Categorized Crawler

Abhilasha Rawat¹, Megha Bansal²
¹Student, Computer Science & Engineering, Maharshi Dayanand University, Rohtak, Haryana, INDIA
²Assistant Professor, Computer Science & Engineering, Maharshi Dayanand University, Rohtak, Haryana, INDIA

ABSTRACT
A focused crawler traverses the web selecting out relevant pages to a predefined topic and neglecting those out of concern. While surfing the internet it is difficult to deal with irrelevant pages and to predict which links lead to quality pages. In this paper, a technique of effective focused crawling is implemented to improve the quality of web navigation. Here we develop a system which will concern on categories when we search something on net then it gives multiple choices but suppose we want only relevant choice then we have to go through multiple. So in that case we divide the choices in categories to make the crawler efficient.

Keyword - Crawler, Search Engine, Indexer, Page Repository, Query Engine.

I. INTRODUCTION

A positive point about the Internet and its most visible component, the World Wide Web, is that there are hundreds of millions of pages available, waiting to present information on an amazing variety of topics. But the negative point about the Internet is that there are hundreds of millions of pages available, most of them titled according to the whim of their author, almost all of them sitting on servers with cryptic names. When you need to know about a particular subject, how do you know which pages to read? you visit an Web Search Engine.

The Internet is a global data communications system. It is a hardware and software infrastructure that provides connectivity between computers. In contrast, the Web is one of the services communicated via the Internet. It is a collection of interconnected documents and other resources, linked by hyperlinks (a hyperlink is a reference or navigation element in a document to another section of the same document or to another document that may be on or part of a different domain) and URLs (Uniform Resource Locator (URL) is a Uniform Resource Identifier (URI) which also specifies where the identified resource is available and the protocol for retrieving it.) The plentiful content of the World Wide Web is useful to millions. Information seekers use a search engine such as Google, Yahoo to begin their Web activity.

On the Internet, a search engine is a coordinated set of programs which searches an index and returns matches to a specified keyword. Search Engine is situated on the computer system connected to Internet. Other alternatives for searching information on the internet are MetaSearch engines and directories.

There are differences in the ways various search engines work, but they all perform three basic tasks:
1. They search the Internet or select pieces of the Internet based on important words.
2. They keep an index of the words they find, and where they find them.
3. They allow users to look for words or combinations of words found in that index.

Search Engine to provide best services regularly index millions of web pages involving a comparable number of distinct terms by employing special software known as Web Crawlers or Spiders to retrieve information on web to prepare up catalog for ready reference. The most important measure for a search engine is the search performance, quality of the results and ability to crawl and index the web efficiently. The primary goal is to provide high quality search results over a rapidly growing World Wide Web.

In a Search a Engine, user sends the query. If related query is in indexed pages then page related top query returned to user.

If required pages not in indexed pages then query is sent to crawler module. Crawler module sends the query to crawlers. Crawler search pages related to query and send those pages to page repository. And also sends the related link back to crawler module. Crawler module when gets these link, it sorts them according to their relevancy and sends them back to crawler. Crawler processes all the links...
till the list is empty and adds the results to page repository. The Indexer indexes the stored data in a particular format. Collection analysis module stores the pages on the basis of their utility. Ranking module ranks the retrieved pages according to their relevance. Retrieved results are sent back to user. These three types of search engines gather their listings in radically different ways.

1) Crawler-Based Search Engines

Crawler-based search engines, such as Google, create their listings automatically. They "crawl" or "spider" the web, then people search through what they have found. If you change your web pages, crawler-based search engines eventually find these changes, and that can affect how you are listed. Page titles, body copy and other elements all play a role.

A human-powered directory, such as the Open Directory, depends on humans for its listings. You submit a short description to the directory for your entire site, or editors write one for sites they review. A search looks for matches only in the descriptions submitted.

3) "Hybrid Search Engines" Or Mixed Results

In the web's early days, it used to be that a search engine either presented crawler-based results or human-powered listings. Today, it extremely common for both types of results to be presented. Usually, a hybrid search engine will favor one type of listings over another. For example, MSN Search is more likely to present human-powered listings from LookSmart. However, it does also present crawler-based results (as provided by Inktomi), especially for more obscure queries.

