Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

"Emerging Technologies in the Computer World", January -2017

Implementing a web crawler in a smart phone mobile application

Abhijeet tawde , Jayesh Patil ,Priya kurandale, Sharique khan

(Department of computer engineering, AISSMS Institute of Information Technology)

Abstract---*Internet users and usage are growing rapidly. These days cause great trouble and effort in the use Side to get the page being searched, which is of concern and Relevant user requirements for the general user approach Search for pages from a large number of available concept hierarchies Use a query to browse the web from an available search engine And receive results based on the search pattern, a few of them The results are related to search, and most are not. Web crawlers play an important role in the search engine Consider the key factors in performance. This paper Including domain engineering concepts and keyword driven Crawling and Dependency Decision Making Mechanism and Using Ontology Concept, to ensure that the best path to improve the crawler performance. This article describes URL-based extraction Keywords, or search criteria. It extracts the URL of the page Include search keywords in their content and consider them the page is just important and does not download the page has nothing to do with the search. It provides high optimality the traditional Web crawler, and can improve the search efficiency is more accurate.*

Keywords--- *Web crawler, keyword, knowledge path, topic specific web crawler, ontology.*

I. INTRODUCTION

The World Wide Web (WWW) having billion web pages and searching documents which is more specific with the user's Requirement is increasingly difficult. The WWW supports dynamic content which is growing increasingly including news, current issues, new technology, financial information, marketing, entertainment, education become widely distributed over a wide area of web. The web crawler mostly downloads only the relevant or specific web pages according to the user requirements rather than downloading all web pages like a traditional search engines. So the basic goal of focused crawler is to select and seek out the web pages that fulfil user's requirement. The link analysis algorithms like page ranking algorithm and other metrics are use to prioritize the URLs based on their ranking and selection policies for downloading most specific web pages. In most social applications available today the data is retained for a limited periodand is normally in free format.. With a supplementary application of world wide search using WebCrawler.There's plenty of opportunity. 5.2 billion people use mobile devices. 30% (1.6billion) use smartphones. There is no existing mobile app consisting of exact functionality that we want the App to provide to the end user. The goal of our paper is to build an application where a user can use a apps data and also if he/she want to search a particular entity they can search it on the web outside its application using aweb crawler implanted in that application .The objective of the proposed system is that a user and search thevarious things in different categories and he/she is also able to search what theother people are talking about that thing in outside world of application.

II. LITERATURE SURVEY

1)SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces

AUTHORS: Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang
In this paper, we propose an effective deep web harvesting framework, namely SmartCrawler, for achieving both wide coverage and high efficiency for a focused crawler. Based on the observation that deep websites usually contain a few searchable forms and most of them are within a depth of three. Our crawler is divided into two stages: site locating and in-site exploring. The site locating stage helps achieve wide coverage of sites for a focusedcrawler, and the in-site exploring stage can efficiently perform searches forweb forms within a site.

2) An Approach to DeepWeb Crawling by Sampling.

AUTHORS: Jianguo Lu, YanWang, Jie Liang, Jessica Chen,

In this paper, we focus on querying textual data sources, i.e., those data sources that contain plain text documents only. This kind of data sources usually provides a simple keywords-based query interface instead of multiple attributes. A naive approach to selecting queries to cover a textual data source is to choose words randomly from a dictionary. In order to reduce the network traffic, queries to be sent need to be selected carefully. The authors used an adaptive method to

select the next query to issue based on the documents already downloaded. A greedy algorithm for set-covering problem is used to select an optimal query based on the documents downloaded so far and the prediction of document frequencies of the queries on the entire corpus.

3) PyBot: An Algorithm for Web Crawling

AUTHORS: Alex Goh Kwang Leng, Ravi Kumar P, Ashutosh Kumar Singh, Rajendra Kumar

PyBot is a Web Crawler developed in Python to crawl the Web using Breadth First Search (BFS). The success of the WorldWideWeb (WWW), which itself built on the open internet, has changed the way how human share and exchange information and ideas.The main purpose of the crawler is to visit pages in the Web and download them in a systematic way for the search engines. A crawler starts with a universal resource locator (URL), explores all the hyperlinks in that page, visit those pages and download the pages. These downloaded pages are indexed and stored for search engines. A search engine will be rated based on its search performance, quality of the results, and its ability to crawl and index the Web efficiently. That is why search engines are doing a lot of research on making a better Web crawler. A crawler can be used for crawling through a whole site on the internet or intranet; it works well for our PyBot.

