



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

Article history:

Received 3 October 2014

Received in revised form 12 June 2015

Accepted 15 June 2015

Keywords:

Web searching

Information searching

Music searching

Music queries

Ecommerce searching

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

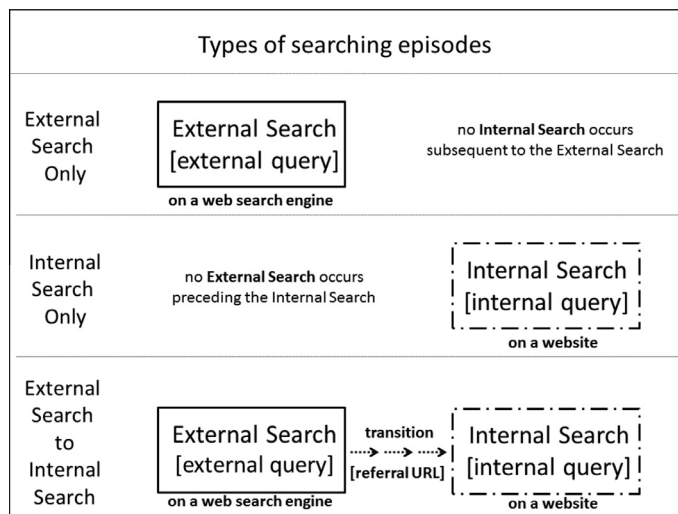


Fig. 1. Comparison of external search (ExS), internal search (InS), and external to internal search (Ex2InS).

information on the internal site, to which the user returns and conducts another InS search. So, there are many possible combinations. However, in this research we are concerned with one such pattern, Ex2InS.

In presenting our research, we begin with a literature review before discussing our research objectives. We follow with our data collection and methodology. We then present our results along with discussion. We end with implications for online businesses and directions for future possible research.

The following are definitions of the key terminology used in this research.

- *Organic traffic*: visitors referred to a website by a major search engine based on clicking on a link in the algorithmic listings on the search engine results page
- *Landing page*: the webpage that a user is directed to after clicking on a link in the results listing on the search engine results page
- *Bounce rate*: the percentage of one-page visits where users arrive and then immediately leaves without taking further action
- *Time-on site*: the duration of a visit to the site
- *Referral keyword (a.k.a., referral query)*: the term(s) that the user typed into the search engine in order to generate the search engine results page from which the searcher clicked on a URL that brought that visitor to the landing page.

2. Literature review

Prior research has either not differentiated or has not investigated the linkage between ExS and InS. In fact, much prior research examining searching on a major search engine (Jansen & Spink, 2005) and searching on specific sites (Wang et al., 2003) has considered them distinct and failed to establish the possible linkage. Our research differs from prior work in three respects. First, we formally distinguish the notions of ExS and InS search sessions. Second, we conceptually incorporate the notions of ExS and InS sessions into an integrated Ex2InS search episode. Third, we specifically investigate the transition inherent in Ex2InS.

We review prior work in the three research domains, which are searching on web search engines, searching on websites, and use of referral URLs. We review each of these bodies of research to understand the linkage between the ExS and InS aspects of this research.

2.1. Searching on web search engines

Researchers have studied web searching in the aggregate, at the session level, at the query level, and at the term level, along with associated behaviors, such as query reformulation and result selection. As with our approach, much of this research has utilized search logs.

At the aggregate level, Jansen, Spink, and Saracevic (2000) performed a comprehensive analysis of searching on the Excite search engine. By mining search logs, the researchers showed that search sessions have few queries and that the queries tend to have few terms. In addition to providing insights, mining web logs can have fruitful implications, and the results can be used for improving a search engine. For example, Web search queries can be used as a form of implicit intent or feedback for the searcher's intended action (Clark et al., 2012). Mining web search logs can also provide insights into the searcher's goals (Jansen, 2006).

Not all people search with the same intent in mind. Previous researchers (Broder, 2002) classified web search queries into three categories: *informational*, *navigational*, and *transactional*. Similarly, Jansen, Booth, and Spink (2008) automatically classified queries into the categories of *informational*, *navigational*, and *transactional* determining that approximately 25% of queries have multiple intents. In related research, Jansen, Booth, and Spink (2009) algorithmically classified the query patterns of web search sessions and found that *reformulation* (i.e., changing a term in the query) is the most popular modification type, followed by *specification* (i.e., changing the query to narrow the scope) and *generalization* (i.e., changing the query to widen the scope). Beitzel, Jensen, Chowdhury, Grossman, and Frieder (2004) classified web queries into 16 different topical categories, reporting that the most frequent category was *other* (16% of the dataset).

