Development of the Search Engine for Product Information in Myanmar

Nann Hla Khaing, Myint Thu Zar Tun
nannhlakhaing@gmail.com,myintthuzartun@gmail.com

Computer University, Maubin

Abstract

Search engines play a central role in knowledge acquisition and knowledge transfer in today's information society. They are large extent responsible for making information on the Internet easily accessible. Web-based information systems can be critical to the success of the system and can enhance its ability to deliver information which contains. It has become more and more difficult for web users to find relevant information in the large information space they are now forced to navigate. In this paper, a specified web search engine that can assist ordinary Internet users to search for products related information is presented. The main objective of this system is to provide high quality, relevant and efficient search results to the user.

Keywords: knowledge acquisition, knowledge transfer, information society, web search engine, web users

1. Introduction

Most people find what they are looking for on the World Wide Web by using search engines like Yahoo!, Alta Vista, or Google. Searching for information with search engines is the most popular Internet activity. There are three types of web search engines. They are as follows:

(i)Crawler-based search engine: A crawler is a program that visits web sites and reads their pages and extracts some information such as title, URL, meta data, links and much more in order to create entries for a search engine database.

(ii)Directory-based search engine: A web directory web sites by subject, related topics and sub-topics that reflect the things the users want to find.

(iii)Meta search engine: In a meta-search engine, the end user submits keywords in its search box and it sends user's requests to several other search engines and/or databases and aggregates the results into a single list or displays them according to their source. [1]

In this paper, the major components of a crawler based search engine and how these components work will be presented. Crawlers are typically programmed to visit sites by following the links to other pages on the site until all pages have been read. The typical web users assume that when they search, the search engine actually goes out onto the Web to look around. Also search engine provides an interface to the users to specify criteria about an item of interest and have the engine to find the matching items. It is used for an information retrieval system designed to find information stored on a computer such as on the World Wide Web or on personal computer. The search engine allows ones to ask for content meeting specific criteria and retrieve a list of items that match those criteria. Search engine also helps to minimize the time required to find information.

Most of the users think that they can easily search the product information in Myanmar online. But they become frustrated when they do not know which site is good and which is not, and how to find this site. To solve these difficulties, they use web search engines for finding their product information. However, they often do not know how to improve their search strategies when they have retrieved too much information or information that is not directly relevant to their queries.

To handle this problem, web masters develop special search engines for finding information. In this system, a specialized web search engine that can assist ordinary Internet users to search for product related information is proposed. It helps users to search for product information that significantly improves its usability and the quality of search results. It also allows the users to search various kind of facts related to their products such as improving for the knowledge and using for encyclopedia and business region.

Most of the work in search engines has focused on system for storing and viewing documents, methods for processing queries and determining the relevant results, and user interfaces for querying, viewing and refining results. At the same time, search engines expose large searchable databases to large numbers of people. These search engines have to assume that searchers would not possess the knowledge to express complex queries. But this product search engine is used the relevant results to be simple and easy-to-use search interface. As a result, this search engine contains product information interfaces in Myanmar and

querying processes to make this system more useable to the users.

This paper is intended to search the product information using the web search engine. There are divided into five Sections to present this system. Section 1 introduces why about to apply the product information in Myanmar. We describe the related work in Section 2. In Section 3, we discuss the technologies of crawler search engine. Section 4 illustrates the architecture of the system with use of database in query for product information in Myanmar. We implement the system interface in Section 5 and the conclusion is follows in Section 6.

2. Related Work

Before there were search engines, there was a complete list of all web servers. The list was hosted on the CERN web server. As more and more web servers went online, the central list could not keep up. [3]

The very first tool used for searching on the Internet was Archie. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however Archie did not index the contents of these sites.

The rise of Gopher led to two new search programs. Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems.

