Implementation of Breadth-First Search Algorithm for Crawling the Websites

Poe Ei Phyu, Soe Yu Maw
University of Computer Studies, Mawlamyine
poeeiphyu77@gmail.com, soeyumaw@gmail.com

Abstract

Around the world, organizations of all sizes have pledged their commitment to build and maintain comprehensive portals to serve their customers employees, trading partners, and constituents. Corporations in all industries, governments at all levels, and nonprofit organizations for all interests have jumped on the portal bandwagon In this paper will create a web-based search portal. To search the websites, administrator must crawl the websites. Crawler and Search engines play a very important role in information retrieval. Rapid increase in web size and dynamic nature of web contents is a challenge to existing crawlers and search engines. A focused crawler aims at selectively seeking out pages that are relevant to a pre-defined set of topics. Numerous crawling algorithms are available to retrieve relevant data on the basis of different metrics, like similarity based on query, forward link count, back link count, page rank, location metrics. In this paper we attempt to present Breadth-First Search web crawler.

Keywords

Portal, Crawler, Content Analysis Theory, Breadth-First Search Algorithm

1. Introduction

Web portals are entry points for information presentation and exchange over the internet. Hence, they require an efficient support communication and information sharing. Current Web technologies employed to build up these portals present serious limitations regarding information search, access, extraction, interpretation and processing. Aside from the search engine standard, web portals offer other services such as e-mail, news, stock prices, infotainment and various other features. Portals intend authenticated portal users; it must have authenticated security which is user name and password. Government portal makes interaction between government web sites and citizens. This system is used as a gateway to content that acts as search engine with precompiled list of related links by topic. In my present system implement with two portions, web crawling and web search portal. Administrator must crawl web sites. Web crawler crawls web sites word by word and save related links in database. When user searches word, web search portal search links include this word. We proposed government search portal which is based on web crawling using Breadth-First Search algorithm.

2. Background Theory

2.1 Content Analysis Theory

Content analysis is a research tool used to determine the presence of certain words or concepts within texts or sets of texts. Researchers quantify and analyze the presence, meanings and relationships of such words and concepts, then make inferences about the messages within the texts, the writer(s), the audience, and even the culture and time of which these are a part. Texts can be defined broadly as books, book chapters, essays, interviews, discussions, newspaper headlines and articles, historical documents, speeches, conversations, advertising, theater, informal conversation, or really any occurrence of communicative language In this guide, we discuss two general categories of content analysis: conceptual analysis and relational analysis. Conceptual analysis can be thought of as establishing the existence and frequency of concepts – most often represented by words of phrases – in a text. For instance, say you have a hunch that your favorite poet often writes about hunger. With conceptual analysis you can determine how many times words such as “hunger,” “hungry,” “famished,” or “starving” appear in a volume of poems. In contrast, relational analysis goes one step further by examining the relationships among concepts in a text. Returning to the “hunger” example, with relational analysis, you could identify what other words or phrases “hunger” or “famished” appear next to and then determine what different meanings emerge as a result of these groupings

2.2 Web-Searching Portals

Portal provides a single point of access

to a variety of content and core service. And ideally offer a single sign on point. Portals give information from diverse sources in unified way. Portals often include calendars an to-do lists, discussion groups, announcements and reports, searches, email and address books and access to news, weather, maps and shopping, as well as bookmarks. The first portals on the web were gateways to content that combined search engines with precompiled lists of related links by topic. The search engine must quickly and accurately provide results to the user. It must handle a variety of the formats and storage locations. Finally, the search result should be stored in the most useful documents visible to the user. There are many types of web searching portals such as Yahoo, Microsoft Network (MSN), American Online (AOL) and FITSTGOV.gov

2.3 Government Portal

There are 4 types of e-Government services that are possible; Government to Citizens (G2C), Government to Businesses (G2B), Government to Employees (G2E) and Government to Government (G2G). In this position paper, we will focus on the Secure Portal possibilities for G2C. Increasingly various governmental agencies are providing services to its citizens via dedicated web sites. As a citizen requiring services from the local, state or federal government, we rely on these separate web sites not only for information but also to avail services like filing tax returns, renewal of license plates/vehicle stickers, communicate with the county for property tax issues etc. Providing a single secure portal that acts as a window to the various services is beneficial not only to the government but also to its citizens. In this //position paper, the need for a secure portal is described, current sample usage of online services by US governmental agencies, an example of an ideal web experience for a citizen to use E-Government services, the challenges and possible solutions to creating secure E-Government Portals for global citizens.