Previous studies have used the concept of states to map the sequence of user-system interactions (Choo, Detlor, & Turnbull, 1998; Ylikoski, 2005). These researchers have typically identified user actions on an information searching system and then classified these actions into states (Penniman, 1975). With this approach, one then can build a state map or matrix of possible moves. Each pattern is a sequence of state changes. This use of states and transitions is a stochastic process from which one can compare patterns of various lengths to test the significance (i.e., to determine what length of pattern predicts arrival at a certain state).

Users' actions on web search engines have also been identified and classified into states in order to provide insights into searcher interaction patterns (Jansen et al., 2009). By classifying users' actions as states, a state matrix can be formulated to map out possible future states based on the previous states. For example, Belkin, Cool, Stein, and Thiel (1995) proposed 16 different information seeking strategies that users employed within an information session. There have also been several research studies that create models of sequences of queries. These models have been built by formulating query sequence graphs (Boldi et al., 2008; Deng, King, & Lyu, 2009).

Such stochastic processes are effective methods for analyzing user searching patterns. Penniman (1975), for example, used this method to examine search–response patterns on a bibliographic database system. Chapman (1981) used this method to compare groups of searchers based on group characteristics by constructing zero- through fourth-order models for each participant group and by statistically testing for intergroup differences. Marchionini (1989) used the state transition approach to investigate the searching behavior of children using an electronic encyclopedia.

Chen and Cooper (2002) conducted state-transition analysis, defining a state as a certain address of the viewed page, clustering users into six groups based on patterns of the states (Chen & Cooper, 2001). Using six clusters of usage patterns, Chen and Cooper (2001) showed that there were statistical differences among the groups. In related research, Chen and Cooper (2002) used 126,925 sessions from an online library system and modeled access patterns using Markov models. In that study, the researchers found that a third-order Markov model explained the majority of the user clusters; they reported that a third-order model described five of the groups, and a fourth-order model described the remaining one cluster.

Su, Yang, and Zhang (2000) investigated n-order models utilizing path profiles of users from a Web log to predict future page requests. Shen, Dumais, and Horvitz (2005) used a Markov model to make inferences of searcher topic interests by using visited uniform resource locators (URLs). Chen, LaPaugh, and Singh (2002) employed the user's history and frequency of access to predict future page requests. Jansen et al. (2009) used a query as state method to map query reformulation on a web search engine.

Synthesizing the outcomes of this line of research, user sessions contain a small number of queries (typical one to three queries). User queries are typically short (one to three terms). Queries are mostly of informational intent; although, there are also substantial percentages of transactional and navigational queries, as well as some queries with mixed intent. The topics that users search for are extremely variable, even when categorized. The number of results that a user clicks on per query is limited.

2.2. Searching on Websites

There have been few studies examining the interactions during site searching, and these studies have mostly been conducted on academic, library, or government websites. Hert and Marchionini (1998) studied the information seeking behavior of users on statistical websites. Chi, Pirolli, and Pitkow (2001) investigated information scent for analyzing website usage. Wang et al. (2003) analyzed a university website transaction log containing four years of data. Katz and Byrne (2003) conducted users studies to determine when a visitor to a site would engage with the site search application versus site browsing. Chau, Fang, and Sheng (2005) studied a state governmental website, finding that users behave similarly when using a general-purpose and a website search engine in terms of the average number of terms per query and the average number of result page views per session. Kruschwitz, Lungley, Albakour, and Song (2013) studied university site logs, reporting that users prefer past users' query suggestions, as opposed to content-generated items. Huntington, Nicholas, and Jamali (2007) studied the search engine log files of a news website, noting that the effectiveness of search engine functionality could be evaluated using the number of searches in a session. Prior work has documents that the range of topics of site search (Wang et al., 2003) is narrower than that of general web searching (Jansen & Spink, 2005), and one would expect this finding given that individual websites are usually focused on a single topic or narrow range of topics.

Concerning other user searching behaviors, previous research (Chau et al., 2005) on what we refer to as InS has found close parallels with ExS on web search engines (i.e., short sessions, short queries). The number of results viewed on InS seems somewhat larger relative to ExS (Bucklin & Sismeiro, 2003). This may be because of the richer and more focused content, which more closely aligns with the user's searching intent.

2.3. Referral links

There have also been research studies examining referral links (i.e., the URL on a search engine result that takes a searcher to a website once the searcher has clicked on the link in the results listing). When a user visits a major search engine (e.g., Baidu, Google, NAVER, Yandex), submits a query to that search engine, and clicks on a link presented on the search engine results page (SERP), the search engine typically forwards information (called the referral URL) to the website pointed to by the link on which the user clicked. One item that the referral URL can contain is the query that the user submitted, which is known as the referral keyword. A typical format for a referral URL containing a referral keyword on the Google search engine for the query *music*, would be: <http://www.google.com/search?&q=music>. If there is no search term, webmasters can use approaches to infer what was the search term.

