Removal of Image Advertisement from Web Page

Hetal R. Parmar
Post Graduate Student, ME Computer Engineering
Thadomal Shahani Engg. College
Bandra, Mumbai University, India

Prof. Jayant Gadge
Head, Computer Engineering Department
Thadomal Shahani Engg. College Bandra, Mumbai
University, India

ABSTRACT

With the phenomenal growth of the web, there is an ever increasing volume of data and information published in numerous web-pages. It is said that web is noisy. A web page typically contains a mixture of many kind of information e.g. main contains, advertisements, navigational panels, copy right blocks etc… for a particular application only part of information is useful and the rest are noise. These all seriously harm web mining. Advertisements and Sponsor images are not much important in surfing.

As there is a need of technique that keep common navigation structure as it is but removes image advertisement and improve surfing efficiency. In this paper a small application HTML Tag Differentiator is created which removes image advertisement using rule based classifier.

General Terms

Data Mining, Web Content Mining, Image Advertisement

Keywords

Image Advertisement, HTML Tag Differentiator, Rule Based Classifier, Web Content Mining

1. INTRODUCTION

Today, most information resources on the WWW are published as HTML or XML pages, and the number of the web pages is increasing rapidly with the expansion of the web, In order to make better use of web information, technologies that can automatically re-organize and manipulate web pages are pursued such as web information retrieval, web page classification and other web mining work.

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. It is necessary to detect and remove noise specially image advertisement which distract the user from web page's actual content or and waste bandwidth. This also reduces reliability of information on web by increasing presence of advertisement[1].

Many Internet sites draw income from third-party advertisements, usually in the form of images sprinkled 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 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 substantially impede their browsing. Other users dislike paying for services indirectly through advertisers, preferring direct payment for services rendered. Finally, some users disagree with the very notion of advertising on the public Internet[2].

2. LITERATURE REVIEW

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

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

a) Banner Advertisement

b) Text Advertisement

c) Video Advertisement

d) Pop-Up Advertisement

e) Interstitial Advertisement

f) Content Sponsoring Advertisement

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

2.1 Learning to Remove Internet Advertisement

AdEater system is a browsing assistant that automatically learns advertisement detection rules and then applies those rules to remove advertisements from internet pages during browsing[2].

AdEater system lakes by below features

1. Some users prefer one sided error (when in doubt leave image in act). There is no any way to bias AdEater in this manner.

2. AdEater 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. AdEater is not incremental system. In incremental system classifier is modified based on update to the training instances.

2.2 Internet Junk buster

The Internet Junkbuster is a lean and mean proxy that is specifically designed to block advertising banners (specified by URL 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].

2.3 Muffin

Muffin is a free filtering proxy for the web written in Java; runs on all platforms. Similar to Junkbuster, 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].

2.4 Web Washer

Web Washer, written by Siemens, is a high quality ad filtering personal proxy, free for personal use. It removes ad banners based on size, gets rid of pop-up windows and stops animated graphics. It only works on Windows[3].

2.5 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 usually share some common contents and presentation styles, while the 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 then 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[4].

3. PROPOSED ARCHITECTURE

As a part of partial implementation 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.

