Chapter 3 Research Model

This chapter illustrates how the structure and model of this research are developed using the literature discussed in chapter two. The research structure is described in the next section. In this thesis, we adopt the prototype method to illustrate the feasibility and validity of the research model.

3.1 Research Structure

The benchmark workload model is run on the web search services to evaluate performance and precision. The literature studied in chapter two helps to identify important generic constructs of the web search algorithms. In chapter three, web-search-specific requirements will be analyzed in more detail to justify the design of the benchmark. This thesis develops a generic benchmark model that is portable and scaleable.

The research model is shown in Figure 3.1. The benchmark consists of a benchmark workload model. The web search benchmark model consists of three models: a page model, a query model and a control model. The page model and query model depend on the generic constructs, operations of generic constructs and constraints requirements. In addition, the control model is created before the generic workload model is generated and executed so as to measure and evaluate web search.

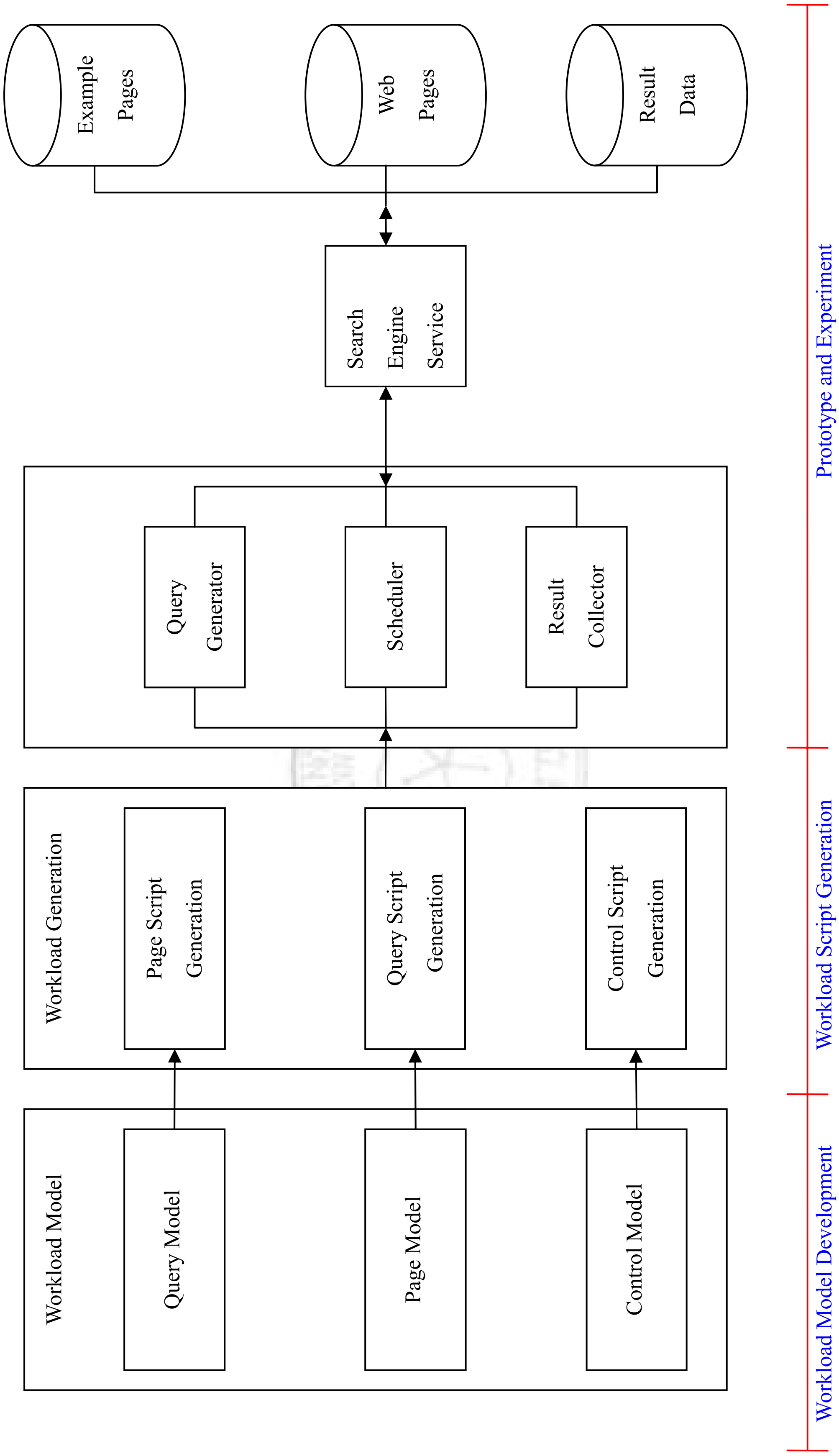

Figure 3.1: Research Structure

3.2 Components of Research Model

In this research, we focus on generic constructs of the web search algorithm. The benchmark model proposed would capture generic constructs and operations of generic constructs of the algorithms studied in chapter two. Designed as a generic construct based benchmark workload model, it is possible to implement the benchmark on different web search algorithm.

