MODELING OF HUMAN WEB BROWSING

BASED ON THEORY OF INTEREST-DRIVEN BEHAVIOR

YANG YANG

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE AND ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO

JUNE 2016

ABSTRACT

The ability to generate human-like Web-browsing requests is essential for testing and optimization of WWW systems. To date, the majority of approaches to modeling of human-like browsing behavior have assumed the use of data mining techniques on collections of previously gathered browsing sequences (i.e., server logs). However, in situations where the amount of data pertaining to previous browsing history is insufficient or unavailable altogether, data mining approaches are generally not suitable. In this thesis a new model of human-browsing behavior – the so-called HBB-IDT model – has been proposed. The model is based on the theory of interest-driven human behavior and does not assume the availability of server-side logs (i.e., previous browsing history). The defining features of the model are: (1) human browsing on the internet is regarded as a dynamic interest-driven process, based on which the URL sequence and the corresponding stay times are generated; and (2) the user's browsing interests are linked to actual characteristics of the visited Web pages, such as their theme, visibility and information quality. Given that the model does not rely on the existence of Web logs, it can be applied more generally than the previously proposed data-mining approaches.

The key assumptions, conceptual motivation, defining algorithms, evaluation framework, as well as the final experimental results pertaining to the HBB-IDT model are presented in this thesis. The experimental results show that the probability of generating human-like browsing sequences is much higher using the HBB-IDT model

than using the pre-set request list model or the random crawling model (two models most commonly deployed in present-day Web test tools and DDoS tools).

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Dr. Natalija Vlajic and Dr. Uyen Trang Nguyen, for their knowledge, guidance, and support. Dr. Natalia Vlajic introduced me to this exciting research area, provided me with a number of ideas and classic studies, and carefully proofread the thesis. Dr. Uyen Trang Nguyen reviewed the key aspects of the research, offered many important suggestions, and contributed to the preparation of the experiments. Their guidance and support were vital to this research.

I would also like to thank the experiment volunteers for taking time out of their busy schedules; they provided critical data for the evaluation of the HBB-IDT model. I would also like to thank my friends, colleagues and other participants for their time, effort, and helpful suggestions.

Most importantly, I would like to thank my family for their love, encouragement, and support during my academic life at York University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

Chapter 1
Introduction

With the increasing reliance on the Internet and Internet-related technologies, Web-based services have grown considerably more important in nearly every aspect of modern life. As a result, the workloads of popular Web sites have also grown substantially, and at the same time these systems have become increasingly more exposed (i.e., vulnerable) to the so-called denial-of-service (DoS) attacks. In order to adequately deal with (i.e., prepare for) excessive work-load conditions - either those caused by legitimate or DoS-induced traffic - Web site owners resort to the use of the so-called performance testing tools. At the heart of these tools are algorithms intended to generate workloads (i.e., workload conditions) similar to those generated by real human visitors to a Web site. In this study, we propose a novel method for modeling of human-like browsing behavior, which aims to overcome the limitations or previous models that have been used for this purpose.

1.1 Motivation and Objectives

With the increasing demand for Website robustness and security, workload testing tools are now widely used to evaluate the performance limits of many Websites as well as the their ability to defend against DDoS attacks. To the best of our knowledge, most performance testing tools are designed deploying the so-called random/probabilistic

models (see [1] to [8]) or server-side log-mining approaches (see [9] to [16]). (In random/probabilistic models human browsing behavior is depicted using the well-known probability distributions (e.g., Zipf and Poisson), while in server-side log-mining approaches the information pertaining to past human browsing is used to model/predict future browsing behavior.) Both methods have certain weaknesses: The output of random/probabilistic models can substantially deviate from actual human behavior, possibly decreasing the efficacy of respective performance testing tools as well as impacting the accuracy of respective results. On the other hand, the server-side log-mining methods require a great deal of server-side logs in order to train the respective models, and before the tests based on these models could take place. Furthermore, given their reliance on historic data, these models are not suitable for many real-world situations, including: 1) workload testing of new Web sites before they have actually been published/mounted, or 2) testing the workload capacity of a Web site in response to some new (i.e., previously unseen) event.

Given the limitations of the random and log-mining approaches, it is important to develop a model that can generate human-like browsing behavior without: a) making any assumptions about the distributions that govern human-browsing behavior, and b) without relying on the availability of past server logs. The goal of the work outlined in this thesis is to develop such a model using the concepts of human interest-driven behavior (see [17] to [21]), and then to evaluate the given model on a real-world Website and against real human browsing behavior. In other words, the main objective

of this study is to determine an approach whereby a Web tester could generate URL-request sequences similar to those that would be generated by real human visitors. The approach assumes that only the page-contents of the target Website are accessible, while the browsing logs on the server are NOT available and/or required.

To accomplish the main objective, the following detailed objectives are set forward in the presented research:

(1) Identify the key factors that drive human browsing and analyze their interactions and interdependencies.

(2) Propose a mathematical model that captures the above.

(3) Develop a prototype simulation software based on the proposed model.

(4) Evaluate the model in a test environment

1.2 Research Contributions

The major contributions of our research include the following:

(1) The concepts pertaining to the general theory of interest-driven human behavior are integrated into a model of human browsing on the Web – which we name the 'model of Human Browsing Behavior based on Interest Driven Theory' (HBB-IDT).

(2) A series of algorithms are proposed to adequately incorporate all the key aspects of the general interest-driven theory into the HBB-IDT model, so that a practical workload testing tool based on the HBB-IDT model can be implemented and evaluated.

1.3 Thesis Structure

The content of the thesis is organized as follows. In Chapter 2 a review of the related literature is presented, and the main advantages and disadvantages of random/probabilistic-based and data-mining-based approaches to the modeling of human browsing behavior are discussed. In Chapter 3, basic assumptions pertaining to our work are introduced, and a conceptual model of human browsing based on the theory of interest-driven behavior is proposed. In Chapter 4, detailed equations pertaining to the HBB-IDT model are derived and explained. In Chapter 5, a prototype software used for the evaluation of the proposed model is introduced. Also, in this chapter, the actual evaluation processes as well as the obtained experimental results are presented. In Chapter 6, the main conclusion of this study and possible directions for future works are outlined.

Chapter 2
Literature Review

Scholars from various field have been interested in studying human browsing behavior in the Web. For example, many approaches have been developed by computer security experts to simulate human browsing behavior based on random/probabilistc models. At the same time, data mining researchers have proposed several server-side log-mining algorithms to identify abnormal requests (i.e., to distinguish between human- vs. bot-generated traffic in case of DDoS attacks). Furthermore, many physicists have looked at this field as well, as they are interested in modeling collective behavior patterns based on the huge datasets from Web logs. In this chapter, the related literature from these fields is introduced and discussed.

2.1 Approaches Based on Random/Probabilistic Models

The simplest way to generate a sequence of Web-browsing requests is by deploying the random/probabilistic model. According to this model, URL requests are generated by randomly selecting pages from the target Web site (or choosing them based on the Zipf distribution), and then assigning a thinking (stay) time to each of the pages according to the Poisson or Pareto distribution [1]. This model has been adopted by many workload test systems and application-layer DDoS hacking tools, such as the well-known ***ab***

(Apache HTTP Server Benchmarking Tool) [2], the open source load testing tool ***Tsung*** [3] and the widely used ***JMeter*** [4].

However, many studies note that the random/probabilistic model is far from accurate in presenting realistic human-generated browsing streams [5] [6] [7]. In Oikonomou et al.'s work (2004), URL sequences generated by automated bots based on random models are compared to the sequences generated by actual humans. The results show that 98.6% of randomly simulated URL sequences have a very low probability (<5%) of occurring in the examined server logs, while the actual human URL sequences have a significantly higher probability of occurring in the examined logs. In particular, approximately 80% of human URL sequences have a probability of occurance greater than 40% [5]. This implies that the Web browsing of common human users does not exhibit unpredictable (i.e., random) behavior.

In [8], Dimitris Gavrilis et al. identified another disadvantage of the random model. Namely, with this type of model, the hyperlink/semantic information of requested pages is not taken into account. Hence, it is possible that subsequently requested pages are neither linked to each other nor contain related information – something unlikely to occur when requests are generated by a real human visitor. The same phenomenon was also observed by Xie et al. in [7], and the authors of this work proposed an Access-Matrix-based approach to detect application layer DDoS traffic (i.e., request sequences) generated by bots deploying random models.

Although the random model has proven to be rather unrealistic, it is nevertheless the most widely used model of Web-browsing behavior, mainly as it is easy to implement and does not rely on the use/existence of server logs.

2.2 Approaches Based on Server-Side Log Mining

To improve accuracy of browsing models, many researches have looked into the use of data-mining techniques to deduce the key characteristics of human browsing from existing server logs.

Schechter et al. [9] proposed a method for predicting HTTP request sequences based on the server-side path profiles, similar to the code branch prediction algorithm of microprocessors. This method first logs each user's visiting history and sets up a profile containing their visited URL sequences and corresponding frequencies. Subsequently this profile is used to predict new users' URL sequences, obtaining rates of prediction accuracy of over 40% in tests.

Zhong et al. [10] created a similar prediction method by profiling browsing-paths from server logs using the N-Gram model. One of the key findings of their work is that for N equal to or greater than 3, the accuracy of prediction improves significantly.

Markov models are also widely used in studies dealing with prediction of human browsing sequences (see [11][12][13][14][15]). In Zukerman et al. [11], the space-Markov model, second-order Markov model and linked space-time Markov model are

used to profile server logs. Awad et al. [12] combined association rule mining with a modified Markov model to predict URL sequences. Given that higher-order Markov models may be too complex to meet real-time prediction requirements, Deshpande at el. [13] designed a method of intelligently selecting different parts of ordered Markov models to profile server logs. Their algorithm can effectively reduce state complexity while maintaining high prediction accuracy.

Other statistical methods such as the Bayes model or distance-based clustering have also been studied in this context. For example, Poornalatha et al. [16] showed that a variant of distance-based clustering algorithm exhibits similar accuracy to Markov models in terms of predicting user requests – but in a significantly shorter amount of time.

Although the above mentioned data-mining approaches generate very realistic URL request sequences, they may not be practical in cases when there is not a statistically significant amount of related server logs, such as in the case of a Web site that has just gone live, or a Web site that frequently changes its structure and content. In such sites, using past log information to make prediction about the users' future browsing behavior is likely to produce suboptimal results, at best.

2.3 Approaches Based on Content Analysis

The content of Web pages is a significant factor impacting the way humans browse the WWW. Consequently, semantic relationships between URL requests of human-

generated browsing sequences have been frequently studied in the past. Moreover, several researchers have incorporated content analysis into basic log-mining approaches to improve simulation performance.

Shen et al. [27] studied the problem of dynamics in Web searches, and he proposed the so-called possibility matrix to depict the way humans transition from one topic to another. Mabroukeh et al. [28] and Hoxha et al. [29] both proposed frameworks for predicting URL sequences based on a pre-built ontology semantic Web. Although the need for an existing Web domain ontology limits the applicability of these frameworks, they demonstrate the feasibility of simulating human-like URL sequences by relying on the knowledge of semantic relations.

Given that different types of users may be interested in different types of Web content, user characteristics may also be helpful in predicting browsing sequences. Examples of works in this area are the studies by Goel et al. [25] and Duarte et al. [26]. Specifically these works looked at the impact of general browsing interest and habits of various groups of users on their overall Web-browsing behavior.

2.4 Research on Human Behavior Patterns

As indicated in the preceding sections, many researchers believe that there exist non-random patterns in human browsing behavior. One way of identifying these patterns is by studying the logs (i.e., historic databases) of past human browsing. However, as

indicated in the first chapter, it is not easy and may even be impossible to obtain sufficient amount of past log-data in many real-life cases.

Nevertheless, the latest research on the nature and mechanisms driving general human behavior offers another way of predicting human browsing in the absence of server logs. Namely, in recent years, several studies in the field of the so-called social physics have attempted to identify the common laws that shape human behavior. Among these studies, the work by Barabási et al. [17] has earned special attention and recognition. In this work it has been argued that human behavioral patterns follow certain laws, which causes them to exhibit non-Poisson distributions. In particular, as Barabási noted, "*In most human-initiated activities, task selection is not random, but the individual executes the highest-priority item on its list.*" Thus, if we can determine the factors that drive/impact human priorities in certain situations/conditions (e.g., in the case of Web browsing), we should be able to more accurately model/simulate human behavior under those situations/conditions.

From the base argument established in [17], researchers have further hypothesized that 'personal interest' is most likely factor driving human priorities and thus overall human behavior ([17] - [21]). Specifically, human behavior seems to be driven by an interplay between 'personal interests' and 'frequency of events/actions'. As stated in [19], "*frequency of events/actions is determined by the interest, while the interest is simultaneously affected by the occurrence of events/actions*". In other words, a new event/action may initially spark interest in the same/similar type of activity and

intensify its occurrence. However, as the overall number of repetitions (i.e., frequency) of this activity increases, the respective interest is likely to subside. This, in turn, will gradually decrease the frequency of activity back down to (near) zero. The interdependence between the frequency of an event/activity and the interest in the given event/activity is illustrated in the below figure.

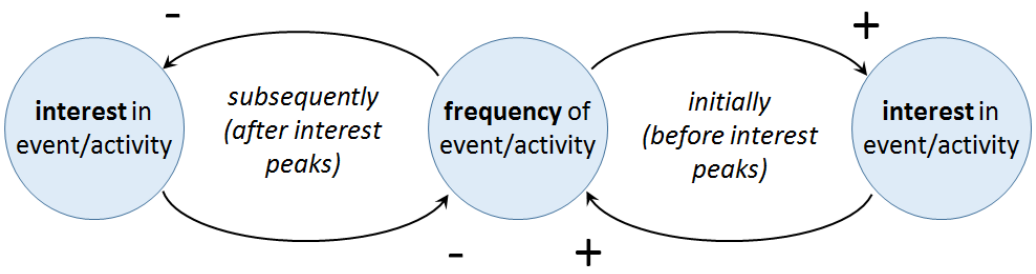


Figure 1. Interdependence between the frequency of an event/activity and the interest in that.

It should be noted, however, that Zhou's theory considers human interests and events as very general and abstract processes/variables. Hence, in order to apply this theory to the modeling of human browsing-behavior it is necessary to develop more detailed (i.e., appropriate) depiction of interest-changing processes, the Web page selection-probabilities as well as individual Web page think-times.

2.5 Summary of the Existing Research

Based on the literature review, the following conclusions may be drawn:

(1) Human browsing behavior exhibits certain common patterns.