2.4 Semantic Web

The Semantic Web (SW) can be defined as an extension of the current web. Here the information is presented in a well-defined manner, better enabling computers and people to work in cooperation. Adding formal semantics to the Web will aid in everything from resource discovery to the automation of all sorts of tasks The Semantic Web initiative calls for the development of a set of application layer protocols that will be able to provide semantic descriptions of web resources. These protocols will be inter-operable and will enable logic-based decision-making in web applications.Semantic Web foundation involves publishing in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML). HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts.

2.5 Crawlers

Web crawlers are programs that exploit the graph structure of the Web to move from page to page. Web crawlers have been to retrieve Web pages and add them or their representations to a local repository. Such a repository may then serve particular application needs such as those of a Web search engine. At the other end of the spectrum, one can have personal crawlers that scan for pages of interest to a particular user, in order to build a fast access cache. . In its simplest form a crawler starts from a *seed* page and then uses the external links within it to attend to other pages. This process is repeated until relevant result is found.

2.6 Semantic Web Crawler

It differs from a traditional web crawler in two regards: the format of the source material it is traversing, and the means of specifying links between information resources. Whereas a traditional crawler operates on HTML documents, linked using HTML anchors, a Semantic Web crawler operates on RDF metadata with linking implemented using the rdfs.

2.7 Focused Crawling

A general purpose Web crawler gathers as many pages as it can from a particular set of URL's. Where as on the other hand, a focused crawler is designed to only gather documents on a specific topic, thus reducing the amount of network traffic and download. The goal of the focused crawler is to selectively seek out pages that are relevant to a pre-defined set of topics. The topics are specified not using keywords, but using exemplary documents. Rather than collecting and indexing all accessible web documents to be able to answer all possible ad-hoc queries, a focused crawler analyzes its crawl boundary to find, the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the web. The performance of a focused crawler depends mostly on the richness of links in

the specific topic being searched, and focused crawling usually relies on a general web search engine for providing starting points.

2.8 Breadth –First Search

Breadth-First algorithm is one of the simplest strategies for crawling. It uses the frontier as a FIFO queue, crawling links in the order in which they are encountered. Figure 1 illustrates the Breadth-First algorithm. The main advantage of breadth first search algorithm is that it can act as a domain specific Crawler. Breadth-first search (BFS) is a general technique for traversing a graph. A BFS traversal of a graph G, visit all the vertices and edges of G. Determines whether G is connected and computes the connected components of G. And then compute a spanning forest of G. BFS on a graph with n vertices and m edges takes o (n + m) time. BFS can be further extended to solve other graph problems. Find and report a path with the minimum number of edges between two given vertices.

2.9 Breadth-First Search Algorithm

```
CurrentPage = RootPage;
NextLevel = ture;
while (NextLevel())
Begin
NextLevelPages = null;
NextNodes.Next = null;
NextNodes.Current = RootPage;
while (NextLevel)
begin
        NextLevel⇓false;
        while (NextNodes.Current!=null)
        begin
NextNodes.current.Process( );
for(inti=0;i<NextNodes.Current.Neighbors.Count
; i++)
begin
        if (not (already process))
        begin
        current=NextNodes.current.Neighbours(i
);
        current.Process();
        if (Current.subNodes!=null)
        begin
        tmpNextNodes = current.subNodes;
        NextLevel = true;
        end;
    end;
end;
NextNodes.current = NextNodes.Next;
end;
if (NextLeve == true)
begin
        NextNodes = mpNextNodes;
        end;
    end;
end;
```

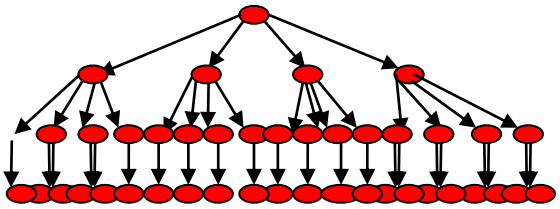

Figure 1: Bredth First Search Crawling

3. Related Work

3.1 Crawling Algorithms

With huge size of web, even large search engines cover only a portion of the publicly-available Internet; a study by Lawrence and Giles in year 2000 showed that no search engine indexes more than 16% of the Web. As a crawler always downloads just a fraction of the Web pages, it is highly desirable that the downloaded fraction contains the most relevant pages and not just a random sample.

