Advanced Search Tools for Online Resources

By: Justin Chase
Faculty Sponsor: Dr. Ted Pedersen

Statement of Problem

The main objective of this project is to develop a tool or set of tools that would allow a user to easily construct advanced search queries for online resources. The primary online resource in mind for this project is the Google[1] search engine, accessed programmatically through its freely available API[2].

Often times when forming a query the search engine will not produce satisfactory results. With difficult to find data, advanced search options become more and more useful. These options, however, can be cryptic and syntactically unwieldy. Additionally, finding the correct results often times requires using the proper keywords, which is not always easy to discover if you do not know them exactly. One aspect of this project will focus on generating keywords to help assist a user with future searches and the other main focus will be to provide a clean and efficient interface for performing advanced searches.

The increased reliance on search engines for online resources increases the need for a tool such as this to help optimize the searching capabilities of users. An efficient utility primarily focused on searching would be enormously useful. The main purpose of this project is to discover and provide techniques to increase the effectiveness of searching for online resources.

Method of Inquiry

Technically speaking, the most interesting aspect of this project will be the generation of keywords for more effective searches. Based on work done in Ted Pedersen's class [CS 5761: Introduction to Natural Language Processing[3]] keywords will be generated by searching through web pages resulting from the previous search and the words found will be tallied and weighted based on the rank of the page they are found in. After filtering the list of resulting words what remains is a list of keywords that are found frequently on these pages. These words tend to be the typical "buzzwords" of the common subject of the resulting pages and may make excellent candidates for further searches.

One other main technical concern is the creation of a plug-in style of interface. New libraries will be able to plug into the framework of the project easily to provide connectivity to new resources. This will help to increase the ideal of easily searching through online resources.

In order to accomplish these goals I plan on making use of Microsoft's .NET[4] technology and coding this application in C#[5] [C-Sharp]. C# .NET has proven to be very

effective in making quality Graphical User Interfaces and though this will primarily be used on Microsoft Windows Operating System's, C# has the potential to be cross platform. I plan to put some effort forward to keep as much functionality as possible into separate libraries to enhance reusability and the ability to "plug-in" new libraries. Additionally I plan on using a distribution known as Mono[6] to use these libraries (not graphical interface components, however) to create these search tools for Linux operating systems as well.

Part of this project will also be to attempt to discover new methods of enhancing searches and providing useful feedback to users in order to refine a search.

Expected Results

The intention of this project is to yield a tool useful in accessing various online resources as well as provide some new ideas about how to refine advanced searches and give useful suggestions to users. I also intend to refine the keyword generating algorithm by reviewing the weighing system and enhancing the filtration methods.

Time Table

This project will be carried out during the spring of 2005 and I plan on working for 12hrs. / week for 10 weeks, through January, February and March.

- Week 1-2 (24 hrs.):
  - Draw up requirements documents and UML diagrams.
  - Create initial framework and plug-in framework.
- Week 3-4 (24 hrs.):
  - Create Google API engine and interface.
- Week 5-6 (24 hrs.):
  - Implement keyword generator.
  - Create other engines and interfaces for other API's if available.
- Week 7-8 (24 hrs.):
  - Port search libraries to a Linux system with Mono installed and create an interface for it.
- Week 9 (12 hrs.):
  - Finish up project and tie up loose ends.
  - Make sure all code is commented and presentable.
- Week 10 (12 hrs.):
  - Prepare final report.

Budget

We are requesting a budget of $1620. This includes a $1380 stipend for working the hours during the spring of 2005 as described above as well as $240 for the following books[7]:

- Cross-Platform .NET Development: Using Mono, Portable.NET, and Microsoft - .NET - M.J. Easton; $50
- Open Source .NET Development : Programming with NAnt, NUnit, NDoc, and More - Brian Nantz; $30
- Microsoft Visual C# .NET Deluxe Learning Edition-Version 2003 - Microsoft Corporation; $80
- Google and Other Search Engines : Visual Quick Start Guide (Visual Quick start Guides) - Alfred Glossbrenner; $15
- How to Do Everything with Google - Fritz Schneider; $17
- Mining Google Web Services: Building Applications with the Google API - John Paul Mueller; $20
- Google Hacks: 100 Industrial-Strength Tips & Tools - Tara Calishain; $18
- Shipping: $10

Relationship to Faculty Sponsors Research

Dr. Ted Pedersen does research in Natural Language Processing and teaches several classes on this topic. The idea of searching through text and generating a set of keywords was discussed in one of his classes and turns out to be an interesting Natural Language Processing problem. Additionally search engines in general provide many different interesting Natural Language Processing experiences and this project intends to work with them heavily. Dr. Pedersen's research and classes correspond closely to the main subject of this project in these categories.

Educational Objectives

As bodies of data grow larger and larger it becomes increasingly valuable to have efficient methods of searching for specific results. The effectiveness of different search techniques varies notoriously. The main objective of this research is to explore some new techniques of offering effective search options and learn how to best represent this intuitively. Searching effectively through text is a vast and important field and the educational and practical benefits of such research seem to be very valuable.

---

[1] http://www.google.com
[2] http://www.google.com/apis/
[3] http://www.d.umn.edu/~tpederse/teaching.html
[4] http://www.microsoft.com/net/
[5] http://msdn.microsoft.com/vcsharp/
[6] http://www.mono-project.com/about/index.html
[7] Prices are taken from amazon.com on 10/07/04