The first Web search engine was Wandex, a web crawler developed in 1993. Another very early search engine, Aliweb, also appeared in 1993. JumpStation which was released in early 1994 used a crawler to find web pages for searching, but search was limited to the title of web pages only. On the first full text crawler-based search engines was WebCrawler, which came out in 1994. It let users search for any word in any webpage, which became the standard for all major search engines. It was also the first one to be widely known by the public. [5]

Soon after, many search engines are appeared. These included Excite, Infoseek and AltaVista. Yahoo! Was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search. [4]

Microsoft first launched MSN Search in the fall of 1998.Around 2000, the Google search engine rose to prominence. As of late 2007, Google was by far the most popular Web search engine worldwide.

In these days, a number of country-specific search engine companies have become prominent; for example, Baidu is the most popular search engine in the People's Republic of China and guruji.com in India. On the other hand, different types of domain-specific search engine become very popular.

3. Technologies of Crawler Search Engine

The term "crawler" refers to the technology that captures information from a web page and follows links to find more information. Crawlers retrieve resources from web servers in exactly the same way that web browsers do: sending a message to the web server containing a request for a specified resource. The most basic function of any crawler is to retrieve web pages using HTTP. Usually a crawler only needs to retrieve the HTML pages on a site, and it can skip the time consuming process of downloading the images. [3]

When a user searches the web using a search engine, he is always searching a partial copy of the real web page. It may be helpful at this phase to review the operation of a search engine from a technical and organizational perspective. Two main parts of the search engine are (i)Crawling and indexing the Web and (ii)Searching the Web

3.1 Crawling and Indexing the Web

Search engine indexing parses and stores data to facilitate fast and accurate information retrieval. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document which would require considerable time and computing power. Most of the data found on websites is stored in HTML and understanding the structure of these HTML document is very important. [2]

Thus, the indexer examines the web page's HTML code and locates the tags within it that facilitate storing indexes in the search engine's database. The indexing can be text indexing or meta indexing. In text indexing, the search engine will harvest all the text of the page, the process it to extract a list of relevant content words from each page. But processing the whole page can be time consuming. Meta indexing works with meta data placed in the different documents and web pages by the web master. The web masters decide which keywords are relevant for the web page and insert these in meta tags which are, in turn, indexed by the search engine. The advantage of meta indexing

is that processing is fast and efficient. For that reason, meta indexing is used in the development of this system.

3.2 Searching the Web

A web search query is a query that a user enters into web search engine to satisfy his information needs. Keyword search is currently one of the most important operations in search engines and numerous other applications. The main features of keyword search are

- Providing natural language search
- Easy to use and flexible

In this system, keyword search technique is used by reducing some of its drawbacks to retrieve high quality, relevant and efficient search results to the end users. Thus, this system can precede the query steps to match the indexes during the querying processing. There are four steps in query processing.

(i)Tokenizing: As soon as a user inputs a query, the search engine must tokenize the query stream. Usually a token is defined as an alphanumeric string that occurs between white space and/or punctuation.

(ii)Parsing: Users may employ special operators in their query, including Boolean, adjacency, or proximity operators; the system needs to parse the query first into query terms and operators.

(iii)Removing stop words: Some search engines remove stop words from the query.

(iv)Creating the Query: Each particular search engine creates a query representation depends on how the system does its matching.

4. Architecture of the System

While product information is often said to be the most sought information on the web, product-related search tools on the web are so rare. Thus, building a specific search engine that will be able to access the product related information will highly support the Internet users who need to find this information. For that reason, this system, which can search the information of the product by using crawler search engine, has been implemented.

In this system, there are two parts: (i) the crawling and indexing task can be done only by the administers who need to click on the crawl button. Then the crawler program will be run which will download the web pages and extracts the indexes to be stored in the search engine's database.(ii) For a user who uses this search engine can search the information by entering just a product name, a phrase, or a sentence that include product name in the home page's query box. If he wants to search more complex query, he can use the advanced search page and if the user is the one who is not so familiar with the product name, he can use the unknown search page.