There have been a number of studies leveraging referral URLs. Stolz, Barth, Viermetz, and Wilde (2006) examined referral queries to determine users' intent for webpages. Downey, Dumais, Liebling, and Horvitz (2008) studied 80 million URL visits, discovering that searchers who have a rare search goal tend to submit a common query to the search engine and then arrive at a page that contains the information, which appears to be a more effective approach than simply submitting a rare query. This finding is similar to the orienteering approach described in Teevan, Alvarado, Ackerman, and Karger (2004).

Several research studies have investigated browsing behavior after the searcher visits a search engine. White and Drucker (2007) analyzed the navigation trails of 2527 participants, finding two classes of users, which are the users who tend to navigate away from the SERP and those who tend to have an exploratory behavior by submitting many queries and browsing numerous sites. Other studies have also identified navigation strategies according to browsing activity. Catledge and

Pitkow (1995) recorded the browsing activity of 107 users and identified three navigation strategies. Ortiz-Cordova and Jansen (2012) used 712,643 web search queries linking referral keywords to customer behaviors on the site. The researchers used k-means clustering to segment the customers based on their search queries and associated behaviors in order to identify the revenue generation potential of each segment.

2.4. Summary of prior work

From a review of literature, there has been considerable research concerning searching on web search engines. This domain of research has explored a wide variety of facets, but it has been limited to behavior occurring mainly on search engines. There has been much less research concerning searching on websites, and the existing research has typically treated this site search as distinct from the searching actions on the search engine, even though it was most likely the actions on the search engine that brought the visitor to the website. In other words, prior work assumes the searcher just appeared at the site, does not investigate how the searcher arrived there, and does not consider whether or not the process of arrival impacts subsequent searching on the site. Finally, there has been limited research on the study of referral queries and site search, as the research that does exist primarily focuses on the site-to-site browsing aspects of the user.

Further, there has been no research that we could locate linking these three domains into an integrated view of ExS2InS episodes; therefore, there are several unanswered questions about how ExS and InS are related. Primary among these is, “Does the searching intent change from Ex2InS?” Similarly, “Do the user’s searching tactics change from Ex2InS?” Also, “Are there patterns in the InS queries based on the ExS referral keyword?” These are some of the questions that motivate our research.

3. Research objectives

The research goal is to identify the relationships between an ExS session and subsequent InS session. We link these two searching sessions into an integrated *external search to internal search (ExS2InS) episode* using the referral keyword contained in the referral link. This referral keyword is the last query submitted by the visitor to the search engine before arriving at a website.

We address the following research objectives:

Research Objective 1: *Classify the intent of external search–internal search episodes using the external search query and subsequent internal search queries.*

To investigate research objective one, we develop and implement a system on a major ecommerce website to record the ExS referral query and subsequent InS queries of a searcher. We develop the search logging system via a script that collects both the ExS referral keywords and subsequent InS queries. We then implement another script to automatically classify both the ExS and InS queries into categories based on keywords and a taxonomy specific to the music domain (e.g., Pachet & Cazaly, 2000). From the results of this classification, we are able to discern insights into searcher intent during integrated Ex2InS episodes by classifying each query as a searching state.

Using the results from this classification of queries into searching states within ExS2InS episodes permits us to investigate research objective 2.

Research Objective 2: *Develop a predictive model for internal search based on the external search referral query.*

For research objective two, we use an n-gram methodology to determine the searching state patterns for each ExS2InS episode (e.g., *external query, internal query 01, internal query 02, ... internal query n*). We use the resulting patterns of this n-gram analysis to formulate a probability transition matrix for each ExS2InS episode. Using this transition matrix, we predict the probability of the next InS state for the aggregate data set.

Our research design is discussed in the following section.

4. Research design

4.1. Data collection site

The ecommerce site we use for data collection is www.BuenaMusica.com (BuenaMusica), a popular Spanish-based entertainment website. At the time of the study, BuenaMusica offers its visitors the ability to play songs on demand, watch music videos, view song lyrics, look up artist information and biographies, check the music-industry latest news, communicate in chat rooms, create their own profile page, and listen to streaming radio (see www.BuenaMusica.com for a complete range of site features). BuenaMusica displays advertising, via Google AdSense, and the business is supported by the revenue it generates from these ads. See Fig. 2 for an illustration of the BuenaMusic website at the time of the study.

Fig. 2. BuenaMusica.com homepage at the time of the study.

