

Research Enterprise Office Search Portal

by

Halida Anis Hilmi

Dissertation submitted in partial fulfillment of
the requirements for the
Bachelor of Information Technology (Hons)
Information Technology

JAN 2004

Universiti Teknologi PETRONAS

Bandar Seri Iskandar

31750 Tronoh

Perak Darul Ridzuan

t

HD

30.37

.H157

2004

- 1) web portals
- 2) IT/IS -- Thesis

CERTIFICATION OF APPROVAL

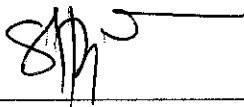
Research Enterprise Office Search Portal

by

Halida Anis Hilmi

A project dissertation submitted to the
Information Technology Programme
Universiti Teknologi PETRONAS
In partial fulfillment of the requirement for the
BACHELOR OF TECHNOLOGY (Hons)
(INFORMATION TECHNOLOGY)

Approved by,



(Ms Syarifah Bahiyah Rahayu Syed Mansoor)

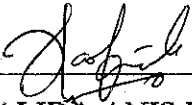
UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

JAN 2004

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references acknowledgements, and that original work contained herein have not been undertaken or done by unspecified sources or persons.



HALIDA ANIS HILMI

ABSTRACT

All the employees in University Technology Petronas need to access information instantaneously in order to enhance their functionality and efficacy. Is it easy to collaborate and gather the right information at the right time? Is all the research within a company documented? Is it easily available to all employees? And what happens when an employee leaves the company?

This project is an analysis of current practices and outcomes of the search portal and the nature of it as they are evolving in most of the organizations. The findings suggest that interest in search engines across a variety of industries is very high, the technological foundations are varied, and the major concerns revolve around achieving the correct amount and type of accurate research and garnering support for contributing to the search portal. Implications for practice and suggestions for future research are drawn from the study findings.

This project focused on the search function. The research is on how to make this search portal useful to the University Technology Petronas (UTP) community that is the UTP staff and lecturers. These search portal solutions are ideal for operations and maintenance manuals that once were reserved for 3-inch thick binders sitting on the shelves of many treatment plants. Moving the manual standard procedures, troubleshooting, theory, alarms, and equipment descriptions to an electronic, web-based solution offers many benefits. For one, the information can be updated and kept current much more effectively because it can be changed in one place and instantly updated at all access points. By developing this search portal, the staff and lecturers will be able to get information fast and efficiently.

ACKNOWLEDGEMENT

Firstly, the author would like to express deepest gratitude to Allah S.W.T for giving the strength, wisdom and patience in order to complete this project as per time given.

The author also would like to indicate that this project would have never completed without the help and support from many people. The author would like to send her deepest gratitude and thanks to Ms. Syarifah Bahiyah, Lecturer in Universiti Teknologi PETRONAS who was her Final Year Project Supervisor for all the guidance, support , good patients , advises and motivation which really helped and gave the author inspiration and strength to complete this project.

Next, special thanks to my beloved parents Mr. Hilmi Mohd Nashir and Puan Hawijah Ahmad and also family members for the morale support, love and encouragement for the author to complete the project.

Not forgotten, to all lecturers in UTP for every suggestion and continuous support; and to all my friends and colleagues in Universiti Teknologi PETRONAS, thank you for being committed with me throughout the project development. Thank you also for sharing experiences, knowledge and brilliant ideas with the author in order to complete the project.

Finally, to everyone who was involved either directly and indirectly, the author sincerely appreciates the efforts. Thank you very much.

TABLE OF CONTENTS

CERTIFICATION	i
ABSTRACT	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vii
CHAPTER 1: INTRODUCTION	1
1.1	Background of Study	1
1.2	Problem Statement	2
1.3	Objectives and Scope of Study	3
1.3.1	Relevancy of the Project	3
1.4	Feasibility of the Project within the Scope and Time Frame	4
CHAPTER 2: LITERATURE REVIEW AND THEORY	5
2.1	What Makes a Search Engine Good?	5
2.2	The BEST Search Engine: Table of Features	7
2.3	Search Engine Features	9
2.4	User Search	10
2.5	Presentation and Ranking	11
2.6	Traditional Search and Retrieval	11
2.6.1	Pre- vs. Post-Coordination	11
2.6.2	Content-based Filtering and Social Filtering	13
2.6.2.1	Pre-Coordination Enables Content-Based Filtering	13

2.6.2.2	Document Annotation and Appraisal Enables	
	Social Filtering	14
2.7	Moving Traditional Information Retrieval to the Web	14
	2.7.1 Introduction to WFTS Issues	15
	2.7.2 Enhanced WFTS	17
2.8	Documents as Web Knowledge Bases	18
2.9	Organizational KMSS design challenges	19
2.10	A Web-based Document KMSS	20
2.11	GOOGLE (http://www.google.com)	21
2.12	A Search Methodology	22
2.13	How to Search Successfully	22
2.14	Search Engine Ranking Methods and Algorithms	25
CHAPTER 3: METHODOLOGY		29
	3.1 Procedure Information	29
	3.2 Tools/ Equipment Required	30
CHAPTER 4: RESULTS AND DISCUSSION		33
4.1	Results – Print Screen	33
	4.1.1 “Login” Page	33
	4.1.2 “About Us” Page	34
	4.1.3 “Search” Page	35
	4.1.4 “Researchers” Page	36
4.2	Discussion – Use Case for Research Enterprise Search Portal	37
	4.2.1 User	37

4.2.2 System	38
CHAPTER 5: CONCLUSION AND RECCOMENDATION	39
5.1 Conclusion	39
5.2 Recommendation	40
REFERENCES	41

LIST OF TABLE AND FIGURES

Table 1	:	What Makes Search Engine Good?
Table 2	:	Search Engine Features
Figure 1	:	System Development
Figure 2	:	Print Screen – “Login” Page
Figure 3	:	Print Screen – “About Us” Page
Figure 4	:	Print Screen – “Search” Page
Figure 5	:	Print Screen – “Researchers” Page
Figure 6	:	Use Case Diagram – User
Figure 7	:	Use Case Diagram - System

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND OF STUDY

Today there are a score or more of "Web location services." A search engine proper is a database and the tools to generate that database and search it. A catalog is an organizational method and related database plus the tools for generating it. They provide news, libraries, dictionaries, and other resources that are not just a search engine or a catalog, and some of these can be really useful [1].

Most of the staff and lecturers in University Technology Petronas are constrained both by time and patience in the course of a search session. They need to get information fast. Today's Retrieval Interfaces pose formidable challenges to the user—he or she must browse, laboriously, one document at a time from the Retrieval Set with only limited data and metadata clues as signposts that might point to a document actually relevant for the problem at hand [2].

Early information technologies were designed to assist managerial and professional workers by processing and distributing vast amounts of information to managers organization-wide. Over several decades systems evolved to systems focusing on providing tools for ad-hoc decision analysis to specific decision makers, and to systems designed to provide updated, often real-time, relevant information to senior and middle managers. These systems each contributed to individual and organizational improvements in varying degrees and continue to be important components of an organization's information technology investment. An emerging line of systems targets professional and managerial activities by focusing on creating, gathering, organizing,

and disseminating the REO "research" such as the information and important data. These systems are referred to as REO Search Portal.

When employees possess the requisite research or information and are able to use it at the right moment, relationships with customers, dealers, suppliers and distributors generally improve. Such workers can make better decisions by increasing the amount of relevant information that they have access to. A search portal introduces the elements of expertise and experience through collaboration capabilities and shortens the time it takes to make better decisions.

1.2 PROBLEM STATEMENT

1.2.1 Problem Identification

For now, the Research Enterprise Office (REO) do not really has any problem with managing the research papers. This is because there are only a few researches done by the researchers. But these are a few problems that usually faced by companies:-

- ❖ There are tools to support the capture, modeling, validation, verification and maintenance of the research papers. However these tools do not extend to supporting the processes for managing the research at all levels within the organization.
- ❖ Difficult to analyze and plan its business in terms of the research it currently has and the research it needs for future business processes.
- ❖ Hard to identify and formalized existing research, acquiring new research papers for future use, archiving it in organizational memories and creating systems that enable effective and efficient application of the research within the organization.
- ❖ Research papers are used in everyday practice by professional personnel who need access to the right information, at the right time, in the right location.