Developing a benchmark requires defining the test workload model first. In this research, we provide a benchmark workload model that evaluates the performance and efficiency of web search according to user requirements. The components of the workload model are based on the generic constructs and operations of generic constructs summarized from the algorithms in the literatures study. The workload model consists of three components: the page model, the query model and the control model. The page model describes a generic page layout structure and the query model defines some important criteria to query the search engines. Table 3.1 illustrates that mapping of generic constructs and workload model. Table 3.2 illustrates that mapping of operations of Generic constructs and workload model The control model defines the variables that used to set up the benchmark environment. Three components define the experimental factors of the benchmark. In addition, the performance metrics must be defined to measure the benchmark results.

Table 3.1: The Mapping of Generic Constructs and Workload Model

Generic Constructs	Workload Model
Term	● Query Model: e.g. Input Keyword
Web page	● Page Model ● Query Model: e.g. PageRank, Authority-Hub
Phrase	● Query Model-TLS

Table 3.2: The Mapping of Operations of Generic Constructs and Workload Model

Operations of Generic Constructs	Workload Model
Count	● Query Model: e.g. VSM, Okapi, TLS
Position	● Query Model: e.g. CDR
Tag	● Page Model ● Query Model: e.g. PageRank, Authority-Hub

3.3 Page Model

Fundamentally, pages are the source of web searching, composed of html format. According the features of an html file, the search algorithms are constructing to improve the effectiveness of search results. The literature review discussed in the second chapter indicates the link structure of pages is an essential factor of web search. Therefore, we divided the page model into two categories—single page structure and multi-page structure (See Figure 3.2).

The page model is described below:

