Towards an Integrated Clickstream Data Analysis Framework for Understanding Web Users' Information Behavior

Yu Chi[1], Tingting Jiang[2], Daqing He[1], Rui Meng[1]
[1]School of Information Sciences, University of Pittsburgh
[2]School of Information Management, Wuhan University

Abstract
Clickstream data offers an unobtrusive data source for understanding web users' information behavior beyond searching. However, it remains underutilized due to the lack of structured analysis procedures. This paper provides an integrated framework for information scientists to employ in their exploitation of clickstream data, which could contribute to more comprehensive research on users' information behavior. Our proposed framework consists of two major components, i.e., data preparation and data investigation. Data preparation is the process of collecting, cleaning, parsing, and coding data, whereas data investigation includes examining data at three different granularity levels, namely, footprint, movement, and pathway. To clearly present our data analysis process with the analysis framework, we draw examples from an empirical analysis of clickstream data of OPAC users' behavior. Overall, this integrated analysis framework is designed to be independent of any specific research settings so that it can be easily adopted by future researchers for their own clickstream datasets and research questions.
Keywords: clickstream data; analysis framework; user information behavior
Citation: Editor will add citation

Contact: yuc73@pitt.edu; tij@whu.edu.cn; dah44@pitt.edu; rui.meng@pitt.edu

1 Introduction
With the benefits of being unobtrusive in data collection, transaction log analysis has been introduced to the Information Behavior (IB hereafter) domain for a long time. Recently, with the explosively growing online information, the unprecedented large-scale logs provide a great possibility to understand people's information behavior based on the traces generated by those people in the real-world settings (Dumais et al., 2014).

However, the IB domain has also witnessed an increasing challenge in extracting and interpreting useful insights from the log files because of the emergence of various kinds of websites and diverse user types. Though more and more studies are trying to explore users' information behavior via transaction log analysis, most of them are mainly using search log analysis, a technique only for revealing search behavior. While clickstream data, which records more general user behavior, including all the requests made by the users during the website visit (Montgomery 2001), is usually overlooked and underexploited.

We believe it is of great significance to collect and analyze users' clickstream data. This is because, as evidence shows, searching is not the only means for human to acquire information (Benevenuto et al., 2009). When humans' information needs are vague or difficult to express, they may rely on browsing, or even serendipitous information encountering (Bates, 2002).

However, while an increasing use in clickstream data in the IB studies in recent years, few methodology resources or guiding frameworks can be found. And commonly, different studies developed their own study procedures, which are often ill-structured for others to follow. Considering that the large-scale clickstream data usually contains errors, redundancies, and needs to be extracted, cleaned and analyzed to serve different research questions, we argue that there is an urgent need for establishing an integrated clickstream data analysis framework for researchers to follow, and to promote a comprehensive understanding of web users' online interactions to the IB domain.

To achieve this goal, this paper introduces a structured clickstream data analysis framework, which was first proposed by Jiang (2010) and validated in two later empirical studies (Jiang, 2014; Jiang et al., 2014). The differences between this paper and our previous attempts mainly lay in three aspects. Firstly, the focus of this paper is introducing the framework. While the analysis in our prior empirical studies was constrained by the certain data fields in a specific type of website, the proposed framework in this paper is independent of any research settings and includes as many possibilities as possible in terms of data preparation, investigation, and interpretation. It is designed to be applicable to wide-ranging research studies with the purpose of understanding user's online information behavior. Secondly, we made necessary revisions to the original framework. For example, we kept the footprint and movement as

two different granularity levels in the data investigation stage but changed track to pathway as a representation of all the requests during one visit in chronological order. And four attributes were defined to demonstrate a path, namely duration, length, width, and capacity. Thirdly, in this paper, we illustrate the practices of each step with a clickstream data analysis of an academic library's OPAC. Readers can compare the practice and results in the OPAC system to our previous work in a social tagging system (Jiang, 2014) and a web portal for real estate (Jiang et al., 2014).

2 Literature Review

2.1 Data collection and analysis methods in information behavior research

The literature tracing the method trends in the IB research states that questionnaires and interviews are two dominant research methods since the late 1990s (Julien et al., 2011; Vakkari, 2008). The questionnaires and interviews are obtrusive methods which allow the investigators to understand user's information behavior as well as their inner factors that are associated with the behaviors (Case, 2012). Greifenneder (2014) analyzed the publications collected from IB related venues between 2012 to 2014. The results showed that interviews and questionnaires are still the major research methods, accounting for almost half of the studies. While at the same time other research methods, including log analysis, are becoming popular and will probably play an important role in the future.

Despite the prevalence, the obtrusive methods suffer from several inherent flaws. The primary potential problem is that the methods may be reactive; the users are likely to perform differently because they are aware of being assessed (Kazdin, 1979). Besides, the unnatural research settings and the usually small scale datasets may hurt the generalizability and representativeness of the outcomes (Martzoukou, 2005).

Transaction logs, in contrast, collect user data unobtrusively in a real online environment. In addition, it provides probabilities to study user information behavior on a large scale with a relatively low cost. Search log is the most well-known type of transaction log (Jansen, 2008). Clickstream data is another type of transaction logs and provides rich user information. We will review the usage of log analysis in the next section.