4)Keyword Focused Web Crawler

AUTHORS:Gunjan H. Agre , Nikita V.Mahajan

In this paper, the keyword focused web crawler has been proposed. The keyword focused web crawler algorithm seeks out the URLs of web pages based on their priority and domain ontology. Also the knowledge path plays very important role in finding relevant web pages. The web crawler Breadth First Search, Depth First Search, Page Ranking Algorithms, Path ascending crawling Algorithm, Online Page Importance Calculation Algorithm, Crawler using Nave Bayes Classifier, Focused Web Crawler, Semantic web Crawler etc. Each technique has its pros and cons. Focused Web Crawler is a technique which uses the similarity major to map relatedness among the downloaded page and unvisited page. This technique ensures that similar pages get downloaded and hence the name Focused web crawler.Semantic web crawlers use lexical database to index web pages. The lexical databases provides with senses which help to predict the relatedness of the web page to the query. The senses can be used in varied ways to investigate the interrelation. This paper proposes a technique which clubs above two technique so as to develop a Semantic Based Focused Web Crawler.

5) Topological Tree Clustering of Social Network Search results

AUTHORS: Richard Freeman

The exploding growth of web content is leading to an information overload, in which the use of web search engines is becoming criticalto finding and retrieving relevant content. Despite the numerous advances in information visualisation, the most popular way of presenting search results still remain ranked lists. In this format, the user generally never looks beyond the first three pages, after which they will rather refine their search query by adding more terms or refining the initial query. On the Web the results returned from web search engines, have been widely studied and Search Engine. Optimisation (redesigning a website to improve its web pages ranking) is still a thriving industry. However on social network websites less investigation has been made due to the complexity of the social ties and groups. This paper deals with methods that organise groups (retrieved by a social network search engine) into a set of virtual folders which are labelled automatically using extracted keywords. A method which clusters group pages dynamically, whilst creating a topology between them in a tree view, is presented in this paper.

III. PROPOSED SYSTEM

In this section, we will discuss the motivations behind our work, design and architecture, architecture description and various problems and challenges faced during the implementation of the proposed approach. The objective of the proposed system is that a user and search the various things in different terms and he/she is also able to search what the other people are talking about that thing in outside world of application. A description of the software with Size of input, bounds on input, input validation, input dependency, i/o state diagram, Major inputs, and outputs aredescribed without regard to implementation detail. The scope identifies what the product is and is not, what it will and won't do, what it will and wont contain.. A database is created to collect and store the input data which will be used by the user to rate among the various parameters as per their liking. The major functions performed are crawling, which will be implemented when the user needs to search for a particular thing beyond the database. The Android smartphone application is programmed to receive the data from the database and provide a proper interface for the user with the backend.