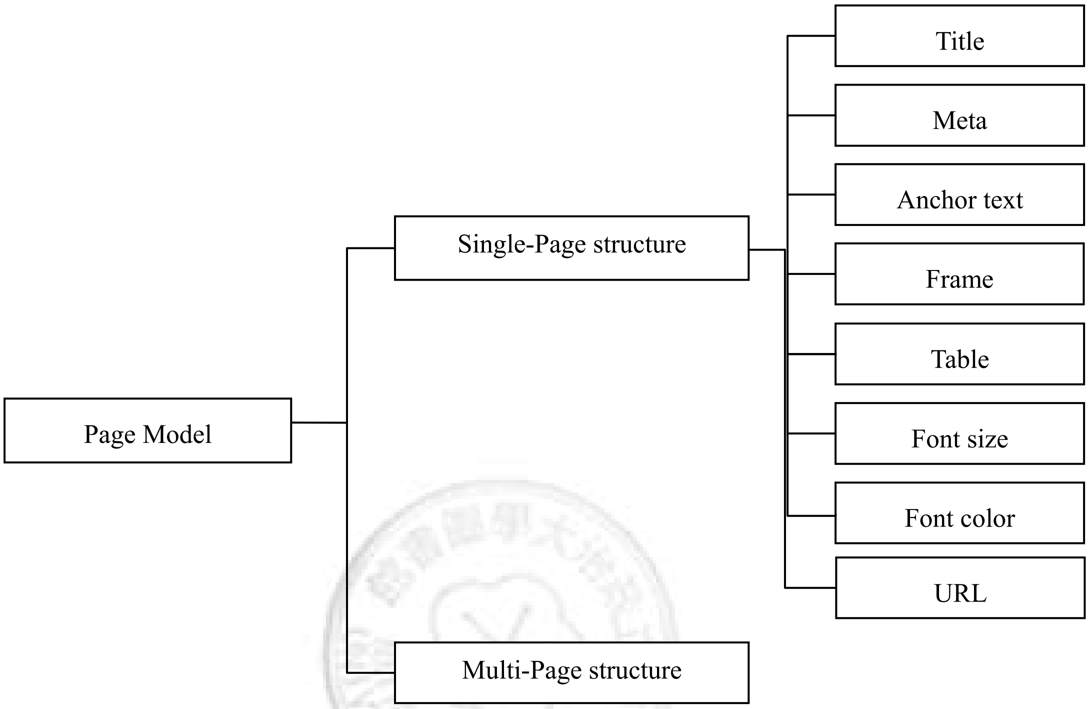

Figure 3.2: The Page Model Hierarchy

3.3.1 Single page structure

Web pages are made up of html tags which a browser can parse that terms can be located in any part of a web page. Therefore, according to the location of the specific tag where query terms appear, we define the single page structure composed of several parts:

(1) Title: <title></title>tag

(2) Meta: <meta keyword="…">, some important meta data about each page, such as size, date and layout structure

(3) URL: words that occur in hyperlinks.

(4) Anchor text – all hyperlinks and their corresponding anchor texts.

(5) Table:<table></table> tag

(6) Frame:<frame></frame> tag

(7) Font size:

●H1～H6 : words with H1~H6 tags

- STRONG: words with font type “bold”, “italic”, “underlined”, etc.
- Large: words with large font size
- Medium: words with medium font size
- Small: words with small font size.

(8) Font color

3.3.2 Multi-page structure

In addition to the term hits in the page layout, the link structure analysis helps us find authoritative pages that offer useful resources among an enormous quantity of web sources. Therefore, finding authoritative pages, such as some authoritative portals, like Yahoo and Msn, etc, will help web search more precisely and is more likely to obtain useful information. The multi-page structure can have many categories such as Portal, Education and government, etc. Tables 3.1 are examples of three categories.

Table 3.3: Examples of three categories

Categories	**Website**
Portal	Yahoo!
	Msn
	Baidu
Education	National Taiwan University
	National Chengchi University
	UCLA
	UC Berkeley
Government	SEC
	TSEC

3.4 Query Model

In this research, we attempt to propose a generic query model for web search applicable to any scenario. Queries are not described based on a specific search engine, but are specified to user requirements. This enables the user to apply them in different scenarios on different search engines. This would help users to evaluate performance of web search with increasing complex queries.

Following the algorithms analysis of web search, a comprehensive set of query specifications are identified. The query model defined is composed of four categories, including query type, link structure, similarity and synonym. Each of them poses different to web search. Users can specify queries according to their requirements, called “user-driven query”. The following section will describe each category briefly. The query model hierarchy is shown in. figure 3.3

.