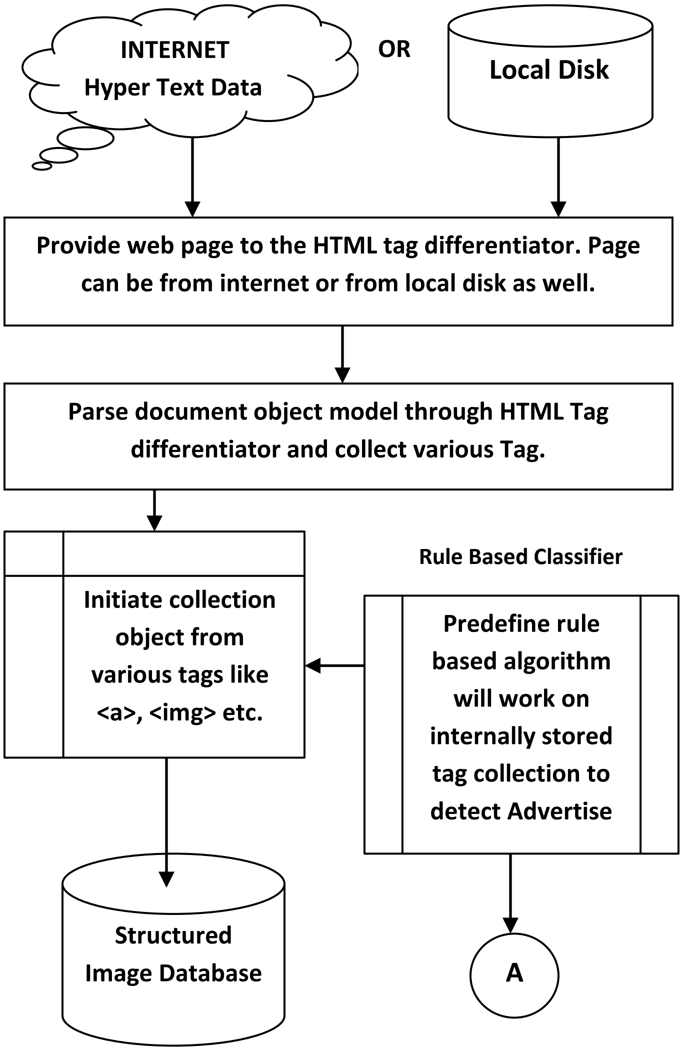

INTERNET
Hyper Text Data
OR
Local Disk
Provide web page to the HTML tag differentiator. Page can be from internet or from local disk as well.
Parse document object model through HTML Tag differentiator and collect various Tag.
Initiate collection object from various tags like <a>, <img> etc.
Rule Based Classifier
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.

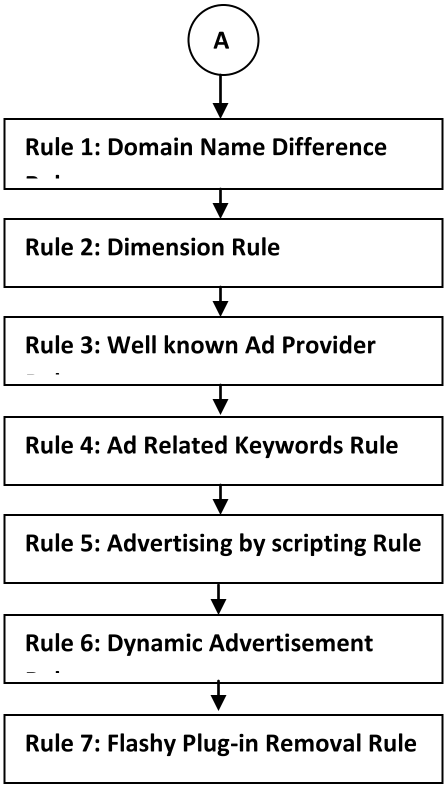

A
Rule 1: Domain Name Difference
Rule 2: Dimension Rule
Rule 3: Well known Ad Provider
Rule 4: Ad Related Keywords Rule
Rule 5: Advertising by scripting Rule
Rule 6: Dynamic Advertisement
Rule 7: Flashy Plug-in Removal Rule

Fig. 2 Rule Based Classifier

4. 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

4.1 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.

4.2 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 well-known 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 they are 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.

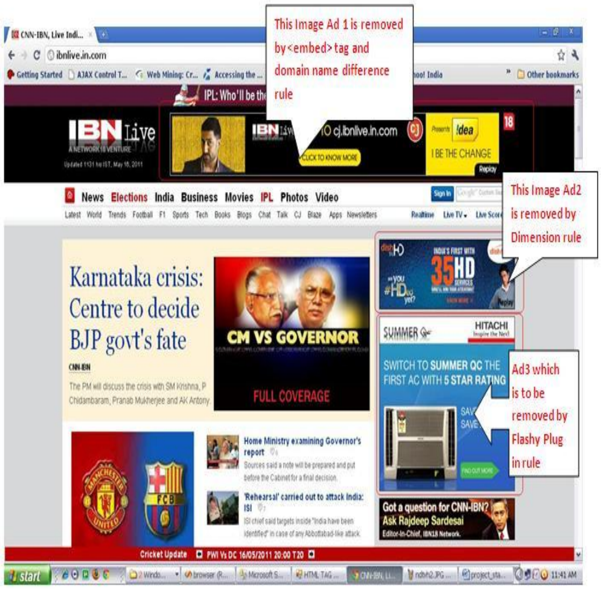
5. RESULTS AND DISCUSSION

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 web-pages, result samples are displayed in below table.

