Web Bots that Mimic Human Browsing Behavior on Previously Unvisited Web-Sites: Feasibility Study and Security Implications

Y. Yang, N. Vlajic, U. T. Nguyen
Dept. of Electrical Engineering & Computer Science, York University, Toronto, Canada
yangcs@cse.yorku.ca, vlajic@cse.yorku.ca, utn@cse.yorku.ca

Abstract — In the past, there have been many attempts at developing accurate models of human-like browsing behavior. However, most of these attempts/models suffer from one of following drawbacks: they either require that some previous history of actual human browsing on the target web-site be available (which often is not the case); or, they assume that 'think times' and 'page popularities' follow the well-known Poisson and Zipf distribution (an old hypothesis that does not hold well in the modern-day WWW).

To our knowledge, our work is the first attempt at developing a model of human-like browsing behavior that requires no prior knowledge or assumption about human behavior on the target site. The model is founded on a more general theory that defines human behavior as an 'interest-driven' process. The preliminary simulation results are very encouraging - web bots built using our model are capable of mimicking real human browsing behavior 1000-fold better compared to bots that deploy random crawling strategy.

Keywords — bot modeling, interest-driven human browsing

I. Introduction

Current day Internet abounds in the number and type of Web bots (software programs used to automate the process of retrieval and collection of Web resources). Some of the Web-roaming bots perform useful jobs and are referred to as 'benign', while others tend to misuse legitimate Internet resources and are referred to as 'malicious'. Examples of benign bots are *spider bots* (used by search engines) and *media bots* (provide updates on weather conditions, news, sports). Examples of malicious bots are *spam bots* (harvest email addresses for the purpose of email spamming) and *click bots* (automate clicks on online ads to fraudulently generate revenue).

Programs designed to mimic/emulate the way humans browse the Internet are a particularly important category of Web bots. Namely, from the perspective of security defenders, these bots are an invaluable tool used in the process of capacity planning and load testing. On the other hand, malicious hackers have an equally keen interest in these types of bots as (e.g.) they can be used to launch powerful hard-to-defend-against DoS/DDoS attacks. Clearly, for both security defenders and malicious hackers, it is critical that the behavior of their human-like acting bots be as close as possible to the behavior of real human visitors to the target system/site.

To date there have been many attempts to develop models that accurately emulate the way humans browse the Web. These models can generally be grouped in two main categories: 1) models that assume the existence of previous browsing history (Web logs) from which it is possible to extract sufficient information/knowledge on how humans browse the target site; 2) models built on the classical assumptions concerning 'web-page think times' (believed to follow Poisson distribution) and 'web-page popularities' (believed to follow Zipf distribution). Unfortunately, both of these models/assumptions are becoming increasingly problematic. For example, many of today's Web-sites are very dynamic in terms of the content they provide and the user populations they attract. Hence, for these sites, log history has very little if any value in modeling/predicting the current or future visitor behavior. On the other hand, the well-known assumption about human browsing behavior being shaped by Poisson and Zipf law is slowly losing ground, as many recent studies contradict this classical hypothesis.

The goal of our work is to propose a radically new approach to the development of human-like behaving bots. In particular, our approach makes no assumptions about specific probability laws governing human Web browsing, nor it relies on any prior log history on which it would be possible to base the modeling of current/future user behavior. The only information that our model relies on is the actual (textual and hyperlink) content of the target site. Using this readily available information, and applying the rules of the so-called 'theory of interest-driven human behavior', our model aims to emulate human-like browsing behavior on new previously unvisited sites (or sites that renew their content and/or structure frequently).

II. Theory of Interest-Driven Human Behavior

The theory of interest-driven human behavior was first introduced in [1] and [2]. According to these two break-through works, many real-world human activities are driven by personal interest, and cannot be treated simply as tasks needing execution. Specifically, human behavior seems to be driven by an interplay between 'personal interests' and 'frequency of events/actions'. As stated in [2], "frequency of events/actions is determined by the interest, while the interest is simultaneously affected by the occurrence of events/actions."

III. Model of Human Browsing Behavior Using Interest-Driven Theory (HBB-IDT)

As previously indicated, the purpose of our research is to develop a model of realistic human-like browsing behavior, given no a priory assumption or knowledge. In general, there are two features that characterize any browsing process,

978-1-4673-7876-5/15/$31.00 ©2015 IEEE

whether generated by a human or bot: 1) sequence of web pages visited, and 2) stay time on each of the visited page.

Now, in our model, we assume that the browsing behavior of a human visitor (i.e., which particular pages he visits and how long he stays on each page) will depend on the visitor's interest in the visited pages. Furthermore, we assume that the user's interest in each visited page is generally composed of two elements: the interest in the page's *theme*, and the interest in the page's *content*.