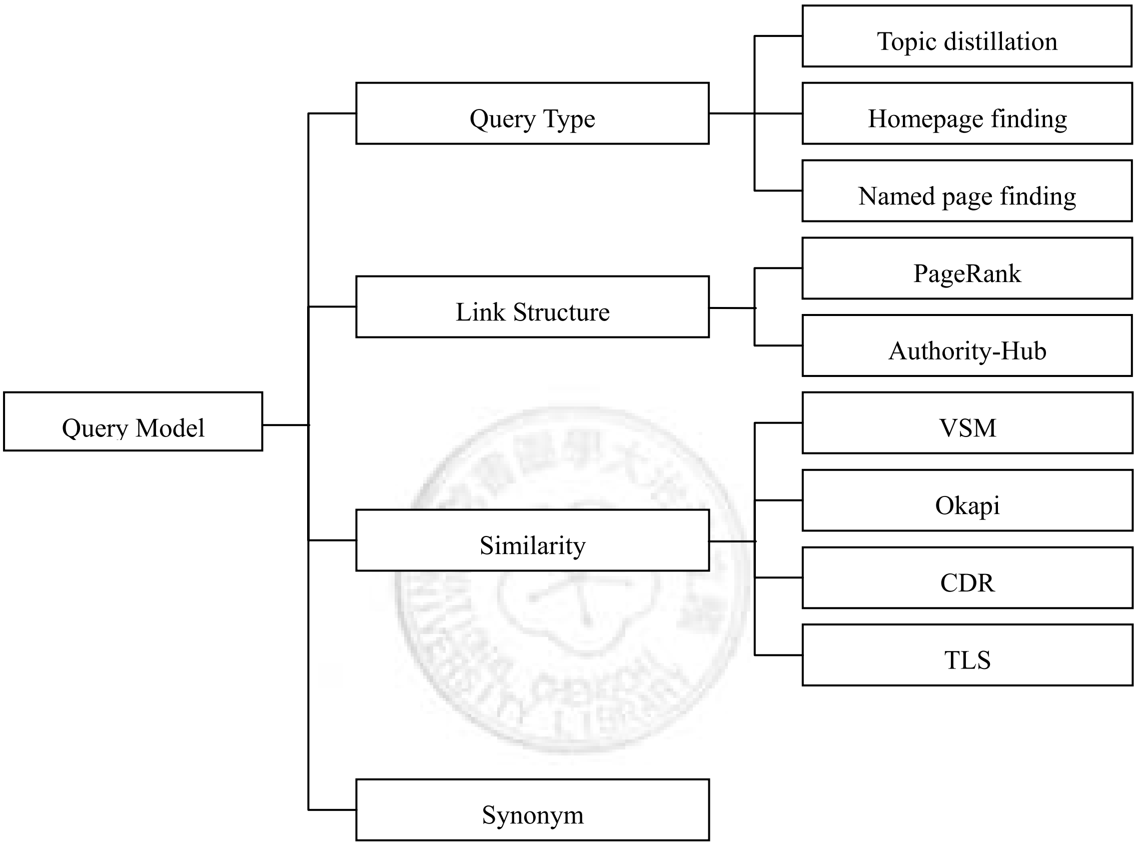

Figure 3.3: The Query Model Hierarchy

3.4.1 Query Type

Searching is viewed as beginning from a user-supplied query (Kleinberg 1998). Thus, we identify the web search as a query, but there is more than one type of query when people conduct searches. An example is finding a web site or related information about one topic. Referring to the overview of the TREC-2004 web track, there are three query types.

(1) Topic distillation (TD)

The query describes a general topic, e.g. ‘electoral college’; the system should return homepages of relevant sites.

(2) Homepage finding(HP)

The query is the name of a site that the user wishes to reach, e.g. ‘Togo embassy’, and the system should return the URL of that site’s homepage at (or near) rank one.

(3) Named page finding(NP)

The query is the name of a non-homepage that the user wishes to reach, e.g. ‘Ireland consular information sheet’, and the system should return the URL of that page at (or near) rank one.

3.4.2 Link structure

Taking advantage of the links between pages helps improve effectiveness of search results, including in-links and out-links. Therefore, in our query model we can specify three kinds of query parameters which represent search techniques based on link structures analysis. This following brief description defines the three parameters.

(1) PageRank

The sequence of search results is showing by computed rank value of search engines.

(2) Authority

A good authority is a page that is pointed to by many good hubs

(3) Hub

A good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

3.4.3 Similarity

In order to get more precise search result, we also include similarity analysis based on four relevance scoring methods - VSM, Okapi, CDR TLS (Longzhuang Li, Yi Shang, and Wei Zhang, 2002). The following will describe each method applied in this research..

