Partial Web Mining Website Mining Individually

Aeesha S. Shaheen

Abstract -The growth of numbers of pages that loaded on the web and the difference of the structures and styles of these pages involves significant challenges. So in this paper we apply web mining to the web site pages in order to benefit from the redundant data that appears in the most pages in the website to decrease the memory size needed to save these pages in each web site alone. The web mining applied by extracting the matched data between the website pages and the home page or the index page then discharges it (the matched data), but the difference data is saved in the server's memory until the page is accessed. After the page is requested by any user, combination applied between the index page file and the difference file for viewing the page on the browser.

Keyword- Pages, data

I. INTRODUCTION

The presence of huge amounts of web data, and massive data warehouses have increased the need to develop tools characterized to analyze the data and extract information and knowledge of them, web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data. The attention paid to web mining, to develop research and software, and web based society.

II. WEB MINING OVERVIEW:

To be able to cope with the abundance of available information, users of the WWW need to rely on intelligent tools that assist them in finding, sorting, and filtering the available information. Just as data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of Web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular in text documents that are published on the Web. Like data mining, Web mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing, and others.

Depending on the nature of the data, one can distinguish three main areas of research within the Web mining community:

Web Content Mining: application of data mining techniques to unstructured or semi-structured data, usually HTML-documents

Web Structure Mining: use of the hyperlink structure of the Web as an (additional) information source.

Web Usage Mining: analysis of user interactions with a Web server (e.g., click-stream analysis). [5][6][10].

III. WEB DATA

The information provided by the data sources described above can all be used to construct/identify several data abstractions, notably users, server sessions, episodes, click-streams, and page views. In order to provide some consistency in the way these terms are defined, the W3C Web Characterization Activity (WCA) has published a draft of Web term definitions relevant to analyzing Web usage. A user is defined as a single individual that is accessing file from one or more Web servers through a browser. While this definition seems trivial, in practice it is very difficult to uniquely and repeatedly identify users. A user may access the Web through different machines, or use more than one agent on a single machine. A page view consists of every file that contributes to the display on a user's browser at one time. Page views are usually associated with a single user action (such as a mouse-click) and can consist of several files such as frames, graphics, and scripts. When discussing and analyzing user behaviors, it is really the aggregate page view that is importance. The user does not explicitly ask for "n" frames and "m" graphics to be loaded into his or her browser, the user requests a "Web page." All of the information to determine which files constitute a page view is accessible from the Web server. [4].

IV. HTTP ARCHIVE

According to new research from HTTP Archive, which regularly scans the internet's most popular destinations, the average size of a single web page is now 965 kilobytes (KB), up more than 30% from last year's average of 702KB. This rapid growth is fairly normal for the internet - the average web page was 14KB in 1995, 93KB by 2003, and 300KB in 2008 - but by burrowing a little deeper into HTTP Archive's recent data, we can discern some interesting trends. Between 2010 and 2011, the average amount of Flash content downloaded stayed exactly the same 90KB but JavaScript experienced massive growth from 113KB to 172KB. The amount of HTML, CSS, and images on websites also showed a significant increase year over year. There is absolutely no doubt that these trends are attributable to the death throes of Flash and emergence of HTML5 and its open web cohorts. [2]

V. PARTIAL WEB MINING

There is thousands of websites that consist of millions of pages and more. Yahoo is now claiming nearly 20 billion pages in its index, making it the largest index on the web. I have it on good authority that Google disputes this, but could not connect to Google tonight. This has the makings of a major pissing match. [8]

There is a lot of confusion about just how many words should go on a web page. Some businesses want to put hardly any words on a page, and others want to load up a page with keywords and information. The answer is two:

Manuscript received on September, 2013.

Aeesha S. Shaheen, Lecturer Assistant Computer Sciences & Mathematics College Mosul University, IRAQ.

For search purposes, Google recommends 250-300 words per page. That's enough text for Google to discern what the page is about and index it properly. The second answer is to put in as many words as it takes to convert the visitor to taking the next step - whether it be placing an order, picking up the phone, signing up for a newsletter, or simply clicking to another page. For some pages that may mean 250 words and for others it may mean 2500 or more!

