Technology PUBLISHED BY MIT **Review**

Friday, July 27, 2007

Building a Better Search Engine

A new natural-language system is based on 30 years of research at PARC.

By Michael Reisman

Powerset, Inc. (http://www.powerset.com/) , based in San Francisco, is on the verge of offering an innovative natural-language search engine, based on linguistic research at the Palo Alto Research Center (http://www.parc.xerox.com/) (PARC). The engine does more than merely accept queries asked in the form of a question. The company claims that the engine finds the best answer by considering the meaning and context of the question and related Web pages.

"Powerset extracts deep concepts and relationships from the texts, and the users query and match them efficiently to deliver a better search," Powerset CEO Barney Pell says.

Even though attempts have been made at natural-language search for decades, Powerset says that its system is different because it has solved some of the fundamental technological problems that have existed with this kind of search. It has done so by developing a product that is deep, computationally advanced, and still economically viable.

Pell says that it's difficult to pinpoint one particular technological breakthrough, but he believes that Powerset's superiority lies in the three decades of hard work by scientists at PARC. (PARC licensed much of its natural-language search technology to Powerset in February.) There was not one piece of technology that solved the problem, Pell says, but instead, it was the unification of many theories and fragments that pulled the project together.

"After 30 years, it's finally reached a point where it can be brought into the world," he says.

A key component of the search engine is a deep natural-language processing system that extracts the relationships between words; the system was developed from PARC's Xerox Linguistic Environment (XLE) platform. The framework that this platform is based on, called Lexical Functional Grammar, enabled the team to write different grammar engines that help the search engine understand text. This includes a robust, broad-coverage grammar engine written by PARC. Pell also claims that the engine is better than others at dealing with ambiguity and determining the real meaning of a question or a sentence on a Web page. All these innovations make the system more adaptable, he says, so that it can extract deep relationships from text.

Powerset chief technology officer Ron Kaplan has led PARC's XLE team since the 1970s and is the author of much of the technology behind XLE that has been licensed to the company. Kaplan says that he and Pell began to collaborate on the idea about two years ago.

Current methods of searching used by more traditional engines focus on isolated keywords and broad but shallow content coverage. This leaves a lot of room for improvement, Kaplan says.

"They are really not getting at relationships," he notes. "The best that they do to approximate relationships are words that are close to other words." He adds that a much deeper level of analysis is required.

Previous attempts have tried to pair some natural-language query processing with standard keyword searches of relevant content. This approach can be seen with some parts of standard search engines like Google, which, if it doesn't understand a user's query, will suggest another phrase or word that it thinks he or she may have meant. Engines such as Google and Yahoo use some components of natural-language search, yet there has not yet been a

full-scale natural-language search engine for consumers. (See "The Future of Search (http://www.technologyreview.com/Biztech/19050/) .") Pell says that this was mainly because the necessary technology was simply not ready. Engines that use natural-language components for aspects of the search, such as iPhrase and EasyAsk, don't process textual content as Powerset does, Pell says, but instead simply query databases for answers to questions. Attempts at full natural-language search, such as that offered by Hakia and Cognition Search, do not cover as rich a representation of concepts or meaning, Pell says.

The company plans to release demo versions of the search engine on its Powerlabs website (http://labs.powerset.com/) , where consumers can test-drive the product beginning in September. User feedback will be taken into consideration as Powerset makes the final product, which is slated for release next year.

"The key challenge is to get the system to the point where people can understand how to use it and get real value out of these systems even though they are not perfect," Pell says. "We are finally at the point where we are going to cross that threshold."

IBM is also in the midst of developing a semantic search engine, code-named Avatar (http://www.almaden.ibm.com/cs/projects/avatar/) , which is targeted at enterprise and corporate customers; it's currently in beta testing within IBM. Project manager Shivakumar Vaithyanathan says that the hardest problems to overcome with natural-language search are finding a way to extract higher-level semantics from large documents while at the same time preserving precision and speed.

IBM's engine is targeted toward searches of internal documents such as e-mail and intranet correspondence. It's designed to be used in cases in which the user seeks to find one particular piece of information that could not be easily located, such as a specific phone number or package-tracking URL that's located in one of thousands of e-mails that a person may have stored on her computer.

Avatar's semantic search seeks to develop "interpretations" of keyword queries that model the real intent behind the query. For example, if the query was "phone number," the search engine would search the thousands of e-mails that a person may receive for the numbers that resemble a phone number. The search engine would provide the user with the useful information he seeks, and not just a keyword entry in an e-mail that contains the words mentioned in the query.

In order to quickly extract all the meaningful information from both the underlying text and the query, Vaithyanathan says, it's necessary to utilize either a lot of computers or a large number of people. Both options are expensive and can be difficult to implement. IBM hopes to find a way to extract meaning in less time and with fewer machines.

"If we do a better job of extracting, then we can do a better job of answering the questions that users give," Vaithyanathan says.