1.2.2 Significance of Project

The search portal is designed to achieve both process results and organizational outcomes. The process improvements involved shortening the proposal time for client engagements, saving time, improving project management, increasing staff participation, enhancing communication, making the opinions of plant staff more visible, reducing problem-solving time, better serving the clients, and providing better measurement and accountability. These process improvements can be thought of as either relating to communication improvements or efficiency gains. The process improvements then, in the minds of the managers, led to cost reduction of specific activities, increased sales, personnel reduction, higher profitability, lower inventory levels, ensuring consistent proposal terms for worldwide clients,

1.3 OBJECTIVES AND SCOPE OF STUDY

- ❖ To have an enterprise-wide vocabulary to ensure that the research is correctly understood;
- ❖ To be able to identify, model and explicitly represent their research;
- ❖ To share and re-use their research among differing applications for various types of users; this implies being able to share existing research sources and also future ones;
- ❖ To create a culture that encourages research or information sharing.

1.3.1 Relevancy of the Project

Upon completion of the project, the output solves the problem stated in the problem statement and achieves the objectives outlined.

1.4 FEASIBILITY OF THE PROJECT WITHIN THE SCOPE AND TIME FRAME

For this project, the time frame given is around 4 months. The author will spend one month to carry out research and writing paperwork. The remaining three months will be allocated in developing the system. The project will be divided into several phases. The author needs to design the website for REO. Next the author needs to develop the search portal by integrating the features that need to be included focusing on the search menu. Later, the author will deal with effectiveness with the website.

CHAPTER 2

LITERATURE REVIEW AND THEORY

2.1 WHAT MAKES A SEARCH ENGINE GOOD?

All search engines consist of three parts: (1) a database of web documents, (2) a search engine operating on that database, and (3) a series of programs that determine how search results are displayed. Because the search engine business is competitive, most search engines also offer additional features that are convenient or fun. The table below shows what can vary within each of the three basic parts in search engines [3]:

Parts of Search Engines	Variables, and their implications for your searches
Database of web documents	<ul style="list-style-type: none"> • Size of database: <ul style="list-style-type: none"> ○ How many documents does the search engine claim it has? ○ How much of the total web are you able to search? • Freshness ("up-to-dateness"): <ul style="list-style-type: none"> ○ Search engine databases consist of copies of web pages and other documents that were made when their crawlers or spiders last visited each site. How often is the database refreshed to find new pages? ○ How often do their crawlers update the copies of the web pages you are searching? • Completeness of text: <ul style="list-style-type: none"> ○ Is the database really "full" text, or only parts of the pages? ○ Is every word indexed? • Types of documents offered: <ul style="list-style-type: none"> ○ All search engines offer web pages. ○ Do they also have extensive PDF, Word, Excel, PowerPoint, and other formats like WordPerfect? ○ Are they full-text searchable? • Speed and consistency: <ul style="list-style-type: none"> ○ How fast is it? ○ How consistent is it? Do you get different results at

different times?

The search engine's capabilities

Search engines let you enter some keywords and search on them. What happens inside? Can you limit in ways that will increase your chances of finding what you are looking for?

- **Basic Search options and limitations:**
 - Automatic default of AND assumed between words?
 - Accepts " " to create phrases?
 - Is there an easy way to allow for synonyms and equivalent terms (OR searching)? Can you OR phrases or just single words?
- **Advanced Search options and limitations:**
 - Can you require your search terms in specific fields, such as the document title? Can you require some words in certain fields and others anywhere?
 - Can you restrict to documents only from a certain domain (org, edu, gov, etc.)? Limit to more than one or only one?
 - Can you limit by type of document (pdf or excel, etc.)? More than one?
 - Can you limit by language?
 - How reliably and easily can you limit to date last updated?
- **General limitations and features:**
 - What do you have to do make it search on common or stop words?
 - Maximum limit on search terms or on search complexity?
 - Ability to search within previous results?
 - Can you count on consistent results from search to search and from day to day?
 - Can you customize the search or display?
 - Is there a "family" filter? Does it work well? Is it easy to turn on or off?

Results display

When search engines return a list of results it "thinks" are what you are looking for. How well does it think like you expect it to think?

- **Ranking:**
 - Are they ranked by popularity or relevancy or both?
 - Do pages with your words juxtaposed (like a phrase) rank highest?
 - Do you get pages with only some of your words, perhaps in addition to pages with them all?
- **Display:**
 - Are your keywords highlighted in context, showing excerpts from the web pages which caused the match?
 - Some other excerpt from the page?
- **Collapse pages from the same site:**
 - If it shows only one or a few pages from a site, does it show the one(s) with your terms?
 - How easy is it to see all from the site?

	<ul style="list-style-type: none"> ○ Can this be changed and saved as your preferred search method?
Other features	Search engine designers try to come up with all kinds of features and services that they hope will allure you to their services.

Table 1 : What makes Search Engine good?

2.2 THE BEST SEARCH ENGINES: TABLE OF FEATURES [4]

	<u>Google</u> http://google.com/	<u>Teoma</u> http://www.teoma.com/	<u>AlltheWeb Advanced</u> Type alltheweb and click <u>Advanced Search</u>	<u>Alta Vista Advanced</u> Type http://www.altavista.co then click <u>Advanced Search</u>
	HUGE. Over 2 billion. Claims over 3 billion but about 1 billion are not fully indexed (i.e., cannot be full-text word searched). Unindexed pages are retrieved if your search matches their titles or match other pages linking to them.	LARGE. Claims to have 1 billion fully indexed, searchable pages, and 1 billion more partially indexed. Strives to become #1 in size.	HUGE. Over 3 billion fully indexed, searchable pages. Sometimes ties for first in tests. Advanced Search worth mastering.	LARGE, but smaller than Google or AllTheWeb. Use the Advanced Search .
	Popularity ranking using PageRank™ . Limit of 10 words per search, excluding OR. Indexes the first 101KB of a Web page, and 120KB of PDF's	Subject-Specific Popularity™ ranking. Suggests terms within results to refine. Suggests pages within results with many links.	No stop words . URL Investigator to find out about a page. Conversion of weights and measures.	Full Boolean searching and powerful Searching within results using SOI BY box in Advanced Search . Basic search provides distracting commercial, paid, and directory entries.
	Yes. Use " " Searches common "stop words" if in phrases in quotes.	Yes. Use " " Searches common "stop words" if in phrases in quotes.	Yes. Use " "	Yes. Use " "

	<p>Partial AND assumed between words Capitalize OR - excludes. No () or nesting. In Advanced Search, partial Boolean available in boxes.</p>	<p>Partial AND assumed between words. Capitalize OR - excludes. No () or nesting</p>	<p>If Boolean expression is selected in Advanced Search, accepts AND, OR, ANDNOT, and ().</p>	<p>AND, OR, AND NOT, NEAR (within 10 words) In Advanced Search, or capitalized in Basic Search.</p>
	<p>Sort of. At bottom of results page, click "Search within results" and enter more terms. Adds terms.</p>	<p>Sort of. Add terms. REFINE pastes suggested sub-topics within results.</p>	<p>Sort of. At bottom of search results. Terms entered will be added to terms previously searched.</p>	<p>Yes. Use <i>Sorted by</i> box under Boolean search box. Sorts and filters search results.</p>
	<p>Based on page popularity measured in links to it from other pages; high rank if a lot of other pages link to it. Fuzzy AND also invoked. Matching and ranking based on "cached" version of pages that may not be the most recent version.</p>	<p>Based on Subject-Specific Popularity™, links to a page by related pages. More info.</p>	<p>Automatic Fuzzy AND. Also seems to use "importance" and links to pages. In Advanced Search, SHOULD INCLUDE gives higher priority to word or phrase in box. Each box read as a phrase. In Boolean Search, rank:word is supposed to rank by that term.</p>	<p>By the terms you specify in <i>Sorted by</i> box under Boolean search box. Relevancy ranked if left blank.</p>
	<p>link: site: allintitle: intitle:</p>	<p>intitle: inurl: site: geoloc:</p>	<p>In Advanced Search, can search within text, title,</p>	<p>title: url: link: host:</p>

	allinurl: inurl: Advanced Search boxes for most of these. Offers special searches.		link name, url, link to the url Also offers commands similar to Google as Special Features.	domain: anchor: text: image: applet:
	No. Search variant endings and synonyms separately, separating with OR (capitalized): <i>airline OR airlines</i>	No. Search variant endings and synonyms separately, separating with OR (capitalized): <i>airline OR airlines</i>	No. Enclose variants in () in top box to create OR search. (<i>airline airlines</i>)	Yes. Use *
	Yes. in <u>Translate this page</u> link following some pages. To English from major European languages.	No.	No.	Yes, to and from English and other languages. Click on <u>Translate</u> following result.

