Integrating Multiple Knowledge Bases within Google Desktop

John Houser and Peter Galvin

Center for Collaboration and Cognition, Colorado State University

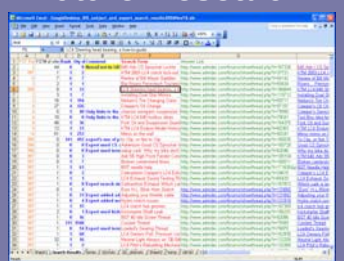
Introduction

Google is clearly the preferred solution when searching for information. However, how does one search for information within proprietary knowledge bases? We believe that we can use the power of Google Desktop (GD) as a repository and search mechanism for local proprietary knowledge by utilizing GD's extended indexing capabilities.

Method

For this research, we decided that an online forum would be the closest real-world example of a large unorganized body of knowledge. Given the nature of GD, we built intermediate tools which allow for easy access into the GD indexing system, including software to "Scrape" and "Parse" the html (downloaded from online forums) and a "Server" that communicates with the GD API.

Future Research



- Continue researching search methods with advanced query options that deliver more valuable solutions to a wider group of unorganized datasets.
- Adapt our tools to take advantage of future GD improvements while continually seeking out better UI techniques with the current version.
- Update the network interface of GDS to use predefined XML formats for a more standard communication method.

Original Post

The 4.5GB we downloaded included 600,000 posts from six forums on Adventure Rider (a global public threaded discussion site dedicated to 41,987 Motorcycle enthusiasts who have freely shared their experiences, knowledge, and adventures).

This screenshot of www.advrider.com includes a link to one of 154,372 threaded discussions which contain a total of 3,415,052 posts.

Parse

A program that searches (using regular expressions) for the information contained within the thread (html) pages and then opens a socket to GDS which receives the parsed information.

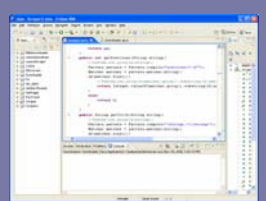
Google Desktop

GD provides a local search engine for files located on your computer. It also has the ability to acquire additional "events" that can vastly expand it's searching capability. For example, we utilized the "email event schema" to add the 600k posts into GD's local index.

The example above shows the results of a search for a KTM motorcycle with the 640cc engine. We were able to limit our search to only the motorcycle data partition by using special keywords embedded in each "event".

Scrape

A simple program that loops through the discussion threads of each forum, downloads the (multiple) html pages associated with each thread, and saves each one as a text file.



GDS

The Google Desktop Server (GDS) acts as a bridge between GD and the information to be indexed. It provides an easy (WIN32) interface that encapsulates the ATL/COM API used to communicate with GD. It allows for up to 10 sockets to connect at any given time with a pop-up window that displays the socket history.

Results

When searching for information which we know exists in the indexed knowledge set; the public threaded discussion Forum Search Tool returned a link to the target information 14% of the time, whereas GD showed a significant improvement by returning a link to the target information 50% of the time.