Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 584–590

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Keyword query based focused Web crawler

Manish Kumar*[a], Ankit Bindal[a], Robin Gautam[a], Rajesh Bhatia[a]

PEC University of Technology, Chandigarh, India 160012

Abstract

Finding information on Web is a difficult and challenging task because of the extremely large volume of data. Search engine can be used to facilitate this task, but it is still difficult to cover all the webpages present on Web. This paper proposes a query based crawler where a set of keywords relevant to the topic of interest of the user is used to shoot queries on search interface. These search interfaces are found on webpage of the website corresponding to seed URL. This helps crawler to get most relevant links from the domain without actually going in depth of that domain. No existing focused crawling approach uses query based approach to find webpages of interest. In the proposed crawler, list of keywords is passed to the search query interfaces found on the websites. The proposed work will give the most relevant information based on the keywords in a particular domain without actually crawling through many irrelevant links in between them.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications.

Keywords: Web crawler; Information retrieval; Focused Web Crawler; Query based crawler.

1. Introduction

A search engine can be defined as a program designed to find information from the World Wide Web (WWW). The search engine produces a result by searching indexed database as per the user query. Typically, the criteria are specified in term of keywords or phrases. The results retrieved are presented in an ordered manner that matches the specified criteria. At the back end, search engines use regularly updated indexes to operate quickly and efficiently. Search engines maintain their database index by searching a large portion of Web. Search engines are different from Web directories as directories are maintained by human editors on the other hand, search engines use crawlers.

* Corresponding author. Tel.: +91-9041858682
E-mail address: manishkamboj3@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.075

A Web crawler is also known as a Web spider or Web robot. These are the automated computer program that browses the WWW recursively by following the hyperlinks [1]. The process of getting data from Web by a crawler is called web crawling or spidering. Web crawlers download the visited webpages so that an index of these webpages can be created. A Web crawler starts with a list of Uniform Resource Locator (URLs) to visit, called the seed URLs. As the crawler starts it get all the hyperlinks in the webpage adds them to a list of URLs to be visited further [2].

This paper proposes a query-based focused crawler using searchable interfaces on webpages. These interfaces expose the backend databases of the website whose seed URL is provided. The proposed work is better than the existing approaches as it does not require following a path to reach to webpages of interest. The proposed crawler shoots a set of queries on seed webpages using our dynamic keyword list. We maintain and optimize keyword list by a learning mechanism and update the list dynamically. The rest of the paper is organized as follows: section 2 represents the literature review of existing work. Section 3 discusses in detail motivation behind the work, design and architecture and implementation details of the proposed work.

2. Background and Related Work

Rather than collecting and indexing all the webpages over Internet, focused Web crawler [3] knows its crawl boundaries. It selectively seeks out webpages that are relevant to a pre-defined set of topics. It finds those links on the webpages that are likely to be most relevant while avoiding the irrelevant region of the Web. An upto date review of the various crawler is presented in [2]. The purpose of a focused Web crawler is to collect all the information related to a particular topic of interest on Web [4]. The study [5] discusses execution plans for processing a text database either using a scan or crawl. The method chosen have a great impact on the execution time and precision. Finding the query interfaces for hidden Web is an active area of research [10]. These interfaces are not used for focused crawling.

The keyword query based focused crawler guides the crawling process using metadata. The keyword data set is used for creating effective queries and the result obtained are feedback to the system. An Indian project for tourism and health named Sandhan [6] which is a multilingual platform is an example of the same. This project aims at identifying the language of a webpage using N-gram method. For the training purpose regional, non-regional and health queries are used. Tang et al. [7] proposed a focused crawling for medical information relevance and quality of the webpages retrieved by the query. They use relevance feedback crawler using query by example. Altingovde et al. [8] constructed a query engine that allows keywords and advanced query on the extracted data. A Web portal that is domain specific is finally made that can extract information from the backend database.

3. Proposed Work

This section discusses the motivation behind the work, design and architecture of the proposed work in detail.

3.1. Motivations

This work can be considered as an extension of our previous work [9]. The last developed crawler involved developing a URL ordering based focused Web crawler. The crawler takes input as the files containing: Indian surnames list, Indian cities and Indian premier institutes names list along with the seed URLs. The basic architecture of our last work is shown in figure 1.

Initially, a DFS crawling technique was applied where the crawler started from the seed URL and keeps on crawling the next URLs linked to the webpage blindly until a certain depth is reached. The number keywords matching the keyword databases present on the webpages are counted. The webpages having maximum number of matched keywords was considered as the most relevant.

In this paper, the above mentioned work has been extended. The top 10 most relevant webpages from each domain collected by the above crawler were chosen. From these webpages a list of priority keywords was generated from the words occurring a maximum number of times in these URLs. The list of the priority keywords thus