This system can be used by all users who want to find the information of the products. By using this system, the user can find the required product information very easily and quickly. The architecture design of the system is presented in Figure 1 and each major component is discussed. In this system, there are altogether three major components which are crawler, indexer and query processor.

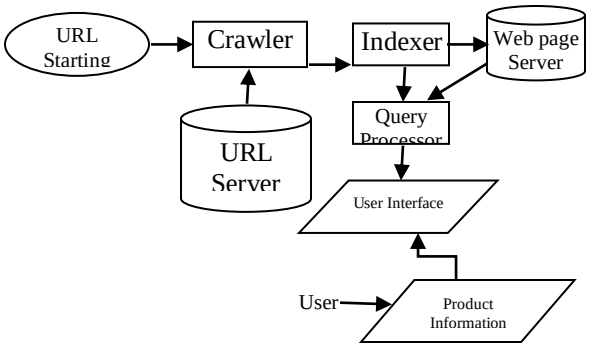

Figure 1. Architecture Design of the System

The crawler has to read the URLs and download the web page only if the page is product web page. It reads the first URL from the web.config file and the rest of the URLs are read from the queue which contains URLs of unvisited web pages. The indexer examines the web page's HTML code and locates the tags within it that facilitate storing indexed in the search engine's database. It extracts the information like URL, title and meta information. This extracted information is stored in the database as indexes when the crawler has finished download the URLs in the queue.

Query processor handles the queries that the user keyed in. It tokenizes the query segments, removes stop words and does its matching against the index database. In database, index table, index value table, type table relating with product table (e.g., rice). In index table, indexNo, sitepath and sitedescription are used. The index value table consists of value and url. In type table, there are typeID and type name and typeID relates to product (rice) table including productID(riceID), name and location as shown in Figure 2.

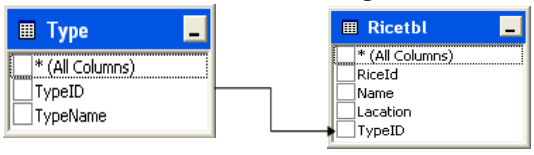

Figure 2. Type Table Links to Rice Table

5. System Interface Implementation

In this system, the crawler has to read the starting URL from the web.config file. The rest of the URLs are read from the queue which contains URLs of visited pages. The crawler checks that the requested URL to be downloaded is the html document. Then the crawler retrieves that html resource from the web server by sending a request message using HTTPWebRequest object. Then the requested URL is responded by sending a web page using HTTPWebResponse object.

The download web page is sent to the indexer for storing indexes in the search engine's database. The indexer examines the web page's HTML code and locates the tags within it that facilitate storing indexes in the search engine's database. The indexer extracts the appropriate information from a web page, such as URL, title and so on. The extracted results, which are called indexes, are stored temporarily in memory. After completing the crawling and indexing process, which means all of the URLs in the queue are downloaded, the indexes stored in the memory are sent to the search engine's database as shown in Figure 3

.

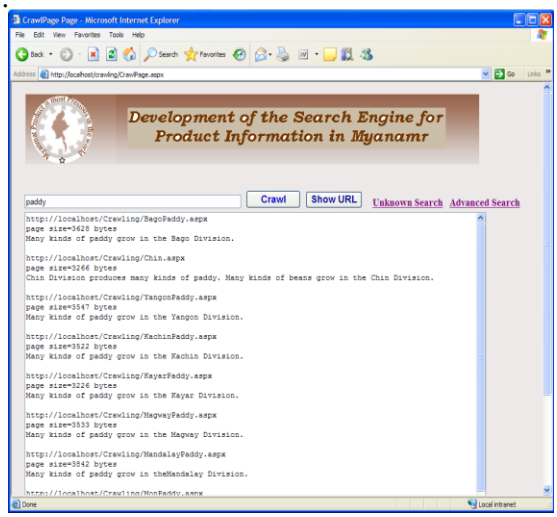

Figure 3. Indexes which will be stored in the Database

When a user accesses this website, this system let the user to search for product information in two ways. The first way is normal search which means he can search the product name in the text box of this page. The search query may be a single word or a phrase, even a sentence can be entered to search for product information. Moreover, if the user types in the keyword "rice", "bean", and "paddy"., the websites concerning "rice", the websites concerning "bean" and the websites concerning "paddy" will be shown in Figure 4.

