Experience: Analyzing Missing Web Page Visits and Unintentional Web Page Visits from the Client-side Web Logs

CHE-YUN HSU, TING-RUI CHEN, and HUNG-HSUAN CHEN, Computer Science and Information Engineering, National Central University, Taiwan

Web logs have been widely used to represent the web page visits of online users. However, we found that web logs in Chrome's browsing history only record 57% of users' visited websites, i.e., nearly half of a user's website visits are not recorded. Additionally, 5.1% of the visits recorded in the web log occur because of unconscious user actions, i.e., these page visits are not initiated from users. We created a Google Chrome plugin and recruited users to install the plugin to collect and analyze the conscious URL visits, unconscious URL visits, and "missing" URL visits (i.e., the visits unrecorded in the traditional web log). We reported the statistics of these behaviors. We showed that sorting popular website categories based on traditional web logs differs from the rankings obtained when including missing visits or excluding unintentional visits. We predicted users' future behaviors based on three types of training data – all the visits in modern web logs, the intentional visits in web logs, and the intentional visits plus missing visits in web logs. The experimental results indicate that missing visits in web logs may contain additional information, and unintentional visits in web logs may contain more noise than information for user modeling. Consequently, we need to be careful of the observations and conclusions derived from web log analyses because the web log data could be an incomplete and noisy dataset of a user's visited web pages.

CCS Concepts: • **Information systems** → **Web log analysis**; *Computational advertising*; *Electronic commerce*; • **Applied computing** → *Online shopping;*

Additional Key Words and Phrases: Clickstream, user behavior, log analysis, user modeling

ACM Reference format:
Che-Yun Hsu, Ting-Rui Chen, and Hung-Hsuan Chen. 2022. Experience: Analyzing Missing Web Page Visits and Unintentional Web Page Visits from the Client-side Web Logs. *J. Data Inform. Quality* 14, 2, Article 11 (March 2022), 17 pages.
https://doi.org/10.1145/3490392

1 INTRODUCTION

Web browsing has become essential in most people's everyday lives. The conventional wisdom is that logs stored by web servers (e.g., Apache or NGINX) or by the history of browsers (e.g.,

We acknowledge partial support by the Ministry of Science and Technology of Taiwan under grant MOST 107-2221-E-008-077-MY3 and MOST 110-2222-E-008-005-MY3.
Authors' address: C.-Y. Hsu, T.-R. Chen, and H.-H. Chen (corresponding author), Computer Science and Information Engineering, National Central University, No. 300, Zhongda Rd, Zhongli District, Taoyuan City, 320, Taiwan; emails: {eleceel, ray941216, hhchen}@g.ncu.edu.tw.

1936-1955/2022/03-ART11 $15.00
https://doi.org/10.1145/3490392

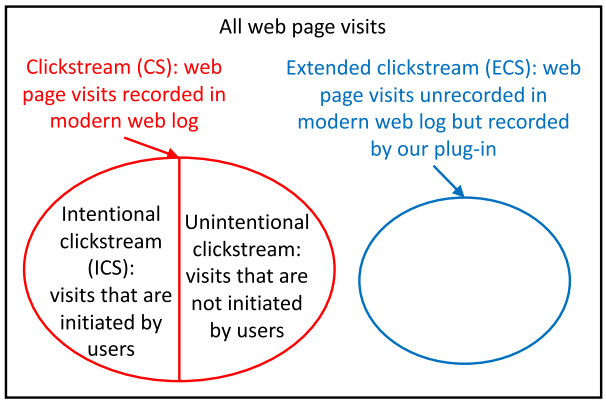

Fig. 1. The relationship between all the online visits of a user, **clickstream (CS)** events, **intentional clickstream (ICS)** events, and **extended clickstream (ECS)** events.

Chrome or Safari) truthfully record a user's web browsing history. We collectively call these logs and records web logs in this paper. Web logs are widely used to analyze users' online behaviors [4, 12, 23, 33]. However, this paper shows that web logs record only approximately half of a user's visited URLs. Additionally, approximately 5.1% of the visited URLs stored in modern web logs are not generated under a user's intentions. As a result, the web log could be a biased collection of a user's web browsing history, so analyses based on web logs may derive biased observations and conclusions.

Figure 1 shows the relationship between all the online visits of a user, the visits recorded in most of today's web log files (particularly, web server logs and browser histories), and some of the visits that are not recorded in the web log but recorded and studied in this paper. We will use the term ***clickstream***, or **CS** for short, to refer to the events recorded in the web log mentioned above. Many researchers use CS events to represent most, if not all, of a user's online visits and use clickstream and web log interchangeably, e.g., [34, 41]. We show that many web visiting activities are not recorded by the web log. We collected some of these missing events, which we call ***extended clickstream***, or **ECS** for short. There could be other page visits that are neither recorded by the CS nor the ECS. However, the goal of the paper is not to collect all browsing activities but to point out that web logs may only represent a small part of a user's online visits. Additionally, among the events in a clickstream, only part of them are initiated by users. These intentional visits in the CS events are collectively called the ***intentional clickstream* (ICS)**. We may use the two terms "web log" and "clickstream" interchangeably in this paper, as in many previous works [34].

We illustrated an example in Figure 2. A user opens a browser and visits website A and then opens another browser window to visit website B. Next, this user opens a link to website C in a new tab, and website C triggers an event that redirects the user to website D. The visits on websites A, B, C, and D are all recorded in the modern web log. However, the user may not be aware of visiting web page D because this visit is triggered by server-side or client-side programs. Assuming that this user decided to switch tabs back to website B and switch windows back to website A, these new visits on websites B and A are not recorded in the web log, even though the user may intend to visit these websites. In this example, since the first four visits on the websites A, B, C, and D are recorded in the modern web log, they are part of the CS events. However, since the visit on D is generated by a redirect event triggered by the server or client code, the user may be unaware of this visit, so only the visits on A, B, and C are part of the ICS events; D is not part of the ICS events. Finally, the "tab switching" behavior and the "go back to the previous page" behavior are not recorded by

