incorporating our XML update methods in the next standard.

AWARDS
I was made AAAI Fellow for my “significant contribution to the development of qualitative reasoning methods, software agent technology, and plan synthesis algorithms.”

I was presented with the WRF / TJ Cable Endowed Professorship.

PUBLICATIONS


REPORT OF INVENTIONS AND SUBCONTRACTS
(Pursuant to "Patent Rights" Contract Clause) (See Instructions on back)

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0990-0982), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4328. Respondents should be aware that no provision of law requires persons to respond to any collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR COMPLETED FORM TO THIS ADDRESS. RETURN COMPLETED FORM TO THE CONTRACTING OFFICER.

1. a. NAME OF CONTRACTOR/SUBCONTRACTOR
   University of Washington
   N00014-96-1-0147

   b. ADDRESS (Include ZIP Code)
   3935 University Way NE, Box 355754
   Seattle, WA 98195-
   19971101

SECTION I - SUBJECT INVENTIONS

5. "SUBJECT INVENTIONS" REQUIRED TO BE REPORTED BY CONTRACTOR/SUBCONTRACTOR "None," so state

<table>
<thead>
<tr>
<th>NAME(S) OF INVENTOR(S)</th>
<th>TITLE OF INVENTION(S)</th>
<th>DISCLOSURE NUMBER, PATENT APPLICATION SERIAL NUMBER OR PATENT NUMBER</th>
<th>ELECTION TO FILE PATENT APPLICATIONS (P)</th>
<th>CONFIRMATORY INSTRUMENT OR ASSIGNMENT FURNISHED TO CONTRACTING OFFICER (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td>(1) UNITED STATES (2) FOREIGN</td>
<td>(1) YES (2) NO</td>
</tr>
</tbody>
</table>

1. EMPLOYER OF INVENTOR(S) NOT EMPLOYED BY CONTRACTOR/SUBCONTRACTOR

<table>
<thead>
<tr>
<th>(1) (a) NAME OF INVENTOR (Last, First, Middle Initial)</th>
<th>(2) (a) NAME OF INVENTOR (Last, First, Middle Initial)</th>
<th>(1) TITLE OF INVENTION</th>
<th>(2) FOREIGN COUNTRIES OF PATENT APPLICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SECTION II - SUBCONTRACTS (Containing a "Patent Rights" clause)

6. SUBCONTRACTS AWARDED BY CONTRACTOR/SUBCONTRACTOR "None," so state

<table>
<thead>
<tr>
<th>NAME OF SUBCONTRACTOR(S)</th>
<th>ADDRESS (Include ZIP Code)</th>
<th>SUBCONTRACT NUMBER(S)</th>
<th>FAR &quot;PATENT RIGHTS&quot;</th>
<th>DESCRIPTION OF WORK TO BE PERFORMED UNDER SUBCONTRACT(S)</th>
<th>SUBCONTRACT DATES (YYYYMMDD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SECTION III - CERTIFICATION

7. CERTIFICATION OF REPORT BY CONTRACTOR/SUBCONTRACTOR (Not required if "X" as appropriate)

<table>
<thead>
<tr>
<th>NAME OF AUTHORIZED CONTRACTOR/SUBCONTRACTOR OFFICIAL (Last, First, Middle Initial)</th>
<th>TITLE</th>
<th>SIGNATURE</th>
<th>DATE SIGNED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zuiches, Carol</td>
<td>Director, Grant &amp; Cont. Svcs.</td>
<td>Signature</td>
<td>6/21/99</td>
</tr>
</tbody>
</table>

DD FORM 882, JAN 1999 (EG)

PREVIOUS EDITION MAY BE USED.
Date: Tue, 26 Jun 2001 16:46:51 -0700 (PDT)
From: Alicen L. Smith <asmith@cs.washington.edu>
To: collins5@u.washington.edu
Subject: Re: our final report (fwd)

---------- Forwarded message ---------
Date: Fri, 22 Jun 2001 13:28:41 -0400
From: Daniel Weld <weld@dakobed.com>
To: "asmith@cs.washington.edu" <asmith@cs.washington.edu>
Subject: Re: our final report

No patents
Thanks

Dan

-----Original Message-----
From: Alicen L. Smith <asmith@cs.washington.edu>
To: Dan Weld <weld@cs.washington.edu>
Sent: Fri Jun 22 09:47:49 2001
Subject: Re: our final report

There may be a few more questions, but so far all I need to know is:

Did you acquire any patents as a result of this research?

If so, please list:

   Inventors' names
   title of invention(s)
   patent (application) number

Thanks,
Alicen