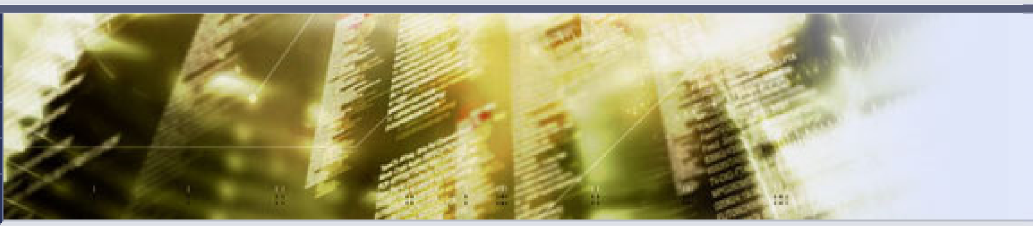


NCSA teams up with U.S. Army to devise smart Web crawling system

News Home
Calendar
Images
Video on Demand
Subscribe to Our Newsletter
Frequently Asked Questions

released 12.05.08

By Erika Strebel

The surface Web is a vast ocean of millions of pages. People search the surface Web every day by putting keywords into search engines like Google or Yahoo! Search.

But when an Army trainer needs to quickly find information to put together a training session, a simple keyword search isn't always the most efficient way to navigate the surface Web. A search for "arms" would pull up everything from 19th century firearms to adjustable rate mortgages. A similar image search turns up a melee of pictures of octopus tentacles, family coats of arms and human appendages. It could take hours to refine a search—hours that a trainer may not have.

To address the need for more focused, faster searching, the Army and private companies teamed up with NCSA researchers.

NCSA's Alan Craig and graduate research assistant Yunliang Liang created a Web crawling system that allows users to create a searchable database of relevant pages and websites, search within that database and choose how to rank the results of that search. NCSA researcher Andrew Wadsworth oversaw the project and aided in designing both the search system and program user interface.

Getting started

In October 2007, the U.S. Army and Vertex Solutions, a software engineering company that specializes in training software, approached NCSA with a proposal involving the creation of an information database for an Army training software prototype called Training Assistant.

"We evaluated several universities and other not-for-profit research groups," says Vertex representative Amanda Palla. "The combination of demonstrable expertise in the area of Web crawlers, the collaborative attitude of NCSA staff, and the proximity to Vertex's Champaign office made working with NCSA an easy decision for us."

Creating the crawler

Craig had previously worked on a surface Web mining project called VIAS. VIAS was a Linux-based system that automatically created databases.

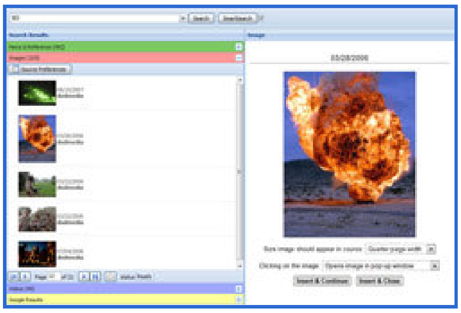
Rather than modify VIAS for the Army's needs, the NCSA team decided to create a new system based on open-source software. The Army wanted to work in Windows and Craig wanted to work with fresh technology.

To begin creating the database of surface Web content, a user needs to provide some keywords, keyword combinations, and several URLs as a starting point for the crawler. From there, the crawler can start searching the surface Web.

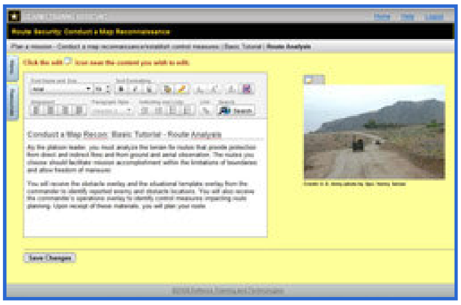
"Our goal was to present a much more focused database of information that we knew would be very pertinent to the needs of the Army trainer," says Craig.

The team started with a list of terms and URLs collected by

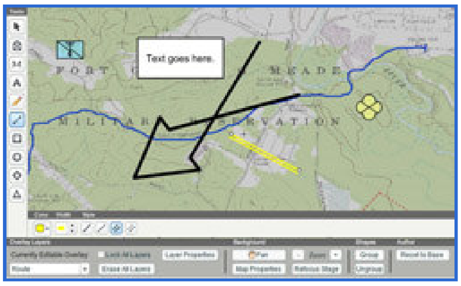
Email this story
Subscribe to our newsletter



The Training Assistant also allows users to search for images within related databases and insert them into their training module material.

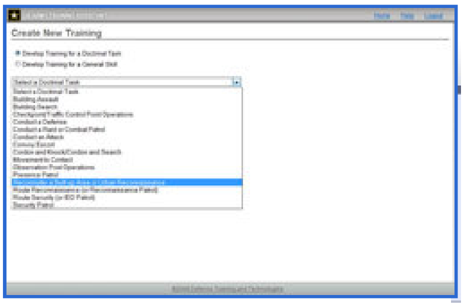


NCSA's team of researchers created a surface Web search system that lets trainers quickly find related text to insert into the software's text editor.



