8-2012

From Clickstreams to Searchstreams: Search Network Graph Evidence from a B2B E-Market

Mei LIN
Singapore Management University, mlin@smu.edu.sg

M. F. LIN

Robert J. KAUFFMAN
Singapore Management University, rkauffman@smu.edu.sg

From Clickstreams to Searchstreams: Search Network Graph Evidence from a B2B E-Market

Mingfeng Lin
The University of Arizona
Tucson, Arizona, USA
mingfeng@eller.arizona.edu

Mei Lin
The University of Hong Kong
Pokfulam Road, Hong Kong
linm@hku.hk

Robert J. Kauffman
Singapore Management University
Singapore
rkauffman@smu.edu.sg

ABSTRACT
Consumers in e-commerce acquire information through search engines, yet to date there has been little empirical study on how users interact with the results produced by search engines. This is analogous to, but different from, the ever-expanding research on clickstreams, where users interact with static web pages. We propose a new network approach to analyzing search engine server log data. We call this *searchstream data*. We create graph representations based on the web pages that users traverse as they explore the search results that their use of search engines generates. We then analyze the graph-level properties of these *search network graphs* by conducting cluster analysis. We report preliminary evidence the presence of heterogeneity among users in terms of how they interact with search engines. This suggests that search engine users may not all benefit from the same functionality in the search engines they rely upon. We also offer additional evidence on the empirical regularities associated with a variety of relevant issues that arise in the business-to-business (B2B) e-market context that we have studied.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval – *clustering, search process, selection process.*

General Terms
Management, Measurement, Documentation, Economics.

Keywords
Big data, clickstreams, data mining, graph theory, keyword search, online markets, search behavior, searchstreams.

1. INTRODUCTION
As e-commerce reaches a higher level of sophistication, increasing revenues from online sales and advertising will be constrained by the existing set of assumptions about the manner in which users browse, select, and examine information. Search engine usage is particularly dominant among the many ways that people now acquire information online. Search engine usage has remained steady in the past decade, and is now recognized as the most popular activity on the Internet. Google, whose main revenue source is search advertising, generated US$36 billion in advertising revenues in 2010, and its power and influence continue to grow. Still little is known about the steps and sequences users take to examine the pages of search results returned.

We call the data that describe a user's steps and sequences in the query process as her *searchstream*. A user's searchstream consists of a variety of different things, including: the websites that she searches and their URLs; the search keywords that she use employs; the links that she investigates along the way; and, in a more general way, the descriptions of the *trajectory of observed behavior* that she demonstrates in the process. A key observation that we wish to offer is that the *economic value* – in particular the *action relevance* in decision-making and in subsequent behavior – of all of the things that characterize a user's searchstream are not clear, based on the relevant research. For example, the search keywords that users key off of, and the rankings of the search results that they receive all remain unclear. To address this, we investigate users' *traversal paths* and their *search results generation* based on an analysis of a large data set from a global-scale B2B electronic market. Our intention is to develop descriptive results of the *empirical regularities* that enable us to categorize user search behavior. We will also discuss several different ways that our new perspective can be leveraged to provide interesting and new perspectives related to the observed searchstreams of real-world e-market users.

Researchers have been able to surface especially insightful evidence based on large amounts of clickstream data obtained from e-commerce systems repositories, agent-based online data collection, Internet screen scraping and other less powerful methods [2, 3, 4]. Other researchers have delved into the different *page categories* that a consumer browses through within a site, as well as *website stickiness* and *viewing duration* [1]. In the search engine services context, the patterns in which users identify potential sellers and their products also have received little attention. Prior research typically considered within-site, category-level analytics, treating multiple *page views* as a category of webpages browsing behavior. An exception is Montgomery et al. [5], who analyzed user search paths on BarnesandNoble.com. They used a multinominal probit model withdata to characterize user search paths, and improve the accuracy of predictive analytics for the expected conversion rate for consumer purchases.

User searchstreams reveal information that is critical to online sellers and advertisers, as well as to the suppliers of search engine services. Graph properties are useful for capturing the micro-level information in the data of the nature we have been discussing. Cycles of search network graph will indicate a user's search breadth, and the lengths of the paths that are traverses will reflect her search depth. By representing all searches as *directed acyclic graphs* (DAGs), we also are able to obtain aggregate-level results. An initial empirical regularity in our research is that nearly all users seem to conduct at least one subsearch. Related to this, users seem to spend a substantial amount of time on the *landing pages* that originated from the search results they click on. We conjecture that the observed

International Conference on Electronic Commerce '12, August 6-8, 2012, Singapore Management University, Singapore.

searchstreams that users demonstrate may be *rational adaptations* to produce greater effectiveness in their search.

2. SEARCH RESULT PAGE DATA

Our data set contains information on the sequence of user steps in browsing the *search result pages* (SRPs) generated by the search engine that is used in the B2B e-market. A key characteristic of the search behavior of users in our data set is that it supports their business objectives. Thus, an assessment of user behavior should indicate core characteristics of *economic behavior* with fewer idiosyncracies compared to data on individual consumer actions. Each user's behavior should be value-driven and reflect the natural heterogeneity of individual differences that arise in a business transaction-making setting.