Table 2: Search Engine Features

2.3 SEARCH ENGINE FEATURES

By far the best service for carefully specifying a search was Open Text. This form has great menus, making a complex Boolean search fast and easy. Best of all, this service permits you to specify that you want to search only titles or URLs. But then there's Alta Vista's little known "keyword" search syntax, now as powerful as Open Text, but not as easy to use. You can constrain a search to phrases in anchors, pages from a specific host, image titles, links, text, document titles, or URLs using this feature with the syntax keyword: search-word. There is an additional set of keywords just for searching Usenet [1].

What could really make engines with large data bases shine, however, would be an improvement in the way they rank and present results. All engines I tested had ranking schemes that were not well documented, based on how many times your search words were mentioned, whether or not they appeared early in the document, whether or not

they appeared close together, and how many search terms were matched. I did not find the ranking schemes very useful, as relevant and irrelevant pages frequently had the same scores.

2.4 USER SEARCH

What can the user do besides typing a few relevant words into the search form? Can they specify that words must be in the title of a page? What about specifying that words must be in an URL, or perhaps in a special HTML tag? Can they use all logical operators between words like AND, OR, and NOT?

Most engines allow user to type in a few words, and then search for occurrences of these words in their database. Each one has their own way of deciding what to do about approximate spellings, plural variations, and truncation. If user just type words into the "basic search" interface user get from the search engine's main page, user also can get different logical expressions binding the different words together. Excite! actually uses a kind of "fuzzy" logic, searching for the AND of multiple words as well as the OR of the words. Most engines have separate advanced search forms where user can be more specific, and form complex Boolean searches (every one mentioned in this article except Hotbot). Some search tools parse HTML tags, allowing user to look for things specifically as links, or as a title or URL without consideration of the text on the page [1].

By searching only in titles, one can eliminate pages with only brief mentions of a concept, and only retrieve pages that really focus on the concept [1].

By searching links, one can determine how many and which pages point at your site. Understanding what each page does with the non-standard pluralization, truncation, etc. can be quite important in how successful user searches will be. For example, if user search for "bikes" user won't get "bicycle," "bicycles," or "bike." In this case, use a search engine that allowed "truncation," that is, one that allowed the search word "bike"

to match "bikes" as well, and would search for "bicycle OR bike OR cycle" ("bicycle* OR bike* OR cycle*" in Alta Vista) [1].

2.5 PRESENTATION AND RANKING

With databases that can keep the entire Web at the fingertips of the search engines, there will always be relevant pages, but how do you get rid of the less relevant and emphasize the more relevant?

Most engines find more sites from a typical search query than you could ever wade through. Search engines give each document they find some measure of the quality of the match to your search query, a relevance score. Relevance scores reflect the number of times a search term appears, if it appears in the title, if it appears at the beginning of the document, and if all the search terms are near each other; some details are given in engine help pages [1]. Some engines allow the user to control the relevance score by giving different weights to each search word. One thing that all engines do, however, is to use alphabetical order at some point in their display algorithm. If relevance scores are not very different for various matches, then you end up with this sorry default. For most uses, a good summary is more useful than a ranking. The summary is usually composed of the title of a document and some text from the beginning of the document, but can include an author-specified summary given in a meta-tag. Scanning summaries really saves you time if your search returns more than a few items [1].

2.6 TRADITIONAL SEARCH and RETRIEVAL

2.6.1 Pre- vs. Post-Coordination

Pre-coordination is defined as fixing the citation order of a subject heading at index time [6], in a card-catalog system this assumes significant time and effort on the part of a cataloger before subject term search is possible. Post-coordination is so named because the keywords are combined at search time; there is no subject term taxonomy specified a priori. Many full text search engines, such as Excite, apply statistical

methods to an unordered vector of keywords. To clarify the difference between pre- and post-coordination, this quote is helpful [7]:

“When concepts are combined or coordinated to form complex subjects, such coordination may be carried out by the indexer or by the searcher. The former is referred to as pre-coordinate indexing and the latter as post-coordinate indexing”.

There have a long-standing debate between Salton [8] [9], a fervent supporter of the virtues of free text search and Blair and Maron [10] [11] who argue that pre-coordinated search is more effective in large document archives. Given a large document archive such as the massive hypermedia repository afforded by the World-Wide Web (WWW) technology, and given ad-hoc user queries, the weaknesses of post coordinate search engines are well-known. They suffer degradation in the following measures:

- precision (fraction of the selected documents which are actually relevant to the user’s information need), and
- recall (fraction of the actual set of relevant documents that are correctly classified as relevant by the text filtering system).
- fallout (fraction of non-relevant documents that are selected).

Pre-coordination systems, by themselves, are encumbered traditionally by the time demands placed on a cataloger to impose a subject term structure and a possible divergence over time between the collective users’ semantics and the original subject term choices.

As the WWW leads to archives growing in size and number, the precision in full-text search grows proportionally worse [11], due in part to the lack of intelligent orderings of keywords [12]. Subject searching exposes users to the difficulties of constructing Boolean queries [13] and forces the users to guess terms in the order in which they were defined by the cataloger, thus again limiting precision. Pre-coordination also limits precision for several [12]: if a catalog omits subdivisions on certain broad topics, only a shallow heading may be guessed by the user, and if a catalog has a topical subject

heading and a subdivision, the user is required to guess the terms in order to locate the document. In either case we encounter a gap between the semantics imposed by the cataloger on the document collection (in the initial taxonomy) and the needs of the information seeker. Note that researchers often call a pre-categorized structure, such as a keyword hierarchy describing nodes (documents). To summarize the debate between pre-coordinate and post-coordinate systems, there is evidence in the literature of user dissatisfaction with both techniques in stand-alone systems.

Retrieval systems on text archives are usually evaluated by two measures, precision and recall. In practice, recall and precision tend to vary inversely; “it is difficult to retrieve everything that is wanted while also rejecting everything that is unwanted” [8]. Precision can be interpreted to mean the probability that a retrieved document will be relevant and recall as the probability that a relevant document will be retrieved [10].

Weaknesses in pre- or post-coordination stand-alone systems have caused a shift in interest to coping with information overload by filtering the document archives either via content or via social mechanisms. The following section reviews those efforts.

2.6.2 Content-based Filtering and Social Filtering

The idea to focus on the *reception* of information via some sort of filtering mechanisms, as opposed to the *generation* of information via a pre- or post-coordinated search [14] [15]. Attention has therefore been focused on Denning’s term *information filtering*. The task is simple: to sort through large volumes of information and present the user with sources of information that are likely to satisfy his or her information requirement [15]. To accomplish this filtering, two major paths have been attempted, *content-based filtering* and *social filtering* which we will now discuss.

2.6.2.1 Pre-Coordination Enables Content-Based Filtering

Content-Based Filtering was introduced by Luhn in 1958 [16]. In its simplest form content filters are straightforward to implement. If a (possibly unmanageably) large document archive is classified by its subject terms and a user creates a personal profile of subject terms, a content filter process can extract a subset of matching documents and better “match the personal bandwidth of the user” [15]. Recent work [17] has concentrated on more sophisticated software treatments to perform intelligent content analysis to generate recommendations and match them to personal user profiles.

A pre-coordinated subject term organization of the documents assists in the creation of content filters; the user need only select from the existing subject term taxonomy to compose a profile.

2.6.2.2 Document Annotation and Appraisal Enables Social Filtering

Social (Collaborative) Filtering, introduced but not implemented by Malone et al. [18] as an alternative to the content-based filtering of their Information Lens system, is an emergent property enabled by user communities ‘marking up’ documents with annotations and appraisals. In this scheme, the representation of the document is based on annotations to that document made by prior readers of the document [15]. Malone et al. speculated that communities of shared interest could be automatically identified in this way, regardless of whether documents’ content could be represented in a way that was useful for selection.

Social filtering systems have been implemented for e-mail, Usenet news, and newswire stories in the Tapestry project [19] which uses a standard client-server protocol. The Tapestry work suggests that a critical mass of users with overlapping interests is needed for social filtering to be effective. GroupLens [20] is another social filtering system, also using a standard client-server protocol, which provides a content and annotation server for Usenet News. The problem of user motivation exists with social filtering systems, since there is no motivation for the first user to annotate anything [15].

2.7 MOVING TRADITIONAL INFORMATION RETRIEVAL TO THE WEB

The Web as a popular hypermedia environment is a natural platform for many organizations to implement IR functionality. Many organizations choose to avoid pre-structuring their documents and simply publish them, in various formats such as ASCII, HTML, MS-Word and PDF to the Web. In this event, the most natural means to conduct search on an unstructured collection is Full Text search. I term FTS on the Web *WFTS*—it inherits the weaknesses of traditional full text search but it also can make use of favorable properties of the WWW hypermedia environment.