3.2 Best-First

The basic idea is that given a frontier of URLs, the best URL according to some estimation criterion or selection metrics is selected for crawling, using the frontier as a priority queue after analysis through went analysis algorithm. In this implementation, the URL selection process is guided by getting from the local database the lexical similarity between the topic's keywords and the source page for the URL. This value has been computed by the crawler. Thus the similarity between a page p and the topic keywords is used to estimate the relevance of all the outgoing links of p. Figure 2 illustrates the Best-First algorithm.

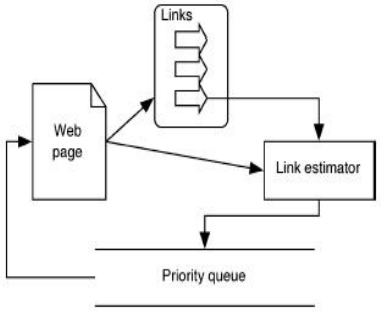

Figure 2: Best First Crawling

3.3 Page Rank

Page Rank was proposed by Brin and Page as a possible model of user surfing behavior.

The Page Rank of a page represents the probability that a random surfer (one who follows links randomly from page to page) will be on that page at any give time. A page's score depends recursively upon the scores of the pages that point to it. Source pages distribute their Page Rank across all of their out links. Formally:

$$PR(p) = (1-\gamma) + \gamma \sum_{\{d \in in(p)\}} \frac{PR(d)}{|out(d)|}$$

where p is the page being scored, in(p) is the set of pages pointing to p, out(d) is the set of links out of d, and the constant < 1 is a damping factor (usually set to 0.85) that represents the probability that the random surfer requests another random page.

4. System Design and Implementation

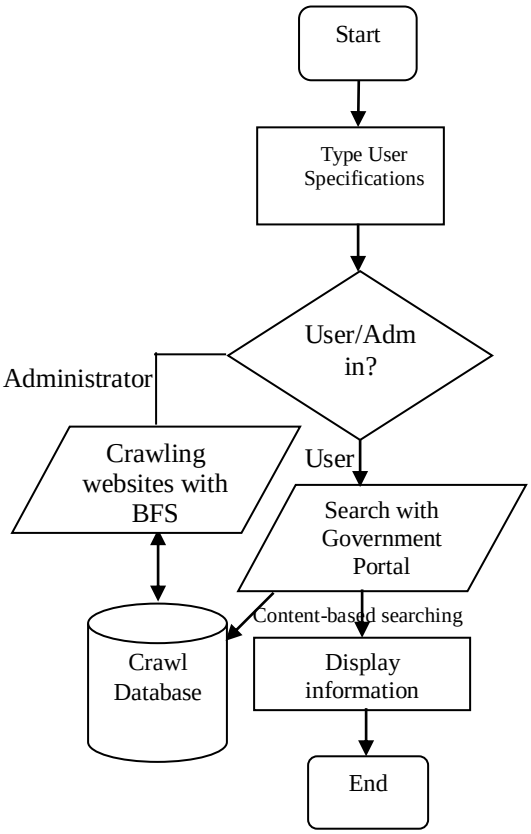

Figure 3: System Design

4.1 Implementation step

In this system implements when user enters this portal, he/she will fill user name and password. And then, he/she can login to government portal. If not, he/she can register to enter this portal. In this system have user and administrator. Administrator must crawl the websites with Breadth-First Search algorithm and save related links in database to search websites by many users. When users enter this portal, he/she want to search severable information about government. Therefore, user don't know websites link. They will type some word to want search information. At that time, portal searches the related links in crawl database and display user.