Concerning web traffic at the time of the study, Alexa.com, a website traffic reporting company, assigned BuenaMusica.com a worldwide traffic rank of 12,824, meaning it receives thousands of visitors per day. At the time of the study, 70.6% of all of BuenaMusica's traffic was referred by major search engines; 25.0% was direct traffic, and 4.3% was referred by other websites. These traffic percentages are typical of many online commercial sites (Blair, 2002; Butlion, 2013). The site is particularly popular in South America, where it is one of the top sites in fourteen different South American countries.

At the time of the study, the Google search engine had indexed 281,000 pages of the BuenaMusica.com domain. Since an InS application is needed with such a large content collection, BuenaMusica has a custom developed InS engine with a search box located at the top right hand corner of every page on the site. A user can submit a query and retrieve results of music elements (e.g., songs, videos, artists) that are hosted within BuenaMusica's extensive content collection. The query suggestion tool offers query suggestions based on the characters entered in the site search box, as shown in Fig. 3.

In summary, BuenaMusica is a popular ecommerce website with tens of thousands of daily visitors. It gets a substantial amount of its traffic from the organic traffic of web search engines. The content collection is nearly 300,000 entries, so a site search is necessary. Given of these attributes, BuenaMusica was a good ecommerce site to use for our study.

There has been an increasing amount of studies examining music searching (Sunghun, 2014). Itoh (2000) reported that subject terms, medium of performance, and genre were most frequently used in initial queries as well as in secondary ones to refine initial queries. Downie and Cunningham (2002) reported that there as information goals become rarer, search sessions become longer. Bainbridge, Cunningham, and Downie (2003) showed that the majority of music queries contain some form of metadata, including the name of the artist or title of the work. Lee and Downie (2004) reported that people use meta-data when searching for music, and they also rely on reviews, ratings, and recommendations. Bainbridge, Dewsnip, and Witten (2005) showed that algorithms using both the pitch and rhythm information of the melodies has better search retrieval performance, with an improvement in both recall and precision.

4.2. Data collection

We collect data over the course of a five-month period, from March 23, 2012 to August 26, 2012. For data collection, we develop a server-side transaction logging system that gathers data including the ExS query that brought the searcher to the

