VITALIZED BI-LEVEL WEB CRAWLER FOR REMOVAL OF REDUNDANT CONTENT IN DEEP WEB INTERFACE

Supriya.H.S[1]

[1]*M.Tech Student, Department of CSE, Bangalore, Karnataka, India.*
supriya.honnur.s@gmail.com

Abstract

Search engine are used to search for appropriate data against trillion web pages, which are stored in several different servers. Normal search engine can search information on Shallow Web. Deep web is huge storage area of hidden information which is not indexed by automated search engines. Challenging job is to locate a deep web. Deep Web can efficiently harvest and explore accurate result for user query very quickly. This paper proposes a vitalized bi- level web crawler to analyze deep web interface and also remove redundant content in its database. In the first level, to stay away from tripping a huge number of pages Web Crawler search for core pages in search engines based on sites. For this web crawler will prioritize highly appropriate ones through ranking the sites for a given query. In the second level, Crawler achieves rapid in-site searching by adaptive link-ranking through excavating most appropriate links. Web is comprehended with several copies of equivalent content or equivalent web pages. Thus the incident of duplicate and near-duplicate content happening on the web will be very frequent. Thus removal of redundant content in deep web will be achieved based on parsing the content of one web page and comparing the pared content with other web page content which can save the storage area and bandwidth for a web crawler to crawl a web page.

Keywords: *Search Engine, Deep Web, Web Crawler, Site Ranking, Duplicate Content.*

---***---

1. INTRODUCTION

Internet becomes a part of our day today life. Every day we will keep browsing internet to get one or other information. Internet consist heap amount of information where each search engine will go through million web pages to get appropriate information for the user query. But how do these search engines get the appropriate words or the indexed webpage in few nanoseconds? Who will keep track of all these information? A web crawler is a automated script which peruses the World Wide Web in a methodical, computerized way, otherwise called a web spider or web robot. Web crawler will scan the web pages, downloads and stores Web pages to serve the search engine. Normally Search engine can search information only on Shallow Web. Deep web is huge storage area of hidden information which is not indexed by automated search engines. Deep Web can efficiently harvest and explore accurate result for user query very quickly. Challenging job is to locate a deep web which is not register with any search engine. Thus vitalized bi-level web crawler to analyze deep web interface and also remove redundant content in its database. In the first level, to stay away from tripping a huge number of pages Web Crawler search for core pages in search engines based on sites. For this web crawler will prioritize highly appropriate ones through ranking the sites for a given query. In the second level, Crawler achieves rapid in-site searching by adaptive link- ranking through excavating most appropriate links. The constraint of Crawler is it consume resources to download web pages it need network bandwidth, it requires memory to maintain private data structures, to select and evaluate URLs it needs CPU time and disk storage to store the text and links of fetched pages as well as other persistent data. As the web documents size is increasing on internet day by day is huge number, the duplication of documents also concurrently increasing on the web, which increases the retrieval time and reduces the accuracy of the retrieved documents. Thusidentifying of duplicate and near-duplicate web pages is necessary. This paper focuses on detection and removal of duplicate of web pages from the dataset. Thus removal of redundant content in deep web will be achieved based on paring the content of one web page and comparing the pared content with other web page content.