4.2 Example case study

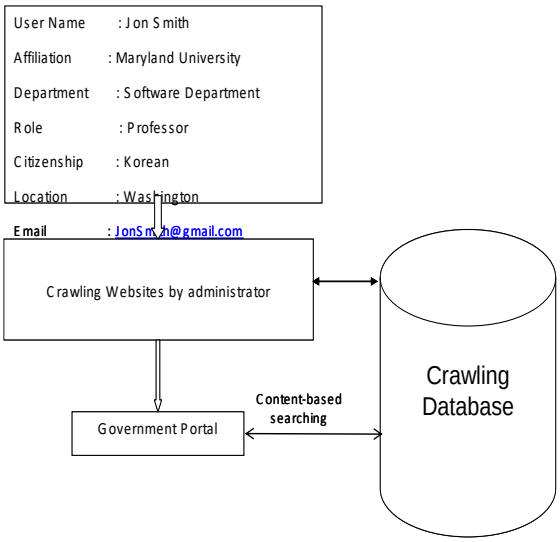

Figure 4: Example case study of the system

4.3 User Interface Design

There are two main program modules. Web crawler is run in the notification are of the system. Searching program is the web based program.

4.4 Web Crawler interface (Windows Program)

Web crawler is running in the notification are of the operating system. It's context menu is shown in figure 6.

Figure 5: Web Crawler Program in notification area

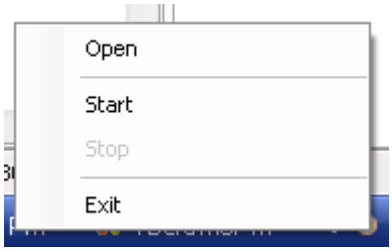

Figure 6: Context Menu of the Crawler

The main form the crawler is shown in figure 7. User can configure the address of the initial link in the Address textbox. To start crawling, user must select "Start" menu item of "Process" main menu.

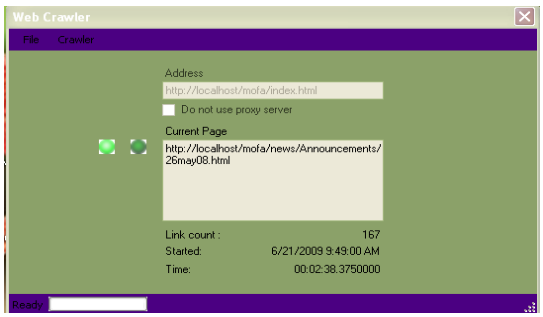

Figure 7: "Start" Menu Item of the Crawler

In web searching portal, there are web crawling and web searching module are included. The starting page of web site is login page. User can login from that page as shown in figure 8.

If the user is new user, he can register as a new user by select "Register" link from login page, see figure 8. New user needs to fill up the registration form as shown in figure 9.

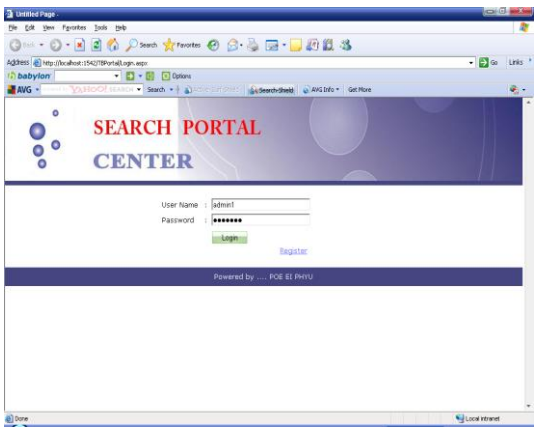

Figure 8: Search Portal Centre Login Page

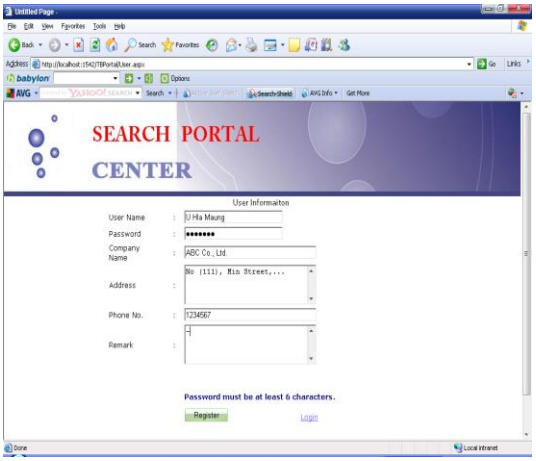

Figure 9: Registration Page of the Web Search Portal

4. 5 Database Design

Database is hosted in the Microsoft SQL Server. For simplicity, there is only one database to store the related link crawling data. The steps for creating the database can be follow.

