DEVELOPING AN ENHANCED SEARCH STRATEGY FOR A LOCAL SEARCH ENGINE

BY

Santos, Cyril Ray

Lo, Maria Cristina

Jiao, Pia Abigail Mitzi

SCHOOL OF ARTS AND SCIENCES
ATENEO DE DAVAO UNIVERSITY

MARCH 2002

TABLE OF CONTENTS

CHAPTER

I. INTRODUCTION 2

1.1 Background of the Study 2
1.2 Statement of the Problem 3
1.3 Objectives 4
1.4 Scope and Limitations of the Study 5
1.5 Significance of the Study 6

II. REVIEW OF RELATED LITERATURE 7

III. METHODOLOGY 24

IV. THEORETICAL BACKGROUND 27

4.1 Existing Search Strategies 27

4.1.1 Link Popularity 27-28
4.1.2 Keyword Frequency 29
4.1.3 Citation Graph Literature 29

4.2 Hilltop Approach 30

4.3 Indexing 30

4.4 Stopwords 31

4.5 Stemming 32

4.6 Synonyms 33

4.7 Search Strategies 33

4.7.1 Directory Search Tool 33
4.7.2 Search Engine 34
4.7.3 Directory with Search Engine 35
4.7.4 Multi-engine Search 35

4.8 Page Ranking 36

4.9 Inverted Index 36

4.9.1 Text Retrieval Using Inverted Indexing 37

5.0 Features and Components of PHP and ASP 38

5.0.1 PHP 4.0 38-40
5.0.2 ASP 40-41

V. RESULTS AND DISCUSSIONS 42

5.1 Comparison of Search Strategies 42

5.1.1 Keyword Frequency 42-43
5.1.2 Link Popularity 44
5.1.3 Citation Graph Literature 45

5.2 Level 2 Citation Graph Literature 46-47

5.3 Framework of the Local Search Engine with the Unenhanced Search Strategy 48-49

5.4 Framework of the Local Search Engine with the Enhanced Search Strategy 50-51

5.5 Criteria for Selecting a Scripting Language 52-54

5.6 Table of Comparison of Scripting Languages 55

5.7 Criteria for Comparing Built-in Features of Scripting Languages 56

5.8 Comparison of Features and Components 57

5.9 Database Structure 58

5.9.1 Inverted Index Table 58
5.9.2 All Pages Table 58
5.9.3 Temp Table 59
5.9.4 Inverted Index Search 59

6.0 Testing Results 60-65

VI. CONCLUSION AND RECOMMENDATION 66

APPENDIX A Codes of the Local Search Engine 67-86

APPENDIX B Screenshots of the Local Search Engine 87-92

APPENDIX C Search Engine Installation Procedure 93

BIBLIOGRAPHY 94 -95

ABSTRACT

With the current scale and growth of the World Wide Web, the importance of being able to search for and locate Web pages effectively and accurately is of prime importance. Currently, the only feasible way a searcher can locate a particular Web-based source is by using a Web search engine. The generic large-scale search engines return results in the thousands, many of which lack relevance to the query; but searchers only tend to look at the first few results, hence an accurate rank is critically important. Additionally, large-scale search engines are unable to scale effectively with the expansion of the Web, especially the local scenario.

Hence, is the need to develop an enhanced search strategy that will improve the accuracy and relevance of search results. With this, the proponents proposed to develop an enhanced search strategy for a local search engine that will mainly address the problem of irrelevant results returned by most local search engines.

In order to achieve the objectives of the proponents, they took necessary steps in developing an enhanced search strategy. After much research and study has been done the framework of the new local search engine was finalized. After developing an enhanced search strategy, a number of tests were taken in order to show that the existing search strategy which is Level 2 Citation Graph addressed and solved the irrelevancy problems caused by existing search strategies.