(2) The patterns of real human browsing behavior differ substantially from the traditionally deployed random models, and instead can be better described by interest changing and semantic relations.

(3) Existing studies on interest-driven human-behavior only reveal macro-level statistical features and are not sufficiently detailed to be directly used for modeling of human browsing-behavior in WWW.

The goal of the study presented in this thesis is to advance the research on modeling of human browsing behavior in WWW by developing an appropriate and sufficiently detailed interest-driven model and then testing this model in a real-world evaluation framework.

Chapter 3
Assumptions and Conceptual Model

In this chapter, we lay the foundation for further discussion by introducing the key assumptions and definition as pertaining to our model of Human Browsing Behavior using Interest-Driven Theory (HBB-IDT). In particular, we identify the major interest-driven factors that shape and influence the human Web-browsing process, and we propose a conceptual state-model depicting this process. In-depth details of our model, including its algorithmic implementation, are presented in Chapter 4.

3.1 Key Assumptions

Because this research is the first attempt to apply interest-driven theory to the modeling of human browsing behavior, we introduce several simplifying assumptions that seem reasonable and justified:

3.1.1 Assumption 1: Consider Content Downloading Only

In general, there are two different types of user activity on the WWW: content downloading (e.g., reading an article or listening to an audio file hosted on a Web site) and content uploading (e.g., posting personal comments to a news-agency or a social-media Web site). Content downloading is the original and most common form of Web browsing. Hence, the work presented in this thesis focuses solely on this form of Web browsing. In other words, the model in this paper considers only static or non-

interactive webpages, which do not allow users to change the content they are reading.

3.1.2 Assumption 2: Consider Text Content Only

Modern Web sites generally consist of text content and multi-media content such as images, video and audio. There is no doubt that multi-media content carries valuable information and has significant effects on human browsing behavior. However, it is rather difficult for computers to automatically understand themes and extract key information from a pure multi-media file. Therefore, the model deployed in our work considers only the text content of each visited Web page and ignores multi-media objects as well as other MIME links (such as PDF) that can be found on the given page. (Please note, in the majority of today's popular Web sites the text content is still far more important than the content carried in their multi-media objects. Hence, the decision to focus on 'text-only' information seems reasonable and statistically well-justified.)

3.1.3 Assumption 3: Full Knowledge of Website Structure and Content

Human Web browsing, i.e., the decision to pick one of many links/Web pages that are referenced on the current Web page, is generally governed by the thematic similarity and/or relevance of the two pages. (We elaborate more on this in Chapter 4.) Now, humans are generally very good at determining the thematic similarity/relevance between the current page and a linked/referenced page 'on the fly' (e.g., by examining

the text contained in the referenced page's link/URL, by examining the physical or logical position of the referenced page's link/URL in the current page, etc.). Unfortunately, automated bots lack this skill (i.e., do not possess such level of intellectual/browsing sophistication). Thus, to improve the odds that our automated bot correctly distinguishes between the thematic similarity/dissimilarity of the current Web page and the pages that it refers to, we assume that the bot knows a priori the full content of all referenced pages. From the implementation point of view, for this to be true, the bot must (be able to) first pre-screen the target site or obtain the site's full content in some other way.

3.2 Definitions of Key Terms

3.2.1 Web Page, Successor and Stay Time

As previously indicated, the purpose of our research is to develop a model of realistic human-like browsing behavior. Two (sets of) features that characterize any browsing process, whether generated by a human or bot, are: 1) the sequence of the Web page visited/requested, and 2) the stay time on each of the visited pages. In this thesis, we employ the following terms and definitions related to the two key features of the browsing process:

(1) ***Web Page.*** A Web page is an HTML document with a unique URL and meaningful text content. A Web page may contain several hyper-links (i.e., URLs) to other HTML documents, multi-media resources, MIME files, CSS and

JAVASCRIPT files, etc. Note that in our work files and embedded objects that do not contain human-readable text (CSS, JSON data, JAVASCRIPT codes, etc.) are not considered a Web page, even though they may also be referenced by a unique URL and their requests/retrievals may also be recorded in server logs. In general, requests to these non-human-readable objects are caused by a request to a Web page that contains or points to them.

(2) ***Web Page's Successor***. During a browsing process, Web page B requested immediately after the currently visited Web page A is called page A's successor.

(3) ***Stay Time on Web Page.*** The stay time on a Web page is defined as the time interval between the user requesting the given page and the user requesting this page's successor. In the literature, this time interval is also commonly referred to as Think Time. Generally, the length of stay time on a Web page is determined by the user's interest in the page's theme and content.

Figure 2 illustrates the above terms using an example of URL logs.

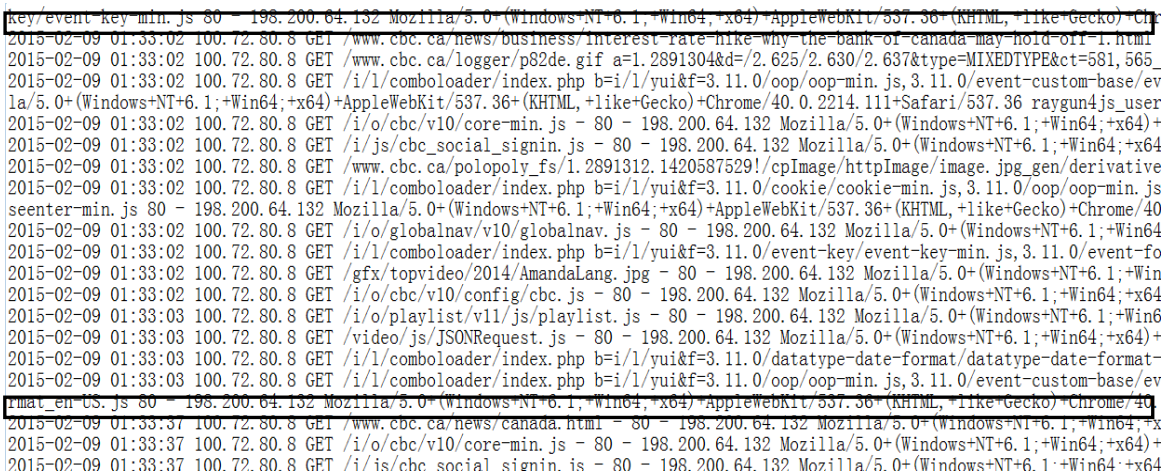


Figure 2. Illustration of Web page, successor and stay time. In the caption of the server logs, only the two highlighted URLs are considered Web pages, while the others are non-textual or meaningless files. The second Web page is the successor of the first one. The stay time on the first Web page is 35 seconds (2015-02-09 01:33:37 minus 2015-02-09 01:33:02).

3.2.2 Theme, Content and Information Quality

In our work/model, we assume that the browsing behavior of a human visitor (i.e., the particular pages that they visit and the length of time they stay on each page) depends on the visitor's interest in each page. Furthermore, we assume that the user's interest in each visited page is influenced by three factors: (1) the user's interest in the page's theme, (2) the user's interest in the page's content, and (3) the content quality of the given page. We define these three concept as follows:

(1) ***Theme***. Theme is the set of main general topic(s), subject(s) or idea(s) conveyed in a Web page. For example, the possible themes of a news-agency page about Apple's stock are Business, Technology and Finances.

Once a user finishes reading Web page *A*, they will select a new Web page to visit choosing from several candidate hyperlinks that appear on page A. Now, given that the user has not opened any of these Web pages yet, thet not know the detailed contents of these Web pages. However, by reading the tip-texts that appear within these hyper-links (as shown in Fig 3), the user is generally able to grasp the major idea, i.e., the main theme, of each 'candidate' page. Then, the user can make a selection based on their interest in each theme.

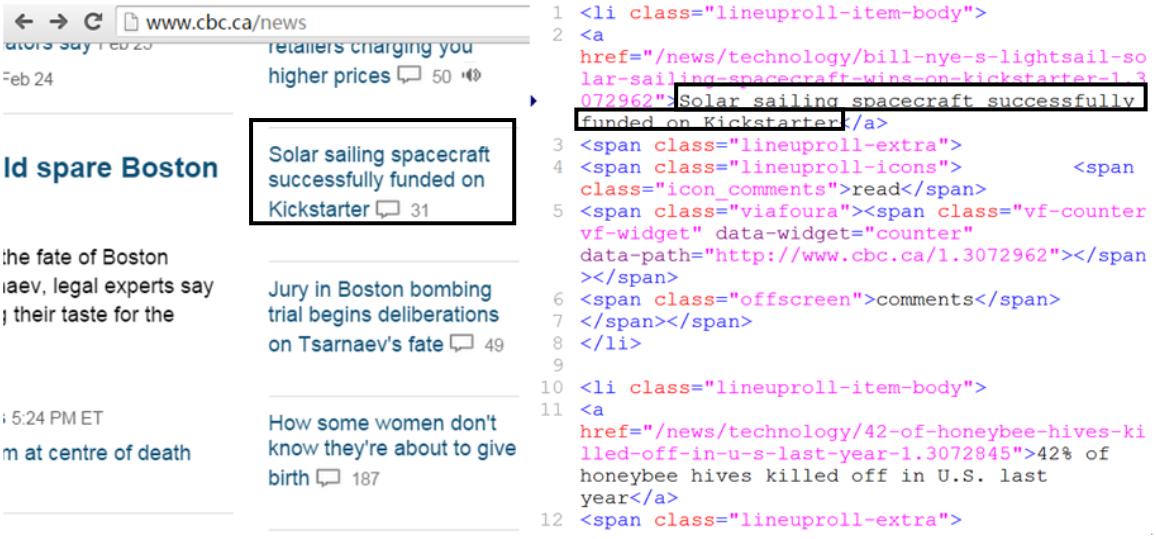


Figure 3. An example of Tip-Texts for a Hyper Link.

(2) ***Content***. The content of a Web page is defined as the key substance of the information provided/found in its text. Webpages belonging to the same theme may (i.e., likely will) provide different content. For example, a news-agency page about Microsoft's stock value would belong to the same theme as a page on Apple's stock (see the definition of *Theme*), but the content of each page would obviously be different.

Once a user selects a Web page based on their interest in the page's theme, the stay time on the given page will depend on their interest in the actual content of the page. Consequently, even though the user may be interested in the theme of a Web page, they may quickly jump/move to another page if the page's content is not attractive enough.

(3) ***Content Quality.*** We define content quality as the actual attractiveness of a Web page's content to the user. It can generally be expected that a Web page with more

information and sufficient but not overly high similarity to the content of previously visited Web page is more attractive to the user, and thus has higher content quality.

Now, during the browsing process, the choice of a particular Web page and the stay time on it are impacted by a fine interplay among the three above mentioned parameters. Namely, once the user decides to open/retrieve a Web page based on their interest in that page's general theme, the stay time on the given page will depend on their interest in the page's actual content as well as the page's content quality. However, in some cases, even though the user may be very interested in the page's theme, they may quickly jump/move to another page if the page's content is not attractive enough (i.e., its content quality is not satisfactory).

3.2.3 Web Page Similarity

In order to measure the 'contextual proximity' (i.e., similarity) between two Web pages, which we refer to as Web page closeness, we establish the following three metrics:

(1) ***Theme Closeness*** depicts the degree to which the themes of two Web pages are similar.

(2) ***Content Closeness*** depicts the degree to which the content of two Web pages is similar.

(3) ***Visibility Closeness*** between the current page and one of its links (i.e., linked pages) depicts the likelihood that the user spots/finds the given hyperlink inside the current page. Obviously, the higher the visibility closeness, the more likely the given link/page will be selected as the successor page.

3.3 Conceptual Model of Human Browsing Behavior

3.3.1 Factors Impacting Human Browsing Behavior

Based on the previous research on interest-driven human behavior discussed in Chapter 2, we have identified users' interest in the themes and content of encountered Web pages as the key factors influencing their browsing in the WWW. We also believe that, although not fully predictable, users' interest in themes and content are expected to evolve and eventually change over time as outlined below:

(1) ***Interest in Themes***. According to the interest-driven theory [19], the change in a user's interest in a particular theme is generally correlated with the user's ***stay time on the given theme***. Namely, when the user first opens a Web page that reflects a new theme, it is reasonable to assume that their interest in this theme is high. Following this, the user is also likely to open other pages on the same/similar theme. However, as the stay time on the same theme increases, the user will gradually become less interested (i.e., bored) with this theme, and they will more likely open a Web page on a (very) different theme.

(2) **Interest in Content**. When a user opens a new Web page, his/her stay time on the Web page will mainly depend on his/her actual interest in the Web page's content. The value of this interest is typically decided by the ***content quality*** of the Web page. Higher content quality generally results in a longer stay time on the page. Based on the discussion in 3.2, we know that the content quality of a page is determined by the page's content length (i.e., amount of provided information) and its ***content closeness*** to previously visited Web pages.

3.3.2 Outline of HBB-IDT Model

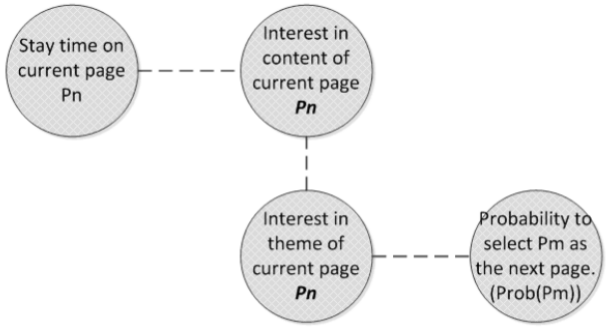
(1) Choices of Key Factors in the Model

Based on the principles and discussions provided in Section 3.3.1, we can draw a simplified model of the interest-driven behavior in Web browsing by humans. Let P_n, P_{n-1} and P_m denote the currently visited Web page, the previously visited page, and one of the candidate pages to visit next, respectively. As shown in Figure 4(a), the interest in the current page P_n and the interest in the theme of P_n are related to the stay time on the page and the selection of the next page to visit. This idea comes directly from the Interest Driven Theory, as discussed in Section 3.3.1.