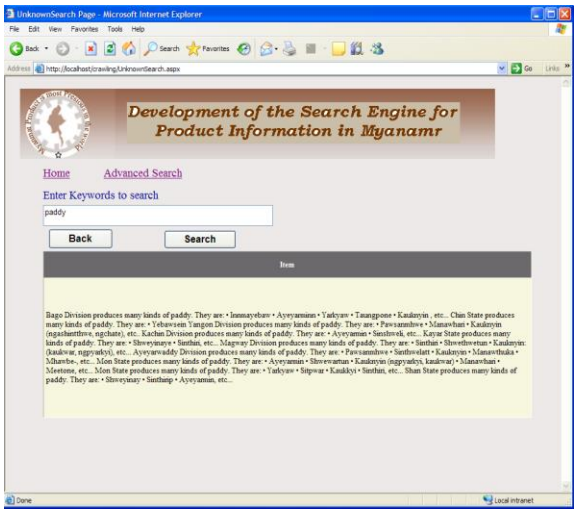

Figure 4. Normal Search of the System

In this page, the user can also see the Advanced Search button under the bottom of the page. If the user wants to search more specific product information, he can click on Advance Search. For example, if the user types the item name as "paddy" and location as"Bago", etc.,. The list of product concerns item name and location as shown in Figure 5.

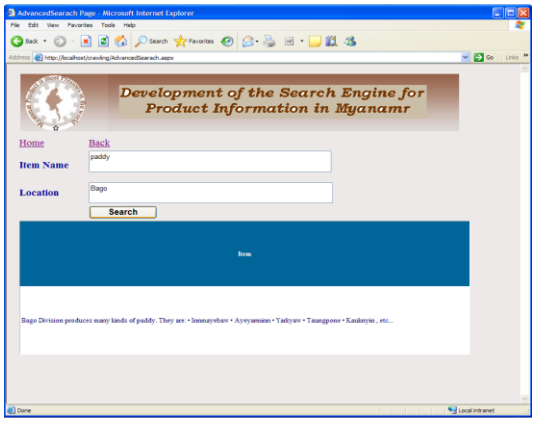

Figure5. Advanced Search of the System

6. Conclusion

Search engines are one of the most exciting fields of today's innovative technologies. Web search engines can help people to find the resources, so that they are looking for, by more clearly indentifying the searcher's intent behind the query. Additionally, the use of the search engine is likely to yield many more result because the storage on the web outdoes man times that of any others. People expect and require information they can trust, delivered in a format that is understandable

and usable. This system is designed to be scalable search engine which searches the product information. The primary goal of this search engine is to provide high quality, relevant and efficient search results to the end users.

References

[1]Baberwal D., Choi B., "Speeding Up Keyword Search for Search Engines", The 3rd IASTED International Conference on Communications, Internet and Information Technology, p. 255-260, 2004.

[2] Brin S., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proceedings of the Seventh International Conference on World Wide Web, p. 107-117, 1998

[3] Cave F., "Crawler Identification and Discovery", An ACAP Technology Working Group Discussion Paper, 17 June 2007.

[4] CERN http
http://www.w3.org/Daemon/Status.html

[5] Crawler – SearchSOA.com Definition http://searchsoa.teachtarget.com/sDefinition/0, sid26_gci211854,00.html

[6] Enterprise Search –Webopedia

http://www.webopedia.com/TEAM/e/enterprise_search.html