I start by introducing WFTS and move on to show examples of studies which research WFTS technologies.

2.7.1 Introduction to WFTS Issues

First I consider the case of Full Text Search on the Web.

The Search Interface

IR researchers have long recognized that users face difficulty in constructing Boolean queries. Thus, some researchers build translators on the front-end which will modify the users' input before it is sent along to the search engine. For example, Lawrence and Giles [21] in the NECI meta search engine project divide searches into *specific expressive forms* or SEFs which are various restatements of the original query. Their example:

'What does NASDAQ stand for?' is transformed into the SEFs 'NASDAQ stands for', 'NASDAQ is an abbreviation', and 'NASDAQ means'. Clearly the information may be expressed in forms other than these, but if the information exists in just one of these forms, it is more likely to satisfy the query. The technique thus trades recall for precision.

Note that the transformation takes place transparently to the user. The NECI engine also parallelizes calls to various major extant engines, for example, Alta Vista, Excite, and others. It starts to return results more quickly than the individual engines by unbuffering

the input and output to the user. When the Retrieval Interface is formed, the core engines which found a given document are indicated. Thus the user has additional useful clues before he or she decides to pursue a document down to the Document layer.

In addition, Shneiderman et al. [22] have worked on an improved user interface to increase the usability of typical search. As they state the problem [22]:

The ideal user interface is comprehensible, predictable and controllable, but many current text-search interfaces — especially on the World-Wide Web — involve unnecessarily complex and obscure features. The result is confusion and frustration for advanced users as well as for beginners, scientists, and students.

Their enhancement to the Search interface helps the user by spelling out the program's interpretation of the inputted keywords and helps the users limit the search, if they so choose, to sections of documents such as the body of a newswire story. This system is integrateable with these efforts.

Enhancing the Document Interface

An interesting approach to improve the visualization of results at the Retrieval interface was presented by Mukherjea and Hara [23]. In this work, *landmark* nodes are identified at a site beforehand. Then, when the users browse the Retrieval list, a graphical representation is shown for each of the documents in the list vis-a-vis landmark nodes. They use the Harvest Information Discovery and Access system to index the site's pages, since Harvest also supplies them with the topology of the pages. This topology is then used to build their 'landscape metaphor' [23].

Cooper and Byrd also focus on a visual interface for retrieved results in their OBIWAN system [2]. They present a fan-out diagram of keyword clusters. Another prong of this research effort is a modification the core search engine with domain-specific vocabulary and specialized handlings of acronyms and proper names. Their empirical study was conducted on management consulting documents at the Giga Group.

Novel visualization work has recently been carried out in a slightly different domain. Rather than operate on documents on their corresponding full-text indices, used as input the transcript of a GroupSystems brainstorming session [24]. To visualize the output by keywords, an AI technique (Kohonen maps) was used. Terms on borders in the Kohonen map with other terms could be inferred to have greater ambiguity.

A more sophisticated treatment of documents has been performed by Phelps and Wilensky [25] on client-side document decomposition. They introduce the term 'multivalent documents': by building Java applets at the client side, they improve the presentation of documents (separating them into text, scanned OCR pages, and other layers). This system does not make use of client-side extensions; rather it relies on a set of lightweight server extensions to capture user annotations and other session statistics. Another interesting possibility is to show the client the documents in the context of the web server's overall structure.

Another tack is to enlist the help of the users to improve the search session over time. In Golovchinsky's work [26], he provides a feedback mechanism whereby users' feedback after a search session is linked to an automated mechanism for building new links. He is addressing the problem that "large hyper-linked collections may overwhelm users with the range of possible links from any node, only a fraction of which may be appropriate for a given user at any time" [26]. In his system VOIR he finds that the new links, brought about by relevance feedback queries, "are more effective than user-specified queries in retrieving relevant information" [26].

2.7.2 Enhanced WFTS

Many researchers are not content to simply transplant FTS to the Web, arguing that useful metadata about documents is a rich source of information discovery as well. For example, the links in a document can be exploited, as well as links pointing to a given document from other sources. Furthermore, the location (URL) of a document contains useful clues, such as its depth in a server's hierarchy or its proximity to other

documents. Since regular WFTS is not exploiting all the clues we can gather in the rich WWW environment, it is worthwhile to examine attempts to improve the situation.

The Power of Metadata

Classic IR theory recognizes that metadata (information about documents) is very useful for document categorization and retrieval. If metadata was encoded in a structured manner, powerful fielded searches would be possible on a given document archive. However, the WWW as it stands today places obstacles in the way of assigning document metadata in a coherent way.

Legacy HTML documents permit limited Metadata with META tags. However these are quite ad-hoc; many legacy documents will be missing the tags altogether and there is no consensus how they should be used. In addition, Adobe Portable Document Format (PDF) documents permit limited metadata, such as Subject, Author and Title.

Note that “HTML’s strength was its simplicity — it combined simple document structuring with presentation information in a readily understandable fashion” [27]. However, this does not permit workflow or the automated handling of documents by software modules, which can be termed “intelligent clients” [27]. Here, eXtensible Markup Language (XML) is useful. It “helps domain authors logically structure their documents consistently” [27].

2.8 DOCUMENTS AS WEB KNOWLEDGE BASES

In contrast to well-structured fielded database, unstructured or semi-structured (template-based) documents represent an increasingly important part of organizational knowledge bases. Documents have the potential to be highly expressive, with embedded multimedia objects. While expressive and strong in presentational markup (rendering) they are often poor in semantic markup (no ontology) making knowledge search and discovery difficult. Note that document repositories represent merely the potential to transfer knowledge to individual readers. Although the literature commonly speaks of

knowledge bases of paper documents, for example Ford's initial knowledge base of more than 30,000 paper pages [28] the labeling of information as knowledge a priori can be misleading. Paper documents do not add to the organizational knowledge base unless two conditions are met: (i) that they are read by one or more readers, and (ii) that the readers, in interacting with the information and data contained in the document, increase their own personal knowledge. If these two conditions are met, the necessary remaining step from the organizational point of view is that individuals form groups to articulate and amplify their knowledge. For the remainder of the KM discussion, when we use the term knowledge base in the document context we are referring to the potential of the document archive to impart knowledge to the recipient, not an intrinsic quality of the collection.

Documents are an interesting area of study; since the WWW facilitates distributed document publishing they are a common component of knowledge bases. The chief factor underlying the ease of document publication has been the near universal acceptance of open network standards (the TCP/IP protocol suite) enabling interoperability in the application-layer protocols, for example HTTP [29]. Still, professional document work products typically incur a high cost of creation in time and effort. Inefficient document bases for storage and retrieval effectively diminish the value of professional and expert time invested in document creation.

This system does not make use of these advanced in-place edit or workflow features, rather, via its lightweight annotation facility it leaves the core document untouched.

2.9 ORGANIZATIONAL KMSS DESIGN CHALLENGES

In addition to technical challenges organizations often lack adequate incentives for knowledge sharing and management. These difficulties are often exacerbated in emerging federalist organizations which are dynamic, team based problem solving structures with distributed authority. These organizations may address a wide variety of problems that limit the usefulness of static ontologies. The first decision business units make is the choice of specific groupware products, such as Notes (Domino) or Intranet

product suites [30], the broader issue is how to organize the documents underlying the groupware product to facilitate knowledge transfer.

As a result it is not surprising that most systems in the past have covered limited domains [28]. As document publishing is simplified, and Intranets link individuals in organizations to rapidly expanding web document bases, the previous problems in the design and maintenance of KMSS become more pronounced. To address some of these problems I will develop this system which provides a flexible architecture and scalable KMSS to support federated organizational structures. Believe this organizational form will increasingly prevail necessitating the design of KMSS technologies to better fit the requirements for managing knowledge within this type of organization.

2.10 A WEB-BASED DOCUMENT KMSS

Typical Web Full Text Search (WFTS) engines which provide post-coordinate search have deficiencies which translate into inadequate support for KM. For example, there is no way to share resource discovery made during the course of an ad-hoc search session for one's future use or between users. There are also extremely limited data and metadata clues to assist the user as he or she traverses the system from the front-end (the Query Layer) to the intermediate layer, which is an array of hyperlinks to base documents (the Retrieval Layer) and on to the bottom layer, the Document Layer. In a typical implementation, the user has no knowledge of others' prior searches or results at the Query front-end and has very few clues at the Retrieval layer. The Retrieval layer might show the document's title or a brief summary, but this is often not enough due to the time-intensive commitment of browsing documents at the Document layer.