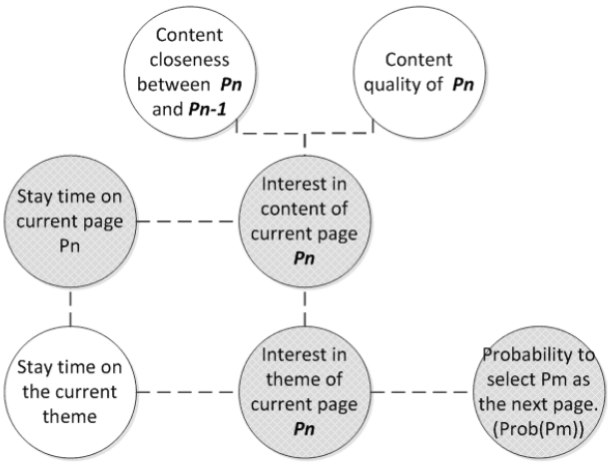
There are also other factors that affect the stay time and the two interest factors mentioned above (see Figure 4(b)). Firstly, the stay time on different pages of the same theme constitutes the stay time on this theme, which then determines the user's interest in the theme. Secondly, the content quality of the current page P_n and the

content closeness between the current page P_n and the previous page P_{n-1} affect the user's interest in P_n as well as the interest in the theme of P_n. Therefore, we added these three factors - the stay time on the current theme, the content quality of P_n and the content closeness between P_n and P_{n-1} - into the model, as shown in Figure 4(b).

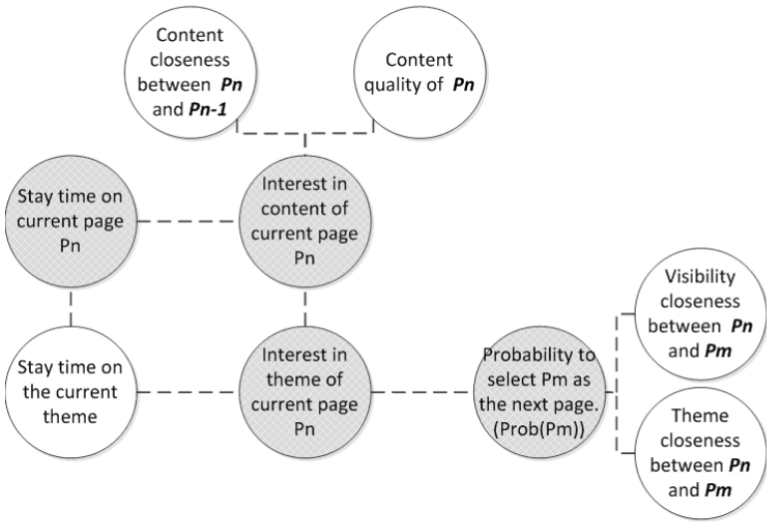
Similarly, since the user's choice of what page to visit next is influenced by his/her interest in the current theme, we must also know the closeness between the current theme and the theme of a candidate page. Thus a new factor, theme closeness, is added to the model. Finally, whether the user can easily find a hyper-link to a candidate page determines the probability of that page to be opened next, so we added one more factor, visibility closeness, to the model. Figure 4(c) shows all the factors we have discussed above. Based on these factors, we built the proposed model of human browsing behavior using the interest-driven theory.



(a) Step 1: Core Factors



(b) Step 2: More Factors Added



(3) Step 3: Final Model With 9 Fctors

Figure 4. Steps of choosing factors in the HBB-IDT model.

(2) Model Outline

The proposed model of human browsing behavior using the interest-driven theory (HBB-IDT) is illustrated in Figure 5. In the diagram, the symbols "+" and "–" indicate positive and negative correlation, respectively. The diagram captures not only the key states and parameters, but also their interdependence and dynamics.

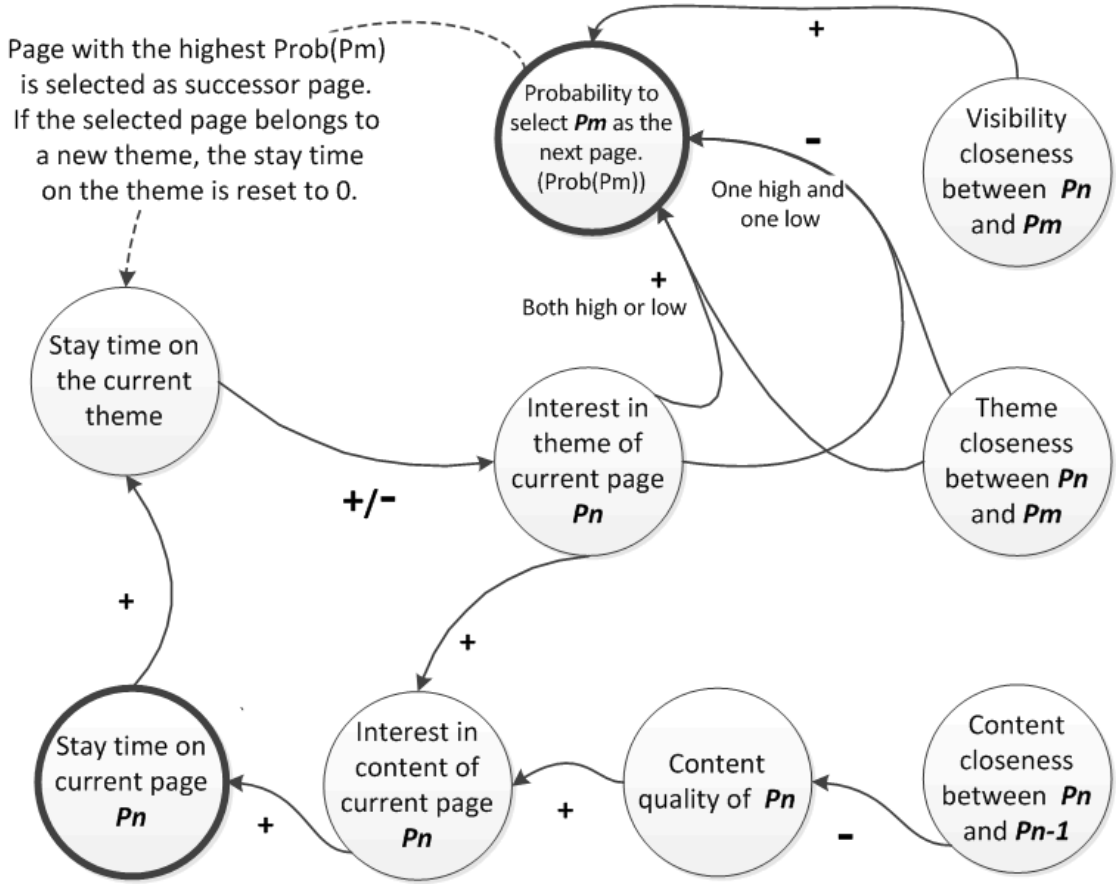


Figure 5. Dynamic relationships illustration, in which "+" means positive influence and "–" means negative influence. P_n is the current Web page, and P_{n-1} is the page visited before P_n. P_m is one of the candidate pages to be visited in the next step.

According to this model, when the user browses Web page P_n, the stay time of the user on this page (the state in the lower-left corner of Fig. 4) is decided by his/her

current interest in the content of P_n. The extent/strength of this interest, however, is determined by two additional factors: the current interest in the general theme of P_n and the quality of the content of P_n. (As discussed, we define content quality of P_n as a function of two factors: content text length and the closeness of content P_n to the previous page P_{n-1}.)

The interest in the current theme also affects the selection of the next page. If the user's interest in the current theme is high, and the theme of a page P_m is very close to P_n's, the probability of the user visiting P_m in the next step is higher. However, if the user's interest in the current theme is low (i.e., they are likely to change themes) and if P_m's theme is different from P_n's, the probability of visiting P_m in next step shall be high too. In other cases, the chance of P_m being chosen is low. At the same time, if the user can easily find a hyper-link to P_m on Web page P_n, the visibility-closeness between P_m and P_n is high, and the possibility of selecting P_m is further improved.

As further shown in the model, user interest in the current theme is mainly affected by their stay time on the theme. This stay time is the sum of the stay-times on consecutively visited Web pages belonging to the current theme, and it is reset if the user visits a Web page with a new theme. Generally, the stay time of the current theme affects the user's interest in a nonlinear way. As revealed by previous studies, the user's interest in a theme is likely to increase (or remain high) for some initial period of time, but after that (i.e., once the user obtains a sufficient amount of information), the interest is likely to decrease.

Chapter 4
HBB-IDT: Key Algorithms and Quantitative Model

In Chapter 3, we have outlined all key driving factors behind our HBB-IDT model and we have described their interaction mechanisms. In this chapter we present a precise analytical depiction of each of those interactions, thus allowing that a detailed (quantitative) model of HBB-IDT be developed and, subsequently, evaluated.

4.1 Analytical Model of Interest in a Theme

According to the model shown in Figure 5, it is quite common that a user looks at several different topics/themes (in succession) during a single browsing session. The transition from the current theme to another is generally correlated with the (cumulative) stay time on the current theme. Namely, when the user first opens a Web page 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. However, 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 Web page on a different theme. We capture this phenomenon in our model by indicating that the interest in the current theme (i.e., theme of currently visited page)

is initially positively but then negatively impacted by the cumulative stay time on the given theme (see Figure 5).

To be actually able to depict the change a theme analytically, let us provide a few clarifications. First, we consider the stay time on a theme to be the sum of stay times on the consecutively visited Web pages corresponding to this theme. In practice, however, it is very difficult to find two Web pages covering exactly the same theme. In some cases, even if there is a single word that differs between two pages, there may be a noticeable difference in their respective themes. Therefore, to decide whether two pages belong to the same theme, we incorporated a pre-set threshold θ into the model. More precisely, we annotate the theme closeness between Web page *i* and *j* by $S(i,j)$, where *i* and *j* are considered to belong to the same theme if $S(i,j) > \theta$ and vice versa. (The exact expression for S(i,j) is provided in equation (3) of the next section.)

Furthermore, let us suppose that the user has browsed *(i-1)* Web pages on the same theme before visiting the current page (i.e., the current page is the i^{th} visited page), and let us annotate the user's stay time on page *(i-1)* by t_{i-1}. Consequently, the user's stay time on the given theme can be depicted by:

$$d_i = \begin{cases} 0 & ,if\ \ i = 1\ \ or\ \ S(i,i-1) \le \theta \\ d_{i-1} + t_{i-1} & ,if\ \ S(i,i-1) > \theta \end{cases} \quad ,0 < \theta < 1 \quad (1)$$

Clearly, from formula (1), if the closeness value for pages *i* and (*i-1)* is less than θ, the current visited page *i* is viewed as belonging to a new theme, and consequently, the stay time on the current theme shall be reset to 0.

Based on the above, the interest in the current theme (i.e., theme of *page i*) can be modeled using formula (2).

$$c_i = Max\left(0,\ \frac{c'-1}{{d_m}^2}(d_i - d_m)^2 + 1\right) \quad (2)$$

In formula (2), c' is the initial interest value on a new theme, which can be preset to a constant value or a random number in the range of *(0,1),* d_m (also a preset constant) represents the time point at which the interest on a given theme reaches its highest value, while d_i represents the actual stay time on the given theme.

Figure 6 depicts the character of the change in the user's interest in the current theme (c_i) relative to the stay time on the given theme (d_i) defined by (2). As discussed in Chapter 3, the value of interest in a given theme should be a non-linear function of stay time, which first increases but then keeps decreasing after reaching its peak value. Based on Figure 6, we observe that expression (2) very much satisfies this (general trend) requirement. Furthermore, from Figure 6, we can also observe that the interest value in a theme is first set to a random value c' when the user opens a Web page corresponding to a new theme. Then, the interest value keeps increasing until the stay time reaches d_m. Once the stay time surpasses d_m, the interest value starts

declining. Clearly, from (2) and Figure 6, if a user's stay time d_i equals d_m, his (her) current interest value on the given theme is at its maximum value 1.

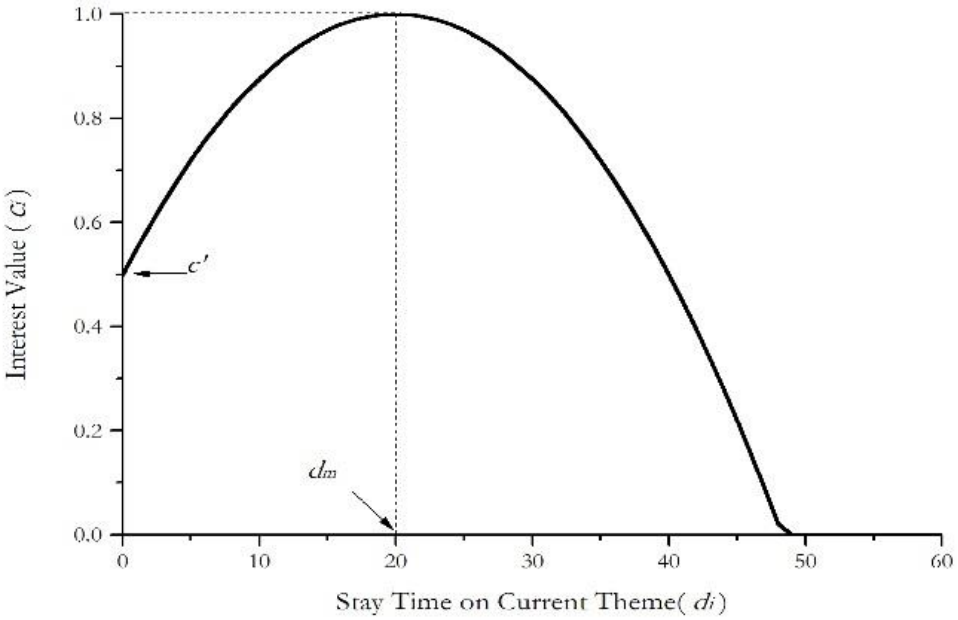


Figure 6. Demo shape of formula (2). In this demo, c' is set to 0.5, and d_m is set to 20.

4.2 Analytical Model of Theme Closeness and Content Closeness

In the model outlined in Figure 5, theme closeness and content closeness are important factors affecting user interest in theme/content. Because we only consider text content in this research (see Premise 2 in Chapter 3), we use computational linguistic methods to quantify the values of these two parameters.

4.2.1 Analytical Model of Theme Closeness

To decide how close the themes of two Web pages are, we examine the text-similarity of their hyperlink tip-texts, page titles and keywords (the content in <META> tags such as <meta name="keywords">).

Generally, in information retrieval and text mining, each term (found in a document) is notionally assigned a different dimension allowing for the entire document (i.e., Web page) to be characterized by a vector, where the value of each dimension corresponds to the number of times that the respective term has appeared in the document. Consequently, based on this model, one of the most practical methods of computing similarity between two documents/Web pages is cosine similarity, which corresponds to the inner product or their respective vector representations (i.e., it measures the cosine of the angle between them) [36].

Given that different words may belong to the same theme – for example, *Computer* and *Automobile* both belong to theme *Technology* – in our model, every word is first mapped into corresponding theme-domains. In the filed of linguistic analysis, there are many word classification hierarchies, such as EuroWordNet Domain-ontology (Vossen, 1998) and SIMPLE domain hierarchy (SIMPLE, 2000). In this study, we choose a widely used category, WordNet Domains Hierarchy (WDH, version 3.2), which was proposed by Bernardo Magnini et al. (2004) and updated in 2007 (see [37] and [38]). This word list classifies 115,424 English words into 161 domains, and each word may be mapped to several domains. The original mapping-corpus of Word Net

Domains can be downloaded from http://wndomains.fbk.eu/download.html, and is illustrated in Figure 7.

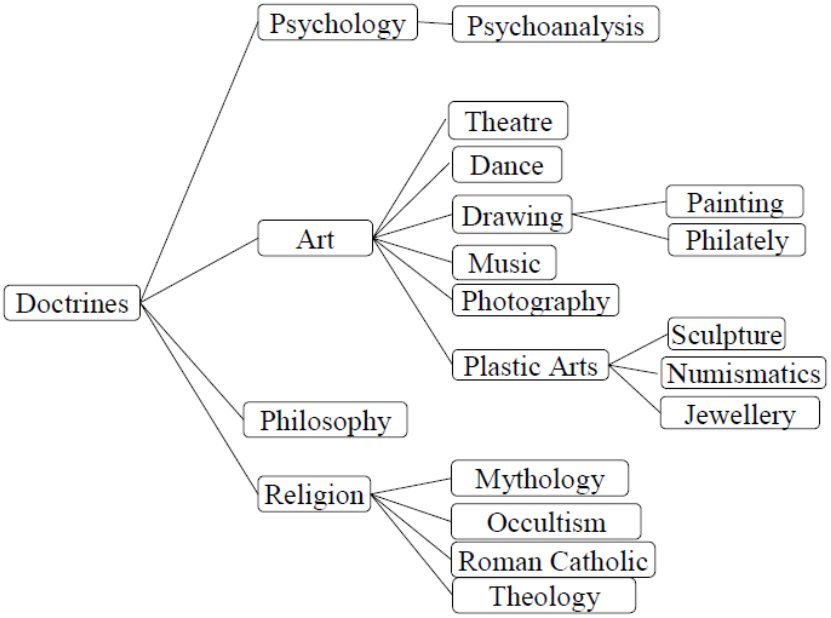


Figure 7. A fragment of the word-domain hierarchy in WDH taken from the literature [38].

To make calculating theme closeness with this hierarchy easier, in our work we have encoded all the domains based on their position in the hierarchy tree. For example, the top-level domain Pure Science is encoded as 04000000, its sub-level domains such as Astronomy, Biology are encoded as 04010000 and 04020000, respectively, and sub-level domains of Biology such as Biochemistry and Anatomy are encoded as 04020100 and 04020200. Using this scheme, all common English words are classified into one or more encoded domains, as shown in Figure 8.

	A	B	C	D
1	id	word	domain_name	domain_code
34514	01909406-n	Porcellionidae	animals biology	04030000 04020000
34515	01909537-n	Porcellio	animals biology	04030000 04020000
34516	01909591-v	caper	factotum	06000000
34517	01909686-n	sow bug	biology entomology	04020000 04030100
34518	01909692-v	hurdle	factotum	06000000
34519	01909823-s	accompanying	factotum	06000000
34520	01909847-n	sea louse	entomology	04030100
34521	01909860-v	dive	factotum	06000000
34522	01909940-n	Amphipoda	animals biology	04030000 04020000
34523	01910079-s	affiliated	factotum	06000000
34524	01910121-v	nosedive	factotum	06000000
34525	01910198-n	amphipod	biology entomology	04020000 04030100
34526	01910279-s	associated	factotum	06000000
34527	01910279-v	duck	factotum	06000000
34528	01910360-n	Orchestiidae	animals biology	04030000 04020000
34529	01910410-v	crash-dive	aviation	05100100

Figure 8. Mapping from Words to Theme-Domains by WDH

We annotate all theme-domains in WDH as a set $H = \{h_j\}$, where h_j is the j_{th} theme-domain in the list. Furthermore, we use *R*(h_j, *i*) to describe the closeness of Web page *i* and theme-domain h_j. In particular, we check every word in the page's title and its key words (as found in the page's <meta> tags), and if the given word also exists in the word list of domains H (e.g., is found in theme domain h_j}, we increase the value *R*(h_j, *i*) by 1.

However, to make *R*(h_j, *i*) more accurate, we also consider the hierarchy-structure among theme-domains in WDH. For example, a word belonging to the domain "Arts (encoded as 01060000 in this study)" and another word belonging to the domain "Dance (encoded as 01060200 in this study)" are to some degree close, although they belong to different (sub)domains. In particular, we compute *R*(h_j, *i*) based on the code-match level.

For example, suppose there are two domains A and B, which are encoded with code values CA and CB, respectively; their code-match level can be obtained by $\frac{Level\ of\ the\ Overlapped\ \ Between\ CA\ and\ CB}{Max(Level\ of\ A, Level\ of\ B)}$, where the term *level* indicates the position of a given domain in the word-domain hierarchy. For example, in Figure 7, domain "Doctrines" locates on the 1st level, domain "Theatre" locates on the 3rd level, and domain "Painting" locates on the 4th level. A domain's level can be determined by checking its encode in WDH, which contains 8 digit numbers and every two digits refer to one level. If two digits referring to a level are both zero, the domain must locate on a higher level. For instance, domain "Biology" is encoded as 04020000, which means that the domain locates on the 2nd level in the hierarchy, since the 5th to 8th digits in its encode are all zeros. Similarly, the level of domain "Biochemistry" is 3 since its encode is 04020200, in which the last two digits are zeros.

To illustrate the above, let us assume a word from page *i*'s is classified into domain 01060200. In that case, *R*(01060200, *i*) will be increased by 1, while *R*(01060000, *i*) will be increased by 0.66. The latter value is derived from the maximum level of domain 01060000 and 01060200 being 3 (our WDH encoding uses two digits to represent one level, thus 01060200 is at the 3rd level in the WDH hierarchy), and the level of their overlapped part 0106 is 2; thus the respective code-match level is 2/3=0.66. Similarly, *R*(01060100, *i*), or any other (suu)domain with a code of the type '0106----', is also increased by 0.66.

Based on the above, we represent each Web page *i* as a vector in an H=161 dimensional space: [$R(h_1, i)$, $R(h_2, i)$, .., $R(h_H, i)$]. Consequently, the theme closeness between two Web pages *i* and *k* (i.e., between their hyperlink tips, titles and keywords) is calculated using the following cosine-similarity formula:

$$S(i,k) = \frac{\sum_{j=1}^{|H|} R(h_j, i) \cdot R(h_j, k)}{\sqrt{\sum_{j=1}^{|H|} R^2(h_j, i)} \cdot \sqrt{\sum_{j=1}^{|H|} R^2(h_j, k)}} \qquad (3)$$

4.2.2 Analytical Model of Content Closeness

As discussed in Chapter 3, two Web pages with a very similar theme may still have different content, which must be considered when computing the content quality of pages. We compute the content closeness of two Web pages *i* and *k* using formula (4), in which $G = \{g_j\}$ is a list of all meaningful words found on that Web page. The term 'meaningful words' refers to nouns and verbs (a total of 115425 words) contained in WordNet 2.0, a lexical database for English maintained by Princeton University (https://wordnet.princeton.edu/).

$$S'(i,k) = \frac{\sum_{j=1}^{|G|} R(g_j,i) \cdot R(g_j,k)}{\sqrt{\sum_{j=1}^{|G|} R^2(g_j,i)} \cdot \sqrt{\sum_{j=1}^{|G|} R^2(g_j,k)}} \qquad (4)$$

(Note that while in (3) |H|=161, here in (4) |G|=115425 allow for far more refined evaluation of how close the content/information found in two pages is.)

4.3 Analytical Models of Interest in a Web Page's Content and Page Stay Time

In our model, change in the user's interest in the current content effectively signifies that the user has decided to open another/new Web page. When a user opens a new Web page, his/her interest in this new content will mainly depend on (i.e., be positively correlated with) his/her interest in the general theme of the Web page as well as the page's content quality, as shown in Figure 3. Higher interest in the content of the current page will ultimately imply longer stay time on the given page, as also indicated in Figure 3.

In our model, the content quality q_i is determined by the page's content length (i.e., overall amount of text provided) and the closeness of the content to the previous page. Longer content means that there may be more useful information on the page, while lower closeness to the previous page means that more new information may be provided by the page. Therefore, longer content and lower closeness generally imply higher content quality.

In our model, we calculate Web page content length by extracting all of its plain text outside of HTML tags and counting the number of respective words. The count of these non-HTML-Tag words is considered to be the length of the Web page. Please note that there are some methods for filtering nonsense texts (see [31] - [35]) that we have not used at this point but might employ in the future.

Now, suppose *L(i)* is the content length of page *i* and L_{max} is the maximum content length found in all pages; the content quality of this page (q_i) is defined as follows:

$$q_i = \begin{cases} \frac{L(i)}{L_{max}} & , i = 1 \\ \frac{L(i)}{L_{max}}\big(1 - S'(i, i-1)\big), i > 0 \end{cases} \quad (5)$$

Based on the above, we use the following formula to calculate (i.e., predict) the user's stay time on Web page *i*:

$$t_i = \frac{L_{max}}{Z} \cdot c_i \cdot q_i \quad (6)$$

In equation (6), besides q_i, L_{max} represents the maximum content length in all Web pages of the target Web site, c_i is the user's interest in the current theme (as discussed and defined in Section 4.1), and Z represents the average reading speed of a human being[1]. Note that by dividing L_{max} by Z, we get the maximum possible stay time a common user is expected to spend on any page in the target Web site. Clearly, from (6), the user is expected to spend a long(er) time viewing page *i* if both his interest in the page's theme and the page's content quality are high(er).

[1] For example, the average reading speed of English-speaking adults ranges from 200 words per minute to 300 words per minute [39].

4.4 Analytical Model of Visibility Closeness

In our model, higher visibility between Web pages *i* and *j* - which we annotate with $V(\ i,j)$ - means that the user can easily see a hyperlink to Web page *j* while browsing Web page *i* and is thus more likely to choose Web page *j* as the next page to visit. We use the formula shown in (7) to calculate the visibility closeness between Web pages *i* and *j*. In (7), *L(i)* is the overall number of characters in Web page *i*'s content, and *LOC(i,j)* is the number of characters appearing before the hyper link to *j*. According to this formula, and as illustrated in Figure 9, the value of *V(i,j)* will be higher when the link to Web page *j* appears closer to the the head of page *i* (i.e., *LOC(i,j)* is closer to 0), and will rapidly decline otherwise.

$$V(i,j) = \begin{cases} 0 & ,if\ \ there\ \ is\ \ no\ \ links\ \ from\ \ i\ \ to\ \ j \\ 0.5 - 0.5\,sin\left(\left(\frac{Loc(i,j)}{L_i} - 0.5\right)\pi\right) & ,if\ \ there\ \ is\ \ a\ \ link\ \ from\ \ i\ \ to\ \ j \end{cases} \quad (7)$$

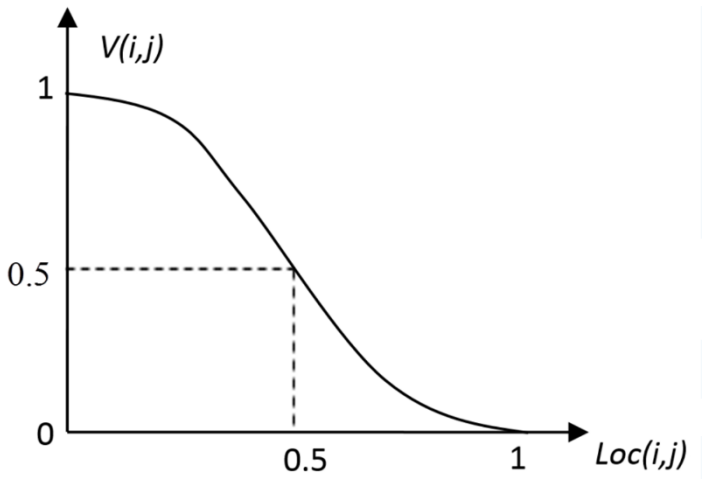


Figure 9. Demo shape of formula (7)

It should be noted that sometimes, there may be several links to page *j* in page *i*. In this case, we simply select the maximum value derived from formula (7) as $V(i,j)$.

4.5 Analytical Model of Probability of (Next) Page Selection

In our model, the first Web page to be visited (Web page *i*=1) can be randomly chosen from all Web pages on the site. Next, the model determines the probability of choosing page *j* as the next page by examining the user's interest in the current theme (c_i), the theme closeness between the current page and page *j* ($S(i,j)$) , and the visibility closeness between the two pages ($V(i,j)$). Now, out of the three parameters, the value of *V(i,j)* is an independent factor that positively feeds into the probability of choosing page *j* as the successor page. (Clearly, links/URLs that are more visually pronounced and 'catchy' have a higher chance of being selected/requested, and vice versa.)

On the other hand, the impact of *ci* and *S(i,j)* on the selection of page *j* is more complicated as it requires that the values of these parameters relative to each other be examined. In particular, if the user's interest in the current theme is high (*ci* is high), and page *i's* theme is very close to page *j's* theme (*S(i,j)* is high), then the probability of visiting page *j* next is high. Similarly, if the user's interest in the current theme is low, and page *i*'s theme is very different from page *j*'s theme, then the probability of visiting page *j* in the next step is also high. In all other cases, page *j* chances to be chosen as the successor page are low(er).

In our implementation of the HBB-IDT model, we use formula (8) to estimate the relative chances of page *j* becoming the successor of page *i* (*E(i,j)*). Ultimately, the page with the highest value of *E(i,j)* will become the actual successor of page *i*.

$$E(\ i,j) = \varphi V(\ i,j) + (1-\varphi)\frac{2c_i \cdot S(\ i,j)}{{c_i}^2 + S^2(\ i,j)} \qquad ,0 \le \varphi \le 1 \qquad (8)$$

In (8), φ is a pre-set parameter that depict the importance/weight of visibility closeness *V(i,j)* relative to the other two parameters. (E.g., for a high φ, we expect the user to be more influenced by the visual organization of a Web page then the elements related to its content/theme, and vice versa.) The combined effect of *S(i,j)* and *ci* on *E(i,j)* is shown in Figure 10. It is obvious from Figure 10 that when the two factors are both high or both low, *E(i,j)* will be significantly higher than when only one of these factors is low.

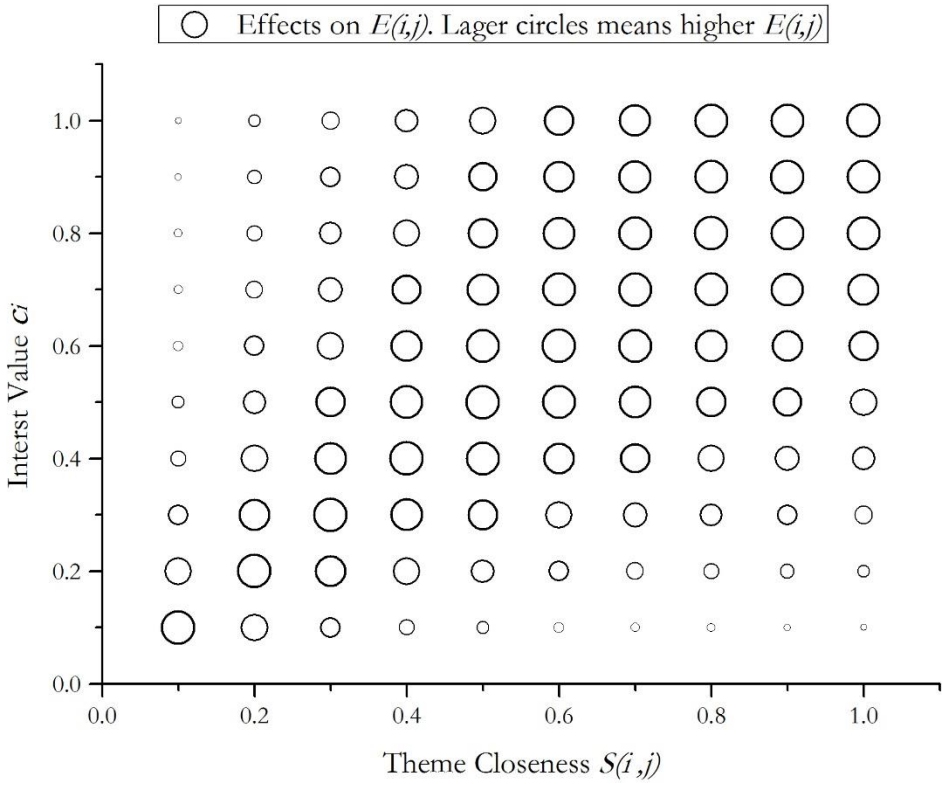


Figure 10. Effects of interest value and theme closeness on selection-priority by formula (8).

From (8), we calculate the probability of page j being chosen as the successor of page *i*- out of all the candidate Web pages (noted by set *W*) – as a normalized value of *E(i,j)* given in (9):

$$P(\ i,j) = \frac{E(i,j)}{\sum_{k=1}^{|W|} E(\ i,k)} \qquad (9)$$

4.6 Model Summary and Discussion

Based on the individual parameter models discussed in the preceding sections, the overall mode of Human Browsing Behavior using Interest-Driven Theory. The following is a brief description of how the model actually operates in real-time:

(1) For a page that is currently browsed/visited by the model, compute the user's interest in the page's (current) theme using formula (1) and (2) from section 4.1 and formula (3) from 4.2.1.

(2) Compute the closeness of the content of the currently browsed page to that of the previously visited page using formula (4), and then obtain the stay time on the current page by formula (5)(6).

(3) For every page linking to the current browse page, compute the closeness of the page's theme to the theme of the current page using formula (3), as well as their respective visibility closeness using formula (7), which was described in 4.4.

(4) Compute every linked pages' possibility of being visited next using formulas (8) and (9) from section 4.5.

(5) Out of all pages linked to the currently visited page the linked-page with the highest probability calculated in step (4) will actually be visited next.

(6) Go back to step (1).

In the end, it is worth pointing that, based on the above descriptions, it is clear that the stay time on a page and the possibility of selecting the next page both depend on the selections of the previous steps. Thus, the HBB-IDT model is essentially a Markov procedure, and as such complies with the general theory of interest-driven human behavior as depicted in works of Barbasi and Zhou.

Chapter 5
Experimental Results

5.1 Evaluation Framework

5.1.1 HBB-IDT Emulation Framework

Based on the model outlined in the previous chapter, we have developed a software framework for the emulation of **H**uman **B**rowsing **B**ehavior using **I**nterest-**D**riven **T**heory (HBB-IDT). The framework can be employed on any target Web site and does not require knowledge of the system logs or prior human behavior on the given site.

The framework is developed in Java and includes the following three components:

(1) **Content Gatherer** crawls and downloads all Web pages from the target Web site. Note that although our model does not require knowledge of the system logs or prior human behavior on the target site, the system does assume that the site's content (i.e., the content of its individual Web pages) is readily known and available.

(2) **Data Analyzer** is responsible for analyzing the content of individual Web pages, determining their respective themes, as well as building the linkage maps of the target site. Data Analyzer's results are stored in a database (we use SQL Server 2014) to support the next functions.

(3) **Human-Mimicking Crawler**. This component is the core of the system. By relying

on the data provided by (1) and (2) and by implementing the functions outlined in Chapter 4, this component aims to (i.e., is capable of) generating browsing sequences on the target Web site that resemble the sequences that would be generated by real human visitors.

5.1.2 Evaluation Software

In addition to HBB-IDT emulation framework, we have also built a software for evaluation of the given framework (i.e., evaluation of HBB-IDT model). In particular, for a given target Web site and a given real browsing sequence on that site, the evaluation software is capable of determining the probability that HBB-IDT model has generated this sequence as opposed to the probability that this sequence was generated by the standard random browsing model. Details of the actual evaluation methods will be introduced in the next section.

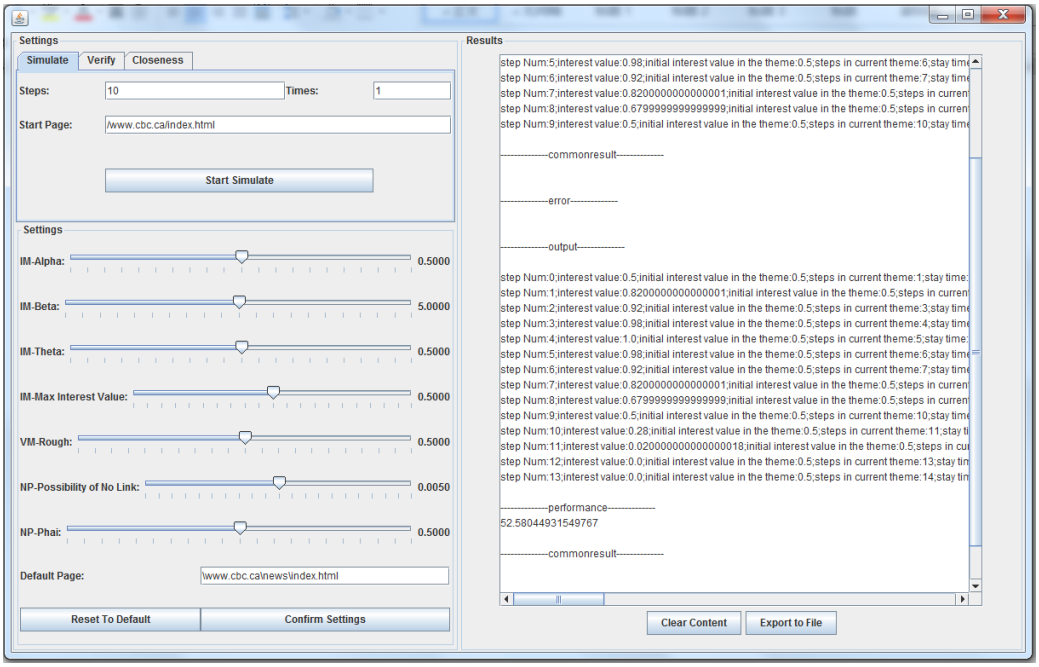


Fig 11. The User Interface of the Evaluation Software

Figure 11 shows the GUI interface of the HBB-IDT evaluation software, together with some running results. Figure 12, on the other hand, shows the architecture of the core classes, including algorithms, simulator components and evaluator components of the HBB-IDT evaluation software.

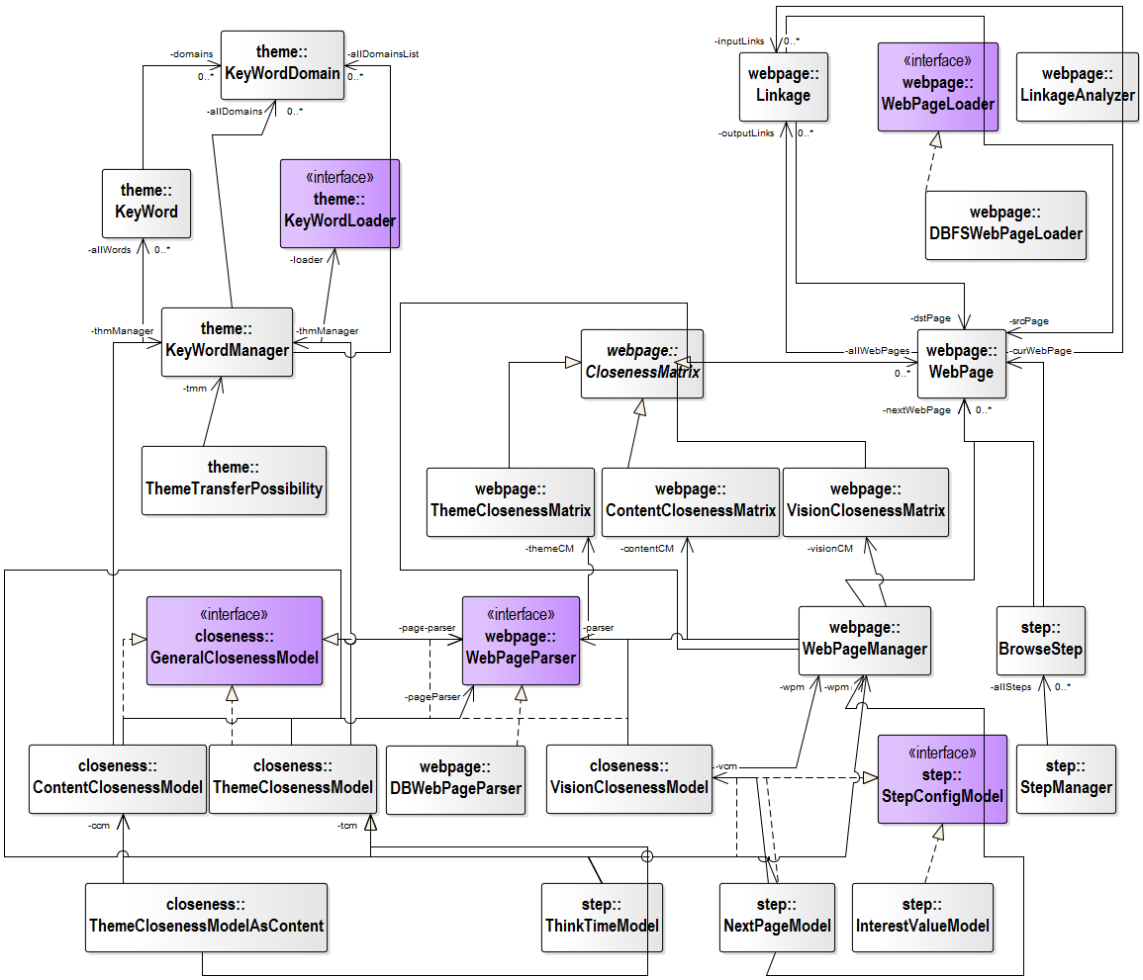


Fig 12. Class Diagram of the HBB-IDT Evaluation Software Framework

(Generated by Enterprise Architecture 12.0)

5.1.3 Simulation Outputs

By running the HBB-IDT emulation framework targeting a test Web site (in our case this was a mirror site of www.cbc.ca - see Section 5.3 for more details), we were able to observe the simulation outputs directly. For an illustration, see Figure 13.

In the simulation shown in Figure 13, the HBB-IDT emulation framework totally lunched 14 Web requests, whose URLs are listed in the column "URL". The column "Step" indicates the sequence number of each request to be visited, and "Stay Time" indicates the for how long the mimic-browser stayed on the Web page.

The outputted URL sequences and stay times in Figure 13 indicate/verify that the HBB-IDT model focuses on meaningful and content-related Web pages most of the time, while other pages, such as advertisements or about pages, are seldom visited. It also changed topics during browsing (in this case, the topics of URL 2-4 are technical news, while URL 1/6/12/14 are social and political news), and the stay time follows a distribution between exponential and power law.

	A	B	C
1	**Step**	**URL**	**Stay Time**
2	0	/www.cbc.ca/index.html	500.000
3	1	/www.cbc.ca/news/world/mexico-missing-students-anniversary-1.html	718.750
4	2	/www.cbc.ca/news/technology.html	923.560
5	3	/www.cbc.ca/news/technology/supermoon-eclipse-apocalypse-1.html	500.000
6	4	/www.cbc.ca/news/technology/everyone-emits-a-unique-cloud-of-microbes-1.html	718.750
7	5	/www.cbc.ca/news/cbc-news-online-news-staff-list-1.html	923.560
8	6	/www.cbc.ca/news/politics.html	997.515
9	7	/www.cbc.ca/sitemap/index.html	838.120
10	8	/www.cbc.ca/connects/index.html	511.928
11	9	/www.cbc.ca/news.html	226.854
12	10	/www.cbc.ca/news/technology/montreal-scientists-score-possible-breakthrough-for-rapid-diagnostic-medical-tests-1.html	79.954
13	11	/www.cbc.ca/news/canada.html	25.592
14	12	/www.cbc.ca/news/canada/ottawa/multimedia/cbc-ottawa-s-news-quiz-for-the-week-of-sept-21-1.html	8.064
15	13	/www.cbc.ca/news.html	2.422
16	14	/www.cbc.ca/news/canada/north/nunavut-s-suicide-inquest-prompts-pleas-for-action-but-progress-slow-1.html	1.009

Fig 13. Part of a Simulation Result on www.cbc.ca

5.2 Evaluation Method

The major goal of our research on the HBB-IDT model is to simulate realistic human-like request sequences. Therefore, it is necessary to identify an evaluation method to determine the degree to which the HBB-IDT model's output is close to the sequences generated by a real human, as well as to compare the outputs of HBB-IDT model with output sequences of other popular Web site workload testing tools.

Currently, there is no universally acceptable way of measuring the accuracy of a human-browsing model. In general, to evaluate the accuracy of a browsing model, it is possible to compare the logs of real humans from the target site with those output by a human-browsing mode using statistical metrics or machine-learning algorithms. However, for this approach to yield accurate results a large number/volume of real logs from the target site would be required, which is often impractical or impossible to obtain. This exactly was the case in our experimentation with www.cbc.ca site .

To compensate for the lack of statistically sufficient amount of real target-site logs, we propose the following evaluation method as a means of comparing the accuracy of the HBB-IDT with that of other models. The basis of this method is the following question: Is it possible for a model to produce a browsing sequence that is identical to a sequence produced by a human? Generally (not strictly), a model that is more likely to generate real sequence can be considered more realistic (i.e., accurate). Thus, by

comparing each model's likelihood of generating a real sequence, it is possible to assess the differences in their respective performances.

Based on the above idea, our evaluation methodology comprises the following procedures:

(1) Select one human-generated URL sequence from the logs collected on our test site (See Section 5.3).

(2) Compute the probability of our HBB-IDT model generating this particular sequence and the expected stay times on each Web page in the sequence.

(3) Compute the probability of other tools/models generating the same sequence.

(4) Compare the two probabilities.

It should be noted that the probability that a site consisting of a large number of Web pages will generate any particular URL sequence is very low (regardless of the employed model) and therefore may not be very informative. Therefore, instead of presenting the two probabilities independently, we compute and present the ratio of the two probabilities, as shown in equation (10).

$$R = \prod_{i=1}^{n} \frac{P(i-1,i)}{P'(i-1,i)} \qquad (10)$$

In formula (10), *R* is the ratio of our model's probability and the probability of another model. *n* is the number of Web pages as found in a provided human sequence.

$P(i-1,i)$ is the probability of moving from page *i-1* to page *i* according to our HBB-IDT model, whereby the pages and order is dictated by the provided human sequence. Based on formula (9) introduced in Chapter 4, we can numerically calculate the value of $P(i-1,i)$. On the other hand, $P'(i-1,i)$ is an equivalent probability corresponding to (i.e., calculated for) the other evaluated model/tool. In this way, we can compute the probabilities to generate a specific real-URL sequence with each model and then compare them relative to each other.

In the following sections, we will illustrate our evaluation process and respective experimental results obtained using the above describe evaluation framework. In particular, Section 5.3 we introduce the test site and real log collection procedure. In Section 5.4, we present our experimental results pertaining to the comparative performance of HBB-IDT relative to other widely used Web site testing tools and models.

5.3 Evaluation Site and Evaluation Procedure

To evaluate our model under real-world conditions, we chose the most popular news Web site in Canada, www.cbc.com as our study case. The reason for this is: (1) the Web site covers a wide range of themes and contents and thus is likely to satisfy a large number of different interests; (2) textual components are the primary source of information in most of its Web pages, which meets our research assumptions; (3) this

is mostly a non-interaction Web site (i.e., most users request and do not posting information), which also agrees with the assumptions of our research.

5.3.1 Test Environment

Clearly, it was not possible for us to gain access to the CBC's human generated logs. Thus, in order to capture the actual human browsing sequences on this site (which would ultimately be needed to evaluate our HBB-IDT model), we set up a mirror-site of the www.cbc.ca/news directory, which will be introduced in detail. The major steps performed while setting up this mirror site include:

(1) Download the Web pages from www.cbc.ca/news using HTTrack Website Copier v3.48 - a popular software that facilitates the creation of Web site mirrors. To avoid downloading of too much content, we have enabled filters in the software setting, including:

a) Do not download news Web pages before 2015.

b) Do not download video and audio files.

c) Do not download special documents such as .doc, .zip, and .pdf.

d) Do not download users' comments posted on each webpage, and disable the comment-functions on the downloaded webpages, so that the mirror site is a pure non-interactive website according to Assumption 1 stated in Section 3.1.1.

e) The maximum depth of links to gather is 4.

f) All links to Web sites other than www.cbc.ca are replaced by a link to an error page, which prompts the visitor to go back to the previous page.

(2) Upload the mirror Web site to Microsoft's cloud virtual server, and configure it in IIS 9.0. The home URL of this site was set to http://cse.cloudapp.net, and the site was accessible from anywhere in the Internet.

(3) To capture the necessary logs and identify different users (even those from the same/shared IP address), an ASP file was placed in the root directory of the Web site and was assigned as the default index page. Then, we set the IIS logging module to catch not only standard information but also the field *CS(COOKIE)*. Therefore, any visitor to the mirror site would be assigned a unique ASPSESSIONID after entering/typing http://cse.cloudapp.net in their browser, and their Session ID would be recorded together with their respective IP address.

By executing the above steps, a mirror of www.cbc.ca/news was finally set up to collect browsing logs generated by our human volunteers.

To consolidate the evaluation, we held the experiment twice during our research, in May 2015 and Sep 2015.

In the first experiment, we began downloading www.cbc.ca at 8:00 pm on May 18th 2015. The procedure lasted 11 hours and was completed at 7:00 am on May 19th 2015. At the end of the procedure, there were 22,740 files (including 13,011 HTML files) on the mirror site, occupying a total of 1.60 GB.

In the second experiment, downloading began at 2:00 pm on Sep 26th 2015, lasted 12 hours and was completed at 2:00 am on Sep 27th 2015. This mirror contains 21,848 files, including 11,421 HTML files, which amounts to a total of 2.18 GB.

5.3.2 Volunteers

As discussed in 5.2, it is/was necessary to collect real-human request sequences to compute the required probability values. Thus, for the purposes of HBB-IDT model evaluation, we solicited several volunteers to browse the mirror site.

All volunteers were from EECS York University and were notified by email with an introduction and consent form. They were asked to enter the Web site and begin browsing the mirror site through the default page http://cse.cloudapp.net/. They were also informed that the experiment was anonymous (the participants did not have to identify themselves).

In the first experiment, four volunteers (excluding the author) attended from 10:00 am to 11:00 pm on May 19th 2015. In the second experiment, there were 14 volunteers (excluding the author) who attended from 8:00 am to 8:00 pm on Sep 27th 2015.

5.3.3 Data Set

According to the IIS logs on the mirror server, there were 5375 URL requests in experiment 1 and 8998 in experiment 2. Based on the logged IP address and ASPSESSIONID data, we were able to group these requests into several sessions, meaning that the same volunteer or BOT sent all requests in one uninterrupted

session. In this way, we extracted 31 sessions in experiment 1 and 17 in experiment 2. Table 1 and 2 show the information for each of these sessions in each experiment, respectively:

Table 1 All Sessions Logged in Experiment 1 for Evaluation

Num	Time	Source IP	ASPSESSION	Page Count	Duration (min)	Request Origin
1	**10:55:43- 12:4:3**	**99.233.3.74**	**CACTRSDT=GBNJFHKBFKAKNBIBHPFJICAF**	**10**	**68**	**Toronto**
2	11:5:12 -20:3:24	66.249.64.177	NO SESSION ID	31	538	Google Bot
3	11:5:14 - 18:47:13	66.249.64.187	NO SESSION ID	24	462	Google Bot
4	11:5:14 - 20:57:56	66.249.64.182	NO SESSION ID	33	593	Google Bot
5	11:13:39 - 11:13:39	66.249.64.177	CACTRSDT=HBNJFHKBKCAMLPHOMNDDHAIK	1	0	Google Bot
6	**11:52:58 - 12:9:53**	**69.46.127.6**	**CACTRSDT=IBNJFHKBFKCIPAHBEMBIMLCC**	**15**	**17**	**Oakville,ON**
7	11:55:59 - 23:0:18	5.9.190.107	NO SESSION ID	6	664	Unknown Bot
8	**13:4:19 - 13:16:18**	**37.24.213.43**	**ACDSQSCS=EEAJIGLBOMFHEAAPHCODHECA**	**7**	**12**	**Europe**
9	14:49:36 - 14:49:42	198.200.64.109	CCCQTTCT=JBADMCMBBFEPLKMAPILOHJKF	2	0	Server Test
10	15:10:48 - 15:10:48	199.212.67.253	CCCQTTCT=KBADMCMBNIIEAICLDNJECDFO	1	0	York U
11	**17:15:23 - 17:28:18**	**66.203.207.67**	**CACQTTDT=MGAFPDNBOAGKAJAGMJFILGHM**	**16**	**13**	**Scarborough**
12	18:57:19 - 18:57:19	24.114.69.40	CADTQSCS=KGOHOPNBAKJFHDEPHLJNMBBE	1	0	Toronto
13	18:59:59 - 19:2:49	37.24.213.43	NO SESSION ID	3	3	Europe
14	20:0:0 - 20:0:0	54.86.138.239	NO SESSION ID	1	0	Amazon Bot
15	20:0:2 - 20:0:2	52.7.18.41	NO SESSION ID	1	0	Amazon Bot
16	20:0:2 - 21:5:23	198.200.64.109	AQSBBADT=NAJHJDDBGKJPJMKAAEILMKPF	2	65	Server Test
17	20:0:10 - 22:59:33	5.9.190.101	NO SESSION ID	4	179	Cxense Bot
18	20:0:15 - 20:4:49	89.145.95.42	NO SESSION ID	3	5	UK

19	20:6:6 - 20:6:6	::1	ASSDDDAT=CLJNHNDBCLHMLJGOBLBPBAAH	1	0	Server Test
20	20:8:7 - 20:8:7	52.0.133.35	NO SESSION ID	1	0	Amazon Bot
21	20:41:46 - 21:26:13	198.200.64.109	CADTQSCS=MGOHOPNBJEJMNANACNPODLJI	4	44	Server Test
22	20:47:40 - 20:47:40	52.7.185.248	NO SESSION ID	1	0	Amazon Bot
23	21:11:18 - 21:11:18	24.212.213.53	CADTQSCS=NGOHOPNBBIDMFOLCPJAONHDM	1	0	Mississauga, ON
24	21:12:19 - 21:12:19	52.7.218.87	NO SESSION ID	1	0	Amazon Bot
25	**22:54:32 - 23:4:34**	**198.200.64.109**	**AABRTTCT=JDDFJLPBNIGJMEJAECKBMMEE**	**17**	**10**	**The Author**
26	2:6:50 - 2:6:50	66.249.78.252	ASQBBBDT=LEIPFJGBPGHLEGNDAEHKGCJO	1	0	Google Bot
27	2:10:30 - 9:49:26	66.249.78.252	NO SESSION ID	4	459	Google Bot
28	2:11:32 - 2:11:51	66.249.67.216	NO SESSION ID	2	0	Google Bot
29	2:11:44 - 9:49:26	66.249.78.246	NO SESSION ID	2	458	Google Bot
30	2:31:29 - 2:31:29	95.211.224.136	NO SESSION ID	1	0	leaseweb bot
31	7:49:31 - 7:49:31	99.251.202.34	ACCTQSCS=FLMHEKDCGFPMNJFJMCLKNJOK	1	0	Fredericton, NB

*** IP locations were looked up on http://whatismyipaddress.com/**

Table 2 All Sessions Logged in Experiment 2 for Evaluation

Num	Time	Source IP	ASPSESSION	Page Count	Duration(min)	Request Origin
1	8:34:37 - 8:34:54	99.227.146.169	SCSBTRRA=JPJJDOODFANHAPHLGPMIFCPI	4	0.3	Brampton,ON
2	8:37:38 - 11:23:15	198.91.172.239	SCSBTRRA=LPJJDOODPIDGKEPCFGNDKNMA	16	166	The Author
3	**8:37:58 - 8:47:54**	**99.227.146.169**	**SCSBTRRA=KPJJDOODLGDLOACLNFPBDAOE**	**14**	**10**	**Brampton, ON**
4	**9:5:27 - 9:34:34**	**99.225.221.240**	**SCSBTRRA=MPJJDOODCGIHHELEFBNPBJKH**	**11**	**29**	**Toronto,ON**
5	9:40:57 - 9:40:57	184.173.183.174	NO SESSION ID	1	0	Softlayer Bot
6	**9:52:49 - 9:57:56**	**65.94.54.180**	**SCSBTRRA=NPJJDOODBINCLFJEACBNCFBH**	**7**	**5**	**Toronto,ON**
7	**10:9:31 - 10:16:44**	**99.225.238.208**	**SCSBTRRA=OPJJDOODDCKHMFPKINCCBHMG**	**14**	**7**	**Toronto,ON**
8	**10:43:13 - 10:58:36**	**216.105.80.21**	**SCSBTRRA=AAKJDOODHMJLPHFGFFIKIEOO**	**24**	**15**	**Toronto,ON**

9	**10:44:36 - 10:53:56**	**99.238.81.13**	**SATBSRQA=DIKBGPADNLHEBBDCOBODNFAE**	**7**	**9**	**Markham,ON**
10	**10:47:4 - 10:59:51**	**99.234.62.15**	**SCSBTRRA=BAKJDOODHOGJLGEPKKKABKGH**	**13**	**13**	**Toronto,ON**
11	**11:2:22 - 11:9:56**	**104.255.14.239**	**SCSBTRRA=CAKJDOODPLPEKEMFLOGBAFDJ**	**5**	**8**	**Toronto,ON**
12	**11:10:8 - 11:13:4**	**99.233.223.79**	**SCSBTRRA=DAKJDOODOIAPILCALBLDPPKE**	**11**	**3**	**Toronto,ON**
13	**11:58:53 - 16:7:34**	**99.238.110.111**	**QCSDRRQB=NMIPPIAALIGOPMHMDOEHBPEO**	**14**	**249**	**Thornhill,ON**
14	12:51:36 - 12:51:39	198.91.172.239	QAQDSRRB=NCBNCPAAACAINFEKGHCCLLNJ	2	0	The Author
15	**13:10:47 - 13:24:49**	**99.234.210.73**	**QAQDSRRB=OCBNCPAAJMAJAKNMEAMFPGDA**	**14**	**14**	**Toronto,ON**
16	**17:42:44 - 17:55:22**	**99.225.235.249**	**QCQBSRQA=COPHEBDAHPLFFBAOFFENLIBJ**	**14**	**13**	**North York, ON**
17	**19:0:8 - 19:15:56**	**70.55.34.47**	**SCRBSQRA=PJLNGKDANIKANJIEFKFOEBFH**	**19**	**16**	**Toronto,ON**

*** IP locations were looked up on http://whatismyipaddress.com/**

In Table 1 and 2, the column "Page Count" indicates how many Web pages were requested during the respective session. It must be noted that this count is much smaller than the count of actual URL/HTTP requests in Web logs because a visit to one Web page can prompt several URL/HTTP requests (typically, 50 requests per page), e.g., images, javascript and css files. The column "Request Origin" shows the geographic locations from which the requests were sent. This information was obtained by performing an IP address lookup using **http://whatismyipaddress.com/**.

According to Table 1, most of the recorded requests in the first experiment were from Web Bots. On the other hand, sessions enumerated 1, 6, 8, 11, 25 (also highlighted as bold in the table) can clearly be associated with human visitors to the site, including volunteers and the author himself. However, in Table 2, there is only one session

(session 5) that is recognized as a BOT, while the other 16 are all from volunteers or the author. The BOT counts in experiment 1 are much higher because the mirror server had been running for 1 month before experiment 1; thus, BOTs from Google and other companies had enough time to find and reach this server. After experiment 1, we shut down the server for 4 months until the day before experiment 2 started. Therefore, few BOTs knew of the existence of the mirror server during experiment 2.

According to the information given in Table 1 and 2 (column *Page Count* and column *Duration*), average counts of Web pages browsed by real-human visitors are 13 (in Table 1) and 12.3 (in Table 2), and average time per session are 24 minutes and 35 minutes respectively.

For a more objective dataset, we abandoned the author-sessions in both of the experiments. Therefore, we identified 4 and 14 human-generated sessions in experiment 1 and 2, respectively, which were used to evaluate our HBB-IDT model.

5.3.4 Additional Data Pre-Processing

Prior to the evaluation, the raw data corresponding to the human-generated sessions (that were collected as described in Section 5.3.3) had to be processed further for the following reasons:

Some re-visited Web pages were not recorded by the logs because of the caching mechanism built into some of the clients' browsers. (For example, for a user with a browser cache, if the user's real browsing sequence is Page A -> Page B -> Page A ->

Page C, the recorded sequence in the server-logs may appear as Page A-> Page B -> Page C, even though Page B may have no direct links to Page C. In this case we can deduce with 100% certainty that Page A was revisited before Page C by checking the value of ‘referrer’ filed in the HTTP request for Page C as captured in the logs.) A similar omission occurs in pages that can automatically redirect to other pages using the <REFRESH> tag.

To automate the process of ‘insert back all omitted URL requests’, as well as the process of removing irrelevant URL (i.e., requests for secondary Web pages), we developed a Visual Basic (VBA) routine. Figure 14 shows snapshots of logs in Experiment 1 before and after this additional processing has been done.

	B	E	I	K	
4	time	cs-uri-stem	c-ip	cs(Cookie)	cs(Referer)
2831	17:04:19	/www.cbc.ca/news.html	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	-
2832	17:04:19	/www.cbc.ca/i/css/v11/scripts.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2833	17:04:19	/www.cbc.ca/i/o/globalnav/v10/css/dropdown.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2834	17:04:19	/www.cbc.ca/i/o/globalnav/v10/css/globalnav.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2873	17:04:19	/www.cbc.ca/i/o/sm/v10/gfx/sprite.png	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2874	17:04:20	/i/o/ticker/sprite.gif	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2875	17:04:20	/www.cbc.ca/i/news/v10/gfx/icon_a-v.png	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2877	17:04:34	/www.cbc.ca/sports/hockey/nhl/russians-may-be-sanctioned-for-o-canada-snub-1.html	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2878	17:04:34	/www.cbc.ca/i/news/v10/css/storypage-print.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2879	17:04:34	/www.cbc.ca/i/sports/v11/css/quicklinks.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2880	17:04:34	/www.cbc.ca/cmlink/7-7.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2881	17:04:34	/www.cbc.ca/sports-content/v11/includes/css/sn-modal.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2882	17:04:34	/fast.fonts.net/cssapi/e1221177-ee49-4844-8771-0dbbef470ab2.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2883	17:04:34	/www.cbc.ca/i/sports/v11/css/goalfeed.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2884	17:04:34	/www.cbc.ca/i/sports/v11/css/sports-video.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2885	17:04:34	/www.cbc.ca/sports-content/v12/plugins/playoffbracket/style.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou
2886	17:04:34	/www.cbc.ca/i/sports/v11/css/nav.css	37.24.213.43	ASPSESSIONIDACDSQSCS=EEAJI(	http://cse.clou

Fig 14.a). A portion of the logs before processing. It can be seen that the request for each actual HTML document (Web page) is followed by tens of secondary URL requests (i.e., requests for non-textual objects embedded in the given page).

	A	B	C	D	E	F
1	stay time	url	Refer	Source IP	Session ID	time
2	15	/www.cbc.ca/news.html		37.24.213.43	ACDSQSCS=EEAJ	0.71133102
3	352	/www.cbc.ca/sports/hockey/nhl/russians-may-	/www.cbc.ca/news.html	37.24.213.43	ACDSQSCS=EEAJ	0.71150463
4	0	/www.cbc.ca/news.html				0.7155787
5	247	/www.cbc.ca/news/canada/british-columbia/ma	/www.cbc.ca/news.html	37.24.213.43	ACDSQSCS=EEAJ	0.7155787
6	0	/www.cbc.ca/news.html				0.7184375
7	37	/www.cbc.ca/news/world.html	/www.cbc.ca/news.html	37.24.213.43	ACDSQSCS=EEAJ	0.7184375
8	27	/www.cbc.ca/news/politics.html	/www.cbc.ca/news/world.html	37.24.213.43	ACDSQSCS=EEAJ	0.71886574
9	41	/www.cbc.ca/news/technology.html	/www.cbc.ca/news/politics.html	37.24.213.43	ACDSQSCS=EEAJ	0.71917824
10		/www.cbc.ca/news/technology/the-internet-s-	/www.cbc.ca/news/technology.htm	37.24.213.43	ACDSQSCS=EEAJ	0.71965278

Fig 14.b). Logs from Fig 14.a) after processing. Only actual visited Web pages are extracted, and missing URLs are added back according to CS(REFERER), as shown in the 4th and 6th row. The stay time for each Web page is also calculated.

5.4 Experimental Set-up and Results

Based on the evaluation method proposed in section 5.2, we calculate the probabilities of the HBB-IBT model generating the real sequences described in 5.3 as well as the probabilities of these sequences being generated by some common Website workload testing tools / models.

5.4.1 Tools/Models Used for Comparison

There are many Website workload testing tools that are used in practice, and they deploy different request-sequence-generating models. By reading their documents or source codes, we were able to classify them into the following categories:

(1) Tools that Use Pre-set Request List Model (PRLM, Sample: *Apache ab*).

Many workload testing tools have adopted this model, including the Apache HTTP server benchmarking tool “ab” [2]. In this tool, the user provides/sets a URL list in the

command line or a configure file before running the tool. The testing tool then generates requests by randomly choosing URL from this list based on a uniform distribution.

In our experiment, to use PRLM-based tools to simulate a request sequence, we listed all Web page files (13011 HTML Web pages on the mirror server in Experiment 1 and 11421 in Experiment 2) in a configure file, and the model randomly selected one at the time. Clearly, with this model, the probability of jumping to Web page *i* from Web page *i-1* – annotated as $P'(i-1,i)$ in formula (10) – is always $\frac{1}{N-1}$, where *N* is the overall number of all pages in the target site. In the case of www.cbc.ca, this value is $\frac{1}{13010}$ in Experiment 1 and $\frac{1}{11420}$ in Experiment 2.

(2) Randomly Crawling Model (RCM, Sample: *JMeter*).

Some popular testing tools also support Web page crawling functions, which means they are able to jump from one page to another provided the later page is linked to the previous one. For example, the widely used tool JMeter provides an HTML-Link-Parser so that it can randomly choose and visit a link in the current page with equal probability (uniform distribution) [4].

Hence, in the case of this model, the value of $P'(i-1,i)$ in RCM is $\frac{1}{n}$, where *n* is the link-count of page *i*.

(3) Record-Replay Model (RRM, Sample: *Tsung*).

To achieve outputs that are more similar to those of humans, tools such as Tsung also provide interfaces to record users' browsing sequences and then replay these sequence in future tests [3]. In particular, in the case of this tool, the user is first asked to visit the

target Web site in a special Web browser which can record users' activities. All pages he/she has visited are recorded into one or several session files. Subsequently, the user can launch a test in the tool, and all request sequences generated by the tool are copies of the previously recorded sessions.

This tool may be good for testing of static Web-sites – where the content of the site and thus the 'most likely' human sequences do not change over time. However, in case of dynamic web-sites - such as new-agency sites, including CBC – the pre-recording of sessions is likely not to yield good/accurate/realistic results. For this reason, we didn't compare RRM with HBB-IDT in this research.

5.4.2 Results

According to the evaluation methods described in 5.2, we compared the probability of generating real sequences using the HBB-IDT model against the probability of those sequences being generated by the PRLM and RCM models (introduced in Section 5.4.1) using formula (10).

5.4.2.1 Comparison of Human-Generated Sessions

(1) URL Sequence Comparison

In the dataset that was collected in Experiment 1, 4 out of 31 sessions were launched by volunteers, whose session numbers were 1/6/8/11. Table 3 shows the results of the comparison for these four particular sessions.

Table 3 Evaluation Results for the Simulation of Real Human Sequences in Experiment 1

Session Num. (Num. in Table 1)	Ratio of HBB-IDT vs PRLM (*R* in Equation 10)	Ratio of HBB-IDT vs RCM (*R* in Equation 10)
1	$3.75 \text{ X } 10^{23}$	131.54
6	$7.87 \text{ X } 10^{32}$	4738.66
8	$1.01 \text{ X } 10^{18}$	515.44
11	$1.36 \text{ X } 10^{51}$	175.82

Based on Table 3, it is obvious that the HBB-IDT is always far more likely to generate real human sessions than the random model. For example, with respect to generating the same request sequence as in Session 1, the probability obtained with the HBB-IDT model is 131.54 times greater than the probability obtained using Random Crawling Model, and $3.75 \text{ X } 10^{23}$ times greater than the probability obtained using Pre-set Request List Model.

In the dataset that was collected in Experiment 2, there were 13 volunteer sessions. The respective results are shown Table 4:

Table 4 Evaluation Results on Simulating Real Human Sequences in Experiment 2

Session Num. (Num. in Table 2)	Ratio of HBB-IDT vs PRLM (*R* in Equation 10)	Ratio of HBB-IDT vs RCM (*R* in Equation 10)
1	$5.74 \text{ X } 10^{5}$	1.99
3	$2.66 \text{ X } 10^{44}$	2262.86
4	$1 \text{ X } 10^{36}$	825.59
6	$4.67 \text{ X } 10^{13}$	7.71
7	$7.13 \text{ X } 10^{50}$	608.66
8	$8.53 \text{ X } 10^{94}$	2676784.97
9	$3.99 \text{ X } 10^{17}$	7.26
10	$1.95 \text{ X } 10^{43}$	2344.59
11	$8.11 \text{ X } 10^{11}$	14.01
12	$5.33 \text{ X } 10^{54}$	13835.93

13	2.70 X 10^{39}	8861.56
15	3.70 X 10^{41}	49010.34
16	5.98 X 10^{54}	37859.40

According to Table 4, the comparison results in Experiment 2 are similar to those collected in Experiment 1, showing that the HBB-IDT has a much higher probability of simulating human browsing sessions compared to the Random Crawling model and the Pre-set Request List Model. In certain extreme cases (for example in session 8 in table 4), the probability of generating a real human browsing sequence using the HBB-IDT is as much as 2.67 million times greater than the probability obtained using the RCM.

(2) Stay Time Comparison

The stay time of each Web page is another output of a simulation model and is used as an important evaluation metric.

Based on the raw logs collected in Experiments 1 and 2, we obtained the time-stamps of each Web page (i.e., the time when a Web page, in a sequence, was actually requested). Based on the obtained values we were able to estimate the volunteers' stay times on each Web page using the formula $d_i = t_{i+1} - t_i$, in which t_i and t_{i+1} are the time stamps of Web page *i* and *i+1*, and d_i is the estimated stay time on Web page *i*. We also computed the expected stay times on these Web pages using the HBB-IDT model, as explained in section 5.2 step (2).

In case of random models such as RCM and PRLM, stay time values are generally random numbers following a uniform distribution. Thus, to evaluate/compare the stay times of these two models, we generated a random number for each Web page in a sequence and used it as this page's respective stay time.

By aggregating the above data, we obtained three stay time vectors: Actual Stay Time (estimated), HBB-IDT Stay Time and Random Model (RCM/PRLM) Stay Time. All vectors are of the same dimensionality, which is the count of URL-requests launched by human volunteers in the respective experiment. Based on Table 1 and 2, the vector dimensionalities are 48 in Experiment 1 and 169 in Experiment 2. Figure 15 shows a fraction of these vectors corresponding to Experiment 2.

After generating/computing the stay time vectors in Experiment 2, we subsequently computed: a) the Pearson correlation between the HBB-IDT stay time vector and the actual stay time vector, and 2) the Pearson correlation between the Random model stay time vector and the actual stay time vector. Figure 15 shows the obtained values of the two correlation matrices. These values indicate that the stay times of the HBB-IDT model are weakly positively correlated to the actual stay times (0.226 on the significant level 0.007), while the Random model values are non-correlated (-0.005 on the significant level 0.95) with the actual stay times. Thus, we conclude that, overall, the HBB-IDT model results in better stay times relative to the random (RCM and PRLM) models.

	A	B	C	E
1	**Visited Web Page**	**Actual ST**	**RCM ST**	**Model ST**
2	/www.cbc.ca/news.html	1	224	100
3	/js.indexww.com/ht/cbc.html	1	693	216
4	/js.indexww.com/ht/cbc.html	1	110	216
5	/www.cbc.ca/news/politics/canada-electio	256	841	100
6	/www.cbc.ca/news/politics/justin-trudeau-	277	600	216
7	/www.cbc.ca/news/business/microsoft-offi	405	306	776.7964
8	/www.cbc.ca/news/canada/british-columbi	164	476	100
9	/www.cbc.ca/news/health/heart-stroke-sat	237	522	100
10	/www.cbc.ca/news/world/hajj-stampede-d	218	417	216
11	/www.cbc.ca/news/technology/a-hidden-o	47	440	100
12	/www.cbc.ca/sports/vote-best-canadian-sp	1	527	100
13	/www.cbc.ca/news.html	1	747	100
14	/js.indexww.com/ht/cbc.html	1	680	216
15	/www.cbc.ca/news/business/microsoft-offi	217	731	776.7964
16	/www.cbc.ca/news/world.html	19	664	100
17	/www.cbc.ca/news/canada/toronto-18-ring	349	583	776.7964
18	/www.cbc.ca/news.html	19	348	999.6636
19	/js.indexww.com/ht/cbc.html	1	699	216
20	/www.cbc.ca/news/canada/newfoundland-	104	477	100

Fig 15. Stay Time Vectors. Actual ST is the actual stay time on each Web page as derived from recorded volunteer sessions, RCM ST is the expected stay time of Random models, and Model ST is the expected stay time of the HBB-IDT model.).

Pearson Correlations

		ACTRUAL ST	HBB-IDT ST
ACTRUAL ST	Pearson Corr.	1	0.22594
	Sig.	--	0.00686
HBB-IDT ST	Pearson Corr.	0.22594	1
	Sig.	0.00686	--

Pearson Correlations

		ACTRUAL ST	RCM ST
ACTRUAL ST	Pearson Corr.	1	0.00529
	Sig.	--	0.95019
RCM ST	Pearson Corr.	0.00529	1
	Sig.	0.95019	--

Fig 16. Correlation matrices of the HBB-IDT and Random Models to Actual Stay times. These results were calculated using Origin Pro correlation analysis tools.

In addition to the results presented in Figure 16, we also compared the frequency distributions of the three stay time vectors. Here the term *frequency* indicates the

occurrences of stay-time values within an interval in each vector. For example, in the vector of Acctual-Stay-Time, there are 3 URL-requests whose stay time values are 0, 42 between 0 to 10, and 23 between 10 to 20. Then the frequency of stay time in interval [0,0] is 3, that of stay time in interval (0,10] is 42, and that in interval (10,20] is 23. As shown in Figure 17, the stay time frequency distribution of HBB-IDT is much closer to that of actual cases than the Random models. Furthermore, we also computed the Cumulative distribution of stay time to make the comparison much clear. Cumulative frequency means the count of all values not greater than a given value. For instance, in the above example, the cumulative frequency of stay time value 20 is 68, which equals *3+42+23*. The cumulative distributions of the three stay-time vectors are shown in the lower part of Figure 17, which also reveals that HBB-IDT is much closer to the actual case.

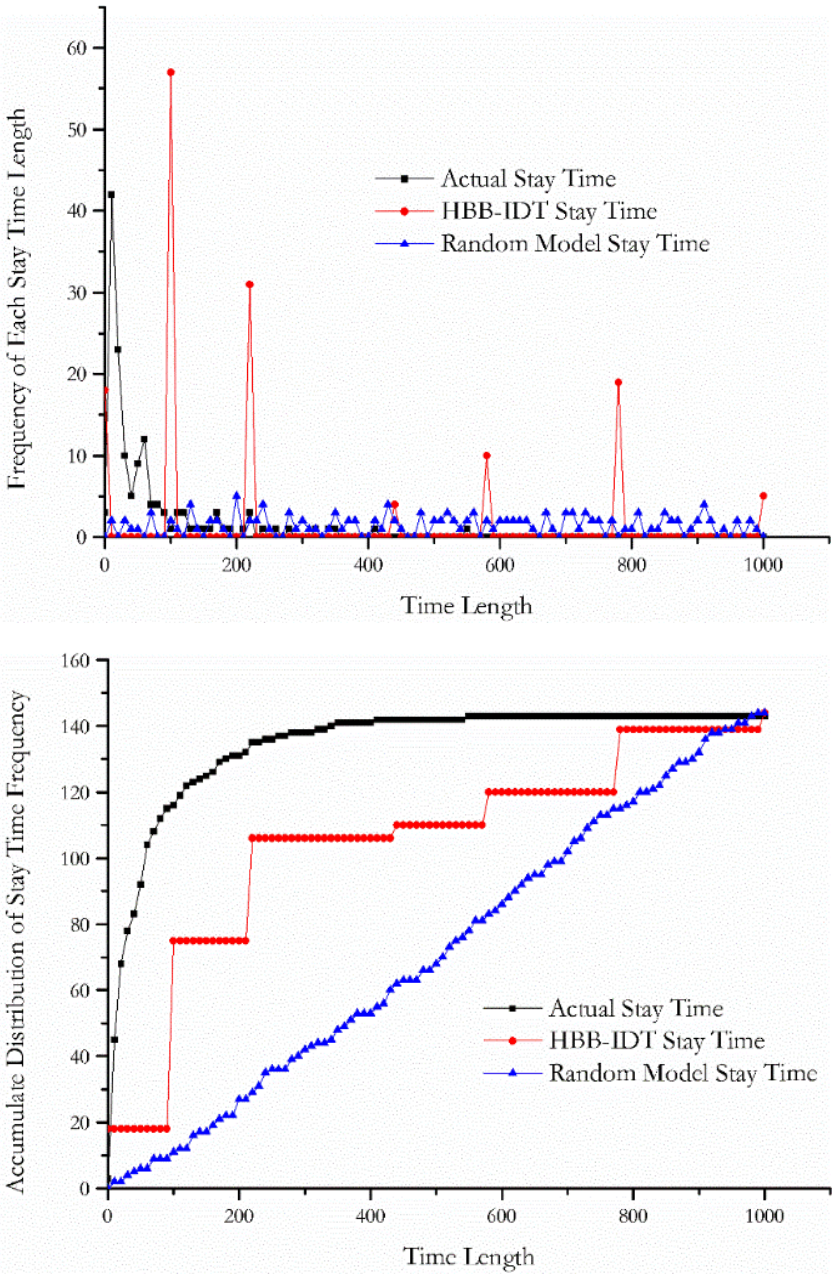


Fig 17. Stay Time Frequency Distributions (upper graph) and Cumulative Frequency Distributions (lower graph).

Because of time limitations, we chose not to address additional details regarding stay time distributions. However, based on the above analysis and figures, it is still obvious that HBB-IDT generates more human-like Web page stay times than RCM and PRLM do.

5.5 Discussions and Limitations

The results of our experimentation with HBB-IDT support the assumption that our model is more accurate than the existing models/tools, such as the Random Crawling Model adopted by JMeter and the Pre-set Request List Model adopted by Apache-ab, with respect to simulating human Web browsing behavior.

However, there are limitations to the experiment and evaluation.

(1) Web pages with titles that contain French characters were not recognized by the simulation software, thus affecting simulation accuracy.

(2) The content of the mirror Web site were not 100% up to date (at least 12 hours were needed to create and upload a mirror), which was a factor potentially affecting/skewing volunteer behavior.

(3) Clearly, the experiments involved a limited number of volunteers. Ideally this number would have been larger in order to increase the overall confidence in the obtained results.

Chapter 6
Conclusions and Future Work

In this thesis, we propose an interest-driven model of human browsing behavior in the WWW. The model is based on the well-known theory of interest driven behavior, and is applied to Web sites that are not visited previously or do not have sufficient statistical information about past user browsing behavior.

In this chapter, we outline open issues and discuss future research directions.

(1) Bi-directional interactions between users and Web sites are not considered in the current model. Currently, our research and experiments focus mainly on non-interactive webpages, such as a university's portal site or a news agency's web site. These websites do not allow users to modify or add to the content they are reading. User behavior may be different on interactive websites because their interest is now extended from reading to communicating with others. In the future, we will extend the current model to take into account user behavior on interactive websites.

(2) The accuracy of the model can be further enhanced by using more advanced semantic and visibility analysis. Several advanced methods [28] – [35] exist that can accurately evaluate the semantics and visibility of webpages. These methods are very complex and powerful, and capable of addressing problems caused by

asynchronous technologies such as AJAX. In the next stage of our work, we will incorporate these methods into our model to enhance its accuracy.

(3) We will also continue to evaluate the model on other real-world Web sites using larger numbers of human participants.

Although the research presented in the thesis is ongoing, we expect that the HBB-IDT model and the initial experimental results will act as a catalyst for broader discussion and ultimately mark a new era in the design and utilization of bots/crawlers mimicking human behavior.

Bibliography

[1] Busari M. and Williamson C., "*ProWGen: a synthetic workload generation tool for simulation evaluation of web proxy caches*". Computer Networks, vol. 38, no.6, pp. 779-794, 2002.

[2] Apache Software Foundation, *ab - Apache HTTP server benchmarking tool*, http://httpd.apache.org/docs/2.2/programs/ab.html, 2014.

[3] Niclausse N., *Tsung's documentation*, http://tsung.erlang-projects.org, 2014.

[4] Apache Software Foundation, JMeter User's Manual, http://jmeter.apache.org/usermanual/component_reference.html#HTML_Link_Parser, 2015.

[5] Georgios O., and Mirkovic J., "*Modeling human behavior for defense against flash-crowd attacks.*", Proceedings of the 2009 IEEE international conference on Communications, pp. 625-630, 2009.

[6] Jung J., Krishnamurthy B., and Rabinovich M., "*Flash crowds and denial of service attacks:Characterization and implications for CDNs and web sites*", Proceedings of the 11th international conference on World Wide Web, pp.293-304, 2002.

[7] Xie Y., Yu S., "*Monitoring the application-layer DDoS attacks for popular websites*", IEEE/ACM Transactions on Networking, vol.17, no.1, pp.15-25, 2009

[8] Gavrilis, D., Chatzis, I., Dermatas E., "*Flash Crowd Detection Using Decoy Hyperlinks"*, Proceedings of the 2007 IEEE International Conference on Networking, Sensing and Control, pp.466-470, 2007.

[9] Stuart S., Krishnan M., and Smith M. D., "*Using Path Profiles to Predict HTTP Requests.*" Proceedings of the Seventh International World Wide Web Conference (Computer Networks and ISDN Systems), pp. 457–467, 1998.

[10] Zhong S., Yang Q., Lu Y., and Zhang H., "*WhatNext: A Prediction System for Web Requests Using N-Gram Sequence Models*", Proceedings of the First International Conference on Web Information Systems Engineering, pp.214-221, 2000.

[11] Zukerman, I., Albrecht. W., and Nicholson A., "*Predicting user's request on the WWW",* Proceedings of the Seventh International Conference on User Modeling, pp.275-284, 1999.

[12] Awad M. A., Khalil I., *"Prediction of user's web-browsing behavior: Application of markov model*". IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 42 no.4, pp.1131-1142, 2012.

[13] Deshpande M., Karypis G., "*Selective Markov models for predicting Web page accesses"*. ACM Transactions on Internet Technology (TOIT), vol. 4, no.2, pp. 163-184, 2004.

[14] Lee C. H., Lo Y. and Fu Y H., "*A novel prediction model based on hierarchical characteristic of web site"*, Expert Systems with Applications, vol. 38 no.4, pp.3422-3430, 2011

[15] Nigam B. and Jain S., *"Generating a new model for predicting the next accessed web page in web usage mining",* Proceedings of the International Conference on Emerging Trends in Engineering and Technology, pp. 485-490, 2010.

[16] Poornalatha G., and Prakash S. R., "*Web Page Prediction by Clustering and Integrated Distance Measure*", Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, pp. 1349 – 1354, 2012.

[17] Barabasi A.-L., "*The origin of bursts and heavy tails in human dynamics*", Nature vol. 435, pp. 207-211, 2005.

[18] Barabasi A.-L., "*The Architecture of Complexity*", IEEE Control Systems, vol. 27, no. 4, pp. 33–42, 2007.

[19] Zhou T., Han X., and Wang B., "*Towards the understanding of human dynamics*", Science matters: humanities as complex systems, pp. 207-233, 2008.

[20] Zhou T., Kiet H. A. T., Kim B. J., Wang B. and Holeme P., *"Role of activity in human dynamics"*, Europhysics Letters, vol. 82, no 2, 2008.

[21] Han X., Zhou T. and Wang B., *"Modeling human dynamics with adaptive interest"*, New Journal of Physics, vol. 10, 2008.

[22] Dezsö Z., Almaas E., Lukács A., Rácz B., Szakadát I.and Barabasi A.-L., "*Dynamics of information access on the web*", Physical Review E, vol. 73, no.6, pp. 066132.1-066132.6., 2006.

[23] Gonçalves B., and Ramasco J. J., "*Human dynamics revealed through Web analytics*." Physical Review E, vol 78, no.2, pp. 026123.1-026123.7, 2008

[24] Liu C., White R.W., and Dumais S., "*Understanding web browsing behaviors through weibull analysis of dwell time*", Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 379–386, 2010.

[25] Goel S., Hofman J. M., and Sirer M. I., *"Who Does What on the Web: A Large-Scale Study of Browsing Behavior"*, Proceedings of the 6th International Conference on Weblogs and Social Media, Pp.130-137, 2012.

[26] Torres S. D., Weber I., and Hiemstra D., *"Analysis of search and browsing behavior of young users on the web"*, ACM Transactions on the Web, vol 8 no.2, pp.1-54, 2014

[27] Shen X., Dumais S. and Horvitz E., *"Analysis of Topic Dynamics in Web Search"*, Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, pp. 1102–1103, 2005.

[28] Mabroukeh N. R., and Ezeife C. I., *"Using Domain Ontology for Semantic Web Usage Mining and Next Page Prediction"*, Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp.1677-1680, 2009.

[29] Hoxha J. and Agarwal S., *"Semantic formalization of cross-site user browsing behavior"*, Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, pp. 488-495, 2012.

[30] Finch, S. P. and Chater, N. *"Boots trapping syntactic categories"*, Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society of America, pp. 820-825, 1992.

[31] Bullinaria J. A. and Levy J. P., *"Extracting semantic representations from word co-occurrence statistics: A computational study"*, *Behavior research methods*, vol 39 no.3 pp. 510-526, 2007.

[32] Harik G. R. and Henzinger M. H., US Patent 7716216 (Assigned to Google), "*Document ranking based on semantic distance between terms in a document"*, 2010.

[33] Gupta S., Kaiser G. E., Grimm P., Chiang M. F. and Starren J., "*Automating content extraction of html documents"*, World Wide Web, pp.179−224, 2005.

[34] Zhai YH, Liu B., *"Structured data extraction from the Web based on partial tree alignment"*, IEEE Trans. on Knowledge and Data Engineering, vol 18, pp. 1614−1628, 2006.

[35] Zunger Y., US Patent 7913163 (Assigned to Google), *"Determining semantically distinct regions of a document"*, 2011

[36] Singhal A., *"Modern Information Retrieval: A Brief Overview",* Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol.24, pp. 35–43,2001.

[37] Magnini B. and Cavaglià G., *"Integrating Subject Field Codes into WordNet",* Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, pp. 1413-1418, 2000.

[38] Bentivogli L., Forner P., Magnini B. and Pianta E., *"Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing"*, Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, pp. 101-108, 2004.

[39] Anderson N. J., *"Improving reading speed"*, English Teaching Forum, vol. 6, Apr-Jun, pp. 2-5, 1999.