External to internal search: Associating searching on search engines with searching on sites

Adan Ortiz-Cordova [a], Yanwu Yang [b], Bernard J. Jansen [c],*

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA
[b] Huazhong University of Science and Technology, China
[c] Social Computing Group, Qatar Computing Research Institute, Doha, Qatar

ARTICLE INFO

ABSTRACT

We analyze the transitions from external search, searching on web search engines, to internal search, searching on websites. We categorize 295,571 search episodes composed of a query submitted to web search engines and the subsequent queries submitted to a single website search by the same users. There are a total of 1,136,390 queries from all searches, of which 295,571 are external search queries and 840,819 are internal search queries. We algorithmically classify queries into states and then use n-grams to categorize search patterns. We cluster the searching episodes into major patterns and identify the most commonly occurring, which are: (1) *Explorers* (43% of all patterns) with a broad external search query and then broad internal search queries, (2) *Navigators* (15%) with an external search query containing a URL component and then specific internal search queries, and (3) *Shifters* (15%) with a different, seemingly unrelated, query types when transitioning from external to internal search. The implications of this research are that external search and internal search sessions are part of a single search episode and that online businesses can leverage these search episodes to more effectively target potential customers.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

It is known that people frequently use multiple information platforms (Hoa, Lin, & Chen, 2012), such as search engines and websites (Kumar & Tomkins, 2010), in order to address a task that involves searching. For example, a substantial percentage of Web users view major search engines (e.g., Google, Baidu, Yandex, NAVER) as entry points to the Web. These users may then traverse from a major search engine to a particular website (e.g., Amazon, TMall, Ulmart, Gmarket), at which point the searchers may refine their search (Obendorf, Weinreich, Herder, & Mayer, 2007). These searches are obviously related via a common task. However, searches on these separate platforms have typically been analyzed in isolation (i.e., treated as separate searching sessions). In reality, given the single underlying task, the searches likely comprise a single search episode that should be examined holistically (i.e., not treated as separate searching sessions) in order to understand the underlying user task. In this research, we examine one aspect of these multi-platform searching episodes, namely the search that brings the user from a major search engine to a particular website.

Understanding this type of searching is important, as the majority of websites rely on search engines for substantial percentages of their traffic, with more than 80% of Web and Internet users employing a search engine as their starting point

* Corresponding author.

E-mail addresses: adan.ortiz@outlook.com (A. Ortiz-Cordova), yangyanwu.isec@gmail.com (Y. Yang), jjansen@acm.org (B.J. Jansen).

(Bucklin & Sismeiro, 2003). This search engine traffic is critical for many online businesses, as traffic from search engines is free and continual, providing potential customers and users. Popular websites typically get crawled and indexed by search engines on a regular basis. So, if a website has a number of pages indexed by the search engines, any of these indexed pages has the potential to appear in the search results and become the landing page for the user. The landing page referred to by a link on a search engine results page becomes the de facto homepage for the website. If a user determines that the landing page does not contain the information he/she is looking for, the user may leave the site and go back to the search engine to look elsewhere. A user bouncing from a website is obviously not good for the online business because it means the potential loss of a sale, a registration, or advertising revenue. As such, it is a standard of web analytics practice to keep the bounce rate (i.e., percentage of visitors arriving at a site and then leaving without taking further action) low (Jansen, 2009).

One way to combat a user bouncing from the landing page is to provide a searching capability via a site search so that users can find the information they are looking for, thereby remaining on the website. As many websites are complex and information-rich, visitors must leverage a site search to find what they want; therefore, site search is an important website information architecture and navigational feature.

We refer to the capability and the action of searching a website as *internal search*, and we refer to queries submitted to the site search service as *internal search queries*. We define internal search as *one or more queries submitted to a site's specific search service in order to find information that is contained on that site*. We refer to the capability and the action of searching using a general purpose search engine as *external search*, and queries submitted during external search are referred to as *external search queries*. Certainly, the view of what is internal or what is external search may vary depending on the perspective of the research.

Specifically, in this research, we investigate the transition between *external search* (ExS) and *internal search* (InS). Both ExS or InS may consist of one or more queries. In this research, we are focusing on those searches where a searcher conducts an ExS and then subsequently conducts InS in continuation of the same search task, although other ExS–InS patterns may exist. In these *external to internal search* (Ex2InS) *episodes*, the searcher conducts the ExS by submitting a query to a major search engine and then submitting follow-on queries to a site search service. So, although ExS and InS can each occur in isolation (i.e., ExS with no InS, or InS with no ExS), this research focuses on combined occurrences, specifically the transition from ExS to subsequent InS usage. Fig. 1 illustrates the ExS, InS, and Ex2InS concepts.

Prior research has not examined the relationship between ExS and InS. In fact, prior research has mainly treated these as distinct searching sessions (Jansen & Spink, 2005; Wang, Berry, & Yang, 2003); though, obviously, they are related because it is the referral query on the web search engine that brings the user to a particular website. Our premise is that InS is a continuation of, at least, the last ExS query. In these situations, there is a connection between the ExS and subsequent InS, and they are each part of the same Ex2InS episode.

Therefore, it is important to investigate the InS behaviors that manifests following ExS in order to develop better InS capabilities, design search personalization features, identify sponsored search keywords, discover user intent, and document missing content that an online business may not have in its data collection. So, online data can support other events, including offline events (Andreas & Pascal, 2013). Investigating Ex2InS as a single search episode may also provide insights that will assist web search engines by offering more granular search results. Therefore, understanding these Ex2InS behaviors has the potential for significant practical impact, in addition to advancing theoretical understanding of user searching.

Naturally, we acknowledge that the search process may be not necessary linear. In fact, Ex2InS may be repeated multiple times (i.e., Ex2InS$_n$). Also, there may be a variety of other patterns, including Ex2In2ExS, where the external site may or may not be the same, or In2Ex2InS, where the user leaves the internal site and uses an external search engine to locate specific
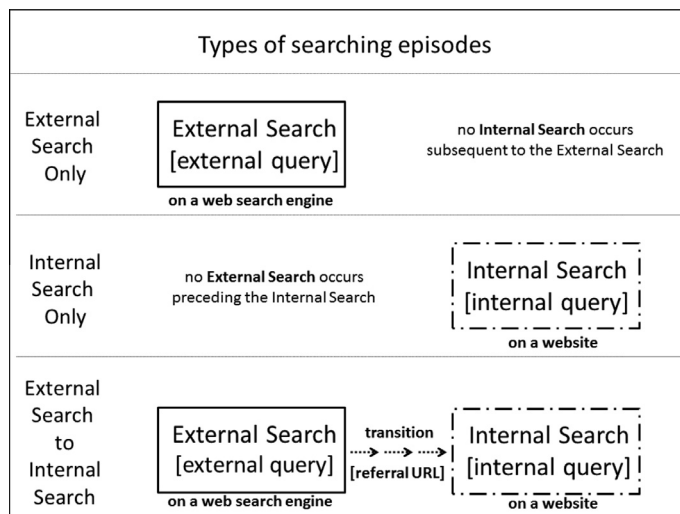

Fig. 1. Comparison of external search (ExS), internal search (InS), and external to internal search (Ex2InS).