2.2 Clickstream data analysis and search log analysis

Because of the rapid increasing of online information and the accompanying rising of the search engines, search log analysis (SLA) becomes an extensively employed method in the information science studies. Such analysis focuses on either a system-side usage examination of the search engine or a client-side behavior understanding of users. The system-side studies can be further divided into the analysis of general-purpose search engines, such as Alta Vista (Silverstein et al., 1999; Jansen et al., 2005) and Excite (Jansen et al., 2000), metasearch engines (Jansen et al., 2007), as well as a searching function in a specific website, such as OPACs (Moulaison, 2011), academic libraries (Han, 2014), and social networks (Teevan et al., 2011). On the other hand, the client-side SLA are even more diverse, ranging from studies on specific search behaviors (Rieh & Xie, 2001; Kato et al., 2013; Teevan, 2006) to certain search user groups (Torres et al., 2010; Tsikrika et al, 2011; Park & Lee, 2013).

Jansen (2006) established an SLA process consisting of data collection, processing and analysis and further broke down the analysis stage into term, query and session level. His work provided methodology foundation and guidance for the following studies. Comparing to search log, clickstream data, however, is mostly found in E-commerce studies. E-commerce websites adopt clickstream data analysis to reveal customers' visiting and purchasing behavior (Moe, 2003; Olbrich & Holsing, 2011), and to measure marketing and merchandising efforts (Aguiar & Martens, 2016; Rutz & Bucklin, 2012; Lee et al., 2010). Information scientists have been ignoring this useful type of data. The next section will review how clickstream data is employed in the existing IB studies.

2.3 Research using clickstream data analysis in information behavior research

We identified two general trends in the research using clickstream data analysis (CDA) in the IB domain. One is to characterize web users' online information behavior in a specific website or platform, and the other is to detect and group web users based on similar behavioral patterns. And these are the two aspects we are going to review the literature.

Firstly, researchers adopted CDA to investigate various types of online user behavior that result in exchanges of information including searching, browsing, serendipitous information acquisition etc., and they conducted the examinations in several types of websites, including OPACs (Villen-Rueda et al.,

2007; Lown, 2008; Asunka et al., 2009) as well as recent web2.0 platforms, e.g., social networks (Benevenuto et al., 2009) and social tagging systems etc. (Jiang, 2014).

Benevenuto et al. (2009) were the first to utilize real clickstream data to analyze user workloads in online social networks. Based on the data collected from a social network aggregator website (an access to Orkut, MySpace, Hi5, and LinkedIn), they reported the frequencies and sequences of user activities. After an in-depth study in Orkut, they found that browsing is the most dominant behavior and accounts for 92% of all users' request. The browsing activity, which called as silent interactions in the paper, can not be studied via publicly available data. Similarly, Jiang (2014) found that browsing by the resource is the most popular user behavior mode and browsing by tag is the most effective one according to a clickstream data analysis in a Chinese-language social tagging system. Besides, users also adopt browsing by user/group, searching, and monitoring during their real interactions. Instead of focusing on a single site or a specific type of sites as the above research, Huang et al. (2007) attempted to characterizing web users' behavior on a macro level; how individual users behave on the Web of as a whole. They proposed a three-dimensional typology of online information behavior, namely Web user's width (i.e., the number of categories of sites explored), length (i.e., the number of sites visited per category), and depth (i.e., the number of pages downloaded per site). An empirical analysis based on the clickstream data obtained from an online panel suggested that the three dimensions of information behavior are highly correlated. To visualize a user's online information behavior, they plotted the values of the three dimensions to a 3D cube.

On the other hand, the large-scale clickstream data also enables the researchers to detect and cluster users that share similar clickstream patterns. The identified groups are critical to personalized recommendation or website design (Ting, et al., 2005). Previously, targeting visitors were usually conducted according to demographic attributes. However, now user groups can be generated from clickstream data and it proves better predictive capabilities (Pai et al., 2014). Additionally, clickstream data analysis helps to detect Sybil accounts, i.e., fake identities in social networks (Wang et al., 2013). Very recently, Wang et al. (2016) proposed an unsupervised clickstream clustering system to capture and visualize user behaviors. In their method, users were represented as nodes and edges were weighted by clickstream similarities, thus they built a clickstream similarity graph. They further tested the effectiveness of the method on two clickstream datasets.

While more and more studies employ clickstream data analysis to understand how users behave on the web in the IB domain, few studies addressed how to conduct the analysis from the original data and commonly, different studies follow their own study procedures. Sen et al. (2006) attempted to propose a structure analysis framework, composing of three levels, i.e., footprint, track, and trail. A footprint is a single clickstream request of a user. A sequence of footprints of a user is represented as a track. And similar tracks are further clustered into a trail. Based on this framework, Jiang (2010) established a tentative clickstream data analysis framework that is more applicable to information behavior research. The three analysis levels were revised as footprint, movement, and track. Later empirical studies gradually modified the framework and tested its effectiveness in different web environments and for solving different research questions. Jiang (2014) investigated users' adoption of information seeking modes in a social tagging system and Jiang et al. (2014) examined the navigation system of a web portal. In this paper, we systematize the analysis process into two standard stages: data preparation stage and data investigation stage. We also remove the features of certain website functions and present in details how to conduct an actual clickstream data analysis from a framework perspective so that further researchers can apply to wide-ranging IB research.

3 Proposed Framework

3.1 An overview of the Integrated Clickstream Data Analysis Framework

Our overarching goal of this study is to contribute a standard clickstream data analysis framework to future information behavior research so that the researchers could apply it to different datasets and to address various research questions. To this end, we intend to introduce the framework independent of any settings. However, in order to make the illustration clear and comprehensible, we will introduce the clickstream data analysis of a Chinese academic library OPAC users' information behavior based on this framework as an example.

