Search Engines With Particular Reference To Multilingual Search

Supervised By:
Prof. Izzeldin Mohammed Osman

Prepared By:
Wegdan A. Elhameed

June 2005
Abstract

- Search Engines are a useful tool to retrieve information from internet.
- Although, internet users speak different languages most of resources are written and published in English.
- This research investigates multilingual search, to help people finding useful contents stored in multiple languages and examines the case of English and Arabic languages.
- The model presented for multilingual search engine is illustrated by a practical example of Sudan University of Science and Technology journal site.
Chapter 1: Literature Review

- Introduction
- DNS (Domain Name Server)
- Web Servers
- Meta Tags
- User Requirements Of The Search Engine
- Search Engine Specification
- Needs To Multilingual Search
- Research Problem
- Research Overview
- Research Goals And Objectives
- Related Works
Chapter 2: Internet Search Engines

- Definitions
- History
- How they work?
- Components
- Query Processing (Query Engine)
- Optimizing Query Execution and Information Retrieval
  - Features and Characteristics
  - Search Engine Architecture
    - Challenges
    - Problems
  - Spamming and Cloaking
Chapter 3: Google Search Engine

- Design Goals
- System Features
- Anchor Text
- Differences Between the Web and Well Controlled Collections
- System Anatomy
- Crawling the Web
- Searching
- Performance
Chapter 4: Intranet Search Engines

- Definition
- Challenges
- How it Work?
- Gathering Documents and Indexing
- Intranet Metadata
- Intranet Search
- Search Problems
- How to make a Good Intranet Search Engine?
- Intranets vs. Internet : Axioms
- Intranets vs. e Internet: Structural Differences
Chapter 5: Multilingual Search Engine Model

- Logical Model
  - Class Diagram
- Use Case Model
  - Use Cases Views
- Dynamic Model
  - Activity Diagram
  - Sequence Diagram
Logical Model
Use Case Model
Crawl Use Case
Dynamic Model
Build a Query Activity

Start

Put search terms

Click search button

Send query

End
Check Relevancy Activity

Start

Send query

parse query

Search

Retrieve results

Compute IR relevancy

Check results ranks

Cache query results

End
Crawls a Page Activity

1. Start
2. Request URLs
3. Build URLs queue
4. Take a URL
   - [Processed]
   - [Not processed]
   - [Go to next URL]
5. Choose crawl algorithm
6. Record URL information
7. Record page information
8. Add links to the queue
9. Compute page rank
10. Record page is processed
11. End
Downloads a Page Activity

1. Start
   - Crawl
   - Send plain request
   - Get a plain response
   - Get header and meta tags

   - [Not exist page]
     - Record URL information
     - Record page information
     - Page is processed
       - End
   - [Exist page (check updated)]
     - [Not updated]
       - Check_update
       - Update URL information
       - Update page information
       - [Updated page]
       - Page is processed
       - End
Processes Query Activity

Start

Submit query

Parse query
Give every keywords wordID
Translate keyword[s]

Search (arabic,english)

[Results found]

Check results

[No results found]

Check results rank
Check results relavancy
Sort results

Send results

Feed back

End
Translate Keyword[s] Activity
Display Results Activity

Start

Receive results

Check browser version
Check results encoding and language
Convert results to HTML

take top k documents

Display results

End
sd Client Build a query

User

Write (search engine URL)

Open (URL)

Put (keywords)

Click (Search Button)

Browser

SearchEngine

send(request)
Check Relevancy Sequence

User

SearchEngine

QueryEngine

IndexServer

send(query)

pass(query)

parse(query)

{by keywords}

search(query)

compute(IRrelevancy)

check(PageRank)

send(results)

send(results)

cache(QueryResults)

view(RelevantResults)
Downloads a Page

Sequence Diagram:

- Crawler
- IndexServer
- WebServer
- URLServer

1. Crawler sends a plain request to IndexServer.
2. IndexServer responds with plain text.
3. WebServer sends headers and meta information to IndexServer.
4. If the page exists, IndexServer sends a message indicating the page exists (sendMsg(PageExist)).
5. If the page exists, IndexServer sends the last page information to WebServer.
6. If the page exists, IndexServer stores the page information.
7. If the page does not exist, IndexServer sends a message indicating the page does not exist (sendMsg(Not exist)).
8. If the page does not exist, IndexServer stores the URL information.
9. URLServer downloads the page.
Indexes a Page Sequence

Process Page:
1. Give the pages a standard format.
2. Breaks the pages in units that will be stored in the index.
3. Normalize the character encoding.
4. Remove the unnecessary HTML tags, extra white spaces, etc.
5. Figure out keywords from Meta tags.
6. Map characters that have the same meaning together.
7. Recognize the document language.

Crawler

URLServer

response(PageInformation)

process(PageInformation)

index (page URL)

index (PageInformation)

IndexServer

sort(index)

WebServer
Ranks a Page Sequence

- Crawler
  - crawl(page)
  - analysis(page link)
  - check(page popularity)
  - check(Link co-citation)
  - compute(page rank)

- WebServer

- IndexServer
  - store(page rank)
Translate Keyword Sequence

- SearchEngine
- QueryEngine
- Dictionary
- IndexServer

Activities:
- pass (query)
- parse (query)
- send (keywords)
- translate (keywords)
Display Results Sequence

User -> Browser: Click (search button)

Browser -> SearchEngine: send(request)

SearchEngine: process(query)

[if found]: send(results)

Browser: CheckBrowser(version, encoding)

[as HTML]: Display [Results]

[not found]: Results

send(FeedBack)
Chapter 6: Case Study

Site description

As a case study of multilingual search, Sudan University of Science and Technology (SUST) journal site has taken to implement this search facilities.

The main site contains two separate Arabic and English sites to be searched separately, this achieved through these search keywords:

- Keyword[s].
- Article title.
- Author articles.
- Categories.
The New Site:
The user inputs his keyword (English keyword or Arabic keyword) in the search box. Then the query is processed in the steps below:

- Connect to the database.
- Search through keywords table for user keywords.
- Find the corresponding meaning to the keyword.
- Send the Arabic keyword to Arabic table to search for it.
- Send the English keyword to English table to search for it.
- Return the results as recordset.
- Display both Arabic results and English results.
- Divide them as five results per page.
Longitudinal Relaxion Time From Maxwell

In this work maxwell equations utilized to derive expression for magentic dipole moment.

Navigate to Page 1.
Chapter 7: Conclusion and Recommendations

- Although internet users differ in their nationalities and languages, search sites do not offer a multilingual search and most of them support one language search (English).
- The research is concerned with multilingual search process to enable users find information they need in both English and Arabic.
- The important points found on search engines area are crawling to the internet sites, indexes the resources found, give every resource a rank, and check information retrieval relevancy.
- In this research we have surveyed the components of search engines and have given an example of a multilingual search. We have implemented the search engine using Active Server Pages, SQL Server database, and ODBC.
Recommendations for Future Researches

Search engines are a non-limited field to study. Here are recommended areas to the future researches and studies:

❖ The process of query execution and optimization.
❖ Results caching to improve the search quality and response time.
❖ More through investigation of crawling algorithms and problems of invisible sites, dead links, and updated links.
❖ Ranking algorithms and information retrieval relevancy.
❖ Browsers encoding and versions and their influence on results displayed to users (frame pages, javascript, and dynamic pages, graphics).
❖ The implementation of a bilingual Arabic/English search engines.
رب اشرح لي صدري
والحمد لله رب العالمين