IV. SYSTEM ARCHITECTURE

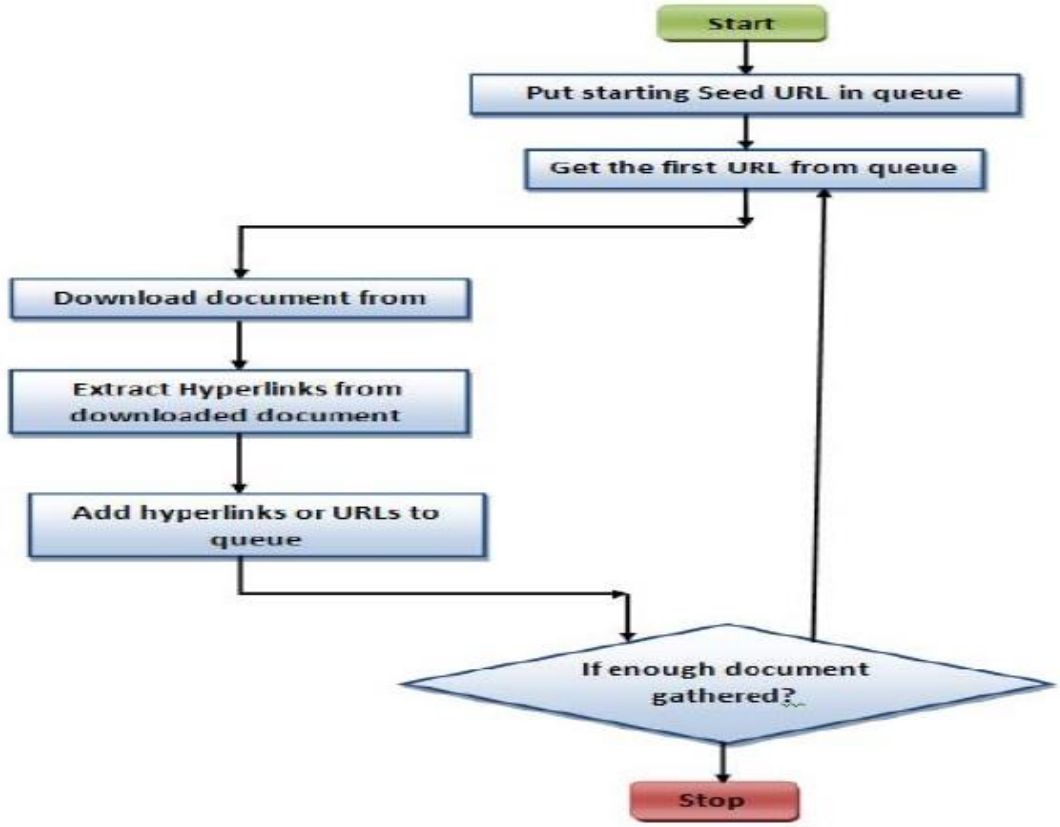

Figure 1. Architecture Diagram of Proposed System

V. MATHEMATICAL MODEL

System Description:
Let S is the Whole System Consists:
A= {U, w, z,Q, R,F,S }
Where,
U is the set of number of users.
U= {u1,u2,.....,un}.
Q is the set of query generated(data to be crawled) from user.
Q={q1,q2,.....,qn}.
R is the set of rating given by users.
R= {r1,r2,..... ,rn}.
Let w is the set of activities.
w={w1,w2,.,wW}
where wi is the ith activity and W is the total number of activities.
Let z is the set of categories.
z ={z1, z2,,zZ}
where zi is the ith categories and Z is the total number of categories.
Let S is the success.
S={Crawled data obtained successfully.}
The data required by the user is crawled successfully. He/she can now successfully rate it as per his or her level of interest.
Let F is the failure.
F={Crawled data obtained failure.}
User did not get appropriate data

VI. CONCLUSION

The proposed concept for web crawling module and an Android smartphone based user interface is designed for people successfully in this study. according to a specific user, he / she use application, according to also include the robot makes the application to spread their search wider areas, giving the user a global search area to take proposals. The main advantage of using keyword focused web crawler asking smart phone is to work smart, effective and does not require important feedback. It reduces the number of retrieved Web pages thus takes less time to crawl as it downloads only the relevant Web pages. The desire is to retrieverelevant web pages and disposal of unsuitable web pages. We have developed ontology based robot with the best way of knowing that retrieves web pages according to relevance decision mechanism. Below measurable advantages have been found of comparing the results with the traditional robots. So a crawler can be implemented in a smart phone application for relevance search.

REFERENCES

[1] Richard Freeman, "Topological Tree Clustering of Social Network Search Results" in Proceedings of the Eight International Conference onIntelligent Data Engineering and Automated Learning (IDEAL'07), Lecture Notes in Computer Science (LNCS 4481), Springer, 16-19 December, 2007, pp. 760-769

[2] SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces,Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin, IEEE Transactions on Services Computing Volume: PP Year: 2015

[3] Susan Dumais and Hao Chen "Hierarchical classification of web content." Proceedings in the 23rd International ACM SIGIR Conference on Research and development in Information Retrieval, Pages 256-263. ACM, 2000.

[4] Susan Gauch, Jason Chaffee and Alexander Pretschner"Ontology Based Personalized Search and Browsing"in UMUAI.

[5] Filippo Menczer, Gautam Pant and Padmini Srinivasan "Topical Web Crawlers: Evaluating Adaptive Algorithms", proceding in the ACM Conference on Research and Development in Information Retrieval , New Orleans, Louisiana, United States, Pages 241 - 249 , 2001

[6] B.Lawrence ,Sergey. "The Anatomy of a Large Scale Hypertextual Web Search Engine Computer Networks and ISDN Systems"1998.

[7] K. Zhang and Y. Guo, X.Q. Cheng, K. Li ,"Crawling Dynamic Web Pages in WWW Forums" ,Department of Computer Engg., volume number 33, in 2007.

[8] A. Agarwal, K.P. Chitrapura, H.S. Koppula, S. Garg, A. Sasturkar, and K.P. Leela, "Learning URL Patterns for Webpage De-Duplication", proce in 3rd ACM Conference Web Search and Data Mining" , page no.381-390, in 2010.

[9] N. Ch. S. N. Iyengar, A. Kannan, and M. Yuvarani " LSCrawler- A Framework for an Enhanced Focused Web Crawler based on Link Semantics Proceeding in ACM, IEEE,WIC and in InternationalConference on Web Intelligence, 2006.

[10] Amritesh Kumar, Debashis Hati, "UDBFC -An effective focused crawling approach based on URL distance calculation", Department of Computer Engineering ,KIIT University, Bhubaneswar, India ,Proceeding in IEEE, 2010.

[11] Ganesh S, Jayaraj M, Aghila G "Ontology Based Web Crawler" Information Technology ; Coding &Computing, volume 2, page 337-341,IEEE, 2004.

