Powerset's Natural Language Wikipedia Search Engine

Tim Converse, Ronald M. Kaplan, Barney Pell, Scott Prevost, Lorenzo Thione, Chad Walters

Powerset, Inc.
475 Brannan Street
San Francisco, California 94107
{converse, kaplan, barney, prevost, thione, chad}@powerset.com

Abstract

This demonstration shows the capabilities and features of Powerset's natural language search engine as applied to the English Wikipedia.

Powerset has assembled scalable document retrieval technology to construct a semantic index of the World Wide Web. In order to develop and test our technology, we have released a search product (at http://www.powerset.com) that incorporates all the information from the English Wikipedia. The product also integrates community-edited content from Metaweb's Freebase database of structured information. Users may query the index using keywords, natural language questions or phrases. Retrieval latency is comparable to standard key-word based consumer search engines.

Powerset semantic indexing is based on the XLE, Natural Language Processing technology licensed from the Palo Alto Research Center (PARC). During both indexing and querying, we apply our deep natural language analysis methods to extract semantic "facts" -- relations and semantic connections between words and concepts -- from all the sentences in Wikipedia. At query time, advanced search-engineering technology makes these facts available for retrieval by matching them against facts or partial facts extracted from the query.

In this demonstration, we show how retrieved information is presented as conventional search results with links to relevant Wikipedia pages. We also demonstrate how the distilled semantic relations are organized in a browsing format that shows relevant subject/relation/object triples related to the user's query. This makes it easy both to find other relevant pages and to use our Search-Within-The-Page feature to localize additional semantic searches to the text of the selected target page. Together these features summarize the facts on a page and allow navigation directly to information of interest to individual users.

Looking ahead beyond continuous improvements to core search and scaling to much larger collections of content, Powerset's automatic extraction of semantic facts can be used to create and extend knowledge resources including lexicons, ontologies, and entity profiles. Our system is already deployed as a consumer-search web service, but we also plan to develop an API that will enable programmatic access to our structured representation of text.