Christopher Hoover (UW/CSE: furby16)
Tatsuro Oya (UW: oya0306, CSE: toya)
CSE 403 Spring 2011
Assignment 1

Introduction and Motivation

Many of the UW's faculty members give undergraduate students the opportunity to participate in their research projects. Since this exposes students to active research projects, gives them college credit, and allows them to develop relationships with faculty that can be important for future research or graduate school, research projects can be vital components of undergraduate study. However, although it is not uncommon for students to participate in such projects, there are very few resources available to help students find research opportunities in the first place. The UW's HuskyJobs service allows students to easily search for internships and full-time jobs, but there is no such service for undergraduate research. Students must either rely on their existing faculty contacts or search faculty home pages to find research opportunities.

In light of this, we propose the development of a search engine website that searches the UW website for potential research areas in response to student queries. With this tool at their disposal, students in any department could locate possible research opportunities quickly and easily. Instead of manually examining faculty home pages, students would be able to search for research positions just like they would search for internships and jobs on HuskyJobs.

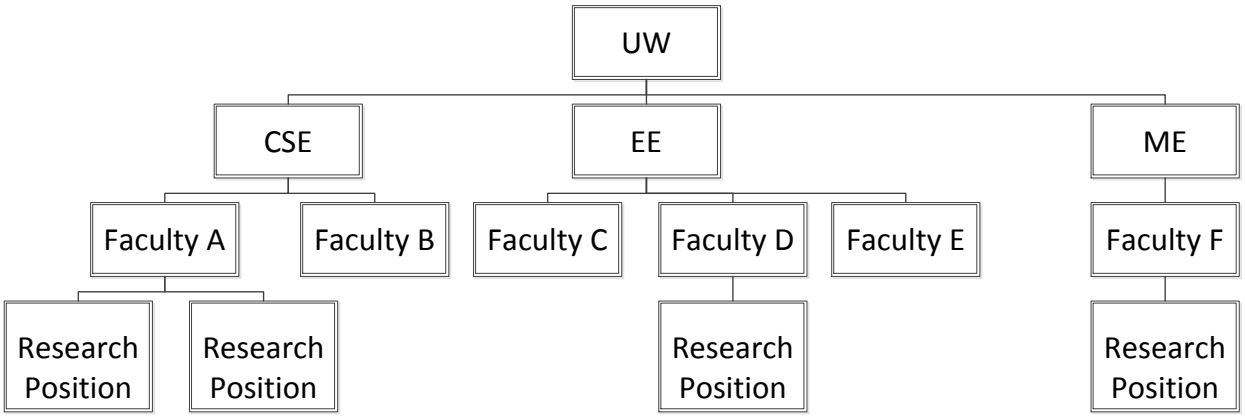
Because this website would be designed to search specifically for undergraduate research opportunities, it would offer superior ease of use and more relevant results than would be possible using the search function already built into the UW website. Adding research position postings to HuskyJobs is another possible alternative to this system, but we believe this would be less desirable as well. HuskyJobs is designed to be a resource for students seeking internships and jobs outside the UW, so adding research positions to its listings would be contrary to its established focus. Research positions are also fundamentally different from internships and full-time jobs, since research often offers course credit rather than financial compensation. Furthermore, with our proposed system, UW faculty would not be required to explicitly manage undergraduate research positions, as they would if such positions were added to HuskyJobs.

Description and Architecture

The website would be composed of three main components: a web interface that allows students to enter search parameters and view results; a search engine capable of navigating the UW website, finding faculty home pages in appropriate departments, and extracting relevant data from them; and a database for storing search results. The search engine itself would be similar to a simple web crawler with a very restricted scope.

The system's operation would be roughly as follows. At regular intervals, the search engine would navigate to each department's website and locate each department's faculty home pages, then use natural language processing to identify faculty that are looking for undergraduate

researchers. The following diagram shows an example of how the search engine would navigate the UW website, starting from the top and moving downward:



The names of all faculty members that appear to offer research positions, along with links to their home pages, their research areas, and their email addresses, would then be stored in the database. When a user submits a query, the system would return results from its database.

One interesting technical aspect of this website is the combination of web, text processing, and database technologies it uses. This project would thus require expertise from all three of these areas, and also require successful integration of these technologies as well.

Challenges

The most significant challenge this project would face is the data extraction from faculty web pages. This task would require sophisticated natural language processing, because the information is not stored in a standardized format. Some faculty might explicitly indicate what research positions in their projects are currently available for undergraduates, while others may only indicate that undergraduate research positions are available without giving further details. For either case, the wording of the information itself, along with its context on the page, is highly variable. Thus, the search engine would need to parse the text on faculty home pages and attempt to extract relevant information from highly irregular data.

Because the natural language processing component is the most technically challenging aspect of this project, additional time would need to be devoted to it during development. This would help to ensure that any difficulties encountered in its design or implementation would be resolved early, before time becomes critical.