Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Prefetching of Data from Hidden Web Using Domain Specific Interface Mapper

Babita Saharawat
Research Scholar
Department of CSE
Manav Rachna College of Engg, India.

Ashok Goyal
Department of IT
Manav Rachna College of Engg, Faridabad, India.

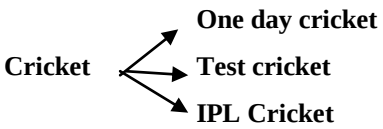
Dr. Komal Kumar Bhatia
Department of CSE
YMCAUST, Faridabad
Faridabad, India.

Abstract— Information searching has becoming one of the most important and popular activities on the Web. These search engines deal only with the surface Web, the set of Web pages directly accessible through hyperlinks, mostly ignoring the vast amount of information hidden behind forms, which is called the hidden Web. Web information is accessed today primarily relies on the search engines. Current search engines cannot make index to the pages which are generated automatically by the back – end of the databases called invisible web or hidden web. The information is hidden behind HTML forms and it is only available on response to user's request. In this paper a system based on domain and keyword specific information extraction is described.

Keywords— DSIM, Information Retrieval, Hidden web, Search engine, Fuzzy matching and Spiders.

I. INTRODUCTION

The rapid growth in the amount of information and the number of users has led to difficulty in providing effective search services for the web users and increased web latency; resulting in decreased web performance. Most of the search engines deal with surface Web only, the set of Web pages directly accessible through hyperlinks, mostly ignoring the vast amount of information hidden behind forms, which composes by the hidden Web. The Hidden Web also called the Deep web, the Invisible Web. Deep Web accessed WWW content that is not part of the surface web, surface web which is indexed by standard search engines. The rapid growth in the amount of information and the number of users has led to difficulty in providing effective search services for the web users and increased web latency; resulting in decreased web performance. Information searching has becoming one of the most important and popular activities on the Web. Most of the search engines deal with the surface Web, the set of Web pages directly accessible through hyperlinks, mostly ignoring the vast amount of information hidden behind forms, which composes by the hidden Web. As compared to the Surface Web, the hidden Web contains a much larger amount of high-quality information hidden behind the databases [1]. It is estimated that there are several million hidden-web sites. Web pre-fetching becomes an important solution where in forthcoming page accesses of a client are predicted, based on domain information. This dissertation will propose an approach for increasing web performance by analysing user domain and perfecting the frequently accessed pages after completing the web structure, so as to provide relevant information to the user**.** As the web is vast resources of information, how to find just the right bit of information that user need or how to provide relevant information to the user from the internet with in a limited time is a big challenge in information retrieval. Hidden web is used for data extraction from web with the help of search engine. Web information can't access without the search engine. Currently used search engines can't make index to the pages which are generated automatically by the back end databases called deep web. The web has now grown as a rich collection of dynamic and interactive services such as image, video conferencing etc. retrieving web pages from remote servers delay can be experienced. One solution is to increase the bandwidth, which will result in increased system cost. To reduce cost and to enhance the performance, cache-based approach is used that is Prefetching. Any user search any query like cricket than result shown like that.



When user wants to search on the web data is given to the search engine in the form of a query. Crawler then goes to the remote server to search matched result. Along with the matched result the crawler can pre-fetch domain specific data that is relevant to the given query and is stored on local server along with the searched data. This data is not indexed thus reducing the searching time if this data is searched again.

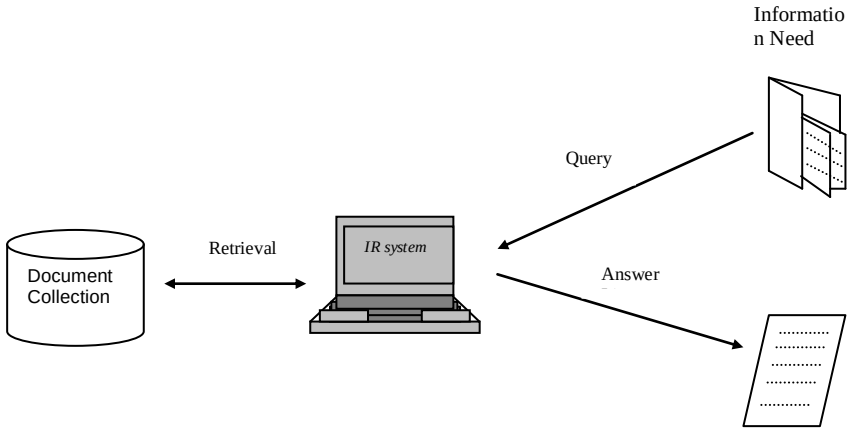
Search Engine: - A web search engine is a software system that is designed to search for information on the World Wide Web or Internet. Search engines were also known as some of the brightest stars in the Internet. Every search engine has basically 3 parts that are

1) Crawling
2) Indexing
3) Searching

Various search engines are available in Internet that are Google, Locus, yahoo, Alta vista, ask, Bing and so on. All search engine accessed only the Surface web not used the Hidden web data.

