www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 5 May 2016, Page No. 16668-16672

Web Harvesting: A Technique for Fast Retrieval of Information from Web

Meenakshi Srivastava[1], *Dr. S.K. Singh*[2]

Assistant Professor, Amity Institute of Information Technology, Amity University, Lucknow, India[1,2]

ABSTRACT

Internet is collection of vast information. Need of searching the web pages for a specific piece of information is a common practice. Search engines are used to search and retrieve the information as per query raised by the user. Making this search process better and fast has always been the area of interest for researchers involved in web mining. The process of searching the web can be improved by Web harvesting. Web harvesting is the process by which specialized software collects data from the Internet and places it into files for end user. It serves a function similar to, but more advanced than, the tasks a search engine performs. Web harvesting is also known as Web scraping. In this article we have explored the field of Web harvesting and emphasized its use for fast and effective retrieval of information from web

Key Words: Web harvesting, Web Archives, Information storage

[I] INTRODUCTION

Web Harvesting stands in name – Web that is Internet which is itself a whole world of information. Harvesting belongs to agriculture harvesting. In agriculture seeds are needed to harvest, whereas in web harvesting information is used to seed and prepare new information. Web harvesting is the process by which specialized software collects data from the Internet and places it into files for an end user. It serves a function similar to, but more advanced than, the tasks a search engine performs. Web harvesting also known as Web scraping, Web harvesting gives the user automated access to information on the Internet that search engines cannot process because it can work around HTML code. Web content harvesting involves the extraction of information by pulling data from both search page results and from a deeper search of the content hidden within Web pages. Web Harvesting is a useful technique for retrieval of Multimedia Information. As Multimedia Information is normally retrieved by keywords, harvesting will play an important role in creating multimedia achieves by retrieving the documents through deeper search. In this manuscript we have discussed the use of web harvesting for web achieving which will facilitate the multimedia information retrieval. Web harvesting is a broad and big concept to learn and analyze. So the content analysis is very tough. In internet not too much content is given. Problems areas of harvesting are not fully described. The basic concern is that nowadays whole data for a particular search is stored anywhere, but if Web harvesting is used then the searched and fetched data of a particular search at user's end.

- Web harvesting is the process by which specialized software collects data from the Internet and places it into files for an end user.
- Web usage harvesting tracks general access patterns and customized usage by Web users. By analyzing Web usage, harvesting

can help to create clarity about how users behave.

- Companies use Web harvesting for a wide array of purposes. Some of the more common data sets compiled are information about competitors, lists of different product prices, and financial data. Data may also be collected to analyze customer behavior [1].

Need of harvesting -

- From the business eye web harvesting gives data that help a business person to analyze the products value in market.
- Harvesting help to create clarity about how users behave. This is another way to improve the function of the Web, but on an end-user level.
- Accurate data needed.
    - Security of data
    - Access effective and efficient data.

In the present article we have discussed the concept of web harvesting, its components, how it is useful, security issues involved in web harvesting.

[II] Components of Web Harvesting:

In this section we have described every useful component in the field of web harvesting.

What is Crawling the live web:

Crawling is a data acquisition method widely used by web archives that capture information for preservation and later access. Besides the textual contents to support search, web archives must exhaustively gather embedded _les to enable the reproduction of the archived pages[3].

Process: Web harvesting is the process by which specialized software collects data from the Internet and places it into files for an end user. It serves a function similar to, but more advanced than, the tasks a search engine performs. Also known as Web scraping, Web harvesting gives the user automated access to information on the Internet that search engines cannot process because it can work around HTML code. The three major types of Web harvesting are for Web content, structure, and usage. Web archives face many challenges related to scalability and information overload because they accumulate previous documents and indexes, unlike web search engines that drop the old versions when new ones are discovered [2]. Web archives already hold more than 181 billion contents and this number continues to grow as new initiatives continue to arise. This data dimension is one order of magnitude larger than the number of documents indexed by the largest web search engine and 150 times more than the con-tent of the Library of Congress.