However due to time constraints, the proponents were not able to fully populate their database. For students who would like to pursue the study, the proponents encourage them to add new features to the local search engine, find better ways to enhance database access and explore on the possibility of enhancing or creating a new search strategy.

CHAPTER I

INTRODUCTION

1.1 Background of the Study

The Internet, particularly the World Wide Web, is a vast source of information that is growing at an explosive rate. As the number of pages on the web grows, so will the number of results that search engines return.

Nowadays, technological revolutions are not isolated to foreign countries only, but have also reached out to many places all over the world including the Philippines. In Davao City alone, there has been a significant increase in the number of businesses, organizations and private homes that have become involved in the web. This consequently causes an increase in the number of local web pages, sites and documents available on the web. Hence, finding information on the Internet is becoming increasingly difficult as the Internet continues to exponentially grow and change.

Search engine developers, specifically in the Philippines, are now faced with the challenge of constructing a search engine that can scale up today's web and at the same time provide high quality search results in a short amount of time. With this, is the need to develop an enhanced search strategy used by search engines here in the Philippines.

1.2 Statement of the Problem

Although search engines gear the Internet to be the most popular product in the market today, there are still some concerns with search engines that have to be addressed. These concerns lie in the lack of relevance that local search engines return to a user's search query. Another concern is the broadness of the scope of most thus, rendering it unable to focus on specific topics in a more localized scene.

The main problem of the study is to develop an enhanced search strategy for a local search engine. The enhanced search strategy will be used in the proposed local search engine prototype.

These are issues that need to be addressed by the proponents in enhancing a local search engine search strategy:

- What are the different search strategies used by existing search engines?
- What are the causes of low relevant results generated by some local search engines?
- What theories or concepts can be used in enhancing a search strategy?
- What are the components of the search engine?
- What scripting language/s should be used in developing a local search engine?

1.3 Objectives

The main objective of the study is to develop an enhanced search strategy that will prove to be more efficient than an existing search strategy of a local search engine.

Specifically, the proponents have the following objectives:

- To be able to identify and understand the different search strategies used by search engines
- To be able to able to determine the causes of low relevant results generated by a local search engine
- To be able to identify the components of a local search engine
- To be able to determine what scripting language/s should be used in developing a local search engine
- To develop a search engine that will use the enhanced search strategy in retrieving relevant pages

1.4 Scope and Limitations of the Study

The study includes the survey of different search strategies used by search engines. Since the goal is to enhance a search strategy that will prove to be more efficient than the existing one in terms of relevance of results, two outputs will be produced.

The first one will be a local search engine implementing the existing search strategy and the other search engine will be implementing the enhanced search strategy. It will also make use of a database wherein submitted or gathered URLs, pages and sites will be filtered, indexed and ranked. The proponents will use MYSQL in creating the database on the basis of results of benchmarks and excellent reviews written by credible persons.

Limitations of the study include the scope of the contents and the speed. The scope of the proposed search engine since will only cater to local web pages, sites and URL's. Speed in terms of indexing and returning of query results will be hampered due to computer hardware limitations.

Since the search engine will only cater to local pages, sites and documents, the group cannot use simple or hybrid crawlers with various strategies in searching for sites because almost all crawlers do not have the capacity to distinguish which site or page belongs to the local or international domain.

1.5 Significance of the Study

The target users of the proposed local search engine are businessmen, students and professionals alike because they have greater need and more access to the Internet. First and foremost, it is beneficial to them because access to local web pages, sites and URL's is just a click away, may it be for business or entertainment purposes.

Businesses and institutions can submit their web site for free advertisement. It can also serve as a guide for people who plan to visit the city and get information regarding the history, geographical information of tourist spots and beautiful places to go.

Another contribution of this study is the presentation of a new framework for a local search engine that is more efficient than the search strategy used by an existing local search engine. It also introduces theories on information retrieval, page ranking, indexing and data mining concepts. Future search engine developers or computer science students can refer to the study if they have plans on developing their own search engine.