A Crawler is a program that retrieves Web pages, commonly for use by a search engine or a Web cache. Roughly, a crawler starts off with the URL for an initial page P0. It retrieves P0, extracts any URLs in it, and adds them to a queue of URLs to be scanned. Then the crawler gets URLs from the queue (in some order), and repeats the process. Every page that is scanned is given to a client that saves the pages, creates an index for the pages, or summarizes or analyzes the content of the pages.

Following is the process by which Web crawlers work:

Download the Web page.
Parse through the downloaded page and retrieve all the links.
For each link retrieved, repeat the process.

II. OBJECTIVE

The plentiful content of the World-Wide Web is useful to millions. Information seekers use a search engine such as Google, Yahoo etc to begin their Web activity. Our aim is to make a search tool that is cost-effective, efficient, fast and user friendly. In response to a query, it should retrieve the most relevant information which has been stored into the database. It should also be portable, so that it can easily be deployed at any platform without any cost and inconvenience. Our goal is to make a Web Search Engine that will retrieve the best matched Web Pages in the shortest possible time.

III. RELATED WORK

A. Architecture & Working of a Web Crawler

Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers, which are all beyond the control of the system. Following is the process by which Web crawlers work:

- Download the Web page.
- Parse through the downloaded page and retrieve all the links.
- For each link retrieved, repeat the process.

In its simplest form, a crawler starts from a seed page and then uses the external links within it to attend to other pages. The process repeats with the new pages offering more external links to follow, until a sufficient number of pages are identified or some higher level objective is reached. Behind this simple description lies a host of issues related to network connections, spider traps, cannibalizing URLs and parsing HTML pages.

The architectural components of a web crawler and their module linkage are shown in figure 1.3

- The URL frontier, containing URLs yet to be fetched in the current crawl.
- A DNS resolution module that determines the web server from which to fetch the page specified by a URL.
• A fetch module that uses the http protocol to retrieve

[Image: Simplified Architecture of a Web Crawler]

- A parsing module that extracts text and set of links from a fetched web page.
- A duplicate elimination module that determines whether an extracted link is already in the URL frontier or has recently been fetched.

IV. PROPOSED WORK

we develop a system which will concern on categories when we search something on net then it gives multiple choices but suppose we want only relevant choice then we have to go through multiple. So in that case we divide the choices in categories to make the crawler efficient. The Search Tool developed by our team, Crawls the Web Pages recursively and stores the relevant data in the database. This data includes Title, Meta Keywords, Meta Title, Meta Description, Body etc of the Webpage. When a query is submitted to the Search Engine, it searches its own database in response to it. In Universal Search, it then lists all the URLs that match the query. In Field Specific Search, a category is also chosen. The Web Pages are then retrieved as per the field chosen. For instance, the user has entered “CAT” as his query. If the user proceeds with Universal Search, all the pages that match the query will be retrieved. But if he chooses a field say Education, only those pages related to CAT as Common Admission Test will be retrieved. Our project also has a provision for its users, to choose how many records per page he wishes to see.

V. CONCLUSION

The categorised crawling is used to divide the more choices into different categories. Through the categorised crawling we can find easily the relevant choice. This is very easy way to find that choice which we want to find.

ACKNOWLEDGEMENT

No paper is created entirely by an individual, many people have helped to create this thesis and each of their contribution has been valuable. I wish to express my deep sense of gratitude to my gratitude supervisors Ms. Megha Bansal, Assistant Professor, Computer Sc. & Engineering Department. I owe their overwhelming debt for helping me to plan and discuss the contents of this work. Their ideas, stimulating comments, interpretations and suggestions increased my cognitive awareness and helped considerably in the fruition of my objective. I remain obliged to them for her help and able guidance through all stages of this paper. Their constant inspiration and encouragement to my efforts shall always be acknowledged. Her observation and comments were highly enlightening.

REFERENCE