

Dedication

It is my pleasure to dedicate this
research to my beloved

Parents,

Husband,

Brothers,

Sisters,

And friends.....

Acknowledgement

I wish to record my thanks and gratitude to my supervisor Prof. Dr. Izzeldin Mohammed Osman, I am grateful to him for entrusting me with the task of preparing this research .I hope his trust has not been misplaced.

I also sincerely appreciate the valuable help and obtained from Ustaz\ Hisham Abdallah Monsoor, Yassir Sedig, and also the assistance of all Computer Science and Information Technology College staff who supplied me with the research resources and gave me their advice.

I acknowledge also with grateful the care of my friends. Finally I wish to record my thanks to my family and husband for their encouragement.

Abstract

How to find needed information from the web is a critical issue in the Internet. Fortunately, search engines are useful tools to retrieve information from the Internet.

Although, Internet users speak different languages most of the resources are written and published in English. This research investigates multilingual search and ways to develop multilingual search engine models. The purpose of multilingual search is to help people finding useful contents stored in multiple languages. This research examines the case of English and Arabic languages.

The research addresses search engines and information retrieval topics such as ranking, crawling, indexing resources, query execution and results relevancy.

The model presented for multilingual search engine is illustrated by a practical example of the on-line site of Sudan University of Science and Technology journal.

مستخلص البحث

مما لا شك فيه ان شبكة الانترنت تعتبر من اكبر المصادر المعلوماتية حيث تزخر بشتى ضروب المعلومات وتصنيفاتها ونجد ان محركات البحث ادوات مفيدة جدا لمستخدمى الانترنت من طلاب وباحثين. وبالرغم من الكم الهائل من المعلومات وتنوع الجنسيات واللغات الا ان معظم معلومات الشبكة العنكبوتية تكتب وتنشر باللغة الانجليزية.

من هنا تنبع اهمية وجود محركات بحث تتمتع بامكانيه البحث متعدد اللغات حيث توفر للمستخدم نتائج البيانات المخزونة بلغات متعددة . هذا البحث يتناول البحث بلغات متعددة ويحاول تطوير نموذج لمحرك بحث متعدد اللغات.والهدف الاساسى من البحث متعدد اللغات مساعدة مستخدمى محرك البحث الحصول على المعلومة المطلوبة بأكثر من لغة ويركز على البحث باللغتين العربية والانجليزية.

ولتحقيق ما سبق ذكره كان لا بد من الاهتمام بالموضوعات المتعلقة بمحركات البحث واسترجاع المعلومات مثل نظام الرتبة للمواقع، الزحف على الانترنت، فهرسة مصادر المعلومات، تنفيذ الاستعلامات وملاءمة المعلومات المسترجعة للإستعلامات .

ونعرض مثلاً عملياً لمحرك بحث باللغة العربية والإنجليزية يقوم ببحث موقع مجلة العلوم والتقانة التابع لجامعة السودان للعلوم والتكنولوجيا.

Table Of Contents

Dedication.....	i
Acknowledgments.....	ii
English Abstract.....	iii
Arabic Abstract.....	iv
Table of Contents.....	v
Table of Figures.....	x
Table of Tables.....	xii
Introduction.....	xiv
Chapter 1: Literature Review	
1.1 Introduction.....	1
1.2 Client And Servers.....	1
1.3 Domain Name Server.....	1
1.4 Web Servers.....	2
1.5 Search Engines.....	3
1.6 Meta Tags.....	3
1.6.1 Meta Description Tag.....	4
1.6.2 Meta Keywords Tag.....	4
1.6.3 Meta Robots Tag.....	5
1.7 User Requirements of the Search Engine	5
1.8 Search Engine Specifications	6
1.9 Needs To Multilingual Search.....	7
1.10 Research Problems.....	7
1.11 Research Overview.....	8
1.12 Literature Review Of Search Engine.....	8
1.13 Research Methodology.....	9
1.14 Risks.....	9
1.15 Related Works.....	9
Chapter 2: Internet Search Engines	
2.1 Search Engines Definitions.....	12
2.2 Search Engine History.....	13
2.3 How Search Engines Work.....	16
2.4 Search Engine Components.....	18
2.4.1 Search Engine crawling.....	18
2.4.2 Search Engine Indexers.....	19
2.4.3 Search Engine Query Engine.....	20
2.5 Search Engine Query.....	20
2.5.1 Basic Techniques for Creating a Search Engine Query.....	20