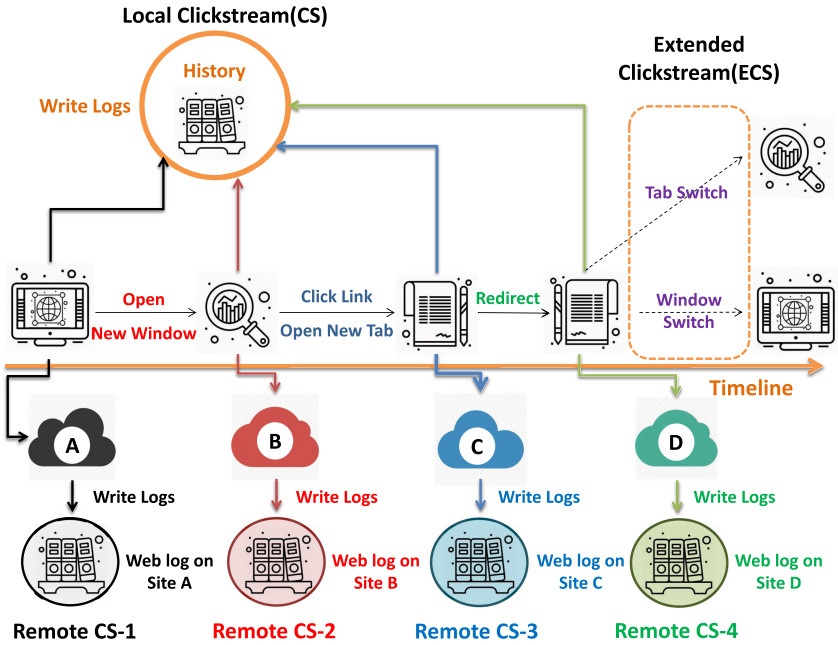

Fig. 2. An example of a user's online visits that include the intentional clickstream events, unintentional clickstream events, and extended clickstream events.

the modern web log, although the user may intend to visit these websites. As we will explain later, we collected and analyzed these events and call them the ECS events. Another example that may cause information missing is the "go back" function implemented in many current browsers [18]. Specifically, web browsers may cache few previously visited websites in the local cache. When a user clicks the "go back" button, the browser renders the page content directly from the local cache instead of sending requests and receiving content from the remote server. Consequently, the new request may only appear in the local history log but not in the remote server log at the **Internet Service Provider (ISP)** or the backend of the website. As we collected users' visits from the client-side (which will be explained later), the visits due to "go back" are recorded in the CS. However, if the logs are collected from the server-side (e.g., from an ISP), these visits would be in the ECS but not in the CS.

To collect users' online web page visits more thoroughly, we developed a plugin for Google's Chrome browser and recruited users to install the plugin by giving e-coupons to these users. We informed all users that we will collect and analyze their online browsing logs. Eventually, we acquired 300+ users and selected 147 users who used this plugin for at least three days as the experimental targets. Although we collected the data from the client-side (i.e., from a Chrome browser plugin), the server-side log also suffers similar issues. Particularly, the server-side log may also contain records not initiated from users (e.g., Remote CS-4, the server-side log on Site *D* in Figure 2) and may also fail to record certain user visits (e.g., tab-switching or window-switching events).

We reported the statistics of the CS events, ICS events, and ECS events in this paper to demonstrate that modern web logs miss a large portion of users' website visits and include few records that are not initiated from users. Additionally, we applied supervised learning algorithms to predict a user's future behaviors based on three different sets of features derived from the CS, ICS, or

ICS plus ECS. The experimental results show that using the ICS or ICS plus ECS better predicts a user's future behaviors, suggesting that ECS may contain extra information not included in the traditional CS. Additionally, the unintentional behaviors in the CS may provide more noise than information, so using either the ICS plus ECS or only the ICS, instead of the complete CS, can be a better choice.

The main contribution of this paper is not the proposal of another new algorithm or another new way to collect users' visited web pages. Instead, we highlight that the clickstream is a partial and biased collection of a user's web page browsing behaviors. We conducted various experiments to support this claim. Since a CS is a biased collection of a user's web page visiting record, we need to be more careful about the observations and conclusions from the studies that assume the CS completely records a user's web page visits. As far as we know, we are the first group to study and report this issue systematically.

The rest of the paper is organized as follows. Section 2 reviews the related work on the analysis of clickstreams. Section 3 explains the data collection strategy and the statistics of the collected datasets. Section 4 shows experiments on predicting a user's future behaviors based on the CS, ICS, and ICS plus ECS. Finally, we discuss our discoveries and future work in Section 5.

2 RELATED WORK

Web logs have been widely used to represent a user's complete online journey [1, 2, 36, 57]. In this section, we review various applications and studies based on log analyses.

2.1 Social Networks and User Modeling based on Log Analysis

One essential study of sociology is to observe or model people's behaviors and interactions among people. However, interpersonal interactions are difficult to observe or quantify directly. Therefore, traditional experiments usually involve questionnaires and lab-controlled simulation experiments, which have obvious disadvantages, such as the fidelity of the simulations, the size of the samples, the measurement of the variables of interest, etc.

As the online social network platforms became popular in the last few decades, people volunteer all sorts of personal information on these platforms, including places they visited, the music they listened to, anniversaries, friend lists, birthdays, job information, relationship status, and many more. For the first time, human activities can be recorded meticulously on a large scale. Unsurprisingly, sociologists, among other researchers, have started to leverage the logs of these platforms to measure users' interactions and behaviors [6, 48]. Studies on this line include person-to-person communication, community and group analysis, friendship formation, information transmission [35, 50, 51], etc. It is shown that simple machine learning approaches can accurately predict an individual's sensitive demographic profiles (e.g., sexual orientation and political views), psychological test scores, and subsequent behaviors [1, 32, 36, 59], given an individual's browsing history or Facebook likes.

Indeed, analyzing users' behaviors on social networks has fundamentally changed the way researchers, especially sociologists, design and conduct experiments. However, if the logs are biased collections, the conclusions derived from these analyses are questionable.

2.2 E-commerce and Recommender Systems Studies based on Log Analysis

E-commerce provides a convenient channel for shopping and may increase the outreach of a business. In addition, since the transactions are digitized, e-commerce retailers heavily use users' visiting logs to conduct various analyses to improve users' satisfaction and stimulate more purchases.

One way to increase direct purchases and improve users' satisfaction is by designing better recommender algorithms to increase the appearance rate of a user's desired items. This line of

research includes content-based approaches [40], collaborative filtering [9, 37, 49, 56], and hybrid methods (using both content-based and collaborative filtering) [5, 16, 17, 42, 45]. Collaborative filtering requires leveraging many users' collective behaviors to determine the relationship between items and users. Since collaborative filtering requires no user labeling or item labeling, it has influenced many works in the last two decades [7, 15, 24, 46]. Many studies have recently used the click-through rate or the conversion rate as the metrics to measure the effectiveness of a recommendation algorithm [8, 10, 36, 47, 60].

While e-commerce retailers highly leverage the web log to discover a user's needs, this paper shows that a web log may only record part of a user's browsing behaviors, and some of the recorded clicks are not initiated by users. Perhaps a quick fix is to consider only user-initiated events, which will be explained further in Section 3 and demonstrated in Section 4.

2.3 Information Retrieval and Search Engines Studies based on Log Analysis

Given a user's query term, search engines rank millions or even billions of objects (e.g., documents, authors, and products) to place the objects that are likely to meet a user's needs at the top of the ranking [11, 13, 21, 22, 25, 28, 30, 38, 43, 53–55]. The quality of a search engine is usually decided by user evaluation. However, since hiring people to test different queries and label the results is costly and labor-intensive, most search engine companies highly leverage online users' responses as the ground truth. For example, if users click the second document given a query, the second document is probably better than the first document for serving the users' needs [39]. As search engines widely use online users' logs to evaluate their ranking algorithms, understanding the biases the web log may have is essential.

Search logs are also used to discover what contents should be added to an enterprise website for a better navigation experience [29, 58]. Particularly, in [29], the authors defined "missing content" as the information that is difficult to obtain via simple page navigation but probably can be retrieved via search. The authors suggested placing such information on a web page that can be easily accessed. Although both [29] and our work studied logs, the "missing information" in [29] is different from our "missing web page visits".

A few previous studies mentioned that the logs might sometimes be distorted [18, 27]. However, they simply described this issue as a limitation of their research but did not conduct systematic studies on this problem. We also found papers studying user behaviors on tab switching [26, 27, 52]. However, these papers primarily investigate parallel browsing behaviors but not the issue of the missing logs.

Most of the research and applications introduced in this section assume that the web logs fully represent users' online web browsing behaviors. However, as we will show later, modern web logs are a biased collection of users' online visits. We believe the researchers in the field of data mining and data analysis and information retrieval should be aware of this issue and be careful with experimental results obtained by analyzing web logs.

3 COLLECTING MISSING WEB PAGE VISITS AND INTENTIONALITY OF WEB PAGE VISITS

This section introduces the data collection method and the statistics of the collected data.

3.1 Data Collection and Preprocessing

We created a plugin for the Google Chrome browser to collect the CS and ECS events from users. We recruited users from the Internet and motivated users to join the experiment by providing e-coupons. We informed all users that the installed plugin collects their browsing behaviors.

Table 1. A List of CS Events, ECS Events, and Intentionality

Type	Event name	Brief description	Intentional?	%
CS	Link	Clicking a link to arrive at the current page	✓	47.3726
	Form_submit	Visiting a page because of a form submission		2.8147
	Auto_bookmark	Clicking through the UI of a browser (e.g., a menu bar) to arrive at the current page	✓	2.5437
	Generated	Clicking a suggested non-URL entry while typing in the address bar	✓	1.6619
	Reload	Reloading a page	✓	1.3892
	Typed	Typing a URL in the address bar to arrive at the current page	✓	0.8231
	Auto_toplevel	Visiting a page because it is a starting page		0.4135
	Manual_subframe	Explicitly requesting a subframe navigation	✓	0.0325
	Keyword	Visiting a page because of the keyword search configuration in the browser	✓	0.0022
	Auto_subframe	Automatically loaded in a non-top-level frame		0.0004
	Keyword_generated	Visiting a page generated by the keyword search functionality	✓	0.0000
ECS	Tab	Visiting a page because of a tab switching	✓	25.5639
	Windows	Visiting a page because of a browser window switching	✓	6.8717
	Blur	Switching to other application windows or closing the browser	✓	5.6950
	Idle	No I/O to the browser for more than 2 minutes	✓	2.4741
	Active	The first I/O after idling	✓	2.3415

Details of CS events are explained in https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/API/webNavigation/TransitionType. The last column displays the occurrence percentage for each event in our collected web log.

As shown in Table 1, this plugin collects a user's various interactions with the browser. The CS events include all the visits recorded in the standard web logs, and the ECS events contain the user and browser interactions that are not recorded in the standard web logs. Based on how a user arrived at a page, we define each record as an intentional or an unintentional visit. The last column of Table 1 displays the occurrence percentage for each event in our collected web logs.

Table 2 shows examples of intentional/unintentional clickstreams and intentional "extended clickstreams". For example, the behavior of "clicking a hyperlink" requires a user to move the cursor to a specific location on a page and click the mouse, so this kind of behavior is regarded as an

Table 2. Examples of the Intentional CS, Unintentional CS, and Intentional ECS Events

	intentional behavior	unintentional behavior
CS	clicks on hyperlinks; URL typing on the navigation bar; clicks on bookmarks	pages loaded in subframes; pop-up windows; page auto-redirect
ECS	tab switching; browser switching	-

Table 3. A Statistical Summary of the Number of Days that the Users had the Plugin Installed

	minimal	first quantile	median	mean	third quantile	maximal
Days	3	70.5	110	41.25	131.5	142

intentional behavior. On the other hand, the pop-up windows are usually triggered by client-side scripts. Therefore, even though the URLs of these pop-up windows appear in the web log, these visits are regarded as "unintentional" behaviors, as they are not triggered by users. We believe that the unintentional events do not exist in the extended clickstreams because all the events in the extended clickstreams are triggered by users.

After obtaining these behaviors, we removed the URLs representing the local hosts. (e.g., 127.0.0.1 and 192.168.0.1) because we thought that this type of URL could be noise. Additionally, we used an online website categorizing service[1] to convert each URL into a category. For example, the URLs google.com, facebook.com, and youtube.com are classified as categories "Search Engines and Portals", "Social Networking", and "Streaming Media and Download", respectively. Eventually, we grouped all the collected URLs into 82 categories.

3.2 Data Statistics

We recruited users from the Internet to install our plugin by providing e-coupons to the users. We informed the users that we would record all their browsing activities. Eventually, we recruited 300+ users; among them, we selected 147 users who used the plugin for at least three days as our experimental targets. The dataset contains users' website visits from Feb. 26, 2019 to Jul. 17, 2019, with 6,623,178 event counts in total.

Table 3 shows a summary of the number of days a user used this plugin. As can be seen, the first quantile is 70.5 days, i.e., 75% of users used the plugin for more than two months, suggesting the numbers reported in this paper are based on users' long-term behaviors.

4 ANALYZING CS, ICS, AND ECS

This section presents the experiments demonstrating the problems with the traditional CS and why the ICS and ICS plus ECS might be better alternatives.

4.1 Statistics of CS, ICS, and ECS Events

When we started this study, we were skeptical of the effectiveness of the ECS events because we thought that the number of ECS events compared to the number of CS events might be very

[1] https://fortiguard.com/webfilter.

Table 4. A Statistical Summary of the Number of Events Per User for the CS, ICS, and ICS + ECS Events

	minimal	first quantile	median	mean	third quantile	maximal
CS	21	11,105	20,381	20,213	35,941	110,580
ICS	21	10,504	19,821	19,181	33,591	103,635
ICS + ECS	73	18,862	36,421	34,937	61,138	178,473

Table 5. A Statistical Summary of the Number of Events Per User Per Day for the CS, ICS, and ICS + ECS Events

	minimal	first quantile	median	mean	third quantile	maximal
CS	7	137	231	157	328	932
ICS	7	128	221	150	311	893
ICS + ECS	24	238	387	270	559	1,758

small and thus unimportant and probably ignorable. Surprisingly, we found that the number of ECS events is very close to the number of CS events, implying that the traditional web logs fail to capture nearly half of a user's online behaviors. Out of the 6,623,178 events collected during the 142 days, the numbers of CS and ECS events are 3,778,777 and 2,844,401, respectively. These numbers suggest that ignoring ECS events seems inappropriate, as the ECS events account for 43% of the total events. The last column of Table 1 displays the occurrence percentage for each event. As shown, tab switching activities, although not recorded in the web log, account for a quarter of the total events.

Tables 4 and 5 show the statistics of the numbers of different event types per user and per user per day, respectively. Since different users may use the plugin for different numbers of days, each number in Table 4 is not simply a multiplication of the corresponding number in Table 5.

The first rows in both tables display the statistical summaries of the clickstream. Current browsers typically only record this type of event (e.g., the "history" recorded by Google Chrome). The second rows show the summary of the events in intentional clickstream (ICS), i.e., the events triggered by users, not by the browsers or the servers. By comparing the first two rows of Table 4, we can see that the unintentional events contribute 5.1% of the total CS events for an average user, suggesting that a user may be unconscious of 5.1% of the recorded events in the traditional web log. We believe including these unconscious events in analyses may introduce more noise than information. The third row shows the summary of ECS events plus the ICS events. As shown, the number of ECS events is close to the number of ICS events.

Based on these numbers, we hypothesize that the standard web log is a biased collection of a user's online web visits. Consequently, analyses based on the web logs containing only the clickstream may reveal only part of a user's online journey. Below, we show two different experiments to support this claim. The first experiment compares the popular website categories ranked by CS, ICS, and ICS + ECS. In the second experiment, we predict a user's future behaviors (the category of a user's following clicked website and time gap before a user's next click) and compare their effectiveness based on different types of logs (CS, ICS, and ICS + ECS).

4.2 Ranking Popular Website Categories

This section shows the ranking of popular website categories. Particularly, we rank the categories based on the events in CS, ICS, and ICS + ECS.

Table 6. The 20 Most Popular Website Categories Ranked by ICS + ECS, CS, and ICS (PDF: Probability Density Function; R1, R2, and R3 are the Rankings of Categories Based on the Numbers of ICS + ECS, CS, and ICS Events, Respectively

Category	ICS + ECS			CS			ICS			
	R1	Count	PDF1	R2	Count	PDF2	R3	Count	PDF3	PDF1/PDF2
Streaming Media and Download	1	1,124,145	17.54	3	558,297	14.77	2	555,767	15.49	1.19
Social Networking	2	938,817	14.65	1	608,224	16.10	1	600,171	16.73	0.91
Search Engines and Portals	3	714,769	11.15	2	559,221	14.79	3	461,377	12.86	0.75
Education	4	570,649	8.9	5	304,364	8.05	5	276,313	7.70	1.11
Information Technology	5	457,948	7.15	6	200,180	5.30	6	189,298	5.28	1.35
Web-based Application	6	391,608	6.11	4	336,886	8.91	4	333,320	9.29	0.69
Games	7	386,895	6.04	7	156,347	4.14	7	152,642	4.26	1.46
Business	8	203,138	3.17	9	108,060	2.86	10	99,250	2.77	1.11
Shopping	9	168,830	2.63	11	94,737	2.51	11	88,601	2.47	1.05
File Sharing and Storage	10	165,818	2.59	10	106,535	2.82	9	105,062	2.93	0.92
Entertainment	11	155,154	2.42	8	117,181	3.10	8	115,618	3.22	0.78
Reference	12	154,226	2.41	12	86,090	2.28	12	80,408	2.24	1.06
Web-based Email	13	115,595	1.80	13	68,741	1.82	13	68,178	1.90	0.99
News and Media	14	100,603	1.57	14	67,278	1.78	14	66,567	1.86	0.88
Newsgroups and Message Boards	15	72,541	1.13	16	35,036	0.93	16	33,127	0.92	1.22
Pornography	16	69,762	1.09	15	42,031	1.11	15	40,939	1.14	0.98
Personal Websites and Blogs	17	68,312	1.08	20	25,497	0.67	20	25,220	0.70	1.61
Instant Messaging	18	63,289	0.99	18	29,973	0.79	18	28,931	0.81	1.25
Auction	19	55,927	0.87	17	33,344	0.88	17	33,078	0.92	0.99
Travel	20	49,540	0.77	19	29,955	0.79	19	25,631	0.71	0.97

Table 6 displays the top 20 most visited categories. As shown, although the top 20 categories of the events in CS, ICS, and ICS + ECS are the same, their rankings are different. The last column in Table 6 shows the ratio between PDF1 (the PDF of ICS + ECS events of a category) and PDF2 (the PDF of CS events of a category). A value of PDF1/PDF2 larger than 1 indicates the traditional web log (i.e., the CS events here) tends to underestimate the popularity of a category. It appears that if we estimate the popularity of a category based on the traditional web log, we may highly overestimate the popularity of the categories "Search Engines and Portals", "Web-based Application", and "Entertainment". Likewise, we likely highly underestimate the popularity of the categories "Information Technology", "Games", "Newsgroups and Message Boards", "Personal Websites and Blogs", and "Instant Messaging". If a web page belongs to the underestimated group, it probably indicates that users tend to put the web page in the background tab and browse the tab from time to time via tab switching. Unfortunately, these later visits are unrecorded by the modern web logs.

We further compare the ordinal association between the ranking of ICS + ECS (i.e., Rank1 in Table 6) and the ranking of CS (i.e., Rank2 in Table 6) based on Kendall's τ correlation coefficient [31]. We used Kendall's τ coefficient instead of the probably more popular Spearman's correlation because Kendall's τ is argued to be more reliable as a ranking-based measure [44].

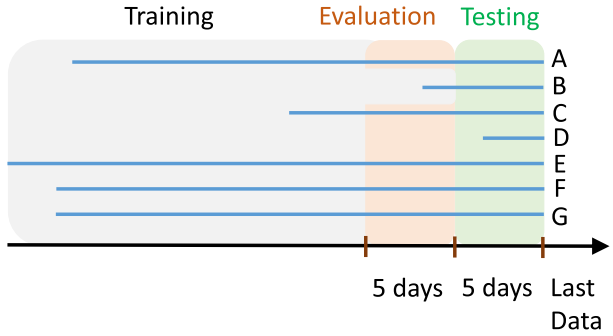

Fig. 3. Splitting the training, evaluation, and testing data. If a user's log is equal to or less than 5 days, all the logs are used for testing (e.g., user D in the figure). If a user's log is between 5 and 10 days, the last 5 days are used for testing, and the remaining are for training (e.g., user B in the figure).

Equation (1) defines the formula of Kendall's τ correlation coefficient.

$$\tau(\ell_1, \ell_2) = \frac{P - Q}{\binom{n}{2}}, \tag{1}$$

where ℓ_1 and ℓ_2 are two orders to compare, n is the number of elements for each list, and P and Q are the number of concordant pairs and the number of discordant pairs, respectively. A pair of observations is concordant if the two observations have the same order in both ℓ_1 and ℓ_2; a pair of observations is discordant if the two observations are ranked in opposite directions in ℓ_1 and ℓ_2. Since every element in ℓ_1 is different, and every element in ℓ_2 is also different, there are no tied pairs; we do not need to consider different variations of Kendall's τ that apply different strategies to account for ties [3].

The Kendall's τ correlation coefficient of Rank1 (the popular ranking using ICS + ECS) and Rank2 (the popular ranking using only CS) is 0.86 with the p-value 2.19×10^{-10} computed by the Mann-Kendall test [20]. These values indicate that among the ranking of the popular categories based on ICS + ECS and the ranking of the popular categories based on the standard CS, only a small portion is discordant, and this result is significant.

4.3 Future Behavior Prediction

This section shows the experimental results for predicting a user's future behavior based on previous events in the CS, ICS, and ICS + ECS. If using the ICS events or ICS + ECS events can yield better predictions than the models that make predictions based on the CS events, it may imply that ECS events contain extra information, and the events appearing in the CS but not in the ICS provide more noise than signal.

To conduct the experiments, we separated the collected data as follows. Referring to Figure 3, for each user, we used the behaviors in the last 5 days as the testing data. If a user's remaining behaviors spanned longer than 5 days, we further divided the remaining behaviors into evaluation data (the last 5 days after excluding the testing data) and the training data (the remaining days) for hyperparameter selection. Based on these rules, the logs of users $A, C, E, F,$ and G in Figure 3 are divided into training, validation, and test datasets; the logs of user B are divided into training and test datasets; the logs of user D is used only for the test dataset. More than 90% of the users used the plugin for more than 30 days in our collected dataset.

Based on the above-mentioned data splitting, we conducted the following two experiments to validate the effectiveness of the ICS events and the ECS events compared to the traditional CS

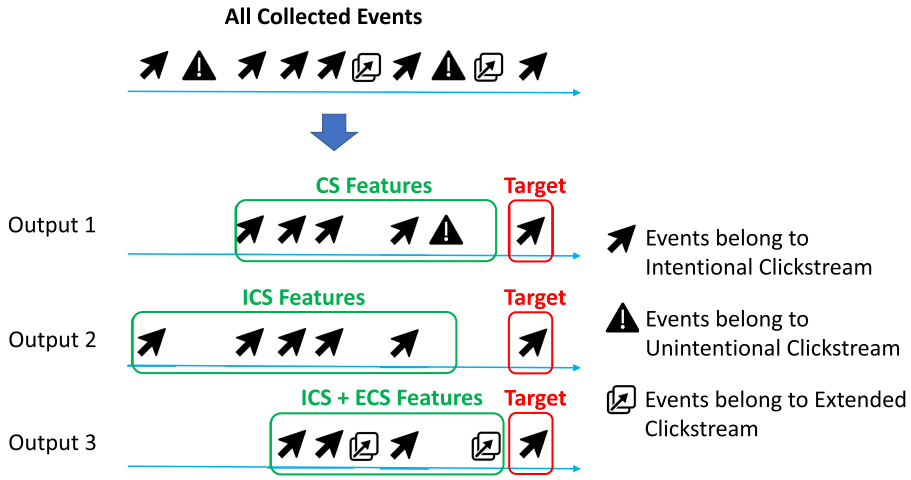

Fig. 4. An example of using five consecutive events to predict the next event, where the five consecutive events may come from clickstream (Output 1), intentional clickstream (Output 2), or extended plus intentional clickstream (Output 3).

events: (1) predicting the category of the next clicked website for a given user; (2) predicting the time gap before the next click of a given user.

4.3.1 Predicting the Category of the Next Clicked Website. The first experiment utilizes the CS, ICS, or ICS + ECS events to predict the category of the next clicked website for a given user.

Figure 4 illustrates these three cases. The top timeline shows all the latest collected events of a user. Output 1 represents the case of using the CS events (particularly the five consecutive events right before the target) to predict the target. The CS events may contain both intentional and unintentional events. Output 2 displays the case of using only the five consecutive intentional events in the clickstream to generate the features. Finally, output 3 demonstrates utilizing the five consecutive events belonging to the intentional clickstreams or the extended clickstreams to create the features.

For each CS (including the ICS) event and ECS event, the generated features include the types (referring to the second column of Table 1 for a list of event types), the URL categories of these events, and their corresponding timestamps.

As there are 82 website categories, this task can be modeled as a multi-class classification problem, which is commonly evaluated via the macro average, micro average, or weighted average of the precision, recall, and F1 scores. We reported the weighted-precision/recall/F1 scores for the following reasons.

First, since the distribution of the URL categories is highly skewed in our case (as shown in Table 6), the macro-precision/recall/F1 scores are dominated by the classes that have fewer instances. For example, the "Travel" category accounts for less than 1% of all ICS + ECS, CS, and ICS events, and the "Social Networking" category constitutes approximately 15% of the events. Thus, if using macro average, correctly predicting one instance in the Travel category would be much more valuable than correctly predicting one instance in the Social Networking category.

Second, micro-precision/recall/F1 scores are less informative because the micro-precision score always equals the micro-recall score, and consequently, the micro-F1 score is also identical because the micro-F1 score is the harmonic mean of the micro-precision and micro-recall. The fundamental reason that micro-precision equals micro-recall can be directly derived from the definition. First,

the numerators of both micro-precision and micro-recall are $\sum_{i=1}^{K} TP_i$ (TP_i is the number of true positives by regarding class i as the positive class and the others as the negative class). Second, the denominators of micro-precision and mico-recall are $\sum_{i=1}^{K}(TP_i + FP_i)$ and $\sum_{i=1}^{K}(TP_i + FN_i)$, respectively ($FP_i$ is the number of false positives when considering class i as positive and the others as the negative class, and FN_i is the number of false negatives when regarding class i as positive and the other classes as negative). Although they look different, $\sum_{i=1}^{K} FP_i$ always equals $\sum_{i=1}^{K} FN_i$, because each incorrect prediction is a false positive for one class and a false negative for another. For example, if c_i is misclassified as c_j, this leads to a false positive for class c_j and a false negative for c_i.

The weighted-precision, recall, and F1 scores are shown in Equations (2), (3), and (4), respectively.

$$P_{\text{weighted}} = \sum_{i=1}^{K} \left(w_i \times \frac{TP_i}{TP_i + FP_i} \right) \times 100\%, \tag{2}$$

$$R_{\text{weighted}} = \sum_{i=1}^{K} \left(w_i \times \frac{TP_i}{TP_i + FN_i} \right) \times 100\%, \tag{3}$$

$$F1_{\text{weighted}} = \sum_{i=1}^{K} (w_i \times F1_i) \times 100\%, \tag{4}$$

where w_i is the ratio of the instances belonging to class i, K is the number of classes, and $F1_i$ is the $F1$ score when regarding class i as the positive class and the others as the negative class.

We applied five popular supervised classifiers, logistic regression, k-nearest neighbors (kNN), random forest, support vector machine (SVM), and XGBoost [14], on the CS, ICS, and ICS + ECS datasets. As we mentioned in Section 3, there are 82 categories in our collected dataset so that a random guess would yield a small accuracy and a small F1 score. Even if we consistently predict the majority category (i.e., "Streaming Media and Download"), the accuracy would be only 17.54%, as the majority class accounts for 17.54% of all events (referring to Table 6).

Table 7 shows the weighted precision/recall/F1 scores of different models using CS, ICS, and ICS + ECS to generate the features. Each number reported in the table is an average of 10 trials. Each row compares one evaluation metric when applying one training model on CS, ICS, and ICS + ECS. We highlight the largest and the second-largest numbers in bold and underscore, respectively.

We observed three things from the table, as described below. First, using ICS or ICS + ECS as the training data usually yielded better Weighted precision/recall/F1 scores than using the traditional CS as the training data, likely because the ICS or ICS + ECS better represent a user's online trajectory. Although the difference may seem small (especially given that CS events account for only 57% of a user's visited websites), we use only the five nearest events to generate the features (as shown in Figure 4). In other words, no matter we use CS, ICS, or ICS + ECS as the experimental data, we used the same number of previous events to generate the features for each prediction. The result suggests that the same number of ICS events (or ICS + ECS events) may contain more information than the same number of CS events. Additionally, according to the full confusion matrices of the results based on ICS events and ICS + ECS events,[2] both ICS and ICS + ECS outperformed CS in almost all categories. Our second observation is that the tree-based models (XGBoost and Random Forest) and support vector machine performed better than the others (kNN and logistic

[2]The full confusion matrices based on ICS and ICS + ECS events are shown in https://github.com/eleceel/DART_analyze_user_behavior/blob/master/xgb_result_one(ics).pdf and https://github.com/eleceel/DART_analyze_user_behavior/blob/master/xgb_result_one(icsecs).pdf, respectively. We show the full confusion matrices online to save the space.

Table 7. Weighted Precision/Recall/F1 Scores of different Models to Predict the Category of the Next Clicked Web Page Trained on the Features Derived from CS, ICS, or ICS + ECS

Training Model	Evaluation Metric	CS	ICS	ICS + ECS
Logistic Regression	Weighted Precision	28.55	**28.96**	28.85
	Weighted Recall	40.25	40.57	**40.73**
	Weighted F1 score	30.16	**30.57**	30.42
KNN	Weighted Precision	72.71	73.09	**73.81**
	Weighted Recall	72.71	72.88	**73.85**
	Weighted F1 score	72.23	72.48	**73.47**
Random Forest	Weighted Precision	**76.74**	76.53	76.60
	Weighted Recall	76.51	76.65	**76.69**
	Weighted F1 score	76.29	76.38	**76.42**
SVM	Weighted Precision	74.39	74.56	**74.65**
	Weighted Recall	74.52	74.54	**74.71**
	Weighted F1 score	74.50	74.55	**74.66**
XGBoost	Weighted Precision	77.10	77.27	**77.41**
	Weighted Recall	77.18	77.43	**77.60**
	Weighted F1 score	77.03	77.29	**77.44**

The highest score of each row is highlighted in bold; the second-highest score of each row is highlighted using underscore. ICS + ECS usually gives the highest score, followed by ECS.

regression), probably because user behaviors are complex by nature. Finally, our third observation can be concluded from the full confusion matrices. The number of observations of a class i is positively correlated with the corresponding accuracy and the true positive rate when regarding the class i as the positive class and the others as the negative class.

4.3.2 Predicting the Time Gap Before a User's Next Click. The second experiment predicts the time gap before a user's next click. We divided the period into five classes: 0 to 5 seconds, 5 to 20 seconds, 20 seconds to 2 minutes, 2 minutes to 20 minutes, and longer than 20 minutes. This division is motivated by the Weber-Fechner law [19], which states that humans' subjective sensation is primarily proportional to the logarithm of the stimulus intensity. As we divide the time gap length into groups, this task can also be regarded as a multi-class classification problem. We can again apply various supervised classifiers and report the results using the weighted precision/recall/F1 scores.

Table 8 shows the result. As before, we applied logistic regression, k-nearest neighbors, random forest, support vector machine, and XGBoost, on the CS, ICS, and ICS + ECS datasets. We also highlight the first and the second largest numbers in bold and underscore for each row, respectively. The results are consistent with our previous experiment: using all the events in ICS + ECS as the training data can yield better predictions than using the ICS events or traditional CS events as the training data, and the tree-based models (XGBoost and Random Forest) and support vector machine again yielded better weighted-precision/recall/F1 scores than the simple models, kNN and logistic regression.

5 DISCUSSION

While the conventional wisdom is that the web logs accurately record a user's web page browsing history, we showed that web logs record only approximately half of a user's web page visits. Moreover, even among the recorded instances, several of them are not intentional visits. Since

Table 8. Weighted Precision/Recall/F1 Scores of Different Models to Predict the Period to the Next Click Trained on the CS, ICS, or ICS + ECS

Training Model	Evaluation Metric	CS	ICS	ICS+ECS
Logistic Regression	Weighted Precision	51.16	51.27	**51.96**
	Weighted Recall	51.18	51.22	**52.00**
	Weighted F1	53.56	53.76	**54.75**
KNN	Weighted Precision	48.74	48.87	**49.95**
	Weighted Recall	48.71	48.83	**49.90**
	Weighted F1	48.71	48.87	**49.92**
Random Forest	Weighted Precision	53.36	53.46	**54.46**
	Weighted Recall	53.31	53.44	**54.44**
	Weighted F1	53.58	53.74	**54.76**
SVM	Weighted Precision	52.26	52.46	**53.53**
	Weighted Recall	52.29	52.51	**53.54**
	Weighted F1 score	52.28	52.48	**53.54**
XGBoost	Weighted Precision	53.53	53.77	**54.74**
	Weighted Recall	53.51	53.77	**54.71**
	Weighted F1	53.50	53.70	**54.77**

The highest score of each row is highlighted in bold; the second-highest score of each row is highlighted using underscore. ICS + ECS usually gives the highest score, followed by ECS.

the web log is a biased collection of a user's online visits, we may need to be more careful about the analyses that use web logs as the ground truth to represent a user's online trajectory. As far as we know, we are the first group to conduct experiments on this issue systematically. These experiments include a simple comparison of popular website categories ranked by CS, ICS, and ICS + ECS, and more complex tasks to predict the category of a user's next clicked website and the period before the next click. All the results show that the ECS contains extra information not recorded by the traditional CS events, and the unintentional clickstreams may contain more noise than information.

Some browsing behaviors may not be recorded by the modern CS or our proposed ECS. For example, several modern e-commerce websites list a fixed number of products on a page, but when a user scrolls down to near the bottom of the page, the page automatically lists more items. As a result, even if a user browses many different items, suggesting that this user is probably highly interested in the current browser shopping, the local web log contains only one record. Even though the logs in CS plus ECS may still miss certain user behaviors, the goal of the paper is to point out that modern web logs are far from a complete list of a user's online visits, not to collect all user behaviors. Listing other missing behaviors is indeed interesting, but the task may become trivial as we enumerate more and more missing behaviors.

As we informed all users that their browsing behaviors are logged and analyzed, the users may modify their behaviors. We did not conduct preventative measures for behavior change for two reasons. First, as shown in Table 3, the median observation period of a user is 110 days. It is less likely that a user can modify the behavior for such a long time. Second, even if a user intentionally adjusts the frequency of visiting particular websites, they probably do not change how they use new tabs and new windows, which accounts for most of the ECS events. However, a more careful analysis should ensure the consistency of a user's behavior. A possible approach is to compare a user's behavior in the first month and the following months, assuming that a user might forget about the tracking after a few weeks. If the logging time is long enough, perhaps we could directly abandon the logs in the first month.

Some readers may also be worried about privacy issues when more user behaviors are collected. But, again, we want to emphasize that the purpose of this paper is to demonstrate that the modern web log is a biased collection of users' online visits, so we should be more careful when deriving conclusions from analyzing clickstreams. Privacy issues are undoubtedly important, but studying user behaviors while preserving privacy is beyond the scope of our study.

REFERENCES

[1] Guo-Jhen Bai, Cheng-You Lien, and Hung-Hsuan Chen. 2019. Co-learning multiple browsing tendencies of a user by matrix factorization-based multitask learning. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 253–257.

[2] Ting Bai, Wanye Xin Zhao, Yulan He, Jian-Yun Nie, and Ji-Rong Wen. 2018. Characterizing and predicting early reviewers for effective product marketing on e-commerce websites. *IEEE Transactions on Knowledge and Data Engineering* 30, 12 (2018), 2271–2284.

[3] Kenneth J. Berry, Janis E. Johnston, Sammy Zahran, and Paul W. Mielke. 2009. Stuart's tau measure of effect size for ordinal variables: Some methodological considerations. *Behavior research methods* 41, 4 (2009), 1144–1148.

[4] Rahul Bhagat, Srevatsan Muralidharan, Alex Lobzhanidze, and Shankar Vishwanath. 2018. Buy it again: Modeling repeat purchase recommendations. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[5] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-adapted Interaction* 12, 4 (2002), 331–370.

[6] Damon Centola. 2010. The spread of behavior in an online social network experiment. *Science* 329, 5996 (2010), 1194–1197.

[7] Hung-Hsuan Chen. 2017. Weighted-SVD: Matrix factorization with weights on the latent factors. *arXiv preprint arXiv:1710.00482* (2017).

[8] Hung-Hsuan Chen. 2018. Behavior2Vec: Generating distributed representations of users' behaviors on products for recommender systems. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 4 (2018), 1–20.

[9] Hung-Hsuan Chen and Pu Chen. 2019. Differentiating regularization weights–A simple mechanism to alleviate cold start in recommender systems. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13, 1 (2019), 1–22.

[10] Hung-Hsuan Chen, Chu-An Chung, Hsin-Chien Huang, and Wen Tsui. 2017. Common pitfalls in training and evaluating recommender systems. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 37–45.

[11] Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, and Clyde Lee Giles. 2011. CollabSeer: A search engine for collaboration discovery. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. 231–240.

[12] Hung-Hsuan Chen, Madian Khabsa, and C. Lee Giles. 2014. The feasibility of investing in manual correction of metadata for a large-scale digital library. In *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 225–228.

[13] Hung-Hsuan Chen, Pucktada Treeratpituk, Prasenjit Mitra, and C. Lee Giles. 2013. CSSeer: An expert recommendation system based on CiteseerX. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. 381–382.

[14] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[15] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. 7–10.

[16] Szu-Yu Chou, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. 2018. Fast tensor factorization for large-scale context-aware recommendation from implicit feedback. *IEEE Transactions on Big Data* 6, 1 (2018), 201–208.

[17] Wei Dai, Qing Zhang, Weike Pan, and Zhong Ming. 2020. Transfer to rank for Top-N recommendation. *IEEE Trans. Big Data* 6, 4 (2020), 770–779.

[18] Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, and Jaime Teevan. 2014. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*. Springer, 349–372.

[19] Gustav Theodor Fechner, Davis H. Howes, and Edwin Garrigues Boring. 1966. *Elements of Psychophysics*. Vol. 1. Holt, Rinehart and Winston New York.

[20] Milan Gocic and Slavisa Trajkovic. 2013. Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia. *Global and Planetary Change* 100 (2013), 172–182.

[21] Liang Gou, Hung-Hsuan Chen, Jung-Hyun Kim, Xiaolong Zhang, and C. Lee Giles. 2010. SNDocRank: A social network-based video search ranking framework. In *Proceedings of the International Conference on Multimedia Information Retrieval*. 367–376.

[22] Liang Gou, Xiaolong Zhang, Hung-Hsuan Chen, Jung-Hyun Kim, and C. Lee Giles. 2010. Social network document ranking. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. 313–322.

[23] Mihajlo Grbovic and Haibin Cheng. 2018. Real-time personalization using embeddings for search ranking at Airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[24] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.
[25] Li-Yuan Hsu, Chia-Hao Kao, I-Sheng Jheng, and Hung-Hsuan Chen. 2021. Toward building an academic search engine understanding the purposes of the matched sentences in an abstract. *IEEE Access* 9 (2021), 109344–109354.
[26] Jeff Huang, Thomas Lin, and Ryen W. White. 2012. No search result left behind: Branching behavior with browser tabs. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. 203–212.
[27] Jeff Huang and Ryen W. White. 2010. Parallel browsing behavior on the web. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. 13–18.
[28] Rolf Jagerman, Krisztian Balog, and Maarten De Rijke. 2018. Opensearch: Lessons learned from an online evaluation campaign. *Journal of Data and Information Quality (JDIQ)* 10, 3 (2018), 1–15.
[29] Harsh Jhamtani, Rishiraj Saha Roy, Niyati Chhaya, and Eric Nyberg. 2017. Leveraging site search logs to identify missing content on enterprise webpages. In *European Conference on Information Retrieval*. Springer, 506–512.
[30] Di Jiang, Yongxin Tong, and Yuanfeng Song. 2016. Cross-lingual topic discovery from multilingual search engine query log. *ACM Trans. Inf. Syst.* 35, 2, Article 9 (Sept. 2016), 28 pages.
[31] Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
[32] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.
[33] Anurag Kumar, Vaishali Ahirwar, and Ravi Kumar Singh. 2017. A study on prediction of user behavior based on web server log files in web usage mining. *International Journal of Engineering and Computer Science* (2017).
[34] Juhnyoung Lee, Mark Podlaseck, Edith Schonberg, and Robert Hoch. 2001. Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data Mining and Knowledge Discovery* 5, 1 (2001), 59–84.
[35] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. 2007. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)* 1, 1 (2007).
[36] Cheng-You Lien, Guo-Jhen Bai, and Hung-Hsuan Chen. 2019. Visited websites may reveal users' demographic information and personality. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 248–252.
[37] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1 (2003), 76–80.
[38] Chao Liu, Fan Guo, and Christos Faloutsos. 2010. Bayesian browsing model: Exact inference of document relevance from petabyte-scale data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4, 4 (2010), 1–26.
[39] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer Science & Business Media.
[40] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*. Springer, 73–105.
[41] Lin Lu, Margaret Dunham, and Yu Meng. 2005. Mining significant usage patterns from clickstream data. In *International Workshop on Knowledge Discovery on the Web*. Springer, 1–17.
[42] Xin Luo, Mengchu Zhou, Shuai Li, Di Wu, Zhigang Liu, and Mingsheng Shang. 2021. Algorithms of unconstrained non-negative latent factor analysis for recommender systems. *IEEE Trans. Big Data* 7, 1 (2021), 227–240.
[43] Masaya Murata, Hiroyuki Toda, Yumiko Matsuura, and Ryoji Kataoka. 2009. Access concentration detection in click logs to improve mobile Web-IR. *Information Sciences* 179, 12 (2009), 1859–1869.
[44] Roger Newson. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal* 2, 1 (2002), 45–64.
[45] Lianyong Qi, Xiaolong Xu, Xuyun Zhang, Wanchun Dou, Chunhua Hu, Yuming Zhou, and Jiguo Yu. 2018. Structural balance theory-based e-commerce recommendation over big rating data. *IEEE Trans. Big Data* 4, 3 (2018), 301–312.
[46] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*. IEEE, 995–1000.
[47] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web*. 521–530.
[48] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762 (2006), 854–856.
[49] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The Adaptive Web*. Springer, 291–324.
[50] Michael Szell, Renaud Lambiotte, and Stefan Thurner. 2010. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences* 107, 31 (2010), 13636–13641.
[51] Jie Tang, Tiancheng Lou, Jon Kleinberg, and Sen Wu. 2016. Transfer learning to infer social ties across heterogeneous networks. *ACM Trans. Inf. Syst.* 34, 2, Article 7 (April 2016), 43 pages.
[52] Maximilian Viermetz, Carsten Stolz, Vassil Gedov, and Michal Skubacz. 2006. Relevance and impact of tabbed browsing behavior on web usage mining. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. IEEE, 262–269.

[53] Zhongyuan Wang, Fang Wang, Haixun Wang, Zhirui Hu, Jun Yan, Fangtao Li, Ji-Rong Wen, and Zhoujun Li. 2016. Unsupervised head–modifier detection in search queries. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 11, 2 (2016), 1–28.

[54] Jian Wu, Kyle Mark Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Suppawong Tuarob, Alexander G. Ororbia, Douglas Jordan, Prasenjit Mitra, and C. Lee Giles. 2015. CiteseerX: AI in a digital library search engine. *AI Magazine* 36, 3 (2015), 35–48.

[55] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality (JDIQ)* 10, 4 (2018), 1–20.

[56] Yi-Che Yang, Ping-Ching Lai, and Hung-Hsuan Chen. 2020. Empirically testing deep and shallow ranking models for click-through rate (CTR) prediction. In *2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 147–152.

[57] Jinyoung Yeo, Seung-won Hwang, Sungchul Kim, Eunyee Koh, and Nedim Lipka. 2018. Conversion prediction from clickstream: Modeling market prediction and customer predictability. *IEEE Transactions on Knowledge and Data Engineering* (2018).

[58] Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. 2005. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 512–519.

[59] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.

[60] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.

Received March 2021; revised August 2021; accepted October 2021