Dynamic Web System

Tarvejahmad Ansari*, Akshay Pillai*, Asif Sayyad*

* Computer Department, Shatabdi Institute of Engineering & Research

Abstract- The World Wide Web is the emerging field available to users to access the contents of the web. This field is parted into two. They are surface web and deep web. The surface web refers to the static and is linked with other pages whereas deep web refers to the web page that is not indexed by the general search engine. The extraction of contents from web pages arise the problem of web page programming language independent. This problem is arising because of the underlying complex structure of the web pages. The extraction of deep web contents from the deep web pages involves both the data record extraction and data item extraction. To evaluate the performance of the extraction process, the method revision is used. To overcome this problem, to build a website application like online forum, which provide location base information in the particular city and user will get all information of his/her interest under one website and need not visit various websites and waste time.

Index Terms- Deep web, Visual Block Tree, Visual Features, Web Data Extraction, Web Mining, Wrapper Generation.

I. INTRODUCTION

Deep Web, as is known to everyone, contains magnitudes more and valuable information than the surface Web. However, making use of such consolidated information requires substantial efforts since the pages are generated for visualization not for data exchange. Structured data objects are a very important type of information on the Web. Such data objects are often records from underlying databases and displayed in Web pages with some fixed templates. In this paper, we also call them *data records*. Mining data records in Web pages is useful because they typically present their host pages' essential information, such as lists of products and services. Extracting these structured data objects enables one to integrate data/information from multiple Web pages to provide have to be found manually or by another system. proposes a method that tries to explore the detailed information page behind the current page to extract data records.

Dynamic Web system is a web application which is useful for the online forums which provide information about given area. Our system provides location base information. User only views information of the area from where he/she login or opens the website. System use Service provider (SP) This system is totally automated and does not require any human interaction to update content of the system. System contains artificial bots which provide live update to system. Information of web system is automatically updated when the owner of the client website update its website contents.
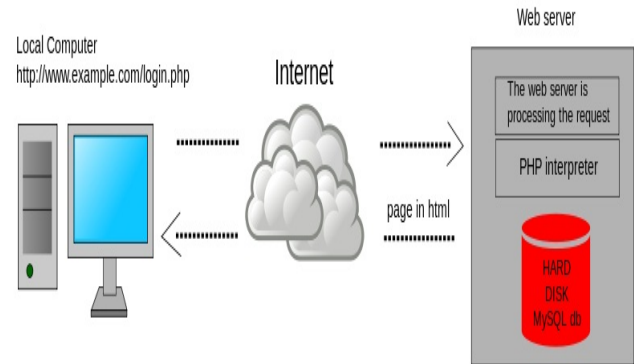

Fig: Architecture of web system

Web pages often contain around the body of an article that distract a user from actual content. Extraction of useful and relevant content from web pages has many applications, including cell phone and PDA browsing, speech rendering for the visually impaired, and text summarization. Most approaches to removing clutter or making content more readable involve changing font size or removing HTML and data components such as images, which takes away from a webpages inherent look and feel. Unlike Content Reformatting, which aims to reproduce the entire webpage in a more convenient form, our solution directly addresses Content Extraction. We have developed a framework that employs an easily extensible set of techniques that incorporate advantages of previous work on content extraction. Our key insight is to work with the Document Object Model tree, rather than with raw HTML markup. We have implemented our approach in a publicly available Web proxy to extract content from HTML web pages.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

System contains artificial bots which provide live update to system. Information of web system is automatically updated when the owner of the client website update its website contains. System use different web mining technology to extract data from remote website at run time. System contain deferent client like city malls, industries, news and hospitals. We also provide facility to compare offers between two malls located in same area. User can easily access all information from this application by just login to system.
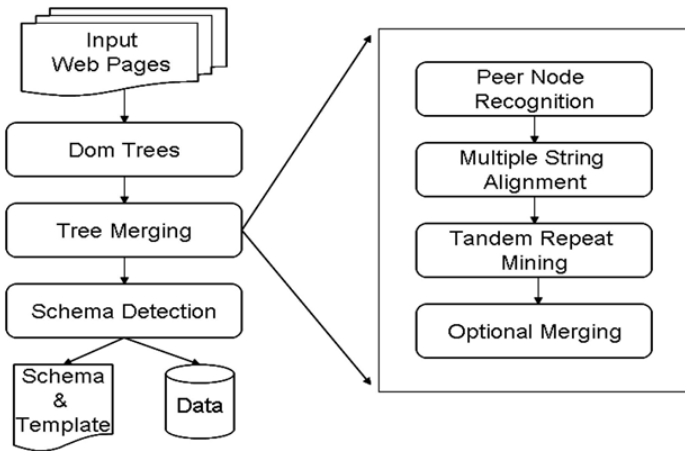
Fig: Internal working diagram