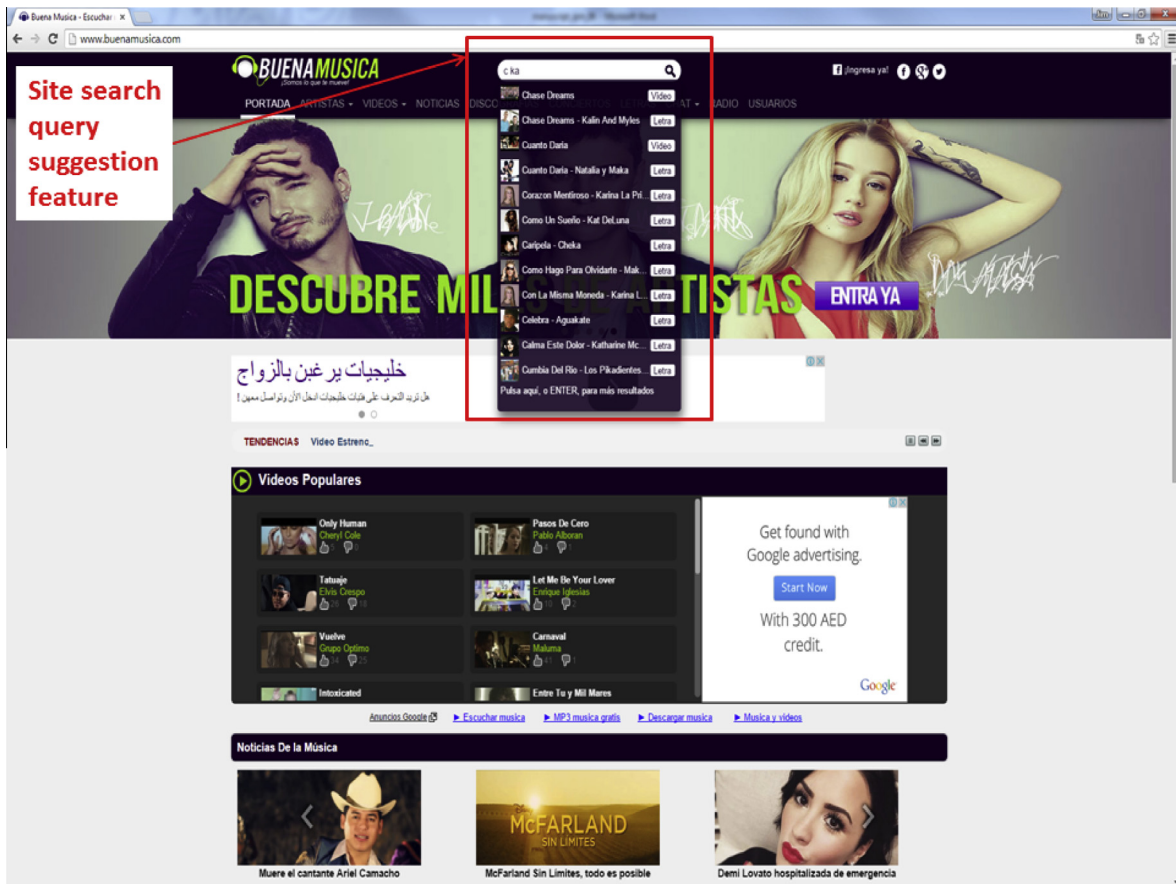


Fig. 3. Site search feature offering query suggestions based on the text entered in the site search box.

site and the queries submitted during InS, storing this data in a relational database. We implement this tracking system alongside the site's core source code files so that it runs every time a browser requests a page.

Since the search logging system ran with every BuenaMusica page view, we could identify new sessions that were referred from an external search engine. When this happens, the logging system extracts the external query from the referral URL, and the script stores the referral keywords from the ExS query in the appropriate database table. In addition, the tracking system also records the subsequent InS queries that were performed in the same ExS2Ins episode, linking the InS queries to the ExS query via the unique session identifier.

We associate the ExS query, InS queries, and session identifier using a foreign key constraint and also generate a unique time stamp for the session. Lastly, the browsed URLs of the pages that the searchers viewed in the session were also recorded.

Before analysis, we had to clean our data to collect only the Ex2InS episodes, as we collected millions of sessions that originated from bots and crawlers. General web searching research has typically used an interaction cutoff for differentiating between human and automated submissions; however, given more complete web analytics data (Jansen, 2009), we can accurately eliminate the non-human traffic and simultaneously identify the sessions of interest for this research.

Instead of an interaction or temporal cutoff, we include only sessions from an ExS that also perform an InS, and the InS queries that were performed were from sessions that originated with an ExS. This allows us to directly link the ExS queries to the InS queries of the same session, while simultaneously eliminating all searches originating from bots and InS from direct traffic. This data cleaning results in Ex2InS episodes (i.e., an ExS followed by an InS).

The entire dataset collected over the five-month period averaged about 2000+ daily sessions that originated from search engines where the visitor subsequently performed an InS. For business commercial confidentiality, we cannot disclose the exact website traffic. However, the 2000+ InS sessions is consistent with industry recognized metrics that 10% of a website's visitors will conduct internal site searches (Kaushik, 2007), though this percentage varies by type of site.

For the data analysis, we had to discard a small number of ExS queries and InS queries that were too ambiguous to be classified. So, our research study utilizes the 295,571 ExS queries (90.2% of ExS queries) and 840,819 InS queries (93.8% of total InS queries) for a total of 1,136,390 queries, both ExS and InS.

5. Research method

Our data analysis approach was to leverage the concept of a query-as-a-state. We then map movements between states to a query transition matrix. By definition, the rows of the matrix are a series of state transitions or changes that are carried out by the searcher over a designated period (i.e., an Ex2InS searching episode). Using this state matrix, a referral query from an ExS and associated website search queries from InS can be labeled and classified into states in order to build a state transition map. By linking these queries from both ExS and InS sessions, one can identify the most frequent patterns as the searcher is referred by a major search engine during ExS and engages in InS. These patterns can then be used to develop predictive patterns for InS on the website (i.e., predict what the searcher is looking for on the website based on the ExS query).

5.1. Taxonomy for classification

Different features can be used for music classification, including reference features such as title and composer; content-based acoustic features such as tonality, pitch, and beat, symbolic features extracted from the scores; and text-based features extracted from the song lyrics. Based on prior work (Dua, Gupta, Choudhury, & Bali, 2011; Inskip, MacFarlane, & Rafferty, 2010; Sheno, 2013), we leverage open coding and the cataloging attributes of the site's content collection to develop a taxonomy of twelve categories for classifying both the ExS and the InS queries from a manual coding of 10,000 InS queries (Ortiz-Cordova & Jansen, 2013, 2014). We then validate our codes with music context experts from BuenaMusica. The derived coding scheme and characteristics are presented in Table 1.