On each web page, try to use a minimum of 250 words for search engine purposes, and then feel free to use as many words as necessary to get your visitor to take the next step. [1][3][9]

The home page (index.html) of Google.com size is 95KB and the size of Google image page 97KB, the difference between the two pages is the picture of the logo of Google and the rest of the page is exactly the same in the two pages but the difference between the index.html page which size 95KB and Google translate page that size is 88KB is more Obvious and the size is less by replacing the image with text and the

Figure (1) Google home or index page

Figure (2) Google image page

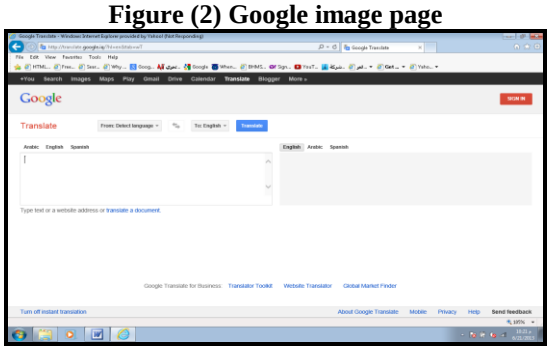

Figure (3) Google translate page

So the idea is to remove most of the redundant data in all the pages that matched the home page more than 35% to decrease the memory that needed to save these pages. The page may be matched the home page much more than this percentage like the previous example the matched percentage between the Google index page and the Google image page is 97%, according to that the total saved memory for each web site alone not less than 35% from the memory needed to save the web site, what can call it Partial web mining

VI. WEBSITE MINING INDIVIDUALLY

Website mining applied for the biggest size websites by save the difference of any page of the website with the most used page or the index page in the same website then the result file manipulated as matrix with row numbers equal to the number of difference rows in the result file column equal to the numbers of characters in each row in the result file. When this page is accessed by the search engine and before displayed on browser must merge the matrix with the home page then view it on browser.

VII. PRACTICAL PART

Partial web mining applied on experimental and simple website called testweb consist of forty pages, eighteen page for the English version of the testweb and eighteen page for the Arabic version of the testweb. The first step is to find the matched percentage of each page in testweb with the index page as shown in the figure (4)

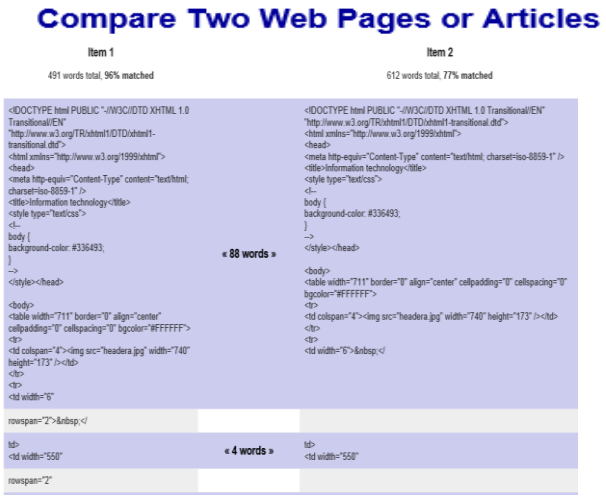

Figure (4) the matched result between 2.html page and index.html page

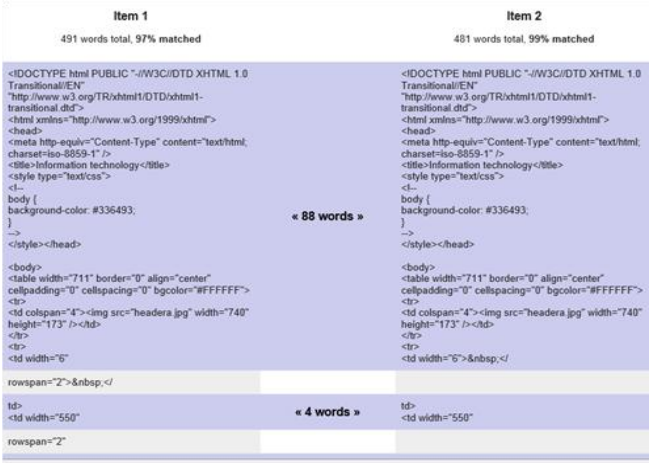

Figure (5) the matched result between 3.html page and index.html page

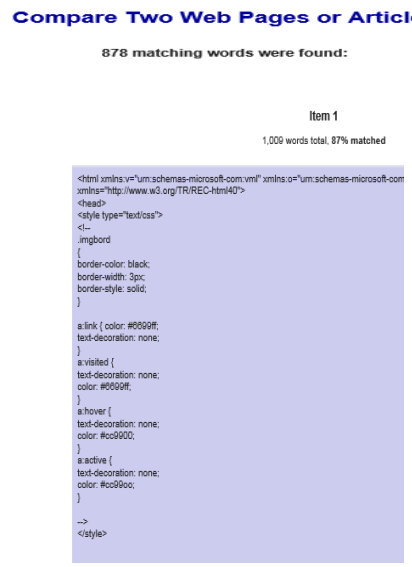

Figure (6) the matched result between 4.html page and index.html page

After finding the matching percentage between all the testweb pages with the home page we obtain table number (1) for the page in English language and table number (2) for the pages in Arabic language

No.	Page name	Matched percentage
1-	2.html	**77%**
2-	3.html	**99%**
3-	4.html	**80%**
4-	5.html	**53%**
5-	6.html	**35%**
6-	7.html	**38%**
7-	8.html	**40%**
8-	aims.html	**67%**
9-	Contactus.html	**93%**
10-	Deadlines.html	**75%**
11-	Form.html	**71%**
12-	Important-date.html	**75%**
13-	Instruction.html	**71%**
14-	Our-college.html	**84%**
15-	Participation.html	**90%**
16-	Preparation.html	**79%**
17-	Scientific-committee.html	**88%**
18-	Topic.html	**80%**

Table no. (1) The match result for English pages

No.	Page name	Matched percentage
1-	2a.html	**74%**
2-	3a.html	**96%**
3-	4a.html	**81%**
4-	5a.html	**44%**
5-	6a.html	**41%**
6-	7a.html	**33%**
7-	8a.html	**35%**
8-	Aims-a.html	**61%**
9-	Contactus-a.html	**93%**
10-	Deadlines-a.html	**75%**
11-	Form-a.html	**82%**
12-	Important-date-a.html	**75%**
13-	Instruction-a.html	**79%**
14-	Our-college-a.html	**80%**
15-	Participation-a.html	**88%**
16-	Preparation-a.html	**79%**
17-	Scientific-committee-a.html	**88%**
18-	Topic-a.html	**75%**

Table no. (2) The match result for Arabic pages

After we get the difference file for each page with the home page, the second step is to convert the difference file into matrix and save it, until the page accessed by any user request led to combine the matrix of difference with the index page as text file then save the file of combination as html to view on the browser.

VIII. RESULTS DISCUSSION

The average of matched percentage for all the pages in table no.(1) and table no.(2) will be 71.5% this result for testweb website. The other hand if the size of testweb is 468MB so when applying the Partial web mining can save 71.5% of 468Mb which equals to 334.62MB for experimental website consist from 38 pages only. We can imagine the size of the save memory if matched pages to home page more than 100 pages or 1000 pages and may be million or more…..

The disadvantage of the Partial website mining is the time spends through the combination between the difference matrix and the index page or the home page.

IX. CONCLUSION

From the above discussion we see it is very difficult to find and extract the useful data or information from www by applying the same method or algorithm because of the huge amount and a huge difference among these pages. The Partial web mining may be the solution some times.

REFERENCES

[1] http://novella.mhhe.com/sites/0079876543/ student_view0/power_google.html

[2] http://tech.slashdot.org/submission/1888512/ average-web-page-is-now- almost -1mb?sdsrc=rel

[3] http://www.andesandassociates.com/ How_many_words_on_web_pa.html.

[4] Jaideep Srivastava, , Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" Department of Computer Science and EngineeringUniversity of Minnesota 200 Union St SE Minneapolis, MN 55455 fsrivasta,cooley,deshpand,ptang@cs.umn.edu

[5] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar," Web Mining - Concepts, Applications & Research Directions Department of Computer Science200 Union Street SE, 4-192, EE/CSC Building University of Minnesota, Minneapolis, MN 55455, USA srivasta, desikan, kumar @cs.umn.edu

[6] Johannes F¨urnkranz, "Web Structure Mining Exploiting the Graph Structure of the World-Wide Web"Austrian Research Institute for Artificial IntelligenceSchottengasse 3, A-1010 Wien, AustriaE-mail: juffi@oefai.at

[7] John Wiley & Sons, Inc. "DATA MINING THE WEB,Uncovering Patterns in Web Content, Structure,and Usage"All rights reserved Copyright 2007.Published by John Wiley & Sons, Inc., Hoboken, New JerseyPublished Simultaneously in Canada

[8] Katherine Andes http://battellemedia.com/archives/2005/08/ how_many_pages_does_yahoo_index.php

[9] McGraw-Hill/Dushkin, "Power Google"Copyright ©2003 Guilford, CT 06437, A Division of The McGraw-Hill Companies.

[10] Miguel Gomes da Costa Júnior Zhiguo Gong, "Web Structure Mining: An Introduction "Department of Computer and information Science Faculty of Science and Technology University of Macau Av. Padre Tomás, S.J., Taipa, Macao S.A.R., China {mcosta, fstzgg}@umac.mo