I.I INFORMATION RETRIEVAL

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. The process begins when a user enters a query into a system and related queries are gives some formal result statements of information. IR query does not uniquely identify a single object in the collection; it has several objects that are match with query with the different degrees of relevancy of matching. IR system always compute a numeric score value on how well each object in the data base matches the query and query rank are according to that values. Top ranking objects are always shown to the user and it is the policy of the every search engine technique to search any query in database.



PROBLEMS IDENTIFICATION

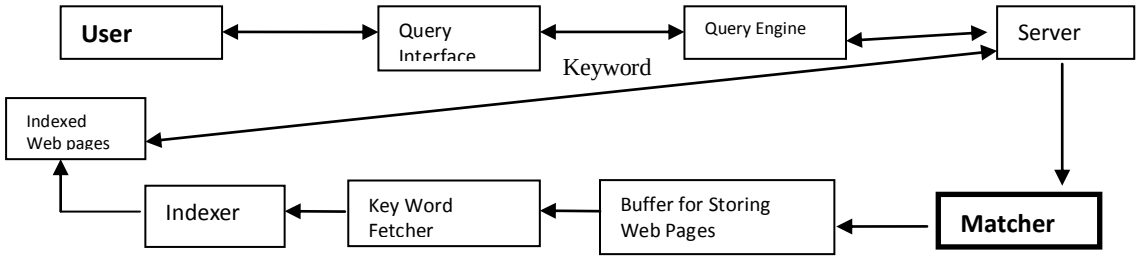
PROBLEM FORMULATION

- In the normal working of search engine, the search engine returns millions of web pages as related to user query result but all the pages search engine first search's its own local database and if there the desired web pages doesn't found then it fetches from www.
- Though all the pages prefetched by the search engine are not related to our work and it is not possible to view all the pages by clicking over the link. Fetching technique is fetch more data from hidden web and stored in local server that means any user search start with finding keyboard but fetched technique also searched and fetched data for mouse and monitor with the keyboard.
- In case of hidden web where the size is more larger than the surface web this would be a difficult task to handle more web pages and store in local database.

OBJECTIVES OF THE STUDY

Our objective is to make a method called prefetching data with the help of DSIM, where data is fetched from www or local database. The search engine return the most relevant pages on the top of the list because it is a normal tendency of searching but data is PREFETCHING from Hidden Web Pages Using (DSIM) with less timing for searching.

II. PROPOSED ARCHITECTURE



Query Interface: - Query Interface is a mechanism in COM (Microsoft's Component Object Model) for determining a known component supports a specific interface to the string (query).

Query Engine: - Query Engine, is a service that takes a description of a search request, evaluates and executes the user request, and returns the results back to the user.

Web server: - Web servers are computers that delivers Web pages to the user. A Web server is a program that, using the client/server model and the World Wide Web's. Hypertext Transfer Protocol (HTTP), serves the files that form Web pages to web users (whose computers contain HTTP clients that forward their requests).

Buffer: - Buffer is a temporary storage area, usually in the RAM. The purpose of most buffers is to act as a holding area, enabling the CPU to manipulate data before transferring it to a device.