Figure 1 presents the major components and the general process of clickstream data analysis. Mainly, there are five key components composing two stages, i.e., data preparation and data investigation. Data preparation is the process of a) collecting and cleaning the clickstream data, and b) parsing and

coding the data according to the research purposes. Data investigation consists of three different levels of granularity: c) footprint level, d) movement level, and e) pathway level.

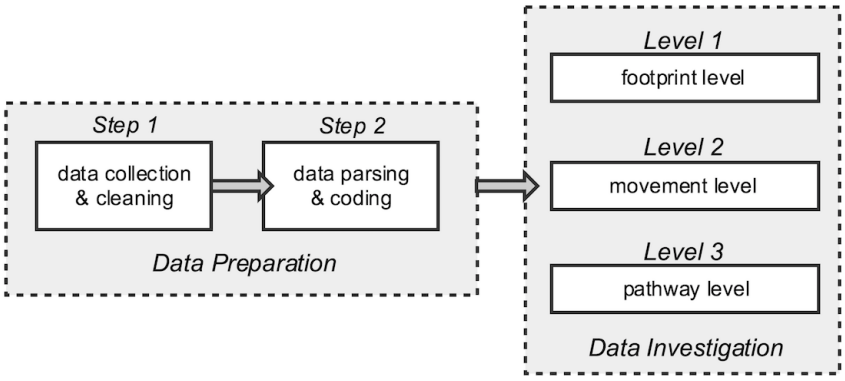

Figure 1. An Integrated Clickstream Data Analysis Framework

3.2 Data Preparation

3.2.1 Step1: Data Collection and Cleaning

When collecting data, the researchers are expected to first make sure their research questions and decide what data to collect. We suggest that there are three pre-requirements to examine after obtaining the data:

a) Is the dataset representative to the research question? Conditions such as the time period, target users could heavily affect the representativeness of the transaction logs. For example, user logs during the holidays are very likely to bias the results of analyzing the website usage on weekdays.
b) Is the transaction log file original? To keep the advantages of unobtrusive data collection method, it is researchers' responsibility to check if the data is completely and naturally collected. Otherwise, the intervention with the data will alter the analysis process as well as the outcomes.
c) Is the user identification information correctly hidden or transcribed? User privacy is a big issue in preparing the data. The transaction log releaser from the website as well as the study researchers should be very careful about protecting users' identification information. Usually, it can be conducted by encrypting user identities, e.g., assigning a new ID with the only purpose of distinguishing one user's interactions from others'.

No matter generated on a server side or a client side, the transaction log files automatically capture information related to each interaction. While some systems may record additional fields (e.g., bytes, DNS), the six fields listed below are the most necessary fields for understanding user behavior in terms of clickstream data:

- User_ID: user's IP address;
- Date: data on which request is made;
- Time: time when request is made in GMT;
- Method: type of client to server request, e.g. "GET" – requesting data from a resource and "POST" – submitting data to be processed to a resource;
- URL: URL of the resource requested;
- Status: HTTP status code returned by the server.

Data cleaning is a critical step in the data preparation. For most log files, there are two main goals in cleaning process. Firstly, the corrupted, duplicated or erroneous records need to be eliminated so that the data after cleaning reflects real user behavior in the certain website. An easy way to quickly identify these records is by sorting each field so that abnormal records would present on the top of, the bottom of, or grouped together in the sorted field (Jansen & Spink, 2006). Aside from them, researchers are encouraged to define redundant data that is not the focus of the study. Removing them would reduce the data size as well as expedite the analysis. Here provides following rules as a reference of redundant records that researchers may take into consideration:

- Failed requests: records whose status codes do not belong to the 200 class (successful requests), such as 404 (not found) and 500 (internal server error);
- External links: records whose URLs start with "http://", such as "http://www.google.com/";

- Requests involving data submission: records whose methods are “POST”;
- Requests for styles, structures, scripts, pictures, and other data: records whose URLs end with “png”, “jpg”, “gif”, “ico”, “css”, and “js”, etc.

3.2.2 Step2: Data Parsing and Coding

Session recognition and identification are usually unavoidable because of their necessity in labeling individual users. Though there exists no universal agreement in session definition, a 30 minutes' consecutive requests is the most adopted one in the literature (Lown, 2008; Chau et al., 2005).

The next step is to identify and remove the automatic session from the non-human requests. Due to the fact that various requests in the logs are actually made by agents such as web crawlers or other non-human programs, the parsing process needs to distinguish them from real user actions. Commonly, a 101 record is adopted as a cut-off threshold (Jansen& Spink, 2006). That means, all sessions containing more than 101 records are recognized as non-human interactions and deleted from the dataset. Matching the user-agent field in the log with the list of popular crawlers, e.g., User Agent String.Com (http://www.useragentstring.com/pages/useragentstring.php) is another effective way to identify the non-human requests. The parsing process can be conducted with either own written programs or SQL queries.

Coding the URL request in the clickstream data is the last step in data preparation. It is to interpret each user request as a meaningful interaction with the system such as reading a document shared by a friend or accessing to a personal homepage. Only with an in-depth understanding of the particular structure of the website under investigation can the researchers build a scheme for the data coding which is supposed to cover all types of interactions within a website. Another issue the researchers need to concern is the granularity of the coding scheme (Niu, 2012). Usually, to extract the representative interactions without missing detailed information, a coding scheme may capture both the higher-level interaction categories and the fine-grained actions in each category.

