

Researchers Use Wikipedia To Make Computers Smarter

Using Wikipedia, Technion researchers have developed a way to give computers knowledge of the world to help them “think smarter,” making common sense and broad-based connections between topics just as the human mind does. The new method will help computers filter e-mail spam, perform Web searches and even conduct intelligence gathering at more sophisticated levels than current programs.

Researchers at the Technion-Israel Institute of Technology have found a way to give computers encyclopedic knowledge of the world to help them “think smarter,” making common sense and broad-based connections between topics just as the human mind does.

The new method will help computers filter e-mail spam, perform Web searches and even conduct electronic intelligence gathering at a much more sophisticated level than current programs, according to researchers Evgeniy Gabrilovich and Shaul Markovitch of the Technion Faculty of Computer Science. The findings will be presented next week in Hyderabad, India during the Twentieth International Joint Conference for Artificial Intelligence.

The program devised by the Technion researchers helps computers map single words and larger fragments of text to a database of concepts built from the online encyclopedia Wikipedia, which has over one million articles in its English-language version. The Wikipedia-based concepts act as “background knowledge” to help computers figure out the meaning of the text entered into a Web search, for instance.

Giving computers this deeper knowledge has been a long-standing problem in artificial intelligence, according to Markovitch. “Humans use a significant amount of background knowledge” to understand text, “but we didn’t know how to have computers access such knowledge,” he said.

Most Web search and e-mail filter programs appear smart by calculating how often certain words appear in two texts, Markovitch explained. “But what is common to all these applications is that the programs that actually do this kind of thing don’t understand text. They treat text as a collection of words, but they don’t understand the meaning of words.”

This shallow understanding is what makes an e-mail spam filter block all messages containing the word “vitamin,” but fail to block messages containing the word “B12.” “If the program never saw “B12” before, it’s just a word without any meaning. But you would know it’s a vitamin,” Markovitch said.

“With our methodology, however, the computer will use its Wikipedia-based knowledge base to infer that “B12” is strongly associated with the concept of vitamins, and will correctly identify the message as spam,” he added.

Or, computers could look at a chunk of text about Saddam Hussein and weapons of mass destruction and know that it is conceptually related to topics such as the Iraq war and U.S. Senate debates on intelligence—even if those terms do not appear anywhere in the original text.

The method also helps computers figure out ambiguous terms—deciding, for instance, whether the word “mouse” refers to the computer device or the fuzzy animal. This can be especially important in translated documents, Markovitch said.

In the near future, the Technion researchers hope to improve their method by adding information from the

Web page links inside Wikipedia articles. They are already pursuing a patent on their work, which they say will be of interest to the intelligence community and Web search engine companies, among others.

Source: American Technion Society

This document is subject to copyright. Apart from any fair dealing for the purpose of private study, research, no part may be reproduced without the written permission. The content is provided for information purposes only.