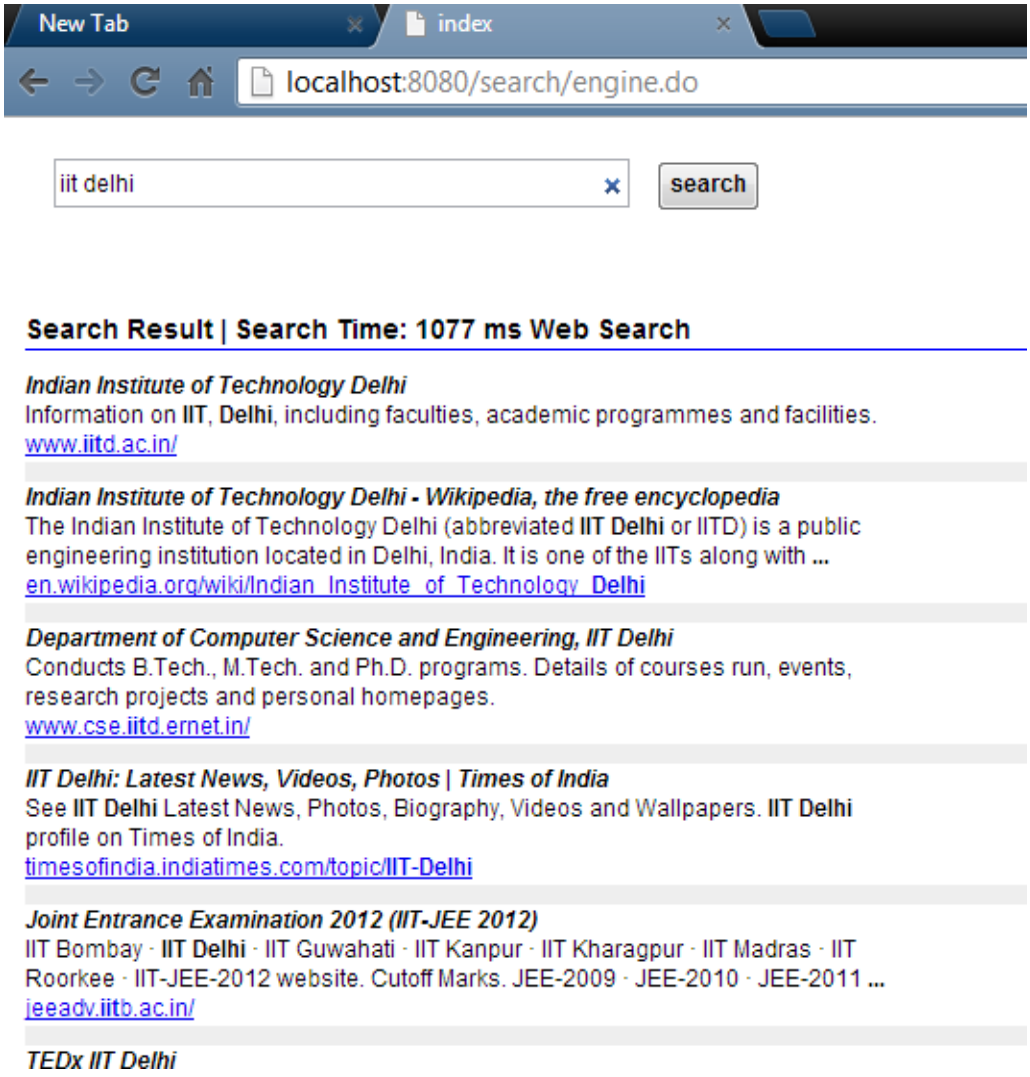
Matcher: - DSIM (Domain Specific Interface Mapper) is an ideal technique for searching the resources according to the domain.

Indexer: - After a page is crawled on web, the next step is to index the page (content). The indexed page is stored in a local database (giant database), from where it can later be retrieved.

