AD-EXTRACTOR TOOL

Developer: Lalit Agarwal

About Ad-Extractor

- A tool to extract and identify advertisements from a given list of webpages.

- Tool extracts both image and textual ads.

- The tool outputs an excel file contain the information of the ads present on the webpages

- The tool was developed as a part of a research study.

Motivation

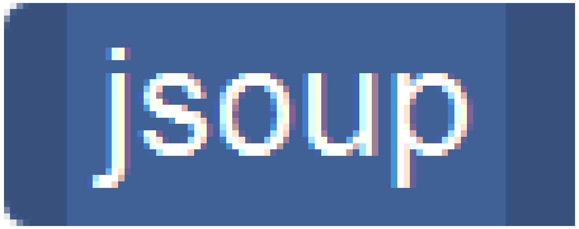
- Understand the nature of advertisements being shown to the users.
- On the basis of data collected, identify any common features to identify and block advertisements on the basis of categories.
- Identify any ads which users find inappropriate or embarrassing.

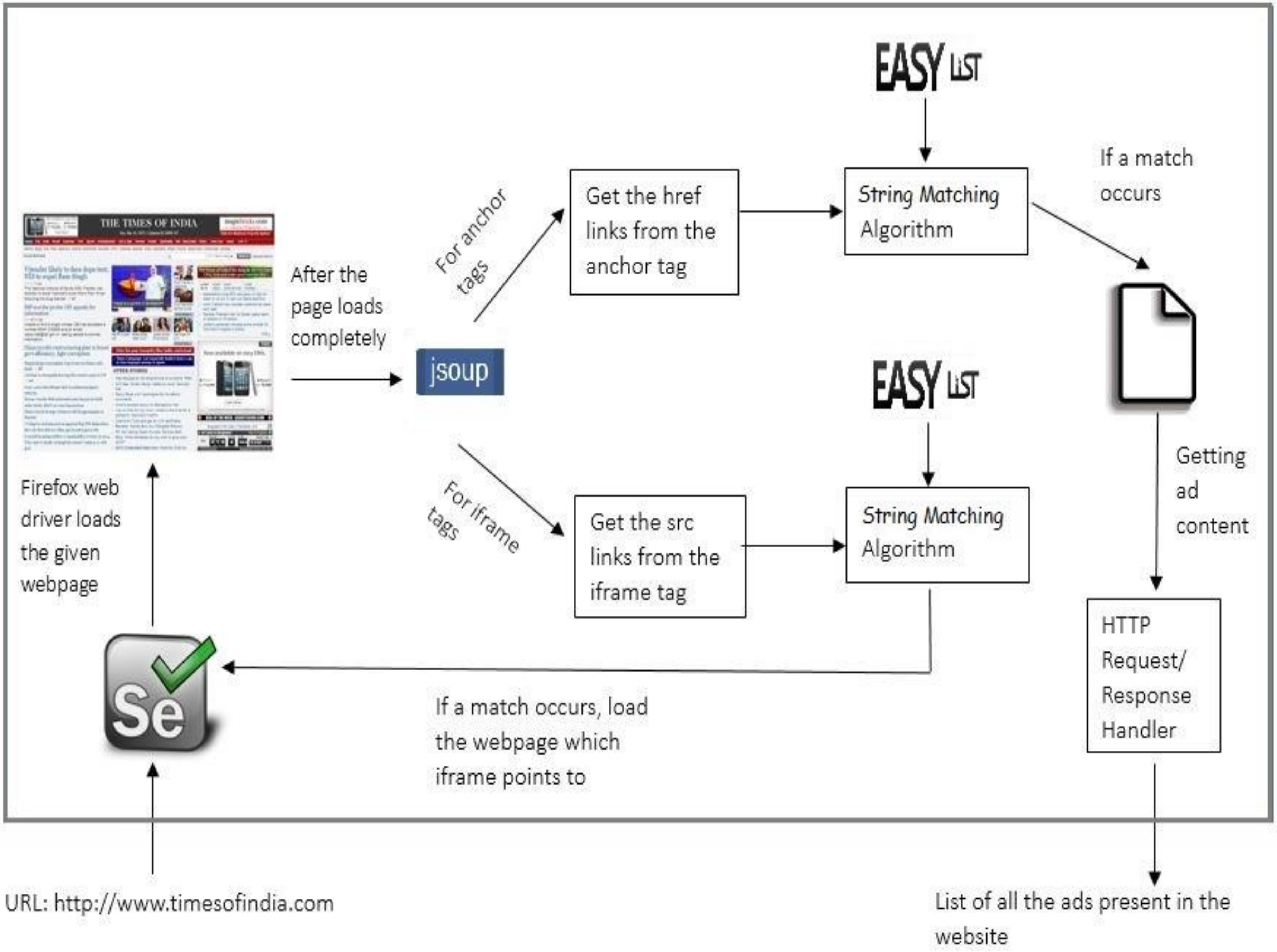
Methodology

- Collected history from user’s browser after their consent.