(1) Vector Space Model (VSM): Scoring is based on the frequency of each term of the query.

(2) Okapi Measurement Method (Okapi): Scoring is based on the frequency of each term of the query and the length of the document.

(3) Cover density ranking (CDR): Scoring is based on the location of an ordered pair term of the query over a document.

(4) Three-Level Scoring Method (TLS): Scoring is based on the frequency of sub-phrase of the query.

3.4.4 Synonym

We hope that search result can include more accurate pages by means of synonyms. First, define synonymous terms of query terms. Then, add these synonymous terms to perform a search query.

3.5 Control Model

After specifying the page model and query model, we need a control model to do executions on the benchmark. The control model defines environment variables to execute the benchmark. These variables are used to set up the execution environment.

- **Steady State**
  The benchmark test must be executed in a steady state, in order to return true performance of the.
- **Test Mode**
  There are three kinds of test mode, cold mode, warm mode, and hot mode. In cold mode, there is no data in the cache. The system cannot retrieve data from the cache directly. Therefore, the performance in cold mode is usually slower than other two modes. In warm mode, the data is left in the cache from prior query. Because of that, the test response time decreases. In hot mode, a query is executed in cold mode first, and then be executed with cache data for several times. The average response time is computed.
- **Test Duration**
  Test duration means time intervals of the benchmark. Each interval must begin after the system has reached steady state and be long enough to generate reproducible throughput results. Each interval must extend uninterrupted for a period of time.
- **Test Sequence**
  Test sequence indicates the order of the queries execute.
- **Number of Repetitions**
  Number of repetitions means execution repeated times of an operation in a test.

3.6 Performance Metrics

Performance metrics are used to measure the execution result. Response time and throughput are two performance metrics often used in evaluation of computer systems.

- **Response time** means time interval between when a request is made and when the response is received by the requester.
- **Throughput** means the number of operations completed by the system per unit time.

Recall and precision are two important measures of evaluation of information retrieval.

However, it is very difficult, if not impossible, to directly apply these measurements to the evaluation of Web information retrieval systems due to the unique nature of the Web.

There is no proper method of calculating absolute recall of search engines as it is impossible to know the total number of relevant in huge databases. This Research followed the method used by Clark and Willett by pooling the relevant results of individual searches to form the denominator of the calculations. The relative recall value is thus defined as (Clarke & Willet, 1997):

- **Relative Recall**

$$\frac{\textbf{Total no. of relevant documents retrieved by a search engine}}{\textbf{Sum of relevant documents retrieved by all three search engines}}$$

The calculation of the Precision requires knowledge of the relevant and non-relevant hits in the evaluated set of documents (Clarke & Willet, 1997). Thus it is possible to calculate absolute precision of search engines which provide an indication of the relevance of the system. In the context of the present study precision is defined as:

- **Precision**

$$\frac{\textbf{Sum of the scores of documents retrieved by a search engine}}{\textbf{Total no. of results evaluated}}$$

- **Relevance criteria**

Referring to the literature (Clarke & Willet, 1997), they adopted the following relevance criteria:

(1) If the page closely matched the subject matter of the query then it given a score of 1.

(2) If a page considered of a whole series of links, rather than the information required, then it was given a score of 0.5 if inspection of one or two of the links proved them to be useful.

(3) Duplicate sites (same URL, same contents) were given a score of 0.

(4) Mirror sites different URLs, same contents) were not considered as duplicates and were scored as if they were unique.

(5) If the system reported a “file not found” or similar message for a particular URL, this suggested that the index had not been updated, and the site was given a 0 score.

(6) If a message appeared which said “There was no response”. The server could be down or is not responding for a particular URL then the page was checked later. If this message repeatedly occurred the page was given a score of 0.

(7) Foreign language pages were often difficult to assess for relevance and were hence replaced by the addition of the next-ranked English page.