Working of crawler :

– Collects contents from the Web, starting from an initial set of addresses

– Iteratively downloads contents and extracts links to find new ones

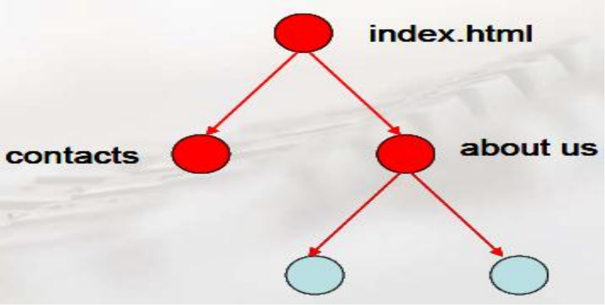


Fig [1] Process of Indexing

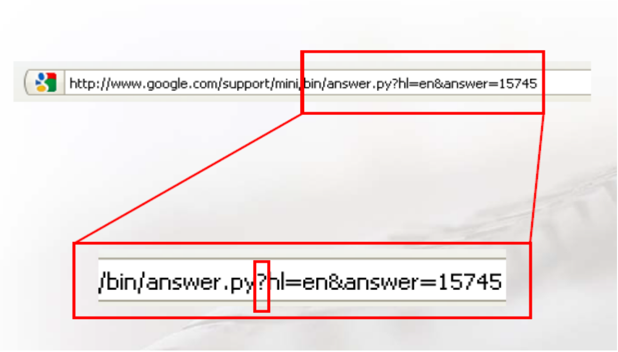
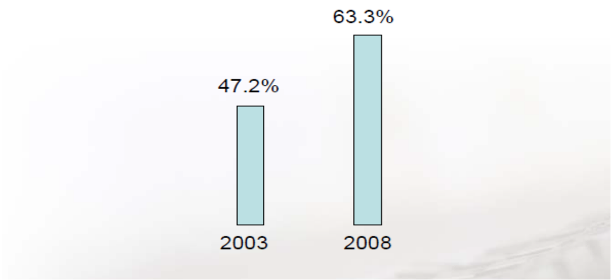

http://www.google.com/support/mini/bin/answer.py?hl=en&answer=15745
/bin/answer.py?hl=en&answer=15745

Fig [2] Identifying the Technical Trend

This is technique by which harvesting is processed. It searched the user search and fetches the content of URL. And process the data and calculation is processed.

DOI: 10.18535/ijecs/v5i5.55



After the Calculation this type of result is given to companies.

[III] AREAS OF HARVESTING:

Area of harvesting is not only limited to pc harvesting, it also search on mobile search for type of browser is used and type of content is searched. Harvesting is a very useful tool for scientific research especially including multimedia information. Harvesting can be very time saving when utilized for searching and storing similar kind of images which are spread all over the web.

[IV] CHALLENGES IN HARVESTING

The amount of data archived is incremental and the waste of resources caused by the storage of duplicates originated by contents that remain unchanged across time is significant[4]. In the contents downloaded in a daily crawl, 46% were du-plicate. In trimestral crawls the level of duplication was 30%.

[V] PROCESS OF WEB HARVESTING

In the process of Web Harvesting, as shown in Fig.[3] following methods are involved

- Web content
- Structure
- Usage.

Web harvesting technique fetches information from HTML code. It goes through all links relates to document. The user's search result and web site response is also monitored. Browser and document viewers for cell phones as well as URL links and their length are also searched.