- We ran this tool on three different set of webpages containing 500, 2500 and 5000 URLs respectively collected from user’s browsing history during the user study.

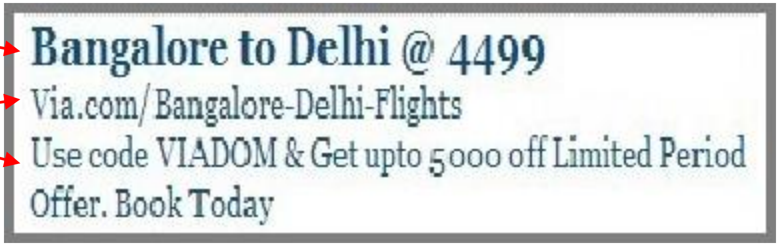
Tools Used



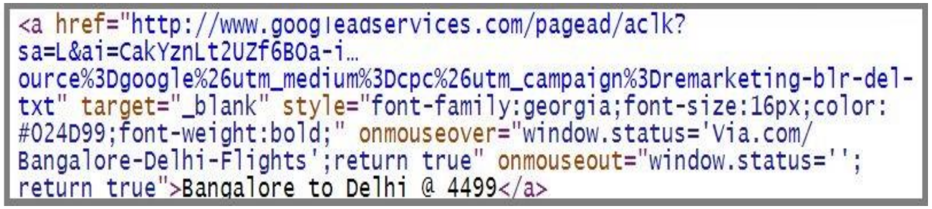

EASY LIST
If a match occurs
Get the href links from the anchor tag
String Matching Algorithm
For anchor tags
After the page loads completely
jsoup
THE TIMES OF INDIA
EASY LIST
Getting ad content
For iframe tags
Get the src links from the iframe tag
String Matching Algorithm
Firefox web driver loads the given webpage
HTTP Request/ Response Handler
If a match occurs, load the webpage which iframe points to
Se
URL: http://www.timesofindia.com
List of all the ads present in the website

Data collected

- Ad-Title
- Ad-Content
- Ad-Display URL
- Ad-Source URL
- Landing Page Title
- Landing Page URL
- Image Source
- URL of the main page
- isThirdParty?
- isIFrame?



Bangalore to Delhi @ 4499
Via.com/Bangalore-Delhi-Flights
Use code VIADOM & Get upto 5000 off Limited Period Offer. Book Today



```
<a href="http://www.googleadservices.com/pagead/aclk?
sa=L&ai=CakYznLt2UZf6BOa-i…
ource%3Dgoogle%26utm_medium%3Dcpc%26utm_campaign%3Dremarketing-blr-del-
txt" target="_blank" style="font-family:georgia;font-size:16px;color:
#024D99;font-weight:bold;" onmouseover="window.status='Via.com/
Bangalore-Delhi-Flights';return true" onmouseout="window.status='';
return true">Bangalore to Delhi @ 4499</a>
```

Simulation Results

- Avg. time taken to get all the ads/webpage: 8 sec
- Avg. time taken to
  - Load a web-page: 4.5 sec
  - Fetch all anchor tags/webpage: 0.5 sec
  - Fetch all iframe tags/webpage: 4.3 sec
- Avg. no. of anchor tags in a webpage: 290
- Avg. no. of iframe tags in a webpage: 4

Results

Data Set	Text Ads	Embarrassing Text Ads	Image Ads	Embarrassing Image Ads
Set 1 (500 URLs)	192	4	156	5
Set 2 (2500 URLs)	1235	29	742	16
Set 3 (5000 URLs)	2587	40	1423	30
Total	**4014**	**73 (2%)**	**2321**	**51 (2%)**