Army trainers can also use the Training Assistant's overlay map editor for training exercises. Trainees can draw and write on maps then compare them to the trainer's map.

Anna Cianciolo, researcher with Command Performance Research, Inc. She worked with Army personnel to define the key terms used to limit the database. The Army supplied websites and other examples of databases so the NCSA team could define exactly what the Army wanted.


Create New Training
Develop Training for a Doctrinal Task
Develop Training for a General Skill
Select a Doctrinal Task
Select a Doctrinal Task
Building Assault
Building Search
Checkpoint/Traffic Control Point Operations
Conduct a Defense
Conduct a Raid or Combat Patrol
Conduct an Attack
Convoy Escort
Cordon and Knock/Cordon and Search
Movement to Contact
Observation Post Operations
Presence Patrol
Route Reconnaissance (or Reconnaissance Patrol)
Route Security (or IED Patrol)
Security Patrol

The Training Assistant interface allows Army trainers to choose the kind of training module they want to design.

"We basically gave it a lot of parameters of what we do and don't care about and where to start looking for info," says Craig.

The process of limiting the database seems straightforward, but it was one of more difficult parts of the project.

"Deciding what goes in or out of the database is a hard question," says Craig. "What goes into the database depends on who you talk to."

Eventually, they decided to create two databases: one with tighter parameters and one with looser parameters. That way, if a relevant page doesn't make it into the tighter database, the crawler will eventually find it.

The NCSA team also spent time tweaking the crawler's search functions to make sure irrelevant data didn't somehow slip into the databases.

"It's amazing what you'll catch," says Wadsworth. "It's like fishing: You never know what you're going to pull out."

The team examined the pages within the database, looking for irrelevant pages or relevant pages not included and determining what other terms needed to be added or excluded from the database parameters.

While the crawling system makes searching the surface Web faster and easier, it also observes online rules and etiquette.

"You can't just go on to MIT's Web site and pound it day and night," says Wadsworth. "You've got to be friendly and nice."

Instead of bombarding a relevant website with continuous hits, the crawler has been programmed to return to a website periodically to collect information.

"These Web crawlers need to comply to a set of rules and etiquette so it doesn't disturb any Web server out there and cause them to go crazy," Wadsworth says. "They can see what's hitting them."

In addition, the crawler works within parameters of the robots.txt file each website has. The file tells crawlers which website directories they may and may not access.

More than crawling

But the NCSA system is more than just a crawler—it also analyzes the results it produces. The crawler has its own ranking system that the user can modify. It uses special algorithms to rank results so the best results come back.

"With the Army we wanted to be in control of what order we present the results in and how we present the results," says Craig.

Unlike Google and other commercial search engines, the crawler allows a user to define how results are ranked.

"The goal is to make it easier for the Army trainers so that they didn't have to go to Google and sift through 9 million results," says Craig.

They can choose to rank results by date, relevance, key term and domain name. In the case of the Training Assistant, the crawler can place results from .mil on a higher priority than those from a .com site.

"It's something easy, fast and all relative," says Wadsworth.

Putting it all together

The NCSA team's Web crawler was integrated with the user interface that Vertex had designed.

Also, because the Army expressed an interest in mining the deep Web—the databases, live newsfeeds and data hidden behind form searches that makes a large segment of the data available online—Craig suggested collaboration with Cazoodle, a software company launched by former NCSA Faculty Fellow Kevin Chang, a professor in the University of Illinois Computer Science Department.

"Alan immediately realized, 'Why reinvent this? Let's go and see what (Cazoodle) has,'" says Wadsworth. "It was so beautiful, we couldn't ask for more."

By combining Vertex's interface with the two Web searching systems within Teaching Assistant, Army trainers using the software can copy text and images from search results of various databases within a single program.

"That's what was so golden about this, the ability to lift information right from the results," says Wadsworth.

Moreover, the different systems multitask during each search. The programs are integrated to allow various components of the Training Assistant to talk to each other and process a user's query. Thus, the Training Assistant can automatically refine a user's search and filter results.

Liang worked with Cazoodle programmer Paul Yuan to ensure that the search systems were compatible with each other and with the final Training Assistant program.

"It's not just a computer program talking to a human," says Craig. "While a human starts the process, there is a lot of integration of various systems to combine the results, rank them, and present them to the user."

Crawling in the future

While the new surface Web crawling system was specifically made for the Army's Training Assistant, Craig wants to make the system more generally accessible.

The NCSA team is working on creating a user-friendly interface for the Web crawling system. Their goal is to create an interface so researchers can easily define their database limits and ranking heuristics without having to know detailed computer programming. They also are working on creating a friendly interface for accessing the resulting databases.

"What we want to do is generalize what we've built and make it broadly applicable to different NCSA communities and a resource for other projects," says Craig.

Team members
Kevin Chang, Computer Science Department
Anna T. Cianciolo, Command Performance Research, Inc.
Alan Craig, NCSA
Yunliang Liang, NCSA
Amanda Palla, Vertex Solutions
Andrew Wadsworth, NCSA
Tim Wentling, NCSA
Paul Yuan, Cazoodle, Inc.

Digg submit del.icio.us Slashdot