- ***Theme*** is the set of main general topic(s), subject(s) or idea(s) conveyed in a web page. E.g., the possible themes of a news page about Apple's stock are Business, Technology and Finances.
- ***Content*** of a web page is defined as the substantial information provided/found in its text. Webpages belonging to the same theme may (likely will) provide different contents. E.g., a news page about Microsoft's stock would belong to the same theme as a page on Apple's stock, but their contents would obviously be different.

During the browsing process, once a user selects a web page based on his/her interest in the page's theme, the stay time on the given page will depend on his/her interest in the actual content of the page. However, according to the theory of interest-driven user behavior, the user's interests in themes and contents of visited pages are expected to evolve and eventually change over time:

(1) ***Change in Theme Interest***. Change of interest in a particular theme is generally correlated with the ***stay time on the given theme***. Namely, when the user first opens a webpage on a new theme, it is reasonable to assume that his/her interest in this theme is high. Following this, the user is also likely to open other pages on the same/similar theme. Nevertheless, as the stay time on the same theme increases, the user will gradually become less interested (i.e., bored) with this theme, and he/she will be more likely to open a webpage on a different theme.

(2) ***Change in Content Interest***. Change in content effectively implies that the user has decided to open another/new webpage. When a user opens a new webpage, his/her stay time on this webpage will mainly depend on his/her actual interest in the webpage's content. The value of this interest is typically decided by the user's interest in the general theme of the webpage as well as the page's *content quality* (which is determined by the page's content length and its *content closeness* to previously visited webpages.)

An overall outline of our model of Human Browsing Behavior using Interest-Driven Theory (HBB-IDT) is provided in Figure 1. The figure captures not only the key states/parameters, but also their interdependence and dynamicity. (In the figure, "+" implies positive and "–" implies negative correlation. P_n is the currently visited webpage, P_{n-1} is the page visited before P_n, and P_m is one of the candidate pages to be visited next.) From Figure 1 it should be clear that if the user's interest on the current theme is high, and page P_m's theme is very close to P_n's, then the probability to visit P_m next is high. Still, if the user's interest in the current theme is low, and P_m's theme is very different from P_n's, then the probability of visiting P_m in the next step will also be high. In all other cases, P_m's chances to be chosen as the next visited page are low.

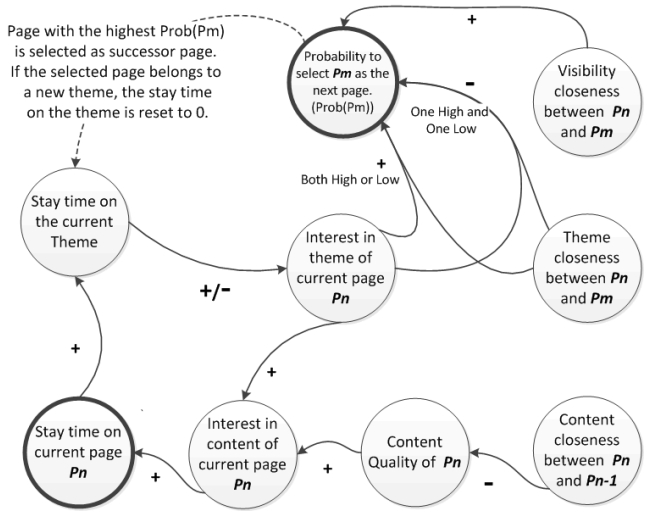

Figure 1 HBB-IDT Model

Due to the space limit, the actual formulas deployed in our model (i.e., expressions for *interest in current theme*, *interest in current content*, *stay-time on current theme*, *page stay/visit time*, *theme closeness*, *content closeness*, v*isibility closeness* and *probability of selecting next page*) are omitted form this paper.

IV. Experimental Results

The software implementation of our HBB-IDT model is built in Java and comprises the following components: **Content Gatherer** (crawls and downloads all web pages from the target website), **Data Analyzer** (analyzes the content of individual webpages and determines their respective themes and linkage maps), **HBB-IDT Crawler** (crawls the target site by implementing HBB-IDT rules). The software/model was tested on a mirror-version of a popular Canadian news-agency site (www.cbc.ca/news). Our preliminary results (Table 1) are very encouraging, as HBB-IDT bots were able to emulate real human browsing behavior 100- to 1000- fold better relative to bots deploying random browsing strategy.

Session Num.	Ability of HBB-IDT vs Random Model to Emulate Human Browsing Behavior
1	220.80
6	5385.00
8	262.00
11	93.84
25	2821164.00

Table 1 Performance of our model vs. random crawl model

References

[1] Barabasi, Albert-Laszlo. "The origin of bursts and heavy tails in human dynamics." Nature 435.7039 (2005): 207-211.

[2] Zhou, Tao, Xiao-Pu Han, and Bing-Hong Wang. "Towards the understanding of human dynamics." Science matters: humanities as complex systems (2008): 207-233.