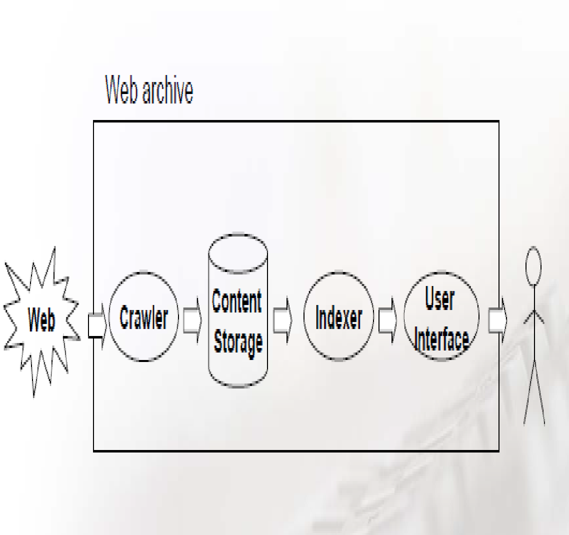
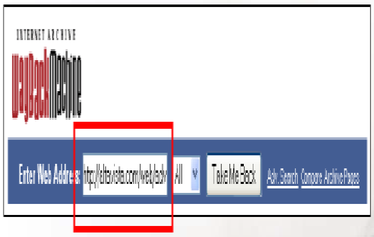


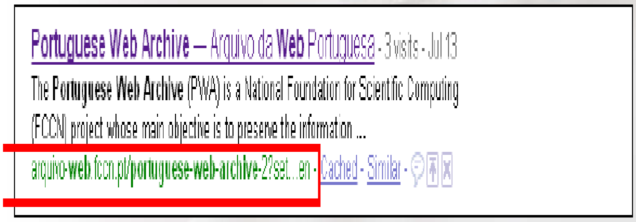
Fig [3] Process of Web Archiving

Harvesting with crawler to provide faster search results:

Search is on user content that is in internet. User type what needed this information is stored and analyzed by harvesting team and URL is source that store it. After storing the information the calculation on harvesting is done. The length of URL search is calculated as shown in example-



http://altavista.com/web/adv



Portuguese Web Archive – Arquivo da Web Portuguesa - 3 visits - Jul 13

The Portuguese Web Archive (PWA) is a National Foundation for Scientific Computing (FCCN) project whose main objective is to preserve the information ...

arquivo-web.fccn.pt/portuguese-web-archive-2?set...en - Cached - Similar -

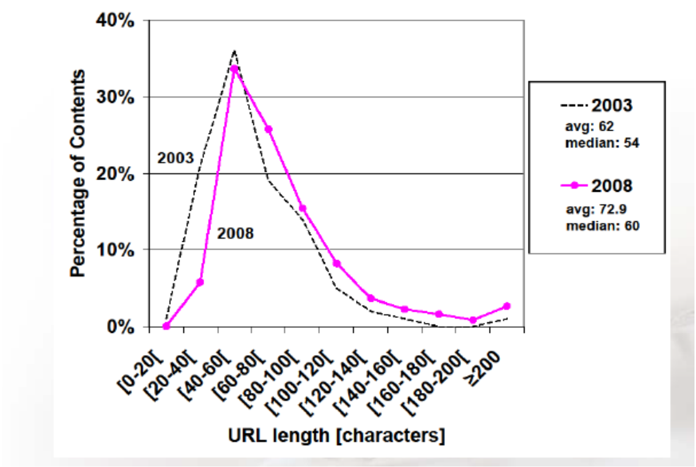


Fig [4] Effect on search category

Challenges Faced in Web Archiving

- Web archiving poses different challenges. The fast growth that web has experienced, combined with its dynamic characteristics, make it difficult to decide what to preserve and keep it up-to-date. The web is very large and still growing every day[5].
- Its decentralized organization results in different policies and adoption of standards by the organizations responsible for the sites. Each different site owner chooses its own standards and builds its site at its own way.
- Pages and websites change, appear and disappear, often with no trace. Its fluid essence results in links that no longer work, whether pointing to the wrong place or no place at all, and often returning a 404 error that web users so well know. This means that at any one time a large amount of links on the internet will be dead or link to the wrong site.
- Internet domain names sometimes disappear or change ownership. Life of a link may be conditioned by link-depth, file format and site owner, as personal web pages have often different lifetimes than institutional sites.
- In the beginning of the WWW pages were essentially static but the evolution of web based technologies made the internet more dynamic.
- Web pages created from dynamic databases are difficult to collect and replicate in repositories. Much database information will be invisible to harvesting robots.[7]
- When a dynamic page contains forms, JavaScript or other active elements, the archived site will lose its original functionality. The simpler and static the site is, the more easily it is harvested.