2. RELATED WORK

To take advantage of the extensive volume data covered in deep web, past work has proposed various strategies and tools, including deep web understanding, hidden web crawlers, and deep web samplers. For all these methodologies, the capacity to crawl deep web is a key task. Yeye He et al. [1] propose a method to crawl deep web information. Deep web crawl is the issue of surfacing rich data behind the web search interface of changed destinations over the Web. Deep web crawling is utilized for different reasons, for example, information integration and web indexing. It is difficult to crawl deep web information. Authors concentrate on entity adapted deep-web sites. The samples for these deep sites are web shopping sites where every element is an item which is connected with rich organized data like brand name, thing name, value, amount etc. Kevin Chen-Chuan Chang et al.[2] propose MetaQuerier method. The Web has been immediately developed by heap searchable online databases, where information are buried behind inquiry interfaces. MetaQuerier framework is utilized for both finding and coordinating databases on the Web. The objective of the MetaQuerier framework to get the deep web methodically and to utilize deep web uniformly. This will offer user with some assistance of querying online database. There are a few difficulties in duplicating of substantial scale information. Chelsea Hicks et al.[4], contrasted with the surface Web, the deep Web contains boundless more data. Specifically, fabricating a summed up web crawler that can file deep Web over all spaces remains a troublesome examination issue. In this research, they highlight these difficulties and exhibit by means of model execution of a summed up deep Web disclosure system that can accomplish high accuracy. Gang Liu et al.[3], proposes crawling and ontology research advances a sort of Deep Web passage programmed revelation strategy. In this research, firstly utilizing the data of particular field Deep Web section structure to build up area ontology, then web structures can be judged by the procedure of the point crawler creeping in the web. In the event that there are structures which are extricated its traits and ascertain the weights from structure's properties and ontology. Download this page when the weights more noteworthy than the altered worth. At long last they utilize test words to analyze the as of now download pages to discover amazing Deep Web section pages. Different from the crawling methods and tools said above, Vitalized bi-level web crawler is a space particular crawler for locating significant deep web content sources. Vitalized bi-level web crawler focuses at deep web interfaces and utilizes bi-level design, which not just orders sites in the first level to filter out immaterial sites, additionally arranges searchable words in the second level. Rather than basically arranging links as relevant or not, Vitalized bi- level web crawler first positions sites and after that organizes links inside a site with another ranker.

3. PROPOSED SYSTEM

To productively and adequately find deep web information sources, Vitalized Bi-Level web crawler is outlined with a two level architecture, site locating and in-site exploring, as shown in Figure. The principal site locating level finds the most significant site for a given point, and after that the second in-site exploring level reveals searchable words from the site.

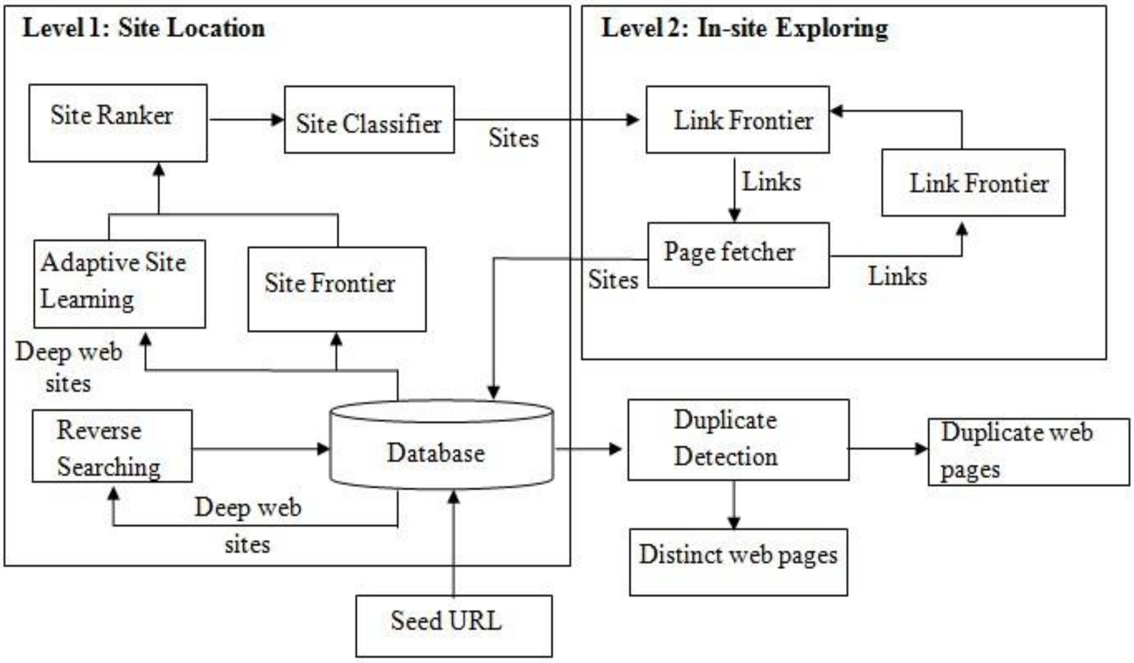


Fig 1: System architecture of Vitalized Bi-level Web Crawler for Removal of Redundant Content in deep interface

