Chapter 1 Introduction

1.1 Research Motivation

With the explosive growth of the World Wide Web, finding useful web pages/sites by using web search engines has become a part of our everyday lives. Web search service is a vital way to find information on the web. However, not every piece of information found is relevant or useful. In fact, the relevance and precision of the search results is much more important than the time a search engine spends to find the information. In order to improve search accuracy, most designers of the web search engines devote to working on search algorithms development and optimization.

Another way to manipulate web search well is to understand the algorithms of search engines especially the ranking algorithm so that users are likely to obtain precise data. Most traditional algorithms are keyword-based and, given a user query, use word frequencies, word importance, document length and other statistical cues to assign potential importance to a document. However, with the emergence of the web many new algorithms for web search have been proposed and are being used in various web search engines today. Many of these algorithms incorporate link-structure of pages in their ranking schemes, and are notably different from the traditional keyword-based document ranking algorithms.

From literature, we realize that there are few open or flexible performance evaluation methods for web search service. And performance measurement and comparison is critical in the advance of web search technology. Hence, a more flexible benchmark approach is needed to benchmark web search service.

1.2 Research Problem

The web search engine service developed in the 1990s. With the popularity of Internet, web search has become an important information technology topic in the information retrieval domain. In order to measure and evaluate the performance and accuracy of web search, a benchmark approach to these new techniques is needed.

However, it is very difficult, if not impossible, to directly apply these measurements to the evaluation of web search engines due to the unique nature of the web such as precision and recall. It is an uncertain issue that how to get precision and recall for evaluating web search. In addition, how to help users find information on web more precisely is also one of our main concerns.

Therefore, we developed a workload model and built a workload generator that is generic construct based. However, we encountered the problem of developing and building, and how to put

generic constructs and measurements to a workload models and building the workload generator?

1.3 Research Objective

There are two pars of the benchmark. One is to develop and build a workload model and design test suites. The other is comparing the precision between the products. In this research, we focus on the development of the workload model. Developing a benchmark requires definition of a workload model. Workload is the core of a benchmark. In this research, we develop a workload model that incorporates web search on internet environment. It tests the performance and the efficiency of web search. In order to capture the semantic aspect of web search, we analyze the key web search algorithms to get the generic constructs of the algorithms. Generic constructs are that

The workload model is designed to be generic-construct-based. The generic model describes the page structure and query structure, which is not tied to a predetermined scenario. The workload model is developed to meet desirable characteristics, or a good benchmark. First, the workload model can be scaled with the complexity of pages, queries and search process. Secondly, the workload model adopts open standards, such as the generic constructs of web search algorithms. Third, the workload model is simple to understand and implement because that generation process is automated.

1.4 Research Limitation

There are several limitations in this research due to time and resource constraints, which are described in the following section.

1. Due to time constraints and infinity of the internet, we can not precisely verify all performance evaluation indicators through our prototype system. Thus, the validity of this research is limited, which can be further improved.
2. In the prototype, the query generator developed thus far is still primitive. It assumes the most functions are supported by the web search service APIs.
3. The experiments just include simple and synthetic tests. Thus, the complexity of experiments in this research is limited, which can be further improved.

1.5 Research Flow

In this research, we studied web search algorithms of web search. Benchmarks on web search are reviewed simultaneously. After that, we systematically analyzed the web-search-specific in order to stimulate a new design. First, we identified research problem, then defined research scope. After literature study, we continued to develop the research model. Then, we implemented a prototype

on the basis of this research workload model to verify the feasibility. Also, an experiment was build to verify the prototype system. Finally, we could get the conclusion.

The research flow is shown in Figure 1.1.

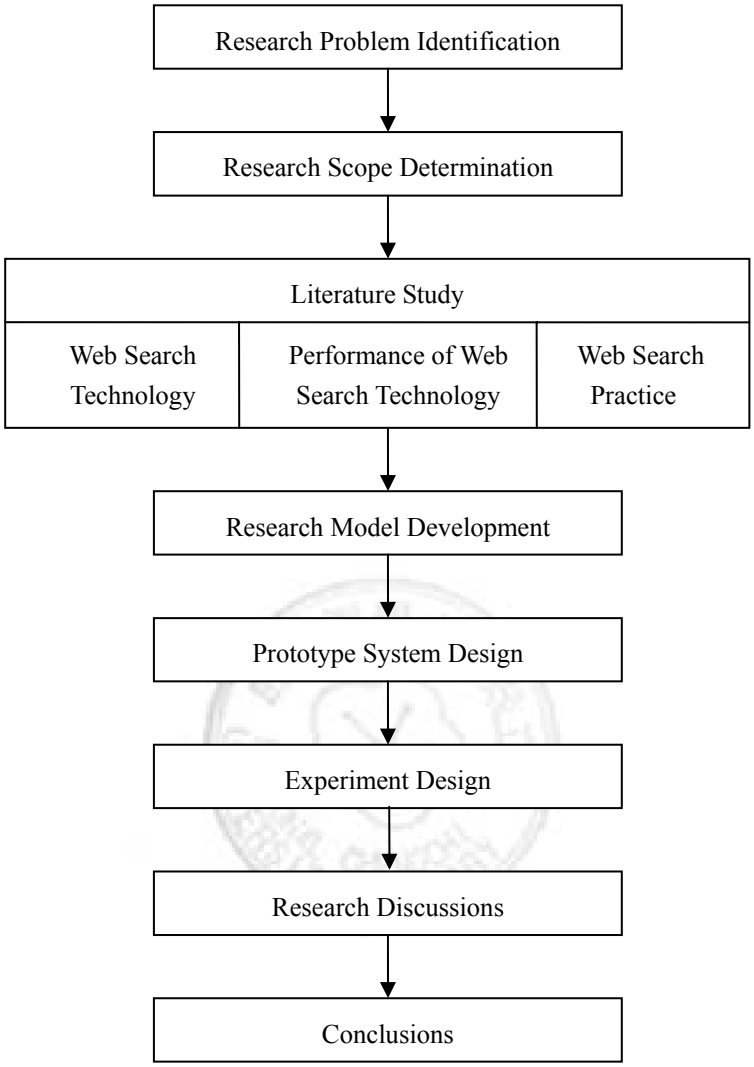

Figure 1.1: Research Flow

1.5 Organization of Thesis

The remainder of the thesis is organized as follows:

(1) In chapter one, it is the introduction of this research that tells why we want to do and what we are going to do about this research.

(2) In chapter two, we will review and discuss related algorithms and benchmarking of web search.

(3) Chapter three presents the model of this research on benchmarking web search.

(4) In chapter four, we will implement the research model of this research on the benchmark

experiment.

(5) In chapter five, we will verify our research model with a specific web search engine.

(6) Chapter six concludes managerial findings, technical findings, summary and Future directions of this research.