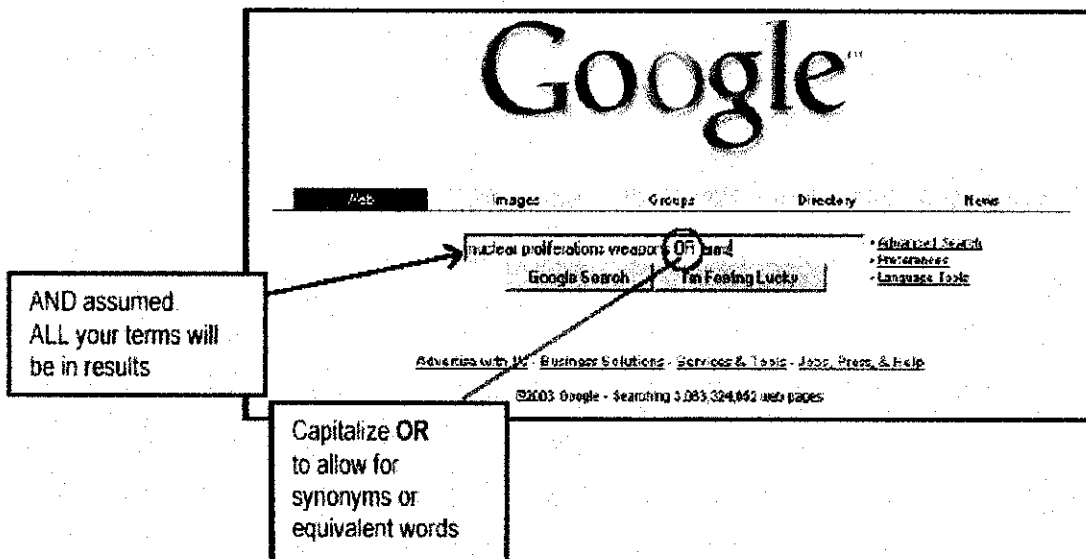
In summary, the system is designed to enable knowledge sharing across business unit boundaries. This system is predicated on the principle that the users and creators of knowledge best know the information relevant to their knowledge management task and

that they can more effectively filter, discover and signal useful knowledge to their peers than an automatic system.

To overcome some of the prior challenges to KMSS design by enabling readers to: actively become secondary authors and co-creators of value who provide document annotations that filter and enhance document content doing away with the artificial author/reader barrier [31].

2.11 GOOGLE (<http://www.google.com>) [34]

- Huge database. Claims over 3 billion, but about 1 billion are not fully indexed and therefore not fully searchable
- Many formats besides Web pages (PDF, Word, Excel, WordPerfect, and more)
- Adequate advanced searching from commands in basic or from Advanced screen. 10 word limit.



SAMPLE SEARCHES:

#1. Keywords and phrases in quotes:

nuclear proliferation

"nuclear proliferation" iran

#2. OR searches in Google:

"nuclear weapons" proliferation OR sales

“nuclear proliferation” ethics OR “ethical issues”

#3. Name of an association, society, company, agency, institution, or person.

“national nuclear control institute”

"nuclear threat initiative"

#4. Search within a site

site:cia.gov "nuclear proliferation"

site:disarmament.un.org "nuclear arms"

2.12 A SEARCH METHODOLOGY

It is recommend that to follow a structured technique such as [34]:

1. **Spell it Out** – Where we need to define the topic, and generate a list of search terms.
2. **Strategize** – Then we need to choose which online tools and resources will work best on our search terms.
3. **Search** - Get online, execute, stay focused, use advanced search features
4. **Sift** – We need to filter the results, and then follow the leads.
5. **Save** – After we have found what we are searching, we need to save it or take notes, organize results, bookmark or share.

2.13 HOW TO SEARCH SUCCESSFULLY

There are several steps to ensure that the search is successful. The steps are [35]:

Step 1: Define Your Topic

Have a very clear idea of your search topic

- Write it down: try to summarize your topic in the form of a question
- Add comments to indicate such things as "I want to find information written since 1990 only" or "I want to limit my search to English language materials only"

Example: What methods can be used to teach children good hygiene?

Step 2: Identify Main Concepts

- Computers do not handle natural language searching very well - they prefer to deal with a search topic one concept or idea at a time
- Divide your topic into concepts: concepts should be meaningful "hard" terms - usually only verbs and nouns
- Your topic may consist of one concept or more than one concept.

Example:

CONCEPT A	CONCEPT B	CONCEPT C
child	hygiene	teach

Step 3: Develop a List of Search Terms

- We need to think of the different ways a writer might express each concept.
Consider:
 - Synonyms (e.g., poor, poverty, disadvantaged, etc.)
 - Alternative spellings (e.g., labor, labour, pediatric, paediatric, etc.)
 - Variant endings (e.g., child, children, childish, etc.)
 - Acronyms (e.g., UN, SARS, etc)
- Some databases have a thesaurus or subject list available to help you develop this list
- It is often helpful to arrange the keywords for each concept in a group

Example:

CONCEPT A	CONCEPT B	CONCEPT C
children	hygiene	teach
child	hygienic	educate
toddler/s	cleanliness	education
preschooler/s	handwashing	instruct

Truncation (Wild Card)

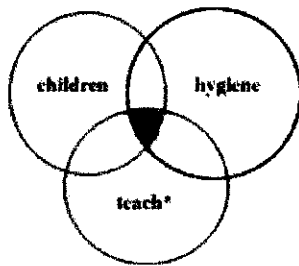
Used in computer searching, truncation is like a wildcard. The symbol used is often the asterisk *, but may be another character. When this symbol is added to the end of a word root, your search will retrieve all possible endings of that word.

Example:

child*: retrieves child, child's, children, etc.

Step 4: Construct Your Search Statement

Boolean operators need to be used to link your search terms together, so the computer system will understand what you are looking for. The most commonly used Boolean operators are **AND** and **OR**.



AND

- Requires that **ALL** search terms be present
- Use to connect different concepts
i.e., to combine your main ideas together
- Use to narrow your search (retrieve fewer results)
- Retrieves the records containing **ALL** terms

Example:

If your search statement is: **children and hygiene and teach***

Your results would contain **ALL** of the following terms:

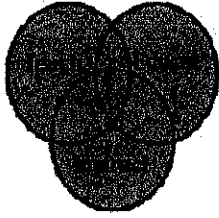
children

hygiene

teach*

(or teaching teaches, etc)





- OR**
- Requires that **ANY** of the search terms be present
 - Use to connect all your synonyms for each concept
i.e., to combine "like" terms
 - Use to broaden your search (retrieve more results)
 - Retrieves the records containing **ANY** term

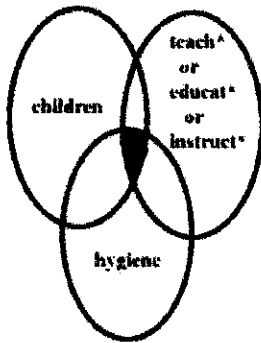
Example:

If your search statement is: **teach* or educat* or instruct***

Your results would contain **ANY** of the following terms:

teach, teaches, teacher, teaching, instruct, instructs, instruction, instructor, instructing
educate, educates, education, etc.

Example:



If your search statement is:

children and hygiene and (teach* or educat* or instruct*)

Your results will contain **both** of the following terms...

children

hygiene

...as well as **one or more** of the following terms:

teach, teaches, teacher, teaching, instruct, instructs, instruction, instructor, instructing
educate, educates, education, etc.

2.14 SEARCH ENGINE RANKING METHODS AND ALGORITHMS

Page Rank

Search engine ranking algorithms are closely guarded secrets, for at least two reasons: search engine companies want to protect their methods from their competitors, and they also want to make it difficult for web site owners to manipulate their rankings. That said a specific page's relevance ranking for a specific query currently depends on three factors [36]:

- Its relevance to the words and concepts in the query
- Its overall link popularity
- Whether or not it is being penalized for excessive search engine optimization (SEO).

Factor #2 was innovated by Google with PageRank. Essentially, the more incoming links your page has, the better. But it is more complicated than that: indeed, PageRank is a tricky concept because it is circular, as follows: Every page on the Internet has a minimum PageRank score just for existing. 85% (at least, that's the best known estimate, based on an early paper) of this PageRank is passed along to the pages that page links to, divided more or less equally along its outgoing links. A page's PageRank is the sum of the minimum value plus all the PageRank passed to it via incoming links [36].