3.3.3 Data preparation in the OPAC research example

The original transaction log file is acquired from the Wuhan University Library's OPAC. It contains 26,732,368 clickstream records collected on the server over a two-month duration in a normal spring semester, from 00:00:00 April 1, 2014 to 23:59:59 May 31, 2014. The file follows the W3C Extended Log Format. It was reduced to six fields –User-IP, Date, Time, Method, URL, and Status. Figure 2 is a snippet from the reduced log file. Then the data cleaning process was completed with Python programs. As a result, 1,882,853 clickstream records remained in the log file, accounting for merely 13.04% of all the records in the original file.

```
USER-IP         Date         Time      Method   URL                                                              Status
116.211.128.23  01/Apr/2014  00:00:09  GET     /F/NY7Q2PY7VX33N4HTIQ31MN5GIJ92UR25V2J7EBSN7IS6J36LTH-12014?func=bor-renew
112.64.235.252  01/Apr/2014  00:00:13  GET     /F/6X2IH2NE2J1IG3SX58RAN9NUP4YMSX6ELT46E6RDDVBNA6INQR-02030 HTTP/1.1" 200
27.17.130.225   01/Apr/2014  00:00:24  GET     /F/I4AYDGVK3DRXNQ8Y9GB3EA5Y6TIDNBHKIPM69PUJ8ARJS3EG93-11904?func=short-jum
180.153.206.30  01/Apr/2014  00:02:10  GET     /F/IJAR4RFMFTN7IV5AXG7YB28A5MNMT7HNTLPGJ1X6JIQXR5VEDJ-29480 HTTP/1.1" 200
180.153.163.206 01/Apr/2014  00:02:11  GET     /F/IJAR4RFMFTN7IV5AXG7YB28A5MNMT7HNTLPGJ1X6JIQXR5VEDJ-29480?func=find-m&fi
101.226.65.107  01/Apr/2014  00:02:15  GET     /F/IJAR4RFMFTN7IV5AXG7YB28A5MNMT7HNTLPGJ1X6JIQXR5VEDJ-29526 HTTP/1.1" 200
101.226.66.179  01/Apr/2014  00:02:18  GET     /F/IJAR4RFMFTN7IV5AXG7YB28A5MNMT7HNTLPGJ1X6JIQXR5VEDJ-29526?func=full-set-
101.226.89.120  01/Apr/2014  00:02:33  GET     /F/IJAR4RFMFTN7IV5AXG7YB28A5MNMT7HNTLPGJ1X6JIQXR5VEDJ-29709 HTTP/1.1" 200
```

Figure 2 A Snippet from an Academic Library OPAC Transaction Log

The server of the OPAC will automatically assign a random session ID to each visit after the first request is made. The same session ID can be found in the URLs of all the records belonging to one visit. Different visits by the same user will be given different session IDs. If different users visit the OPAC at different times with the same computer in the library, different session ID will be assigned though their IP address will be identical. This mechanism offered great convenience for accurate session identification and helped extract 654,598 sessions. The next step was to determine which sessions were non-human ones, and a cut-off of 101 records was adopted. In this way, a total of 653,994 human sessions were obtained. More than 70% of the IPs have only one session.

The final preparation before data analysis was translating users' page requests into tangible actions, such as performing a simple search or checking personal borrowing records. This study created a data coding system based on the OPAC's hierarchical structure and individual pages' functions. There are five major page categories, i.e. the OPAC homepage (H), search pages (S), search result pages (R), resource detail pages (D), and personal library pages (L).

3.3 Data Investigation

Figure 3 provides a simplified yet intuitive illustration of the relationships among the three key levels in the data investigation process which are footprint, movement, and pathway. A user enters the website and

lands on Page 1, executes a series of clicks and reaches Pages 2, 3, 4, and 5 in sequence, and finally exits the site from Page 5. Each click arouses a movement, the changing of location from one page to another, and leaves a footprint, a mark showing one's presence on a page. A pathway takes shape as a result of the click series, composed of all the movements arranged according to the time they occur.

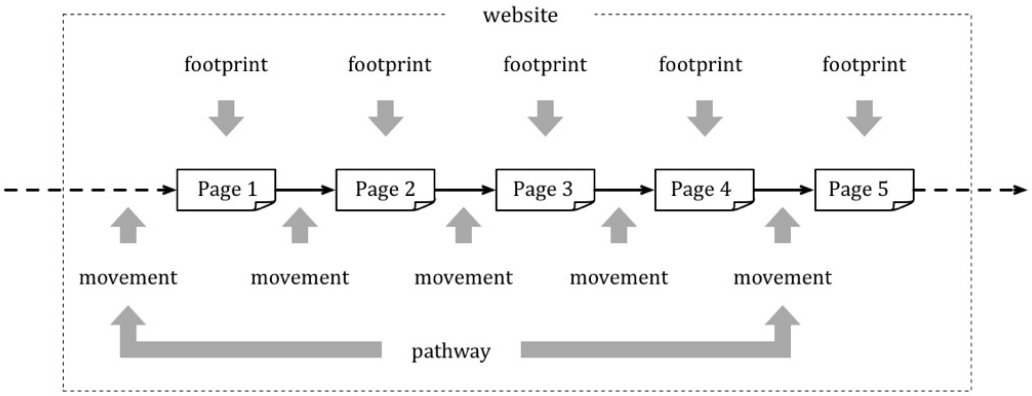

Figure 3. An illustration of the relationships among footprint, movement, and pathway

For a particular footprint which is the ith (i>= 1) page one requests within a site, it is denoted with F_i. The movement leaving that footprint can be represented as M_i: $F_{i-1} \rightarrow F_i$. If one requests j (j >= 1) page(s) during a visit, then the pathway of the visit can be represented as P_j: $F_1 \rightarrow F_2 \rightarrow \cdots \rightarrow F_{j-1} \rightarrow F_j$. In practice, footprints, movements, and pathways were easily recognizable from the processed log file, so data investigation could be conducted at all three levels.