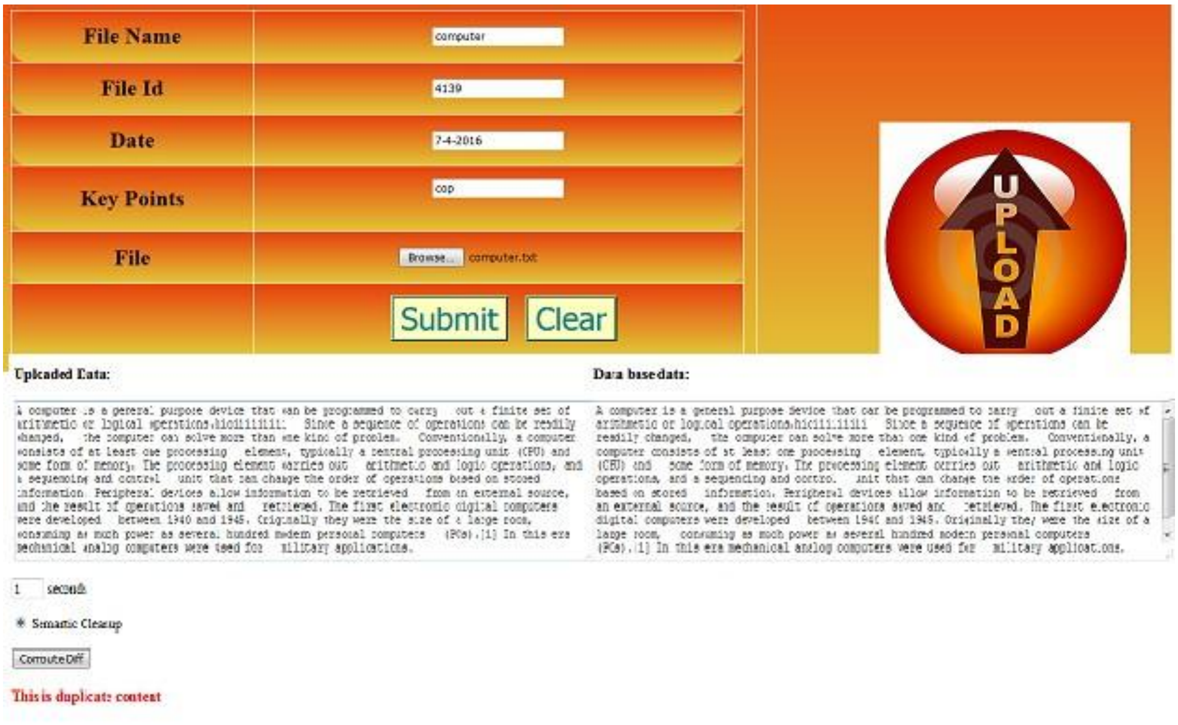
In particular, the site locating level begins with a seed set site to database. Seeds site are contestant sites given for Vitalized Bi-Level web crawler to begin crawling, which starts by taking after URLs from picked seed sites to explore different pages and different domains. At the point when the quantity of unvisited URLs in the database is not exactly a threshold the crawling process, Vitalized Bi-Level web crawler performs "reverse searching" of known deep sites for main pages and feed these pages back to the webpage database. Site Frontier brings home page URLs from the site database, which are positioned by Site Ranker to organize highly appropriate sites. The Site Ranker is enhanced during crawling by an Adaptive Site Learner, which adaptively gains from components of deep sites found. To accomplish more precise results for a focused crawl, Site Classifier orders URLs into significant or unessential for a given topic as indicated by the home page content. After the most applicable site is found in the first level, the second level performs proficient in-site exploration for unearthing searchable forms.