[VI] APPROACHES FOR WEB ARCHIVING

Three main different approaches are taken when considering web archiving:

- Deposit, in which web documents or parts of sites are transferred into a repository;
- Automatic harvesting, in which crawlers attempt to download parts of the web.
- The Internet Archive follows this approach; selection, negotiation and capture, in which web resources are selected, then theirinclusion is negotiated with the site owners and finally it is captured. TheNational Library of Australia follows this approach [3].

These approaches do not have to be taken in separate. Combinations of them may be used, in order to achieve a more successful result. Traditional IT processes just aren’t built to access, transform, and load all this unstructured, real-time data. Most BI and analytics tools utilize data owned by the Enterprise (e.g. Financial, Sales, Manufacturing, Customer Surveys, Market Share, and Pricing data).[8][9] While valuable, this data is often outdated, stale, biased, and incomplete. Live social media is layering mountains of information on top of Web search data. Traditional BI tools just aren’t built for this. Most BI and analytics tools utilize data owned by the Enterprise (e.g. Financial, Sales, Manufacturing, Customer Surveys, Market Share, and Pricing

DOI: 10.18535/ijecs/v5i5.55

data). While valuable, this data is often outdated and stale. As a result, executives are now demanding that CIOs deliver real-time reports on key market activities across product lines [10][11]. They simply can't trust manipulated P&L and marketing reports anymore. So you need to extract Web data, but it's not easy. There are a lot of vendors out there claiming all kinds of capabilities [6].

[VII] SECURITY ISSUE IN WEB ARCHIVING

Security in web harvesting is an important issue because it fetched data which is important for organization[7]. Many times companies hire web harvesting for fetching and then buy storage for storing that data[8][9].

- Authorization
- Authentication
- Confidentiality
- Integrity
- Availability

[VIII] CONCLUSION

In this paper we have presented the importance of Web harvesting and the process through which searched content can be stored. As web is changing its content daily, new features, technologies, concepts are coming up on regular basis harvesting will provide a great support for creating archives for the valuable information. Day by day web content increases and changes its nature so web harvesting provides a great support for storing these data and maintaining its content via creating achieves. Web harvesting is helpful in storing more data, easiness in retrieval and provides fast search. We have found that this technique is very useful for an end user as it stores the search done by users. Collecting and archiving the web poses a number of challenges that have to be overtaken in order to perform a well-succeeded web collection. Different countries and organizations have perceived the importance of preserving valuable web contents and started efforts in order to do so.

REFERENCES

[1] www.archive.pt

[2] Portuguese web achieve.

[3] Meenakshi Srivastava, Dr S.K. Singh, Dr. S. Q. Abbas "Web Archiving: Past Present and Future of Evolving Multimedia Legacy", International Advanced Research Journal in Science, Engineering and Technology Vol. 3, Issue 3, March 2016. DOI DOI 10.17148/IARJSET.2016.3309

[4] PANDORA. [On-line] URL: http://pandora.nla.gov.au/

[5] Internet Archive [On-line] URL: http://www.archive.org/

[6] Joao Miranda "Web Harvesting and Archiving" [On-Line] URL: http://web.ist.utl.pt/joaocarvalhomiranda/docs/other/web_harvesting_and_archiving.pdf

[7] Web Archiving [On-Line] URL: https://en.wikipedia.org/wiki/Web_archiving

[8] Internet Archive [On-Line] URL: https://en.wikipedia.org/wiki/Internet_Archive

[9] Molly Bragg, Kristine Hanna "The Web Archiving Life Cycle Model"[On-Line]

URL:https://archiveit.org/static/files/ archiveit_life_cycle_model.pdf

[10] IIPS[On-line]URL: http://netpreserve.org

[11] Web archiving initiatives [On-line] URL: https:// en.wikipedia .org/ wiki/ List of Web archiving initiatives