Table 1. Computational Result

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.ibnlive.in.com	05	05	00	03
www.indianrail.gov.in	01	01	00	00
www.pharmabiz.com	03	03	00	00
www.jeevansathi.com	01	01	00	00
www.bestwebbuys.com	00	00	00	02
www.myiris.com	06	06	00	00
www.niit.com	00	00	00	01
www.jobs.oneindia.in	02	01	01	00
www.realestate.com	02	02	00	01

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


This Image Ad 1 is removed by<embed> tag and domain name difference rule

This Image Ad2 is removed by Dimension rule

Ad3 which is to be removed by Flashy Plug in rule

Fig 3: Before removal of image advertisement

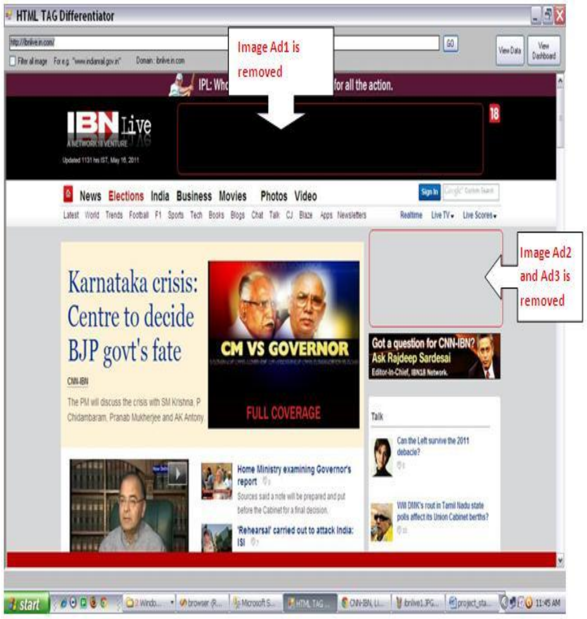

HTML TAG Differentiator
Image Ad1 is removed
Image Ad2 and Ad3 is removed
IBN Live
Karnataka crisis: Centre to decide BJP govt's fate
CM VS GOVERNOR
FULL COVERAGE
Got a question for CNN-IBN? Ask Rajdeep Sardesai
News Elections India Business Movies Photos Video
View Data
View Dashboard
GO
start

Fig 4: After removal of Image advertisement

6. CONCLUSION

In order to remove image advertisement from web page HTML TAG Differentiator system is developed which automatically detects and removes Image advertisements from web pages using Rule based classifier. To accomplish the following objectives seven rules are defined and 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 diversify 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 day by day technology is changing, a way to insert image advertisement is also changing. To cope up with new technology, there is a need to find new method to detect advertise on web by analyzing a view source of web pages and convert that into to well defined rule. That leads to achieve maximum hit rate to find image advertise.

7. REFERENCES

[1] Web Mining
http://www.cs.ualberta.ca/~tszhu/webmining.htm

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

[3] Filtering the Web using WebFilter

http://www-math.uni-paderborn.de/~axel/NoShit/

[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] Viktor Krammer "An Effective Defense against Intrusive Web Advertising," IEEE Sixth Annual Conference on Privacy, Security and Trust, pp. 3-14, 2008.

[6] Neil C. Rowe, Jim Coffman, Yilmaz Degirmenci "Automatic Removal of Advertising from Web-Page Display," ACM JCDL'02, pp.406, July 13-17, 2002.

[7] Web Banner - Wikipedia
http://en.wikipedia.org/wiki/Banner_advertising

[8] Editorial: Special Issue on Web Content Mining SIGKDD Explorations. Volume 6,Issue 2 - Page 1, 2003

[9] Why Block Ads?
http://feh.net/adblock/adblock.shtml