Links of a site are stored in Link Frontier and comparing pages are fetched. Note that site locating level and in-site exploring level are commonly interwoven. At the point when the crawler finds new site, the site's URL is embedded into the Site Database. The conventional crawler tails all recently discovered links. Conversely, Vitalized Bi-Level web crawler endeavors to minimize the number of visit to URLs, and in the meantime boosts the quantity of deep sites. To accomplish these objectives, utilizing the links that are present in downloaded web pages are mostly not enough. This is because that a site more often contains a small number of links to different sites. To address this a crawling strategy is proposed ie., Reverse search the thought is to exploit existing web indexes, for example, Google to discover focus pages of unvisited destinations. This is conceivable in light of the fact that web search engine rank website pages of a webpage and focus pages have a tendency to have high positioning qualities. In Vitalized Bi-Level web crawler, decides the topical importance of a site in light of the content of its home page. At the point when another site comes, the home page content of the site is extracted and parsed by expelling stop words and stemming.

Duplicates are undesirable for many types of data. These include databases, mailing lists, file systems, email and image data. Normally we can find redundancy is in two categories. The first is exact replication of web pages, when two web pages contain the same content. The second, web pages which are very similar called as near-duplicated web pages ie., the pages which are similar and must be more than the threshold value. This paper focuses on detection on and removal of duplicate of web pages from the dataset. Thus removal of redundant content in deep web will be achieved based on paring the content of one web page and comparing the pared content with other web page content. The basic idea is to reduce each document to a set of strings and then calculate the similarity of those sets. To reduce a document to a set of strings, extracts the set of all substrings from a text of uploaded file compare those with database content which is already present in database for that keyword. Calculation of similarity is a simple measure for expressing similarity of sets. Thus duplicate detection and removal will save the bandwidth and storage database also save the searching and sorting time for web crawler.

4. RESULTS

Vitalized bi-level web crawler will crawl the web pages efficiently and faster way compared to other web crawlers. For the redundant content detection during uploading of file by admin, newly uploaded file content are parsed and compared with already parsed and stored content web page in database. If it discovers any duplicate content of web page it will displays the message of duplication.



File Name	computer
File Id	4139
Date	7-4-2016
Key Points	cop
File	Browse... computer.txt

Uploaded Data:

A computer is a general purpose device that can be programmed to carry out a finite set of arithmetic or logical operations.h1c1111111 Since a sequence of operations can be readily changed, the computer can solve more than one kind of problem. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU) and some form of memory. The processing element carries out arithmetic and logic operations, and a sequencing and control unit that can change the order of operations based on stored information. Peripheral devices allow information to be retrieved from an external source, and the result of operations saved and retrieved. The first electronic digital computers were developed between 1940 and 1945. Originally they were the size of a large room, consuming as much power as several hundred modern personal computers (PCs).[1] In this era mechanical analog computers were used for military applications.

Data base data:

A computer is a general purpose device that can be programmed to carry out a finite set of arithmetic or logical operations.h1c1111111 Since a sequence of operations can be readily changed, the computer can solve more than one kind of problem. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU) and some form of memory. The processing element carries out arithmetic and logic operations, and a sequencing and control unit that can change the order of operations based on stored information. Peripheral devices allow information to be retrieved from an external source, and the result of operations saved and retrieved. The first electronic digital computers were developed between 1940 and 1945. Originally they were the size of a large room, consuming as much power as several hundred modern personal computers (PCs).[1] In this era mechanical analog computers were used for military applications.

1 seconds

● Semantic Cleanup

This is duplicate content

Fig 2: Duplicate content detection

In the Vitalized bi-level web crawler's first level ie., site locating phase the site rank is carried based on the site frequency. When web pages are visited frequently its rank will increase as the rank of web page increases crawler will find it as more relevant web page for the corresponding keywords.