We then leverage this coding scheme to automatically classify the entire dataset. We began the classification by leveraging BuenaMusica's extensive database of songs, artists, and videos in order to look for exact matches, for which we did not apply any normalization. We iterate through each ExS and InS query and perform an exact match comparison against values within BuenaMusica's database, which is classified by music elements (i.e., artists, videos, songs, genres). If the expression evaluated to true, then our script labeled the query with the appropriate classification, and the query was also flagged so that it is not reclassified again. For example, a term such as *shakira* would be labeled as (A) for an artist. A query term such as *gasolina* would be labeled as (C) since it is a song. Although there is probability that a query could be classified into more than one category, given our use of exact match criteria against the database and the fairly clearly defined jargon of the music domain, the possibility of a query receiving more than one classification is small. A manual evaluation of 500 queries located no queries that were candidates for dual labeling. We also recruited two independent reviewers to manually code 500 randomly selected queries. The intra-rater agreement based on Cohen's kappa between the two raters was 0.82, which is quite high and further validates our coding approach. The major disagreements between the raters occurred with classifying queries as using *broad terms* versus some possible more specific but still broad query classification (e.g., *navigational*, *artist with additional terms*, *song name with additional terms*, etc.).

After performing these exact match classifications, we then enumerate an array list of words that were of similar type of content (i.e., stemming, synonyms, common misspelling) derived from a manual evaluation of the unclassified queries using a modified snowball technique. Many of these array terms are based on prior research presented in Jansen et al. (2008). The script automatically iterates through the arrays in a hierarchical order: *transactional*, *navigational*, and *informational*.

Transactional: queries are looking for resources that require another step to be useful. The **navigational** queries are looking for a specific website. The **informational** queries are meant to obtain data or information in order to address an information need, desire, or curiosity.

Identification of transactional queries was primarily via term and content analysis, with identification of key terms related to the transactional domain of e-commerce. Transactional searching queries containing terms related to broad collections of songs and lyrics. Transactional queries are also those with 'obtaining' terms concerning lyrics or music (e.g., free music). Transactional queries could also contain 'download' terms (e.g. download, get) or 'interact' terms (e.g. buy, chat, etc.).

Table 1

Coding scheme and characteristics used to classify external search (ExS) and internal search (InS) Queries.

Classification code	Classification	Query example	Query example (translated from Spanish)
A	Appearance of the term "artist" or an artist name, with no additional terms	Shakira, pitbull	Shakira, pitbull
B	Use of broad terms	Videos, musica de,	Videos of, music of, lyrics, songs
C	Appearance of the term "song" or a song name	Amor confuso	Confusing love
G	Use of a music genre	Rock, salsa, hiphop	Rock, salsa, hiphop
I	Use of informational terms	Discografia, biografia,	Discography, biography, news, album
L	Appearance of the term "lyric"	Letra	Lyric of
N	Use of navigation terms	.com, http, www,	.com, http, www, buenamusic
Q	Appearance of the an artist or song name with additional terms	Usher more	Usher more
S	Use of social terms	Chat, perfil, amigo	Chat, profile, friend
T	Use of transactional terms	Bajar, escuchar, gratis	Download, listen, watch, free, search
V	Appearance of the term "video" or a video name	Video de romeo	Romeo video

Transactional queries could also be queries with video, songs, lyrics, images, and multimedia or compression file extensions (e.g., mpeg, zip, etc.).

Navigational queries typically contain the company, business, or organizational name. They are also queries containing domains suffixes. Examples would be www.buenamuscia.com or music.com. Some navigational queries were quite easy to identify, especially those queries containing portions of URLs or even complete URLs. We also classify the company and organizational name as navigation queries, assuming that the user intended to go to the company's website.

Informational searching with natural language terms or queries containing informational terms (e.g. list, playlist, etc.). With the generally clear characteristics of navigational and transactional queries, informational queries became the catchall by default.

For each of these classifications, we develop databases of key terms relating to each. We employ this database of key terms in our automatic classifier. Transaction terms such as *download*, *list*, *free*, or *watch* were in an array labeled Transactional. For example, a query such as *download music* is classified as a transaction (T) since the user is explicitly stating their intended transaction, "download". Navigation terms such as *.com*, *http*, or *www* were in an array labeled Navigational. Likewise, a term such as *biography of shakira* is classified as informational (I) since it can be deduced that the user is looking for biographical information about that artist. A term such as *music of shakira* is classified as broad (B). Additionally, if the query contains explicit music industry-oriented terms such as *video*, *song*, *artist*, or *lyrics* (along with synonyms), the query was labeled as such. For example, a query such as *cancion original de feliz cumpleta* (translation: *original happy birthday song*) is labeled as song (S) since it contains the term *song*. This labeling approach is similar to the one taken in Zhou, Cummins, Halvey, Lalmas, and Jose (2012), so the approach is validated by prior work.

