

Project Proposal for CS249 Course Project

Dean Lee, and Richard Sia

October 19, 2003

1 Objective

To implement a system that discovers search interfaces and classifies the search interfaces automatically.

2 Motivation

From reading of [1], it seems like the results were obtained and verified manually. Thus it would be helpful to automate the process of search interface discovery and classification. Also, it provide a resource for hidden web researchers when they want to crawl and do experiments on hidden web.

3 Approach

The project can be divided into two phases. Phase 1 involves the development of a crawler that identifies search interfaces based on HTML tags. Specifically we would look for action tags that will redirect the current page to a cgi script and form tags [2]. The crawler will be initially trained by crawling through the set of 560 sites provided by [1]; then the crawler will discover new search interfaces. Phase 2 involves the classification of search interfaces. Specifically, the interfaces will be classified into several predetermined domains(e.g. Cars, Hotels, Airlines, etc). The implementation will be the integration of existing classifier tools(e.g. Weka). The particular classifications might be K-means, Bayesian classifiers, C4.5, etc. Using the system, we want to carryout experiments like How many hidden web sites exist on the web? How are the topics distributed?

4 Schedule

While we do not have any specific dates set, we plan to spend 3-4 weeks on the crawler implementation, then 1-2 weeks for the integration of classification tools; if time permits, we will expand/improve our system.

References

- [1] Kevin Chen-Chuan Chang, Bin He, Chengkai Li, and Zhen Zhang. Structured Databases on the Web: Observations and Implications. Technical Report UIUCDCS-R-2003-2321, UILU-ENG-2003-1708, University of Illinois, Urbana Champaign, February 2003.
- [2] Jared Cope, Nick Craswell, and David Hawking. Automated Discovery of Search Interface on the Web. In *Proc. of 12th Australasian Database Conference*, Adelaide, Australia, 2003.