Although this is circular, mathematical algorithms exist for calculating it iteratively. In one final complication, what I just said applies to "raw PageRank." Google actually reports PageRank scores of 0 to 10 that are believed to be based on the logarithm of raw PageRank (they're reported as whole numbers). And the base of that logarithm is believed to be approximately six [36].

Anyhow, there are about 30 sites on the Web of PageRank10, including Yahoo, Google, Microsoft, Intel, and NASA. IBM, AOL, and CNN, by way of contrast,

were only at PageRank 9 as of early in 2004.

Further refinements in link popularity rankings are under development. Notably, link popularity can be made specific to a subject or category; i.e., pages can have different PageRanks for health vs. sports vs. computers vs. whatever. Supposedly, AskJeeves/Teoma already works that way.

It is believed that Inktomi, Altavista, et al. use link popularity in their ranking algorithms, but to a much lesser extent than Google. Yahoo, owner of Inktomi, Altavista, Alltheweb, is rolling out a new search engine, which reportedly includes a feature called Web Rank. More on how that works soon [36].

Keyword Search

Most search engines handle words and simple phrases. In its simplest form, text search looks for pages with lots of occurrences of each of the words in a query, stopwords aside. The more common a word is on a page, compared with its frequency in the overall language, the more likely that page will appear among the search results. Hitting all the words in a query is a lot better than missing some [36].

Search engines also make some efforts to “understand” what is meant by the query words. For example, most search engines now offer optional spelling correction. And increasingly they search not just on the words and phrases actually entered, but they also use stemming to search for alternate forms of the words (e.g., speak, speaker, speaking, spoke). Teoma-based engines are also offering refinement by category, ala the now-defunct Northern Light. However, Excite-like concept search has otherwise not made a comeback yet, since the concept categories are too unstable [36].

When ranking results, search engines give special weight to keywords that appear:

- High up on the page
- In headings

- In **BOLDFACE** (at least in Inktomi)
- In the URL
- In the title (important)
- In the description
- In the ALT tags for graphics.
- In the generic keywords metatags (only for Inktomi, and only a little bit even for them)
- In the link text for inbound links.

More weight is put on the factors that the site owner would find it awkward to fake, such as inbound link text, page title (which shows up on the SERP -- Search Engine Results Page), and description [36].

CHAPTER 3

METHODOLOGY

3.1 PROCEDURE IDENTIFICATION

To relate technical design and social impact is a common theme, not only in the CSCW literature, but more broadly in the argument that systems development as a tool to measure IT constructs is important. For example, a more general argument to support a system development in Information Systems research can be found in Nunamaker, Chen, and Purdin's essay [32]. They state, "Concepts alone do not ensure a system's survival. Systems must be developed in order to test and measure the underlying concepts. Systems development is therefore a key element of IS research" [32]. So, Nunamaker, Chen, and Purdin bridge the gap between technological research (the 'concept') and the social implications, the 'impact' [32], by exploring its use in the field.

Stated that in the system architecture phase, must develop a "unique architectural design for extensibility, modularity." and in the design phase, one solution must be chosen from the alternatives. Finally, in the prototyping build phase, it is necessary to "gain insight about the problems and the complexity of the system" [32]. These steps were undertaken in the development of this system.

Figure 3-1 shows an adaptation of Nunamaker et al.'s essay in the system setting.

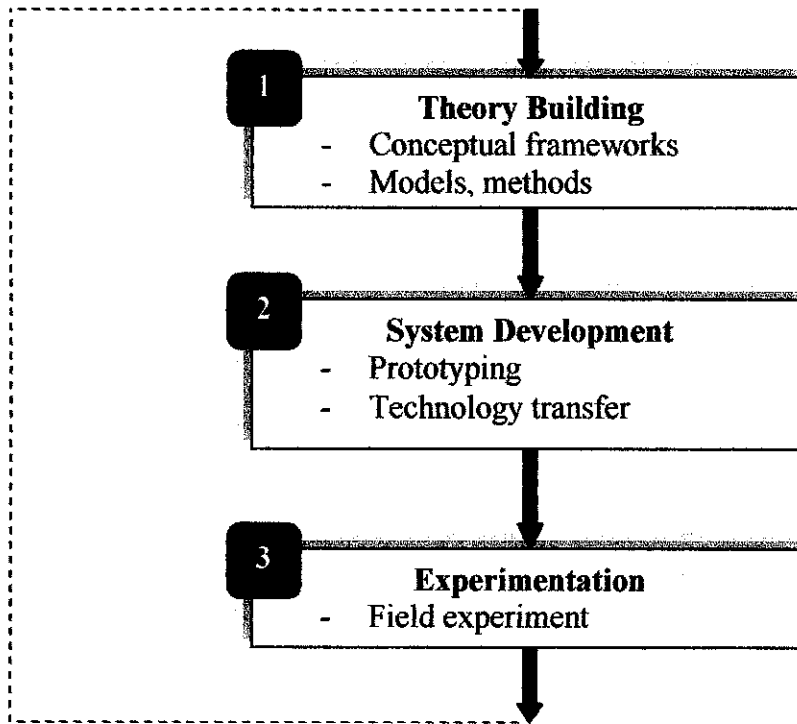


Figure 1: System Development

3.2 TOOLS / EQUIPMENT REQUIRED

- **Macromedia Dreamweaver MX**

Dreamweaver MX 2004 is the professional choice for building web sites and applications. It provides a powerful combination of visual layout tools, application development features, and code editing support, enabling developers and designers at every skill level to create visually appealing, standards-based sites and applications quickly. From leading support for CSS-based design to hand-coding features, Dreamweaver provides the tools professionals need in an integrated, streamlined environment. Developers can use Dreamweaver with the server technology of their choice to build powerful Internet applications that connect users to databases, web services, and legacy systems.

- **Adobe Photoshop 7**

This is an image editing program used for editing color of images, retouching proofs, adding and creating special affects to images. It can be used in Web pages, PowerPoint presentations, and word processing documents.

- **Macromedia Flash MX**

Macromedia Flash MX 2004 is the industry-standard tool for creating effective rich content across desktops and devices. Designers and developers use Macromedia Flash MX 2004 to accelerate projects while maintaining a high degree of creative control. Jump-start projects with templates and components, and take advantage of the vast Macromedia online resource library.

- **My SQL**

MySQL is a relational database management system, which means it stores data in separate tables rather than putting all the data in one big area. This adds flexibility, as well as speed. The SQL part of MySQL stands for "Structured Query Language," which is the most common language used to access databases. The MySQL database server is the most popular open source database in the world. It is extremely fast and easy to customize, due to its architecture. Extensive reuse of code within the software, along with a minimalist approach to producing features with lots of functionality, gives MySQL unmatched speed, compactness, stability, and ease of deployment. Their unique separation of the core server from the storage engine makes it possible to run with very strict control, or with ultra fast disk access, whichever is more appropriate for the situation.

- **PHP**

Self-referentially short for PHP: Hypertext Preprocessor, an open source, server-side, HTML embedded scripting language used to create dynamic Web pages. In an HTML document, PHP script (similar syntax to that of Perl or C) is enclosed within special PHP tags. Because PHP is embedded within tags, the author can jump between HTML and PHP (similar to ASP and Cold Fusion) instead of having to rely on heavy amounts of code to output HTML. And, because PHP is executed on the server, the client cannot view the PHP code. PHP can perform

any task that any CGI program can do, but its strength lies in its compatibility with many types of databases. Also, PHP can talk across networks using IMAP, SNMP, NNTP, POP3, or HTTP. PHP was created sometime in 1994 by Rasmus Lerdorf. During mid 1997, PHP development entered the hands of other contributors. Two of them, Zeev Suraski and Andi Gutmans, rewrote the parser from scratch to create PHP version 3 (PHP3).

CHAPTER 4

RESULTS AND DISCUSSION

4.1 RESULTS – PRINT SCREEN

4.1.1 “Login” Page



Figure 2: Print Screen - “Login” Page

This is the print screen of the “Login” page. At this page user who would like to access this system will have to log in first. Without having access, the user cannot proceed to go to the search portal.

