



University of Waikato researchers are opening the way for 'intelligent computing' by expanding the world's largest machine-readable knowledge base.

Project boosts data growth

The world's largest formalised knowledge base has just got a whole lot bigger, thanks to the joint efforts of researchers in computer science and philosophy at the University of Waikato.

Started 25 years ago in the United States, the Cyc artificial intelligence project aims to develop a comprehensive collection of common-sense knowledge that will ultimately allow computers to perform human-like reasoning.

This knowledge might include that a person can't be the capital of a country, or that you can't work at a job before you were born – obvious to humans but not to a computer.

Up to now, almost all the information held in Cyc has been loaded manually, but a summer research project by two Waikato students has succeeded in 'mining' data from Wikipedia and automatically transferring it to Cyc.

As a result of their work, some 45,000 new concepts have been automatically added to Cyc, increasing the number of common-sense knowledge concepts by 50 per cent – something that's never

been achieved before in Cyc's history, the researchers say.

Sam Sarjant, who's currently doing a doctorate in artificial intelligence in the Department of Computer Science, and Michael Robinson, who's completing separate degrees in cognitive science and Chinese, were recipients of two of the 70 Summer Research Scholarships awarded by the University of Waikato last year. The scholarships are worth up to \$4000 and students complete their research over the 10-week summer study break.

Mr Sarjant was also one of 14 winners of a Top Achiever Doctoral Scholarship, this week, for his doctoral research into artificial intelligence. The Tertiary Education Commission-funded scholarship provides Mr Sarjant with \$93,000 over three years. Five Waikato University students were awarded scholarships in the announcement made on Monday.

During the Summer Research Scholarship, Mr Sarjant dealt mainly with the programming side of the project, which was supervised by Dr Catherine Legg

in the Department of Philosophy and Religious Studies. Mr Robinson helped with the programming and data collection, and worked with Dr Legg on running an evaluation of the algorithm.

"We expanded on a 2008 project finding mappings between Cyc and Wikipedia categories," explains Mr Sarjant, "We created an algorithm to automatically find mappings between, say, the Cyc concept 'dog' and the Wikipedia article on dogs, which included a double checking mechanism using reverse mappings."

The task of mapping Wikipedia articles to categories in Cyc was made more difficult because the structure of the two knowledge bases is quite different, says Mr Sarjant. "Wikipedia's structure is more associative, so for example there are links from 'dog' to 'flea' to 'kennel' to a TV programme called 'Puppybowl'. In Cyc, the categories form a hierarchical structure, so 'Spot' is a 'boxer', a 'boxer' is a 'dog', and a 'dog' is a 'canine'."

Using Wikipedia Miner soft-



“Our automation has enormous potential to save money and time in adding knowledge to Cyc.”

Dr Catherine Legg

ware developed by fellow Waikato PhD student David Milne, Mr Sarjant built in special techniques to ensure correct mappings to Cyc by scanning words in the first sentence of the Wikipedia article and the article’s information boxes, noting which words were capitalised and which weren’t, and analysing the links back to other articles.

“The beauty of the technique we’ve developed is that as Cyc gets bigger, its increased background knowledge can be used to make it bigger again – a sort of bootstrapping if you like,” says Mr Sarjant.

“We’ve shown that by running the algorithm through the material a second time, we can increase the number of concepts by another 25 per cent.”

Mr Sarjant will present the research project’s findings at an international conference on web intelligence in Milan later this month.

“To date the Cyc knowledge base has been largely handcrafted by so-called ‘ontological engineers’, most of whom have PhDs in philosophy, so our automation has enormous potential to save money and time in adding knowledge to Cyc,” says Dr Legg.

“And expanding Cyc in this way will make it much more useful. Some possible applications include making computers able to learn by reading human-generated text such as newspapers, making web searches more intelligent.”

■ Network: www.cyc.com
 Story supplied by University of Waikato



Master mapper: Computer Science Phd student Sam Sarjant, relaxing with dogs Riley and Desmond, is all about finding smart ways to categorise concepts in a knowledge database.
 Photo: KATRINA BIELESKI