We understand that some songs might have the terms *artist* or *video* in their actual name and doing this type of classification might incorrectly label some queries. In order to gauge how prevalent this was, we ran a query against BuenaMusica's database, which, as previously described, contains thousands of music elements. Our query yielded no results with song names that had a *video* string in them. There was only one song with the string *artist* in the song name; therefore, it appears our labeling method was highly appropriate.

Lastly, for any queries that were not yet classified, we perform a reverse match query against BuenaMusica's collections of music elements. Our script iterates through all artist and song names in the database and compares each one against the unlabeled queries. Any queries that have terms that included a song or artist name are labeled as (Q). For example, say we retrieve an artist such as "Shakira" from the BuenaMusica artist database and ran it against the queries from our searching episodes. An InS query such as *shakira addicted to you* is classified as (Q) since the string contains an artist name, Shakira, in it along with additional terms. We repeat this process for songs and classified any queries that matched a song name (Q). Essentially, (Q) means that the query contained a long-tailed string that matched an artist or song name anywhere in the string.

With this process, we classify more than ninety-eight percent of Ex2InS episodes in the database. The remaining did not match any of our classification properties (~2%). In total, we successful classified 295,571 Ex2InS episodes.

This analysis allows us to address research objective 1.

5.2. N-gram analysis

Once we had each query classified, we then address research objective 2. With classifying the query types as states and using the n-grams approach, we had a systemic way of identifying the relationship between the ExS queries and the InS queries belonging to the same Ex2InS episode. The n-grams allow us to do this matching and to also build predictive models for user's InS sessions.

N-grams is a stochastic model that mathematically describes the sequence of states through which searchers progress via transition probability matrices. The value in each cell of a transition probability matrix is the probability of going from the row state to the corresponding column state; therefore, a transition probability matrix describes and predicts a pattern of movements through the state space represented by the matrix. Conceptually, for this research, the probability matrix is a map of user query types beginning with an ExS query type and then the sequence of InS query types. An analysis of user behavior in this manner not only describes a search state (i.e., the particular query classification) at a given point in the sequence, but it also predicts which states are most likely to follow one another (i.e., what particular query classification state will come next).

With our probability transition matrix, we could calculate the state-transition pattern for each episode (i.e., the number of states and transitions for a given episode) and a hash table (Hs) for the entire dataset (i.e., all Ex2InS episode patterns). We use the derived n-grams to develop the probabilistic patterns. N-gram patterns are a probabilistic modeling approach used for predicting the next item in a sequence and are $(n - 1)$ order Markov models, where n is the gram (i.e., subsequence or pattern) from the complete sequence or pattern. An n-gram model predicts state x_i using states x_{i-1} , x_{i-2} , x_{i-3} , ..., x_{i-n} ($n - 1$). The probabilistic model is then presented as: $P(x_i | x_{i-1}, x_{i-2}, x_{i-3}, \dots, x_{i-n})$, with the assumption that the next state depends only on the last $n - 1$ state, which is, again, an $(n - 1)$ order Markov model.

For example, in a given session, a searcher might use a broad term query (B) during ExS at a major search engine and then use the site search engine to submit a series of InS queries for a specific artist (A) and then a specific song (C). That Ex2InS episode would have an n-gram of BAC (i.e., broad → artist → song). By taking this n-gram pattern approach, we can link the ExS queries and the InS queries, classify each query by type, build a state transition matrix, and identify the overall searching

Table 2

Example of session log with queries for search episodes denoted by states to form query state patterns.

Ex2InS episode	Ex2InS query state pattern
1	ABC
2	ABCDE
3	ABCDE
4	A
5	AE
6	AC

Table 3

Example of n-gram hash table showing likely next state and prediction accuracy.

Given the predictive pattern	The next likely state is	With a prediction accuracy of (%)
AB	C	100
BC	D	66
CD	E	100

pattern for the external-internal search episode. By combining all Ex2InS patterns in the dataset, we can probabilistically predict the next state given the current state.

For an aggregate example, consider a search log of Ex2InS patterns consisting of six individual records, each of which represents a user's Ex2InS containing state query patterns, as shown in Table 2. Each state (A, B, C, D, or E) represents a query.

Using these six sessions in the log, a second order model (i.e., two states to predict the third) would generate the hash table in Table 3.

For this example, Tables 2 and 3 show that the accuracy for prediction varies from 66% to 100%.

Using this approach, we create a table consisting of the record identifications and the corresponding Ex2InS episode states. We use this table to conduct our analysis of how the query type changes as the searcher shifts from ExS and engages in InS. We export this table from our database as a CSV file and then import the data into SPSS to conduct the analysis.

We discuss the results next, with preliminary results from this research presented in Ortiz-Cordova and Jansen (2013,2014).

6. Results

6.1. Descriptive results

Before addressing our research questions, we present descriptive results from our analysis of the Ex2InS data set.