Our data represent user-initiated *search sessions.* In each, the user's page visits are numbered sequentially. We observe their search queries, timestamps for their actions, the number of clicks they made on each SRP, and the sequence in which each user steps through the results pages. We employ the structure of a *directed graph* to analyze users' search path. Each search graph originates from the first SRP after a query is submitted. Thus, the *nodes* of a network search graph are the pages the user visits after she starts with the first SRP. The *edges* of the graph show the specific paths the user has taken. There may be one or more edges of paths that are indicated, as a user returns to the first SRP, which is the starting point for all of her sub-search behavior, and a direction of the user's exploration.

The number of search paths reflects a user's search breadth. From the page of origin, the user also may initiate sub-searches related to the current search. We define a user's *search depth* by the extent to which they examine the details of the search results, rather than the number of SRPs they go through. We also use graph properties to rule out contaminated data entries, and to describe the search path characteristics. For instance, when a graph is not a directed and acyclic, this implies that the user returns to the page of origin to start a sub-search.

3. PRELIMINARY RESULTS

Since applying searchstream network analysis methods to search logs is a brand new area with no prior literature, our current investigation is data-driven and exploratory. We report preliminary results based on cluster analysis methods. See Table 1.

Table 1. Summary Statistics of Network Search Graphs

Graph Property	Population False	Population True
Is_Dag	98.6%	1.4%
Is_Eulerian	16.6%	83.4%
Weakly_Connected	0.0%	100.0%

Aggregate statistics show that nearly all of the searches that the B2B e-market's search functionality users made are not DAGs. So most users almost always revisit SRPs or refine their searches. They exhibit economic behavior in their search, but the information that they gather the first time through a given SRP may not be sufficient for them to implement whatever stopping rule they have for their search to support decision-making for product or service purchase. The majority appears to click on other links at least once, and visit destination pages outside of the SRPs. Users spend more than 70% of the time outside of the SRPs they visit to examine details of their search results.

We conducted cluster analyses to identify different patterns in ways users interact with the search engine. From the raw input of 440,376 search sessions over a two-week period, Data Mining Plug-In for MS Excel identified ten different clusters. The size of clusters ranged from 4,821 to 147,107 sessions. Some of the most informative characteristics of each cluster include: number of nodes, network search graph density and Eulerian graph.

Across the ten clusters, most averaged between two and four nodes in the searchstreams. An extreme instance is Cluster 8. 94% of its network search graphs are non-Eulerian, indicating that search was not efficiently conducted by the users. See Table 2. Cluster 8 also has the lowest density of any of the network search graphs. These sessions often feature long periods of time that the user spent outside of main SRPs in the searchstream. This indicates exhaustive browsing and exploring.

Table 2. Characteristics of Three Different Clusters

Cluster Number	Size (of All Sessions)	Avg. Num. of Nodes	Avg. Density	Eurlerian
Cluster 8	1.6%	16.6	0.1	6%
Cluster 2	26%	2	1	100%
Cluster 6	4%	6	0.25	17%

One potential application of our network search graph approach of searchstream data is to identify search engine designs that will be able to minimize the size of this cluster. In our data set, Cluster 8 accounts for only about 1.6% of all sessions. A tentative conclusion might be that the B2B e-market's search engine engenders fairly efficient user search behavior.

A second interesting cluster suggests just the opposite: a cluster characterized by highly efficient, but probably completely uninformative and unsuccessful user search behavior. This is Cluster 2, wherein virtually all search graphs have a density of 1. This suggests that the search user does not click on anything before terminating the session. Even though this case may appear trivial, it turns out that this kind of searchstream behavior actually constitutes the second largest cluster: about 26% of all network search graphs are like this. The implication is that, while not all search sessions are informative, they still create a traffic burden for the search server. Search engine practitioners need to better understand how to reduce the size of this cluster.

Many other clusters seem to fall somewhere in between these extremes that we have described. We are in the process of refining the cluster analysis and assessing other potential observations to provide clear and compelling economic interpretations of the observed behavior. Other data analytics approaches that will be required are likely to include: quasi-experimental data selection, data stratification and other approaches that involve post-cluster analysis exploratory research actions.

4. REFERENCES

[1] Bucklin, R. E., Sismeiro, C. 2003. A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research* 40, 3, 249-267.

[2] Johnson, E. J., Moe, W. W. Fader, P. S., Bellman, S., and Lohse, G.L. 2004. On the depth and dynamics of online search behavior. *Management Science* 50, 3, 299-308.

[3] Moe, W.W. 2006. An empirical two-stage choice model with varying decision rules applied to Internet clickstream data. *Journal of Marketing Research* 43, 4, 680-692.

[4] Moe, W. W., and Fader, P. S. 2004. Dynamic conversion behavior at e-commerce sites. *Management Science* 50, 3, 326-335.

[5] Montgomery, A. L, Li, S., Srinivasan, K., and Liechty, J. C. 2004. Modeling online browsing and path analysis using clickstream data. *Marketing Science* 23, 4, 579-595.