4.1.2 “About Us” Page

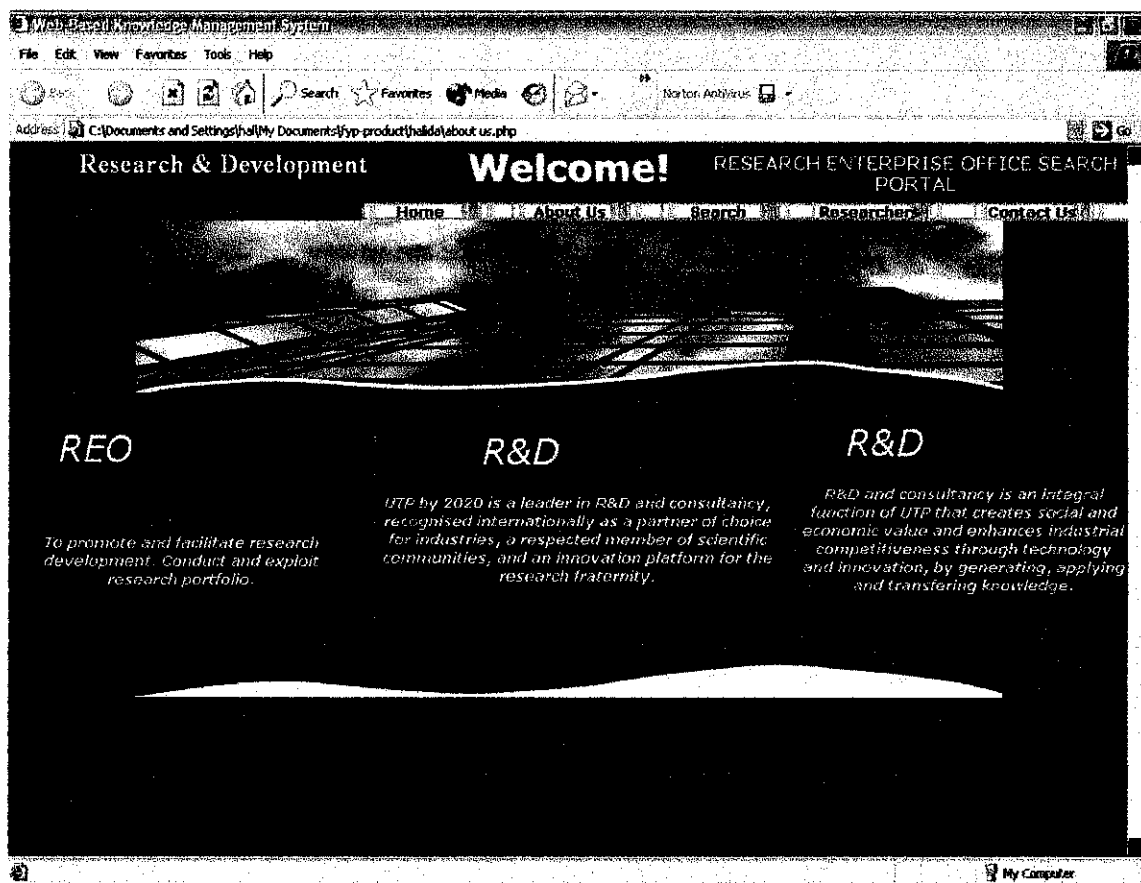


Figure 3: Print Screen - “About Us” Page

This page will describe about Research Enterprise Office (REO) background. Information provided is the REO objective, R&D Vision, and R&D Mission.

4.1.3 "Search" Page



Figure 4: Print Screen - "Search" Page

This page is the most important page. User will be using this page to search for research papers they want. Firstly, the user needs to choose a search type. There are three search types that are by title, by author and by category. Then they need to enter their search term.

4.1.4 “Researchers” Page



Figure 5: Print Screen - “Researchers” Page

This page will be displaying all the researchers name, contact number, E-mail, and programme.

4.2 DISCUSSION – USE CASE FOR RESEARCH ENTERPRISE OFFICE SEARCH PORTAL

4.2.1 User

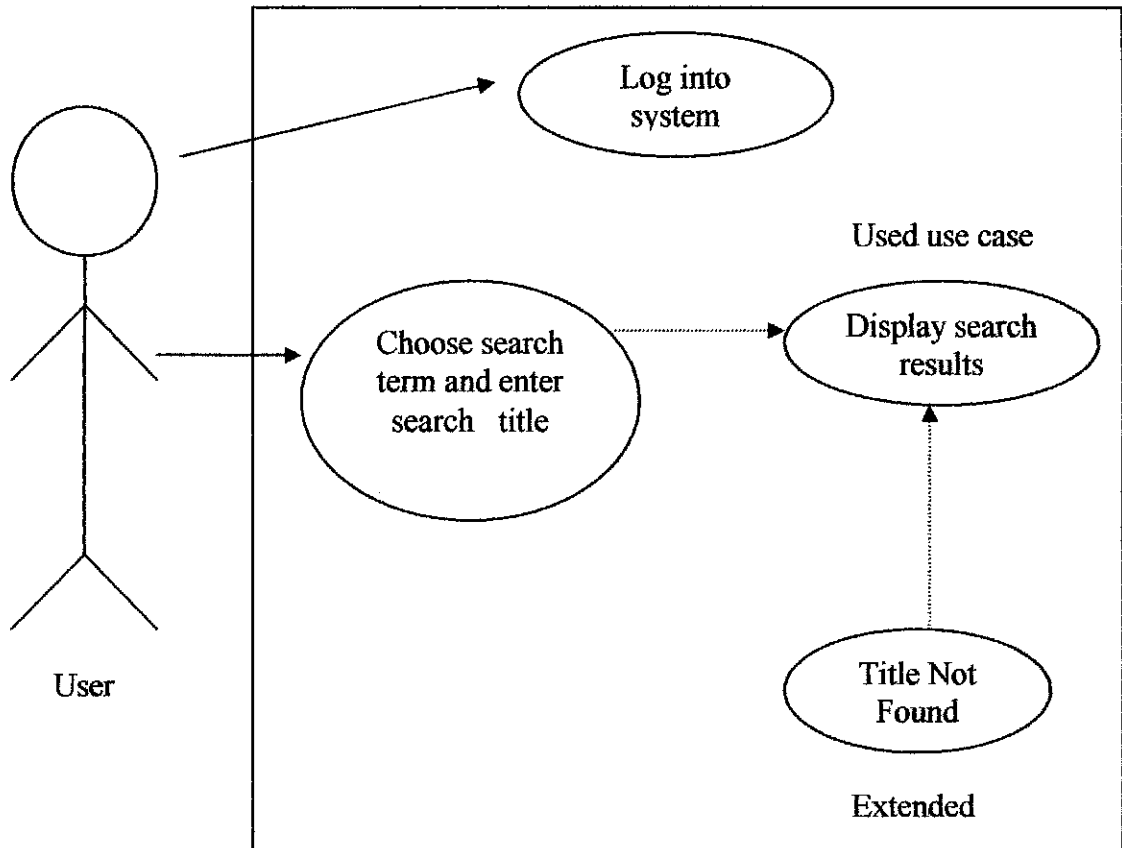


Figure 6: Use Case Diagram – User

Actors: (i) User

Description: User which are staff or lecturers of UTP (Universiti Teknologi Petronas) log in into the search portal. He or she choose a search term and enters the title that he or she would like to search. Then the system will display the search results. If the system found no match for the title entered, the system display an error message and allow the user to enter a different title to search again.