2.5.2	Basic Structure Of A Query.....	20
2.5.3	Logical Operators Search.....	21
2.5.4	Keywords and Key Phrase Search.....	21
2.6	Optimizing Query Execution.....	22
2.6.1	Caching Query Results.....	22
2.6.1.1	The Benefits Of Caching Query Results.....	22
2.6.1.2	Replacements Algorithms.....	22
2.6.1.3	The Limits of Caching.....	23
2.6.1.4	Static Caching.....	23
2.6.1.5	Dynamic Caching.....	23
2.6.2	Improve Query Throughput.....	24
2.6.3	Improve Websites Ranking.....	24
2.6.3.1	How Do Search Engines Rank Pages?.....	24
2.6.3.1	Get A Good Rank.....	24
2.7	Information Retrieval.....	25
2.8	Search Engine Features.....	26
2.9	Search Engine Characteristics.....	26
2.10	Search Engines Challenges.....	27
2.11	Search Engines Problems.....	28
2.12	Search Engine Standard Architecture.....	29
2.12.1	Standard Layers.....	29
2.12.2	Standard Querying.....	30
2.13	Search Engine Optimization.....	30
2.13.1	Optimizing Frames.....	31
2.13.2	Optimizing JavaScript.....	32
2.13.3	Optimizing Dynamic Pages.....	33
2.14	Cloaking	34
2.14.1	How Does Cloaking Work?.....	34
2.14.2	Cloaking Advantages.....	35
2.14.3	Cloaking Disadvantages.....	35
2.15	Search Engine Spamming.....	36
2.15.1	Classification of Spamming Techniques.....	36
2.15.2	Avoid Search Engine Spamming.....	39
Chapter 3:Google Search Engine		
3.1	Hypertextual Search Engine Architecture.....	40
3.2	Design Goals.....	40
3.3	System Features.....	41
3.3.1	Page Rank.....	42
3.3.2	Intuitive Justification.....	42
3.4	Anchor Text.....	42
3.5	Differences Between the Web and Well Controlled Collections	43
3.6	System Anatomy.....	43
3.6.1	Google Architecture.....	44
3.6.2	Data Structure.....	45
3.6.2.1	BigFiles.....	45
3.6.2.2	Repository.....	45

3.6.2.3 Documents Index.....	46
3.6.2.4 Lexicon.....	46
3.6.2.5 Hit Lists.....	46
3.6.2.6 Forward Index.....	47
3.6.2.7 Inverted Index.....	47
3.7 Crawling the Web.....	48
3.7.1 Indexing the Web.....	49
3.8 Searching.....	49
3.8.1 Ranking System.....	50
3.8.2 Feed Back.....	51
3.9 Performance.....	51
3.9.1 Storage Requirement.....	51
3.9.2 System Performance.....	52
3.9.3 Search Performance.....	53
Chapter 4: Intranet Search Engines	
4.1 Intranet Definition.....	54
4.2 Why Companies Need Intranets?.....	54
4.3 Intranets Search Engine Challenges.....	54
4.4 Intranet Search Engines.....	55
4.5 How Intranet Search Engine Works?.....	55
4.5.1 Gathering Document for Indexing.....	55
4.5.2 Indexing Contents.....	56
4.6 Search through Massive Amounts of Data.....	58
4.7 Intranet Meta Data.....	58
4.8 Important Features for Intranet Search.....	59
4.8.1 Search Functionality.....	59
4.8.2 Intranet Search Scaling.....	59
4.8.3 Search Administration.....	60
4.8.4 Search Log Analysis.....	60
4.9 How to make a Good Intranet Search Engine?.....	60
4.10 Problem of Intranet Search.....	61
4.10.1 Intranet VS. Internet.....	61
4.10.1.1 Intranets vs. Internet: Axioms.....	62
4.10.1.2 Intranets vs. Internet: Structural Differences.....	63
Chapter 5: Multilingual Search Engine Model	
5.1 Introduction.....	64
5.2 Modeling?.....	64
5.3 Unified Modeling Language.....	64
5.4 Enterprise Architecture.....	64
5.5 Logical Model.....	65
5.5.1 Class Diagram.....	65
5.6 Use Case Model.....	75
5.7 Dynamic Model.....	88
5.7.1 Activity Diagram.....	88
5.7.2 Sequence Diagram.....	97