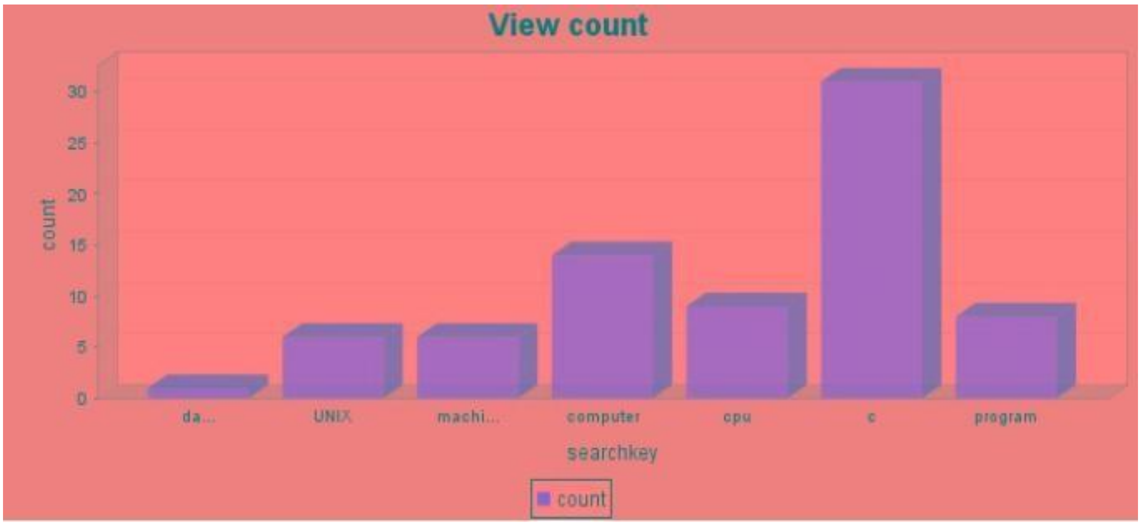


Fig 3: Graph view of ranked sites

The above figure will give a graph view of ranked sites. The site with high rank will indicate the most relevant content for corresponding keywords search.

5. CONCLUSION AND FUTUREWORK

Vitalized Bi-Level web crawler is proposed for productively assembling deep web interfaces. Vitalized Bi-Level web crawler has two level: site locating and in-site Exploring. Deep sites are reverse searched by Vitalized Bi-Level web crawler for core pages, when the threshold value is greater than number of unvisited URLs in the database during the crawling process. The task of site frontier is to fetch web-page URLs from the site database. If un-visited sites are there those sites are handed to site frontier and are arranged by site ranker. Adaptive Site Learner improved the Site Ranker during crawling. Duplicate detection and removal of redundant content in deep web will be achieved based on paring the content of one web page and comparing the pared content with other web page content. Future work of this paper carried out as instead of search for keyword can give any input type like image and search for appropriate images. For duplicate detection, the keywords positions are not consider in the proposed system. To locate the near duplicate detection of web pages can use the positional filtering. Thus it increase the accuracy rate in locating near duplicate and eliminating them.

REFRENCES

[1]. Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, and Nirav Shah. URL Template Generation in “Crawling deep web entity pages”. Dec 2013

[2]. Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. MetaQuerier: Architecture & Techniques in “Toward large scale integration in Building a metaquerier over databases on the web”, july 2005.

[3]. Gang Liu, Kai Liu, Yuan-yuan Dang, crawling and ontology in “Research on discovering Deep web entries” IEEE -2011.

[4]. Chelsea Hicks, Matthew Scheffer, Anne H.H. Ngu, Quan Z. Sheng, “Discovery and Cataloging of Deep Web Sources” IEEE IRI 2012, August 2012.

[5]. Zilu Cui, Yuchen Fu, TF-IDF statistics in “Deep Web Data Source Classification Based On Query Interface Context” 2012.

[6]. Z. Bar-Yossef, I. Keidar, and U. Schonfeld, Basic Heuristics in “Do not crawl in the DUST: Different URLs with similar text” 2007.

[7]. A. Dasgupta, R. Kumar, and A. Sasturkar, . “De-duping URLs via rewrite rules” Aug 2008.