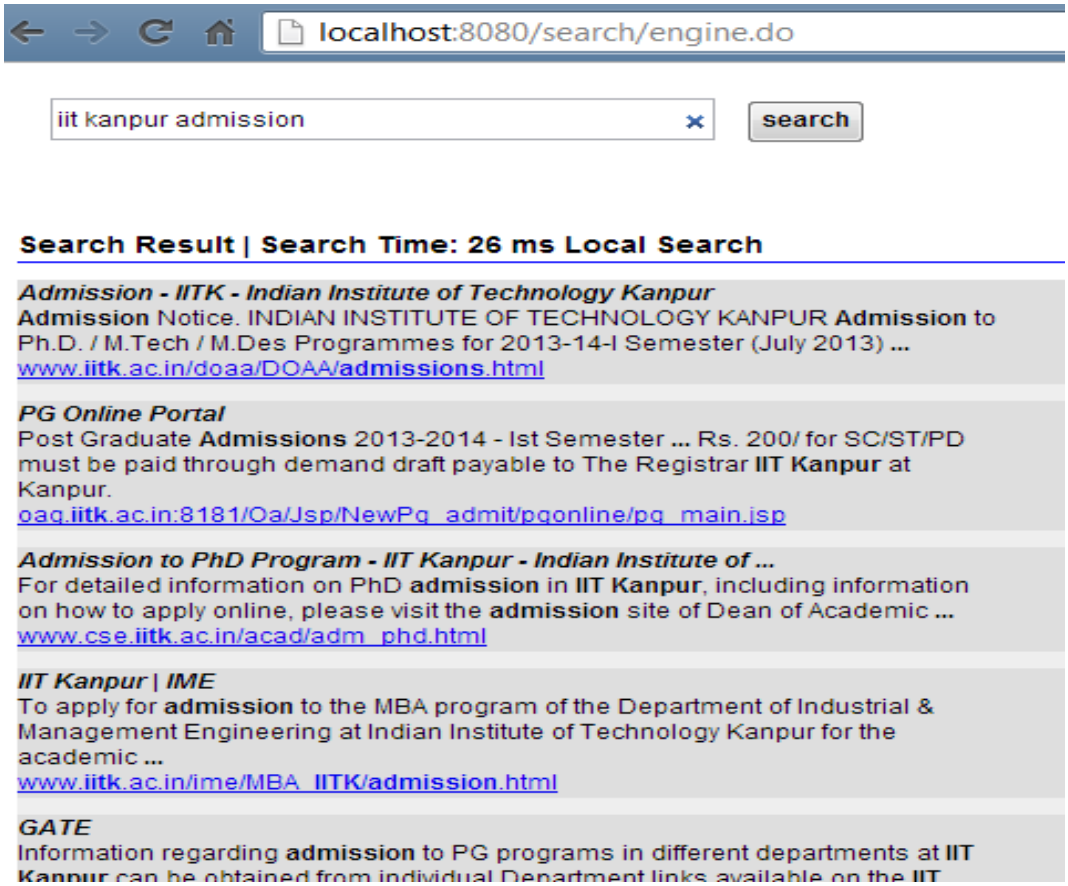
III. RESULT ANALYISIS

It is a domain independent system, with define some related keyword in the database. The developed tool provide more valuable information from the hidden web databases at one single location that will provide effective search environment to the user. The results obtained were encouraging. Hidden web is used for data extraction from web with the help of search engine. Web information can't access without the search engine. Currently used search engines can't make index to the pages which are generated automatically by the back end databases called deep web. The web has now grown as a rich collection of dynamic and interactive services such as image, video conferencing etc. retrieving web pages from remote servers delay can be experienced. One solution is to increase the bandwidth, which will result in increased system cost. To reduce cost and to enhance the performance, cache-based approach is used that is Prefetching. Any user fired query on search engine, it will start analysis first time from the web, local data base which is used prefetch technique behind the search with some association rules.

Search result first step:-At first time any user fired a search query, it will fetch data from web or remote server with 1077milliseconds. That means1000 milliseconds=1sec, it will take approximate 1.077seconds time to search from web.



Second Result step: When we again fired predefined domain value on search engine it fetch data from local server with very less timing.



IV. CONCLUSION AND FUTURE

In a conclusion we can say that DSIM quickly identifies the regions in the interface repository comprising of important fuzzy mappings. The tests conducted on Domain-specific Hidden Web DSIM indicate that it efficiently search the hidden web pages. It further improves by discarding the less important mappings as fuzzy mapping uses a comparestringfuzzy as a selection parameter. Hidden web pages are used searching faster and easy by DSIM for better result in few seconds. DSIM stands for Domain Specific Interface Mapper Technology. The proposed DSIM finds the semantic mappings between the components of different web interfaces of the same domain i.e. all the interfaces belong to the same domain such as origination domain, education domain and airline domain.

REFERENCES

[1] Bhatia, Komal Kumar and Sharma, A.K (2008) A Framework for Domain-Specific Interface Mapper (DSIM), IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.12, December.

[2] Khetwat, Saritha and Dharavath, Kishan (2011) Domain and Keyword Specific Data Extraction from Invisible Web Databases, Eighth International Conference on Information Technology: New Generations.

[3] "Anatomy of a Large-Scale Hyper textual Web Search Engine", Sergey Brin and Lawrence Page, Stanford,CA 94305,

[4] BrightPlanet. Com, The deep Web: Surfacing hidden value.

WEBGROPLY

1) http://www.en.wikipedia.org/wiki/Deep_Web retrieve on 22.04.2013
2) http://infolab.stanford.edu/~backup/google.html retrieve on 22.04.2013.
3) http://www.ijcaonline.org/volume15/number4/pxc3872579.pdf retrieve on 22.04.2013
4) http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5945400 retrieve on 22.04.2013
5) http://brightplanet.com.