III. EXISTING SYSTEM

The web page is displayed regularly in a two-dimensional media, it made users to browse the contents of the web page. A promising research direction is opened where the visual features are utilized to extract deep web data automatically. It also utilizes some non-visual information. The non-visual feature includes the same type of font, frequently occurring symbols and data types are also used. Since the web pages displayed consist most of text and images, web page layout and font are considered as visual information. The fonts are determined by its size, face, color, frame, etc., These visual features are important for identifying special information in the pages. To perform this, the features used are position, layout, appearance and content. The position features describes the location of the data region on deep web page. The layout features describes how the data records in the data region is typically arranged. The appearance features which captures the visual features with in data records. The content features indicate the regularity of the contents in the data records. The proposed system extracts both the structured as well as the unstructured pages.
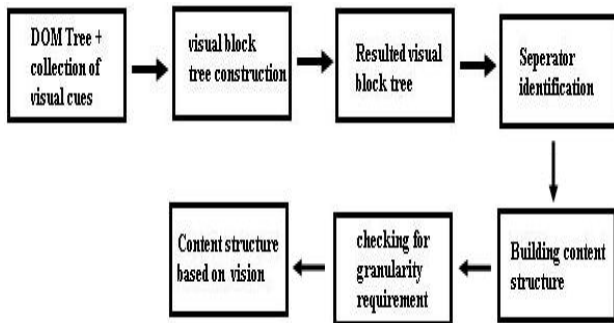
Fig: Flow diagram

IV. PROPOSED SYSTEM

The World Wide Web (WWW) has been an excellent medium for a number of benefits of mankind. It has facilitated communication between people located in any part of the world.

It has provided access to unlimited information in various genres. Apart from a number of services, numerous kinds of entertainment factors can also be found on the Web. But to search information of our interest we have to visit various web applications and thus 100s of web pages. This consumes lots of time. Therefore we are introducing such a system called Dynamic Web System which gives information related to various items under one web application so that you can get the information of your interest on one click. Dynamic Web System is a web application which is useful for the online forum which provides information about given area. In this web application user will get the information of the city from where user logins. The information includes the offers on various items like Mobiles, Electronics, Lifestyles, Kids(Toys), Health related items, Gifts and current news in the city. Here the user need not redirect to other client web application, he gets the readymade information of his interest on our web application itself. Also user gets information related to all kind of material under one web application.

User Friendly: The existing approach dense a very tedious job to the user in order to search a specific information while he is travelling or varying from city to city so taking this work under consideration the proposed system will act user friendly and will also act intellectually in order to give information to the user which he expects and also the user does not need's to move and hunt here and there for that specific information this will definitely make information retrieval very easy and handy to user.

*Dynamic Content Retrieval :-*The which a user will search will be displayed to him dynamically now a days this information is basically provided statically to the user which is the time consuming process and also the information is not updated to a greater extent thus using this app user can retrieve the data dynamically

*Time Utilizing Application:-*As seen before in static system data retrieval requires more time as compared to static system this will definitely help user to find the required data and save time

*No need to surfing on net for information:-*Using this application there is no need to find open number of tab's hunt for information which is required by user the machine will automatically act as a smart machine and will provide the most Frequent data to the user.
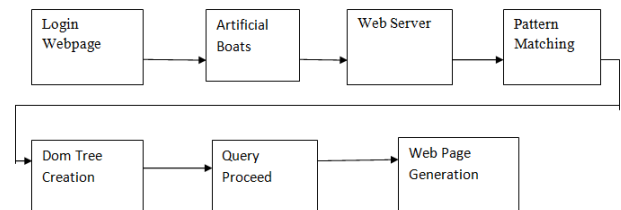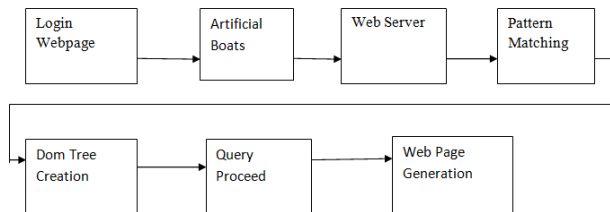
Fig: Block diagram

V. IMPLEMENTATIONS

System use different web mining technology to extract data from remote website at run time. System contain different

client like city malls, news and hospitals. User can easily access all information from this application by just login to system. System to access the area information.