4.2.2 System

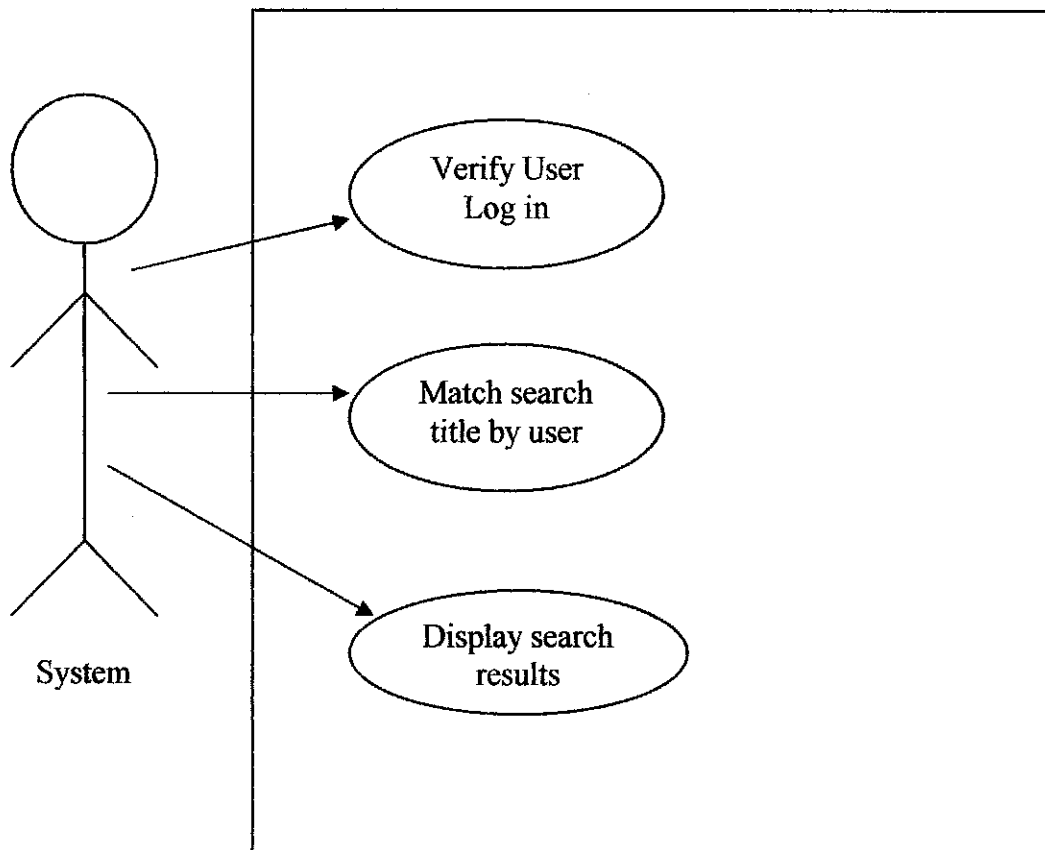


Figure 7: Use Case Diagram – System

Actor: (ii) System

Description: The system verify the user log in. The system match the search title entered by user with the database. The system display the search results.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The capability of web technology to integrate information across and between enterprise-wide systems will become more commanding and easier to use. Furthermore, future Internet standards for integration on an application level and a network level will ensure that the technology is reliable and extensible for all systems. Then, it is only a matter of the public to plan and use the integrated information of these systems in a manner that they can honestly be referred to as the REO Search Portal.

Different engines have different strong points; use the engine and feature that best fits the purpose of this system. One thing is obvious; the engine with the most pages in the database IS NOT the best. Not surprisingly, we can get the most out of the engine by using our head to select search words, knowing the search engine to avoid mistakes with spelling and truncation, and using the special tools available such as specifiers for titles, images, links, etc. Believe that very soon the Web will evolve standards, such as standard categories, ways of automatically classifying information into these categories, and the search tools to take advantage of them that will really improve searching. This system used search by title, author, and category. By searching only in titles, author, and category, one can eliminate pages with only brief mentions of a concept, and only retrieve pages that really focus on the concept [1].

5.2 RECOMENDATION

Here are some suggestion for future work for expansion and continuation:

- ✓ Allow the researchers to upload their research papers themselves
 - As for now, the administrator is the one who is responsible to upload the research papers. If the researcher has the authorization to upload their research papers themselves, it would be much easier and more effective.

- ✓ Allow user to gives comments on the research papers
 - After the user have read or gone through the research papers, they will be given the chance to give their comments about the research. They can also give some recommendation to other user who would want to read the research.

- ✓ Allow access to user that is outside from UTP
 - As for now, the search portal is only for the use of staffs and lectures in UTP only. In the future, it would be possible if the search portal can be used by users outside UTP. This can be used for example by the staffs in KLCC and all other staffs in Petronas.

- ✓ Add features such as bulletin boards or discussion boards
 - To make the search portal more interactive, we could add some new features such as the bulletin boards or discussion boards. By adding this feature, user can interact with other user in the search portal. In the bulletin board, latest research papers that have been uploaded can also be announced.

REFERENCES

- [1] Bruce Grossan (bruce@singu.lbl.gov) (February 21, 1997) Search Engines—What they Are, How They Work, and Practical Suggestions for Getting the Most Out of Them
- [2] Cooper, J.W. and Byrd, R. J. (1998). OBIWAN—a visual interface for prompted query refinement. In *Digital Documents*, volume 2, pages 277–285. 31st Hawaii International Conference on System Sciences, IEEE.
- [3] UC Berkeley - Teaching Library Internet Workshops. *What Makes a Search Engine Good?*
- [4] UC Berkeley - Teaching Library Internet Workshops. *The BEST Search Engines*
- [5] Borghoff, U.M. and Pareschi, R., (1999), "Information Technology for Knowledge Management", *Journal of Universal Computer Science*, Vol. 3, Issue 8, p.:14-16. Also at http://www.iicm.edu/jucs_3_8/information_technology_for_knowledge/paper.html
- [6] Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5):331–340.
- [7] Hunter, E. J. and Blakewell, K. G. B. (1991). *Cataloguing*. Library Association Publishing, Ltd., London, England, 3 edition.
- [8] Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM*, 29(7):648–656.
- [9] Salton, G., Allan, J., and Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2):97–108.
- [10] Blair, D. C. and Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3):289–299.
- [11] Blair, D. C. and Maron, M. E. (1990). Full-text information retrieval: Further analysis and clarification. *Information Processing and Management*, 26(3):437–447.
- [12] Bodoff, D. and Kambil, A. (1997b). Pre-coordination plus post-coordination: The case for partial coordination. Technical Report IS-97-14, New York University, Information Systems Department, New York, NY.

- [13] Borgman, C. L. (1984). *The User's Mental Model of an Information Retrieval System: Effects on Performance*. PhD thesis, Stanford University.
- [14] Denning, P. J. (1982). Electronic junk. *Communications of the ACM*, 25(3):163–165.
- [15] Oard, D. W. (1997). The state of the art in text filtering. *User Modeling and User-Adapted Interaction*. in press.
- [16] Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development*, 2(4):314–319.
- [17] Mooney, R. J., Bennett, P. N., and Roy, L. (1998). Book recommending using text categorization with extracted information. In *Proceedings, AAAI-98 Workshop on Recommender Systems*, Madison, WI. AAAI.
- [18] Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., and Cohen, M. D. (1987). Intelligent information sharing systems. *Communications of the ACM*, 30(5):390–402.
- [19] Goldberg, D., Nicholas, D., Oki, B. M., and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.
- [20] Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.
- [21] Lawrence, S. and Giles, C. L. (1998a). Context and page analysis for improved web search. *IEEE Internet Computing*, 2(4):38–46.
- [22] Shneiderman, B., Byrd, D., and Croft, W. B. (1998). Sorting out searching: A user-interface framework for text searches. *Communications of the ACM*, 41(4):95–98.
- [23] Mukherjea, S. and Hara, Y. (1997). Focus + context views of World Wide Web nodes. In *Proceedings, Eighth ACM Hypertext Conference*, pages 167–176, Southampton, UK.
- [24] Chen, H., Nunamaker, J. J., Orwig, R., and Titkova, O. (1998). Information visualization for collaborative computing. *IEEE Computer*, 31(8):75–82.
- [25] Phelps, T. A. and Wilensky, R. (1996). Toward active, extensible, networked documents: Multivalent architecture and applications. In *Proceedings of the 1st ACM International Conference on Digital Libraries*, pages 100–108, Bethesda, MD. ACM.

- [26] Golovchinsky, G. (1997). What the query told the link: the integration of hypertext and information retrieval. In *Proc. 8th ACM Conference on Hypertext*, New York, NY. ACM, ACM.
- [27] Rada, R., Cargill, C., and Klensin, J. (1998). Consensus and theWeb. *Communications of the ACM*, 41(7):17–22.
- [28] O’Leary, D. E. (1998). Enterprise knowledge management. *IEEE Computer*, 31(3):54–61.
- [29] Baldwin, C. Y. and Clark, K. B. (1997). Managing in an age of modularity. *Harvard Business Review*, pages 84–93.
- [30] Ginsburg, M. and Duliba, K. (1997). Enterprise-level groupware choices: Evaluating lotus notes and intranet-based solutions. *CSCW: The Journal of Collaborative Computing*, 6:201–225.
- [31] Watters, C., Conley, M., and Alexander, C. (1998a). The digital agora: Using technology for learning in the social sciences. *Communications of the ACM*, 41(1):50–57.
- [32] Nunamaker, J. F. J., Chen, M., and Purdin, T. D. M. (1991). System development in information systems research. *Journal of Management Information Systems*, 7(3):89–106.
- [33] UC Berkeley - Teaching Library Internet Workshops. *The BEST Search Engines*
- [34] www.navigators.com
- [35] Memorial University Library-February 19, 2004