3.3.1 Investigation Level 1: Footprint Level Investigation

Given the illustration of footprint in Figure 3, it is represented by clickstream record and stands for the real action hidden behind the quest. With the code scheme at hand, researchers could run programs or use database functions, e.g., SQL commands to translate each clickstream record into meaning footprint according to its URL. The footprint level investigation takes the footprint as the element unit.

This level investigation is conducted to provide information about how users' footprints distributed among different page categories and how they distributed within each category. The former could help shape a basic understanding of users' general usage of the website, and the latter tells about users' preferences when using specific functions offered by the system.

To achieve this, the footprint distribution among page categories needs to be calculated. Meanwhile, for the critical category, a further computing of the footprint distribution within each category enriches the understanding of users' behavior.

In the example of the OPAC research, a total of 2,091,904 footprints were created during the two months. Figure 4 is a treemap that demonstrates their distribution among the five major page categories. Nearly half of these footprints were left on search result pages (R, 48%), followed by personal library pages (L, 25%), the OPAC homepage (H, 13%), resource detail pages (D, 12%), and search pages (S, 2%). The most prevalent form of user-system interaction was viewing search results, which echoes a previous study (Lown, 2008). Also, users frequented their personal libraries in order to make better use of both online and offline services. Nevertheless, the visit of the OPAC homepage was not as common as expected. The viewing of search results did not proportionally lead to many resource detail views, and the use of various independent search interfaces was surprisingly low. Figure 4 demonstrates footprint distribution within each category as well. For example, R1 represents requesting result pages derived from the simple search on the Wuhan University library homepage. D1 means requesting bibliographic information of a book. Other sub-categories are not listed here since they are not the focus of this work.

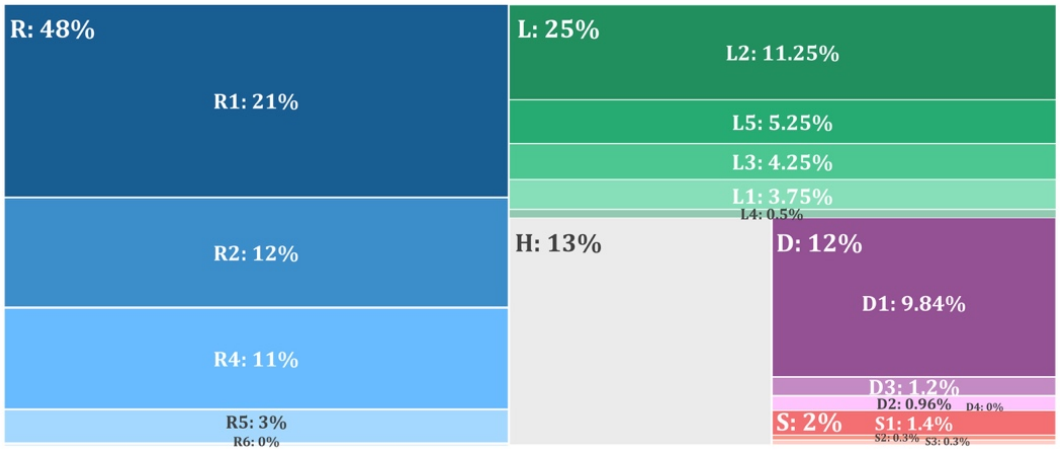

Figure 4. Footprint distribution among page categories and within each category

3.3.2 Investigation Level 2: Movement Level Investigation

A movement is extracted by linking two sequential footprints within a session. It means that the user transfers between two actions. The movement level analysis highly depends on the research question to identify one key footprint or several key footprints within the website. After determining key footprint, the investigation focus could narrow down to the movements that either departing from it or arriving at it.

In the example of the OPAC research, the footprint level analysis shows that the search result pages were the most visited page category, accommodating nearly half of the footprints ever left. The movement level analysis concentrated on the preceding movements and succeeding movements which are immediately associated with the R footprints. Among the 653,994 sessions, only 301,928 of them contain one or more R footprints. A Python program was used to identify and capture the first R footprints in all sessions and then counted their adjacent footprints by type to create a movement map where each arrow represents a type of movements with the width proportional to the percentage (Figure 5).

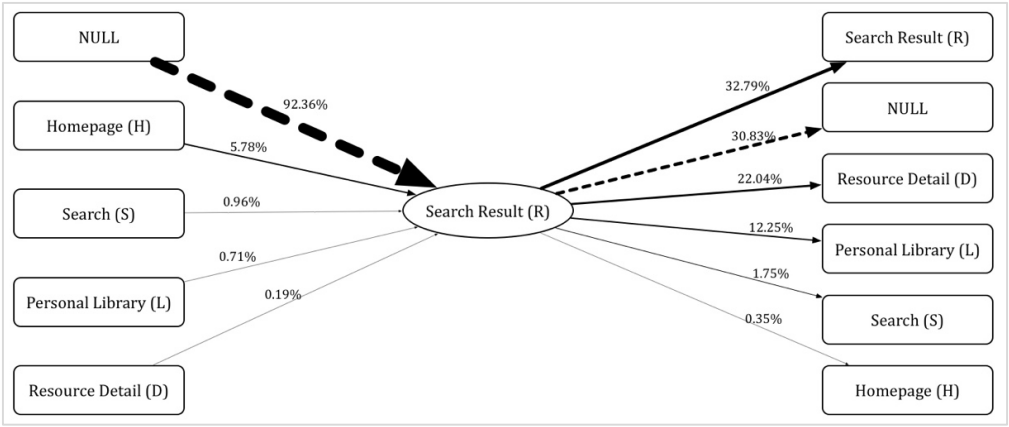

Figure 5. Preceding movements (left) and succeeding movements (right)