VI. MODULES

1. Location extraction

Location extraction is one of the most important module in this project which will automatically extract the location of the strings or in demand requirements of the user this use no of parameters and algorithms into it extraction module will extract the data thorough service provide which have and unique ip address and also a new working module called rss feed can also be used for data extraction paper studies structured data extraction from Web pages, e.g., online product description pages. Existing approaches to data ex- traction include wrapper induction and automatic methods. In this paper, we propose an instance-based learning method, which performs ex- traction by comparing each new instance to be extracted with labeled instances

2. Artificial bots

Artificial bots are the interacting bots which will act upon the data which is searched by the user and will also act upon that specific data in order to sort the data artificial bots also play a vital role in decision making which they decide and on their own and provide the required data to the user. Decision making is the required need to user which will help bot's to extract the required data from no of web pages which a user extracts by surfing which makes a tedious and time computing for him.

3. Dom parser

The Document Object Model (DOM) is a cross-platform and language-independent convention for representing and interacting with objects in HTML, XHTML and XML documents. The nodes of every document are organized in a tree structure, called the DOM tree. Objects in the DOM tree may be addressed and manipulated by using methods on the objects. The public interface of a DOM is specified in its application programming interface (API).

4. Dom Tree Generation

Content extraction is particularly useful for the visually impaired and blind. A common practice for improving web page accessibility for the visually impaired is to increase font size and decrease screen resolution; however, this also increases the size of the clutter, reducing effectiveness. Screen readers for the blind, like Hal Screen Reader by Dolphin Computer Access or Microsoft's Narrator, don't usually automatically remove such clutter either and often read out full raw HTML. Therefore, both groups benefit from extraction, as less material must be read to obtain the desired results.

5. User interface

A dynamic web page is then reloaded by the user or by a computer program to change some variable content. The updating information could come from the server, or from changes made to that page's DOM. This may or may not truncate the browsing history or create a saved version to go back to, but a dynamic web page update using Ajax technologies will neither create a page to go back to, nor truncate the web browsing history forward of the displayed page. Using Ajax technologies the end user gets one dynamic page managed as a single page in the web browser while the actual web content rendered on that page can vary.

6. Web Page Generation

The most notable difference between static and dynamic models of a system is that while a dynamic model refers to runtime model of the system, static model is the model of the system not during runtime. Another difference lies in the use of differential equations in dynamic model which are conspicuous by their absence in static model. Dynamic modeling is exible as it can change with time as it shows what an object does with many possibilities that might arise in time.

VII. FUTURE SCOPE

- The member should be provided with the quick and easy access to information.
- The member can redirect to product details.
- The members are provided with the online newspaper reading in English.
- Can download information from the web.

VIII. CONCLUSION

Dynamic Web system is a web application which is useful for the online forums which provide information about given area. User only views information of the area where he login or open the website. Many web pages contain excessive clutter around the body of an article. Although much research has been done on content extraction, it is still a relatively new field. Our approach, working with the Document Object Model tree as opposed to raw HTML markup, enables us to perform Content Extraction, identifying and preserving the original data instead of summarizing it. The techniques that we have employed, though simple, are quite effective. As a result, we were able to implement our approach in a publicly-available Web proxy that anyone can use to extract content from HTML web pages for their own purposes.
.

REFERENCES

[1] C.-H. Chang, M. Kayed, M.R. Girgis, and K.A. Shaalan, Survey of Web Information Extraction Systems, IEEE Trans. Knowledge and Data Eng., vol. 18, no.10, pp. 1411-1428, Oct. 2006.

[2] N. Kushmerick, D. Weld, and R. Doorenbos, Wrapper Induction for Information Extraction, Proc. 15th Intl Joint Conf. Arti_cial Intelligence (IJCAI), pp. 729-735, 1997.

[3] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Fact or _ction: Content classi_cation for digital libraries. Smart Media Institute, Department of Computer Science, University College Dublin. August 11, 2002.

[4] A. F. R. Rahman, H. Alam and R. Hartono. Content Extraction from HTML Documents. Document Analysis and Recognition Team (DART). BCL Computers Inc. August 11, 2002.

[5] Wolfgang Reichl, Bob Carpenter, Jennifer Chu-Carroll, Wu Chou Language Modeling for Content Extraction in Human-Computer Dialogues

[6] K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiman, S. Teufel. Columbia Multi-document Summarization: Approach and Evaluation

AUTHORS

First Author – Tarvejahmad Ansari, B.E(Computer), Shatabdi Institute of Engineering & Research. t4tarvej@gmail.com

Second Author – Akshay Pillai, B.E(Computer), Shatabdi Institute of Engineering & Research. Akshaypillai24.7@gmail.com

Third Author – Asif Sayyad, B.E(Computer), Shatabdi Institute of Engineering & Research.