Chapter 6: Case Study	
Case Study.	106
Chapter 7: Conclusion and Recommendations	
7.1 Conclusion.....	110
7.2 Recommendations For Future Research.....	111
References.....	112
Appendix: Case Study program.....	A

Table of Figures

Table of Tables

Table 1:	Google Statistics and Storage Requirements.....	52
Table 2:	Browser Connections.....	66
Table 3:	Browser Methods.....	66
Table 4:	Crawler Connections.....	67
Table 5:	Crawler Attributes.....	67
Table 6:	Crawler Methods.....	68
Table 7:	Dictionary Connections.....	68
Table 8:	Dictionary Attributes.....	68
Table 9:	Dictionary Methods.....	69
Table 10:	IndexServer Connections.....	69
Table 11:	IndexServer Methods.....	70
Table 12:	Page Connections.....	70
Table 13:	Page Attributes.....	70
Table 14:	QueryEngine Connections.....	70
Table 15:	QueryEngine Attributes.....	71
Table 16:	QueryEngine Methods.....	71
Table 17:	Results Connections.....	72
Table 18:	Results Attributes.....	72

Table 19:	Results Methods	72
Table 20:	SearchEngine Connections.....	72
Table 21:	Server Connections.....	73
Table 22:	Server Attributes.....	73
Table 23:	URLServer Connections.....	73
Table 24:	URLServer Methods.....	74
Table 25:	WebServer Connections.....	74
Table 26:	WebServer Methods.....	74
Table 27:	Browser Connections.....	76
Table 28:	Crawler Connections.....	77
Table 29:	Crawler Connections.....	77
Table 30:	IndexServer Connections.....	77
Table 31:	IndexServer Connections.....	77
Table 32:	QueryEngine Connections.....	78
Table 33:	SearchEngine Connections.....	78
Table 34:	Server Connections.....	78
Table 35:	Server Connections.....	78
Table 36:	URLServer Connections.....	79
Table 37:	WebServer Connections.....	79
Table 38:	Crawl Connections.....	79
Table 39:	DisplayResults Connections.....	80
Table 40:	DownloadingPage Connections.....	80
Table 41:	Indexing Connections.....	81
Table 42:	ParsingQuery Connections.....	82
Table 43:	ProcessingQuery Connections.....	82
Table 44:	RankingCheck Connections.....	83
Table 45:	RedownloadingPage Connections.....	84
Table 46:	RelevancyCheck Connections.....	84
Table 47:	Requeting Connections.....	85
Table 48:	Searching Connections.....	85
Table 49:	SearchingIndex Connections.....	86
Table 50:	TranslatingKeyword Connections.....	87

Introduction

A great number of search engines can be used to retrieve relevant information on the internet. This research focuses on multilingual information retrieval to help searchers find useful contents and documents.

The researcher developed model based on the characteristics of search engines and the architecture of the client-server system.

Chapter1: the research starts by giving a general overview of internet, search engines and then turns to the research problem, research overview, research goals and objectives, research methodology, research risks, and the related works.

Chapter2: In chapter two the research gives more details in search engines. It starts by defining them; gives their historical background, how they work and their components. It then goes to an important issue, which is considered as the corner stone of the search process, it is query processing and how it can be optimized particularly by caching of queries results using replacement algorithms. The step after processing a query is information retrieval and how to display the results to the user. Then it turns to describe the features, characteristics and the architectures of the search engines as well as the problems and challenges they face. Lastly this chapter makes a little focus on cloaking and spamming techniques.

Chapter3: this chapter talks about Google as a hypertextual search engine. It describes design goals, system features, anchor text, Google architecture and data structure, crawling the web, searching and ranking system, and finally performance including storage requirement, system and search performance.

Chapter4: chapter four describes the intranet search engines start with definition, how they work, collecting information to index it, search an intranet, intranet metadata, important features for intranet search and the process of developing a successful intranet. Then the chapter turns to a comparison

between internet search engine and intranet search engine from axioms and structural points of views.

Chapter5: this chapter describes the proposed multilingual search engine model. The language used to build this model is UML (Unified Modeling Language). The main designed models are logical model (class diagram), use case model, and dynamic model (sequence and activity diagrams).

Chapter6: this chapter describes the multilingual search engine case study.

Chapter7: this chapter contains conclusion and recommendations for future researches and studies.

Lastly a list of references and an appendix containing the case study program are included.