On the left of the movement map are preceding movements. Surprisingly, 92.36% of the preceding movements were also entering movements. In other words, an overwhelming proportion of all the visits to the OPAC started with viewing a search result page. But one must execute a search before getting results. It was noticed that in most cases (89.13%) users were actually directed to the OPAC search result pages from the Wuhan University library homepage (R1) which was however not logged as a within-OPAC page category, therefore the search result page became the first page requested in a session. It's possible that most users deemed the OPAC an organic part of the library website and the simple search interface on the homepage a default entry into the OPAC.

Succeeding movements can be found on the right of the movement map. Almost one-third (30.83%) of the succeeding movements were the exiting movements. It can be explained in two ways: users found what they wanted on the first search result pages and left the system; or they thought they wouldn't find what they wanted in the system and gave up. Another one-third (32.79%) of the succeeding movements left R footprints. They represent the situations in which the user's information need was not

satisfied by the first search result page and further actions were taken, such as navigating to the next page or reformulating the query. The last one-third reflect the natural next steps after result set examination, either clicking through to view more resource detail (22.04%) or utilizing personal libraries (12.25%) to save result items or reserve resources for borrowing.

3.3.3 Investigation Level 3: Pathway Level Investigation

The pathway level investigation aims at detecting important patterns of users' information seeking processes. The boundaries of a pathway are already visible in the log file thanks to data parsing, and it can be extracted by concatenating all the clickstream records in a session. Four attributes were introduced to describe a pathway: duration, length, width, and capacity. For a pathway, its *duration* is the time difference between the generation of the first and last footprints, and its *length* the total number of footprints generated. They are analogous to Jansen & Spink's (2006) session duration and length. The *width* of a pathway refers to the quantity of key footprints (e.g., R in the OPAC clickstream analysis) left along the way, and the *capacity* that of resource achievements (e.g., D in the OPAC clickstream analysis). These footprints are closely related to resources and may reflect the efficacy of resource finding. Together the four attributes draw a comprehensive picture of users' navigation through the website.

Therefore, the researchers first need to calculate the values of the four attributes and descriptive statistics of them could be obtained. Then to give an in-depth analysis of users' pathway, the relationships among the four attributes could be obtained through a correlation analysis. For instance, in the OPAC study, we found that, as a whole, all pairs of attributes are significantly correlated ($p < .01$). However, a strong correlation can be only found between length and width ($r = .911$, $r \geqslant .8$). The more the pages requested during a visit, the more the search result pages viewed, which conforms to users' primary purpose of using the OPAC and the fact that the search result pages attracted the largest proportion of user footprints.

In addition, to characterize the pathways along which users navigate through the website, another method to present the pathway level investigation is identifying and visualizing the typical pathways. Sankey diagram is an appropriate technique which is helpful for observing similar pathways. Figure 6 is a visualization of all the distinct short pathways (i.e., 1<length<=5) altogether so as to reveal users' navigation patterns in the OPAC step by step. The open source tool Sankey Diagram Generator (http://sankey.csaladen.es/) was used to create the visualization (Figure 6) in which the width of the flow branches is proportional to users' visiting traffic volume. It should be viewed from left to right, with the sequence of the footprints in a pathway indicated with the numbers (1 to 5) preceding the page category codes.

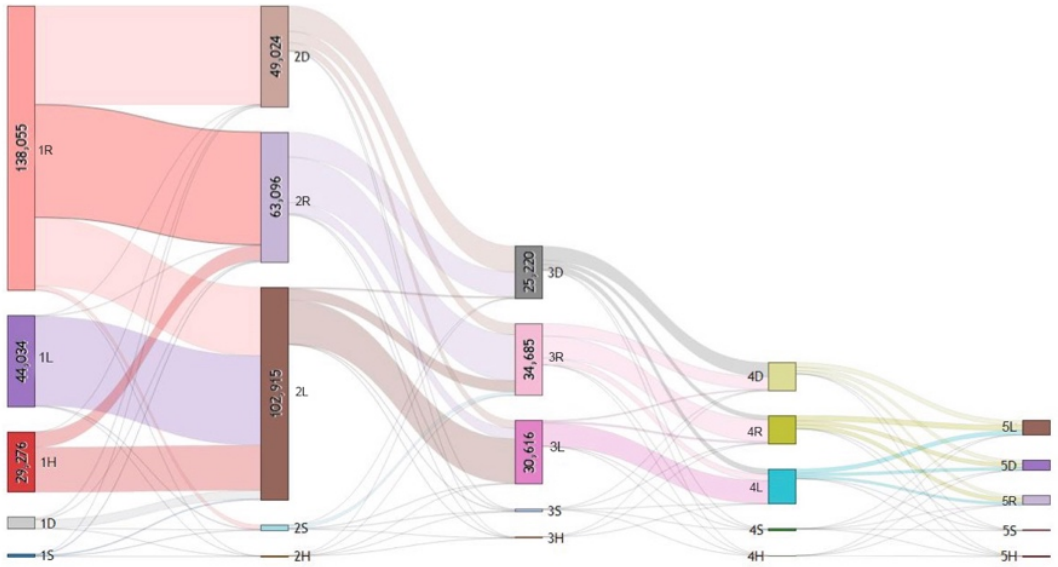

Figure 6. An Example of a Sankey Diagram of the Users' Pathway

4 Discussion

4.1 Benefits

The first obvious benefit of this clickstream data analysis framework is that it provides an approach to understand web users' information behaviors based on large-scale real user transaction logs. Though search log analysis has been adopted widely, given the rich information contained in users' clickstream, now studies get the possibilities to obtain a more comprehensive understanding of online information behavior, rather than only focused on its subset, i.e., search behavior.

The second benefit comes from the investigation process which consists of three different levels of granularities. It enables the researchers and the websites to understand users' behavior with both details and overall trends. This understanding will help to build a realistic and holistic picture of the interactions between the users and the websites, and further, facilitate the websites to grasp the actual usage and get design implications.

Thirdly, aside from contributing a standard easy-to-follow analysis framework to those who have data but don't know how to interpret it, we intentionally make the framework flexible and can be easily modified based on research questions. That means we only describe what key footprint, key movement is in the framework, but it is the research questions and the website features who decide what they are in a particular study. For example, in the current OPAC study, the movement level investigation addressed the movements around R (search result pages) footprint. While if it was a study aiming to reveal how OPAC users were directed to book details, D (resource detail pages) would be identified as the key footprint and a movement level investigation centered on D footprint would be conducted according to the research question.

4.2 Limitations

However, it is also critical to know the limitations of this framework. First of all, as with all the log analysis method, it fails to capture any contextual information in which each request was logged (Sheble & Wildemuth, 2009). Contextual information mainly refers to users' personal information, including implicit characteristics, demographics, motivations, satisfaction levels, etc. (Niu, 2012). These pieces of information are critical in telling a whole story about users' online information behavior. Despite the rich information in clickstream data, it is difficult to directly link users' behavior to the reasons behind. Therefore, considering this inherent limitation of the unobtrusive data collection method, we suggest researchers interested in human factors, if possible, to complement clickstream data analysis with an obtrusive data collection, such as user experiments and surveys.

Additionally, one should recognize that the current framework is primarily designed for investigating users' information behavior within one specific site, rather than exploring their online behavior as a whole. And last but not least, the framework's applicability in other kinds of websites needs to be verified with more empirical studies.

5 Conclusion

Clickstream data provides great opportunities for comprehensively understanding web users' information behavior. To change its long-time being underexploited situation, we offer an integrated clickstream data analysis framework to the information scientists who are interested in using this data source to uncover online information behavior in real-world settings. The framework is conveyed in great details with standard procedures as well as results from an empirical analysis. In this way, we expect future researchers could apply this framework to different websites' clickstream data without much difficulty. The three different investigation levels of granularity allow the analysts to probe into the data from different angles. Besides, by describing key concepts and considerations in each step, rather than strictly rule the practices, we want to emphasize that the framework is flexible and demands the researchers to identify the research interests and determine the focus of the analysis.

This proposed framework is a good starting point for clickstream data analysis in IB domain, however we acknowledge that the practical log analysis heavily relies on the specific features of a website and its particular data fields. The current framework is insufficient in capturing all the possibilities. For example, if the web server doesn't assign a unique user ID as in the OPAC logs, researchers need to figure out alternative ways to identify the sessions generated by different users from the same IP, such as combing the browser information. We expect to refine the current framework to help researchers make decisions at each step. Experience from more empirical studies of different types of websites will be included for polishing the framework in the future.

It is also notable that clickstream data suffers from the inherent shortcoming of the transaction log, the missing of contextual information which is critical in understanding the reasons behind users' behavior. Therefore, for the researchers with this need, other obtrusive data collecting might help to complement. Overall, we hope this work could help to contribute a comprehensive understanding of web users' information behavior in the IB domain.

6 References

Aguiar, L., & Martens, B. (2016). Digital music consumption on the internet: evidence from clickstream data. *Information Economics and Policy*, 34, 27-43. Chicago

Asunka, S., Chae, H. S., Hughes, B., & Natriello, G. (2009). Understanding academic information seeking habits through analysis of web server log files: The case of the teachers college library website. *The Journal of Academic Librarianship*, 35(1), 33-45.

Bates, M. J. (2002). Toward an integrated model of information seeking and searching. *The New Review of Information Behaviour Research*, 3, 1-15.

Benevenuto, F., Rodrigues, T., Cha, M., & Almeida, V. (2009, November). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference* (pp. 49-62). ACM.

Case, D. O. (2012). *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Pub Limited.

Chau, M., Fang, X., & Sheng, O. R. L. (2005). Analysis of the Query Logs of a Web Site Search Engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363-1376.

Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI* (pp. 349-372). Springer New York.

Greifeneder, E. (2014). Trends in information behaviour research. *Information Research*, 19(4).

Han, H., Jeong, W., & Wolfram, D. (2014). Log analysis of academic digital library: User query patterns. *iConference 2014 Proceedings*.

Huang, C. Y., Shen, Y. C., Chiang, I., & Lin, C. S. (2007). Characterizing Web users' online information behavior. *Journal of the American society for information science and technology*, 58(13), 1988-1997.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 28(3), 407-432.

Jansen, B. J. (2008). The methodology of search log analysis. In Bernard J. Jansen, Amanda Spink (Ed.) Handbook of research on Web log analysis (pp. 100-123). *Information Science Reference*, New York, NY.

Jansen, B. J., Spink, A., & Koshman, S. (2007). Web searcher interaction with the Dogpile. com metasearch engine. *Journal of the American Society for Information Science and Technology*, 58(5), 744-755.

Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, 56(6), 559-570.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42 (1), 248-263.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*, 36(2), 207-227.

Jiang, T. (2010). Characterizing and Evaluating Users' Information Seeking Behavior in Social Tagging Systems. *Doctoral dissertation, University of Pittsburgh.*

Jiang, T. (2014). A clickstream data analysis of users' information seeking modes in social tagging systems. In M. Kindling, E. Greifeneder (Ed.) *iConference 2014 Proceedings*. iSchools, Illinois.

Jiang, T., Chi, Y., Jia, W. (2014). Exploring users' within-site navigation behavior: A case study based on clickstream data. *Chinese Journal of Library and Information Science*, 4, 005.

Julien, H., Pecoskie, J. J., & Reed, K. (2011). Trends in information behavior research, 1999–2008: A content analysis. *Library & Information Science Research*, 33(1), 19-24.

Kato, M. P., Sakai, T., & Tanaka, K. (2013). When do people use query suggestion? A query suggestion log analysis. *Information retrieval*, 16(6), 725-746.

Kazdin, A. E. (1979). Unobtrusive measures in behavioral assessment. *Journal of Applied Behavior Analysis,* 12(4), 713-724.

Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. In *Applications of Data Mining to Electronic Commerce* (pp. 59-84). Springer US.

Lown, C. (2008). A transaction log analysis of NCSU's faceted navigation OPAC. *Master's Paper. University of North Carolina at Chapel Hill.*

Martzoukou, K. (2005). A review of Web information seeking research: considerations of method and foci of interest. *Information Research*, Volume 10 Number 2.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, 13(1), 29-39.

Montgomery, A. L. (2001). Modeling purchase and browsing behavior using clickstream data. In *Presentation at the UC Berkeley Fifth Invitational Choice Symposium, Monterey*, CA. Retrieved from: www. andrew. cmu/edu/user/alm3/presentations/choicesymposium2001/montgomery. pdf.

Moulaison, H. L. (2011). OPAC queries at a medium-sized academic library. *Library Resources & Technical Services*, 52(4), 230-237.

Niu, X. (2012). Beyond text queries and ranked lists: Faceted search in library catalogs. *Doctoral dissertation, The University of North Carolina at Chapel Hill.*

Olbrich, R., & Holsing, C. (2011). Modeling consumer purchasing behavior in social shopping communities with clickstream data. *International Journal of Electronic Commerce*, 16(2), 15-40.

Pai, D., Sharang, A., Yadagiri, M. M., & Agrawal, S. (2014, October). Modelling visit similarity using click-stream data: A supervised approach. In *International Conference on Web Information Systems Engineering* (pp. 135-145). Springer International Publishing.

Park, M., & Lee, T. S. (2013). Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, 31(1), 123-140.

Rieh, S. Y., & Xie, H. (2001, November). Patterns and sequences of multiple query reformulations in web searching: A preliminary study. In *Proceedings of the Annual Meeting-American Society for Information Science* (Vol. 38, pp. 246-255). 1998.

Rutz, O. J., & Bucklin, R. E. (2012). Does banner advertising affect browsing for brands? clickstream choice model says yes, for some. *Quantitative Marketing and Economics*, 10(2), 231-257.

Sen, A., Dacin, P. A., & Pattichis, C. (2006). Current trends in web data analysis. *wghu*, 49(11), 85-91.

Sheble, L., & Wildemuth, B. (2009). Transaction logs. *Applications of social research methods to questions in information and library science*, 166-177.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999, September). Analysis of a very large web search engine query log. In *ACM SIGIR Forum* (Vol. 33, No. 1, pp. 6-12). ACM.

Teevan, J., Adar, E., Jones, R., & Potts, M. (2006, August). History repeats itself: repeat queries in Yahoo's logs. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 703-704). ACM.

Teevan, J., Ramage, D., & Morris, M. R. (2011, February). # TwitterSearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 35-44). ACM.

Ting, I., Kimble, C., & Kudenko, D. (2009). Finding unexpected navigation behaviour in clickstream data for website design improvement. *Journal of Web Engineering*, 8(1), 71-92.

Torres, S. D., Hiemstra, D., & Serdyukov, P. (2010, July). Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 847-848). ACM.

Tsikrika, T., Müller, H., & Kahn Jr, C. E. (2012). Log analysis to understand medical professionals' image searching behaviour. *Studies in health technology and informatics*, 180, 1020.

Vakkari, P. (2008). Trends and approaches in information behaviour research. *Information Research*, 13(4).

Villen-Rueda, L., Senso, J. A., & de Moya-Anegón, F. (2007). The use of OPAC in a large academic library: A transactional log analysis study of subject searching. *The Journal of Academic Librarianship*, 33(3), 327-337.

Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., & Zhao, B. Y. (2013). You are how you click: Clickstream analysis for sybil detection. In *Presented as part of the 22nd USENIX Security Symposium* (USENIX Security 13) (pp. 241-256).

Wang, G., Zhang, X., Tang, S., Zheng, H., & Zhao, B. Y. (2016, May). Unsupervised Clickstream Clustering for User Behavior Analysis. In *SIGCHI Conference on Human Factors in Computing Systems.*