Figure 10 shows the data table that stores Link address, related word and related sentence. Searched word, sentence and related links are stored in this table.

When users have been searched, portal showed result pages in Fig 11. These pages include text and related link. They can go, they want to see web pages by clicking.

TWord	TLink	TText
words	http://localhost/...	and perform a ti...
made	http://localhost/...	Despite the...
rights	http://localhost/...	Copyright ...
its	http://localhost/...	Hence, the ...
well	http://localhost/...	...
increase	http://localhost/...	Similarly...
durable	http://localhost/...	In this cont...

Figure 10: Database Design

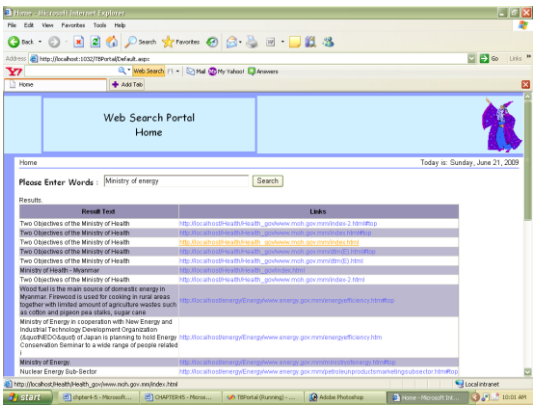

Figure 11: Search Portal Result Page

5. Overview of the System

Web portals are entry points for information presentation and exchange over the internet. Hence, they require an efficient support communication and information sharing. Current Web technologies employed to build up these portals present serious limitations regarding information search, access, extraction, interpretation and processing. Aside from the search engine standard, web portals offer other services such as e-mail, news, stock prices, infotainment and various other features. Portals intend authenticated portal users, it must have authenticated security which is user name and password. Government portal makes interaction between government web sites and citizens. This system is used as a gateway to content that acts as search engine with precompiled list of related links by topic. The dynamic vast nature of web increases the importance of web crawler to get relevant information. A crawler is a program that retrieves and stores pages from the Web, commonly for a Web search engine. This system implements with two portions, web crawling and web searching. Administrator must crawl web sites to give the result for searching the government information by the user. Web crawler crawls web sites word by word and saves related links in database. When users search words, web search portal searches links include word in database by using content analysis theory. This system applies the Breadth-First Search algorithm for crawling the web sites of government portal.

.

6. Conclusions

The dynamic vast nature of web increases the importance of web crawler to get relevant information. In this system we attempted to provide an idea of focused crawling implement with government portal based on BFS crawling. This approach is particularly useful in open environment in which interacts government to government, government to business, government to citizens. In this system, creating a user profile is more secure and then understanding and acting according to that individual or entity's assigned role and authority. So, it is possible to allow the right people access to the right information on the portal.

References

[1] B.B. Hariri, A. Hassan, H. Sayyadi and J. Mohsen "A Method for Focused Crawling Using Combination of Link Structure and Content Similarity", Web Intelligence Research Laboratory, Computer Engineering Department, Sharif University Of Technology, Tehran, Iran

[2] D. Leigh, "A Semantic Web Crawler", leigh@ldodds.com, February 2006.

[3] D. Oberle, "Semantic Community Web Portals Personalization"

[4] D.Riz, and D. Schaffer, J.J.Townsend, "Building Portals, Intranets and Corporate Web Sites Using Microsoft Server"

[5] http://www.gslis.utexas.edu/~palmquis/courses/content.html

[6] http://www.webont.org/owl/1.1/xml_syntax.html

[7] Win Lai Hnin "Developing Web Portal for Tourism Information Service"

[8] Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki.