Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 4, Issue, 04, pp.212-215, April, 2012

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

ELIMINATING OF PICTURE ANIMATION FROM WEB SHEET

[1,*] Sivakumar, P. and [2]Parvathi, R.M.S.

[1]Department of Computer Science and Engineering, Sengunthar College of Engineering, Tamilnadu, India.
[2]Department of Computer Science and Engineering, KSR College of Engineering, Namakkal, Tamilnadu, India.

ARTICLE INFO

Article History:
Received 27th January, 2012
Received in revised form
29th February, 2012
Accepted 24th March, 2012
Published online 30th April, 2012

Key words:
Image Advertisement,
HTML Tag Differentiator,
Rule Based Classifier,
Web Content Mining.

ABSTRACT

By the unusual growth of the web, there is an increasing volume of data and information published in numerous web-pages. From this we understood that web is noisy. A web page contains a mixture of many kind of information e.g. mainly contains, advertisements, navigational panels, copy right blocks etc… in a particular application only part of information is useful and the rest are noise. These all mischief web mining. Advertisements and Sponsor images are not much important in surfing. We need a technique that to keep common navigation structure as it is but remove image advertisement and improve surfing efficiency. In this paper a small application HTML Tag Differentiator rule based is created which removes image advertisement.

INTRODUCTION

These days, most information resources on the WWW are published as HTML or XML pages, and the number of the web pages is increasing with the expansion of the web, To make enhanced use of web information, technologies that can automatically rearrange and manipulate web pages are pursued such as web information retrieval, web page classification and other web mining work. By the growth of information sources available on the World Wide Web, then it is necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. It is essential to detect and remove noise in image advertisement which distract the user from web page's actual content or and waste bandwidth. It reduces reliability of information on web by increasing presence of advertisement.[1] Many web sites draw income from third-party advertisements, images are scattered throughout the site's pages. If judged to be interesting or relevant, users can click on these so-called “banner advertisements”, jumping to the advertiser's own site [2]. Some users also prefer not to view such advertisements. Images tend to dominate a page's total download time, so users connecting through slow links find that advertisements considerably slow down their browsing. Some users dislike paying for services, instead of preferring direct payment for services rendered. Finally, some users disagree with the very view of advertising on the public Internet [2].

LITERATURE REVIEW

Web advertising is to attract potential customers to the advertiser's Web site and/or by placing promotional content to strengthen brand recognition and a link on other Web sites.

Some examples of popular ad types that are currently found on the web.

1) Banner Advertisement
2) Text Advertisement
3) Video Advertisement
4) Pop-Up Advertisement
5) Interstitial Advertisement
6) Content Sponsoring Advertisement

Different approaches for Image Advertisement Removal can be matched with proposed system.

Learning to Remove Internet Advertisement

Ad Eater system is a browsing assistant, that learns advertisement detection rules automatically and then applies those rules to remove advertisements from web pages during browsing [2]. It has features as

1. Some users prefer one sided error (when in doubt leave image in act). There is no any way to bias Ad Eater in this manner.
2. Ad Eater system classifies any image as advertisement or non-advertisement, but there is no any confidence in classification e.g. what is the % of confidence to classify any image as advertisement or non-advertisement.
3. Ad Eater is not incremental system; classifier is modified based on update to the training instances.

Internet Junk buster

The Internet Junk buster is a proxy that is specifically designed to block advertising banners (specified by URL

***Corresponding author:** sivakumarphd2010@gmail.com

regular expression matching) and cookies. It gets the job done remarkably well. Runs on UNIX, Windows NT and Windows 95 and is free, including the source code [3].

Muffin

Muffin is a free filtering proxy for the web written in Java; runs on all platforms. Similar to Junk buster, but more flexible, portable and powerful. It supports several "filters", one of which can delete images based on their width/height ratio (banner ads) and another one allows modifying the incoming HTML stream using a simple language, allowing for stripping other ads [3].

Web Washer

Web Washer, free for personal use, is a high quality ad filtering personal proxy, written by Siemens. It removes ad banners based on size, gets rid of pop-up windows and stops animated graphics.

Eliminating Noisy Information in Web Pages for Data Mining

They propose a noise elimination technique based on following observation:

In a given web site, noise block share some common contents and presentation styles, main content blocks of the pages are often diverse in their actual contents and/or presentation styles. Based on this observation, they propose a tree structure called Style tree, to capture the common presentation styles and actual content of the pages in a given web site. By sampling the pages of the site, a style tree can be built for the site, which can call the Site Style (SST). They Introduce an Information based measure to determine which parts of the SST represent noises and which part represent the main contents of the site. The SST is employed to detect and eliminate noises in any web page of the site by mapping this page to the SST. The proposed technique is evaluated with two data mining tasks, web page clustering and classification.

PROPOSED ARCHITECTURE

Partial implementation of a proposed system Removal of Image Advertisement from Web page, HTML Tag Differentiator is being developed. The Image Advertisement removal consists of 4 major steps. The following fig 1 shows the proposed architecture.

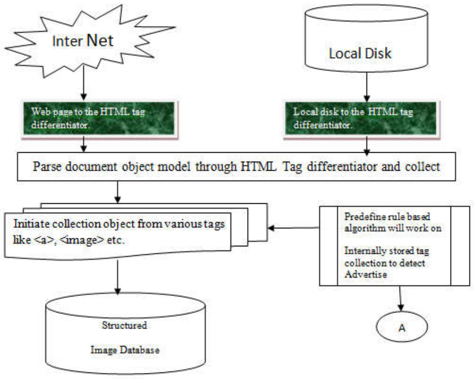

Inter Net
Local Disk
Web page to the HTML tag differentiator.
Local disk to the HTML tag differentiator.
Parse document object model through HTML Tag differentiator and collect
Initiate collection object from various tags like <a>, <image> etc.
Predefine rule based algorithm will work on
Internally stored tag collection to detect Advertise
Structured Image Database
A

Fig 1. HTML Tag Differentiator Architecture

The major 4 steps carried out in HTML Tag Differentiator

Step 1: Mechanism for Detecting Image from web page. The proposed system starts with Web Page parsing process. In this step, the web page is traversed and Document Object Model [DOM] structure of web page is obtained.

Step 2: Mechanism for extracting properties of Image and storing in database. System will collect information about all properties of Image inside <IMG> tag. To achieve accuracy we consider <IMG> tag which is inside <A> tag that is responsible for diversifying users from main content. System will collect properties like Name of Image file, Alt text, Source Url, Height of Image, Width of Image, Aspect Ratio of Image etc.

Step 3: Mechanism for detecting Image is Ad or Non-Ad. This part of system functionality is core performance piece of the system. The actual detection of image will carried out here. Rules are applied to the images to decide that image is ad or non-ad. These rules collected through various theoretical and practical references and observation respectively. Here seven rules are
Defined

Step 4: Mechanism for Removal Advertisements. On execution of third step system will decide whether image is advertise or not and according to that result system will remove that image from the given page.

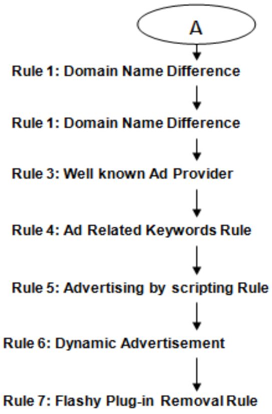

A
Rule 1: Domain Name Difference
Rule 1: Domain Name Difference
Rule 3: Well known Ad Provider
Rule 4: Ad Related Keywords Rule
Rule 5: Advertising by scripting Rule
Rule 6: Dynamic Advertisement
Rule 7: Flashy Plug-in Removal Rule

Fig. 2. Rule Based Classifier

IMPLEMENTATION METHODOLOGY

The proposed system is implemented in C# .net as a front end and SQL Server 2005 as back end. In the proposed System, image advertisement are removed from web page, this has two phases.

Image Advertisement Detection
Image Advertisement Removal

Image Advertisement Detection

This phase is implemented in two parts.

The HTML Tag
Differentiator Rule based Classifier

The HTML Tag Differentiator

The HTML Tag Differentiator parses web document or web page through it and differentiates tags with help of Document

Object model. This system will open a web document either from local disk or Internet. The steps are implemented to detect the image advertisement.
Step 1: Provide web page link.
Step 2: Parse page through HTML Page Differentiator.
Step 3: Create collection object of each tags.
Step 4: Fetch Collection object for <A>, <IMG>, <IFRAME> etc.
Step 5: All relevant feature or properties of tag supplied as an input of Predefine Rule based image Classifier.
Step 6: Follow all rule function to decide whether image is an advertisement or not.
Step 7: Set Ad flag true or false accordingly.
Step 8: Insert all feature or properties of image tag like Src, alt, height, width, aspect ratio, to the underlying structured database for future use.
Step 9: Remove an advertise Image from web document.

Rule based Classifier

In Rule based classifier, seven various rules are define that leads system to decide whether image is an advertise image or not. It’s long learning process. Rules to remove Image Advertisement are as follows:

Rule 1: Domain Name difference Rule

Image stored on different location. If first part of the image URL is different from web page URL then it considered as Image Advertisement. Relevant or irrelevant images are differentiate from their URL name.

Rule 2: Dimension Rule

Block ad banners based on their size. Certain image dimensions are strong clues for ads like 468 X 60 pixels banners e.g. 150 X 500 pixels , 120 X 600 pixels, 160 X 600 pixels

Rule 3: Well-known Ad Provider Rule

Block Content that comes from Well-known ad providers. This rule is implemented by matching content URLs against the Domain names of wellknown ad providers.

Rule 4: Ad Related Keywords Rule

Block images based on ad-related keywords in their URL. Good clue words and phrases were obtained from a study of random commercial web pages. Examples “ad”, “Free”, “now”, “buy”, “join”, “shop”, “click here”, “advertisement”, “soon” etc.

Rule 5: Advertising by scripting

One of the applications of <script> tag and web scripting language like JavaScript is to incorporate advertisement into the web page from well known advertise provider like AdSense, AdChoice etc.

Rule 6: Dynamic Advertisement Rule

The <INS> tag is used to indicate content that is inserted into a page and indicates changes to a document. Clients that aware of this tag may choose to display content inside this tag differently or not at all depending on what theyare designed to do. INS is semantic tag describing something that is inserted to the text after the text was already published.

Rule 7: Flashy Plug-in Removal Rule

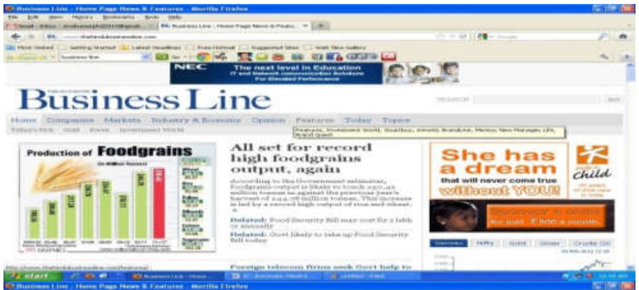
<EMBED> puts a browser plug-in in the page. A *plug-in* is a special program located on the client computer that handles its own special type of data file. The most common plug-ins are for sounds and movies. The <EMBED> tag gives the location of a data file that the plug-in should handle. The <object> tag is used to include objects such as images, audio, videos, Java applets, ActiveX, PDF, and Flash.

RESULTS AND DISCUSSION

HTML tag differentiator is tested with different categories of 50 web sites. In these web sites total 142 image-advertisements are found in ordinary web browser and 139 image advertisements are removed from proposed tool HTML Tag Differentiator browser. For random webpage, result samples are displayed in below table.

Website Name	Total Image Ad Present in Ordinary Browser	Total Image Ad Removed in Project Browser	Ads Classified as Non Ads	Non Ad Classified as Ads
www.freshersworld.com	05	05	00	03
www.yahoo.com	01	01	00	00
www.chennaionline.com	03	03	00	00
www.studyguideindia.com	01	01	00	00
www.studyabroaduniversities.com	00	00	00	02
www.maplandia.com	06	06	00	00

The following fig 3 and fig 4 shows the result before and after Image advertisement removal.


NEC
The next level in Education
Business Line
Production of Foodgrains
All set for record high foodgrains output, again
She has a dream
that will never come true
without YOU!
Child

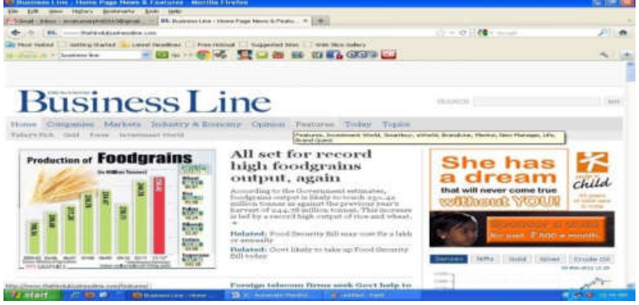

Business Line
Production of Foodgrains
All set for record high foodgrains output, again
She has a dream
that will never come true
without YOU!
Child

Fig 4: After removal of Image advertisement

CONCLUSION

To remove image advertisement from web page, HTML TAG Differentiator system is developed. It is automatically detects and removes Image advertisements from web pages using Rule based classifier. To accomplish the following objectives seven rules are implemented.

Detect Image advertisement, Bring confidence in Image Advertisement detection task, Remove Image advertisement, Improve efficiency of Surfing. HTML TAG Differentiator system achieves 97% of accuracy. Proposed system using web content mining helps to remove Image advertisement from given web page which prevents user to be expand from same. It prepares structured database of website and its image advertisement properties which is useful for further research. Our results show that the proposed system is highly effective. As technology is changing, a way to insert image advertisement is also changing. To cope up with new technology, to detect advertise on web by analyzing a view source of web pages there is a need to find new method and convert that into well defined rule. That leads to achieve maximum hit rate to find image advertise.

REFERENCES

[1] Bar-Yossef, Z. and Rajagopalan, S., Template Detection via Data Mining and its Applications, In the proceedings of Eleventh World Wide Web conference (WWW 2002), May 2002.

[2] Learning To remove Internet Advertisement,Nicholas Kashmerick 3rd Int. Conf. of Autonomous Agent, 1999

[3] Lan Yi et al "Eliminating Noisy Information in Web Pages for Data Mining" SIGKDD .03, August 24-27, 2003,

[4] Lan Yi, Bing Liu, Xiaoli Li "Eliminating Noisy Information in Web Pages for Data Mining," ACM SIGKDD '03, pp.,1-10,August 24-27, 2003.

[5] Thanda Htwe "Cleaning Various Noise Patterns in Web Pages for Web Data Extraction," International Journal of Network and Mobile Technologies ISSN 1832- 6758 Electronic Version VOL 1 / ISSUE 2 / NOVEMBER 2010.

****