Embarrassing Ads

Text Ads

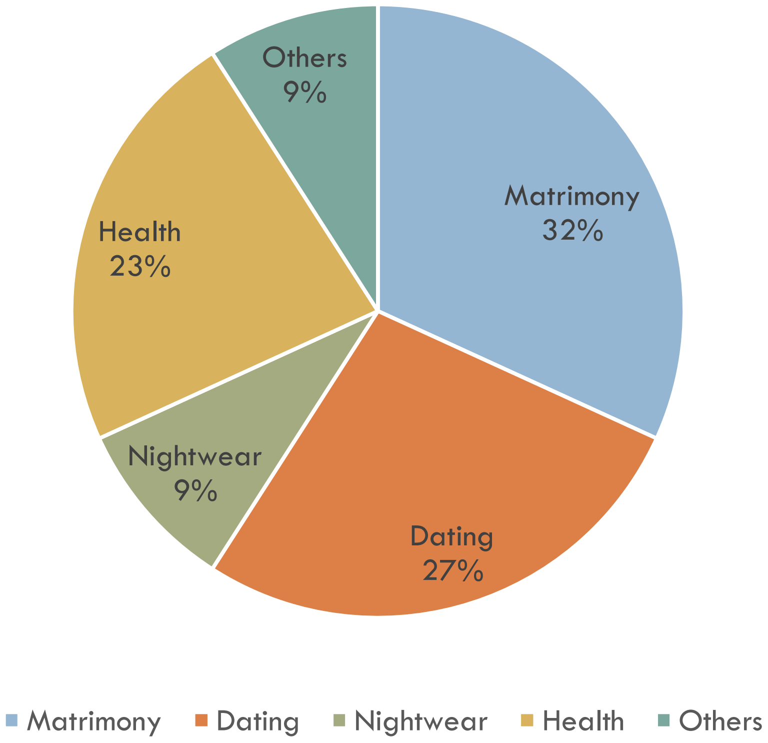
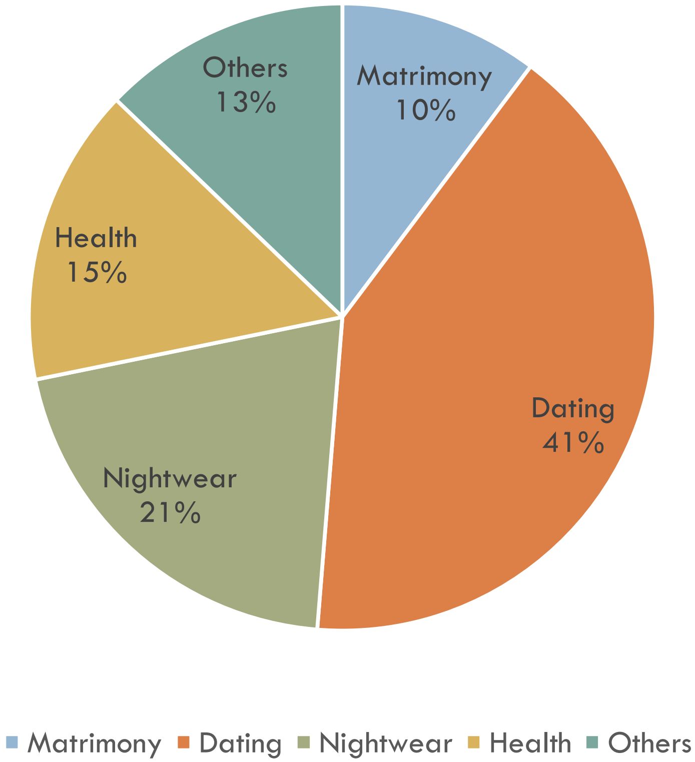


Image Ads



Limitations of the tool

- The tool only identifies textual and image ads. It does not identify flash ads.
- Ads are loaded using Javascript, the tool waits for the entire webpage to load before it can extract the ads.
- Headless browsers tool which can extract ads loaded using Javascript are currently not available.

Acknowledgement

I would like to thank Dr. Saurabh Panjwani, Dr. Sharad Jaiswal and Dr. Nisheeth Shrivastava (Bell Labs, India) for their constant feedback. The tool would not have been possible without their guidance and support.

References

[1] F. Roesner, T. Kohno, and D. Wetherall. Detecting and Defending Against Third-Party Tracking on the Web. In Proc. of NSDI, 2012

[2] J. Mayer and J. Mitchell. Third-party web tracking: Policy and technology. In Proc. of IEEE Symposium on Security and Privacy, 2012.

[3] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In Proc. SOUPS, 2012.