Among the 295,531 ExS queries, 27,870 (94.29%) originate from Google with 13,546 (4.58%) coming from Bing, and 3318 (1.12%) coming from Yahoo!.

We examined the query length of the ExS queries; the results are shown in Table 4.

Most ExS queries were 1, 2, or 3 terms in length. So, it does not appear that the ExS queries are unique from the overall set of web queries. It also reinforces that findings from this research are potentially generalizable to other Ex2InS contexts, as the user searching behavior is in line with prior works investigating user searching (Baeza-Yates & Ribeiro-Neto, 1999).

We examined where the ExS queries directed searchers upon their arrival at the website. Of the 295,571 ExS queries, 169,292 (57%) led to the site's main page as the landing page, while 126,279 (43%) ExS queries referred to some other page

Table 4

External search (ExS) queries length with occurrences and percentage (Average = 2.763).

Query length	Occurrences	Percentage (%)
1	71,954	24.34
2	114,810	38.84
3	48,703	16.48
4	24,780	8.38
5	14,560	4.93
6	9143	3.09
7	5268	1.78
8	2853	0.97
9	1577	0.53
10	866	0.29
>10	1057	0.36
	295,571	100

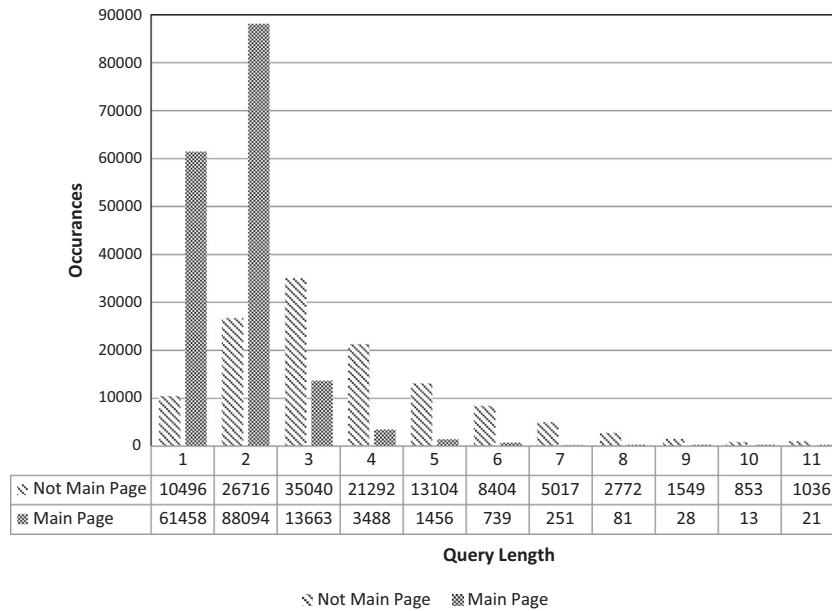


Fig. 4. External search (ExS) queries by query length referring to main page or not main page of site.

Table 5

Internal search (InS) queries length with occurrences and percentage, (Average = 2.249).

Query length	Occurrences	Percentage (%)
1	296,998	35.32
2	299,262	35.59
3	116,381	13.84
4	60,628	7.21
5	33,847	4.03
6	17,358	2.06
7	8545	1.02
8	4208	0.50
9	1758	0.21
10	896	0.11
>10	930	0.11
	840,819	100.00

on the website as the landing page. Generally, as the query length increased, the occurrences of landing on some page other than the main page increased, as shown in Fig. 4.

Of our 295,571 Ex2InS episodes with each containing an ExS query, 47.25% contain one InS query, 22.92% contain two InS queries, 10.76% contain 3 InS queries, and 19.04% contain 4 or more InS queries. So, the usage engagement of site search generally followed that of general web searching (Jansen & Spink, 2005).

Concerning the 840,819 InS queries, 473,898 (56%) were formulated by the searcher, while 366,921 (43%) took advantage of the query service on the InS site, which offers suggested queries to the searcher.

From Table 5, similar to the ExS queries, the InS queries were also composed of 1, 2, or 3 terms. Given that search habits are relatively stable (Jansen & Spink, 2005; Wang et al., 2003), this commonality would be generally expected.

We investigated whether or not the use of the query suggestion tool differed from natural formulated queries; the results are reported in Table 6.

As shown in Table 6, the queries based on the query suggestion tool were shorter than those queries formulated naturally by the searcher. However, a Spearman's rank correlation test indicated a strong, positive monotonic correlation between the naturally formulated queries and ones formulated with the use of the suggestion tool ($p < 0.01$), with a rho value of 0.99.

6.2. Site searching strategies

Addressing research objective 1, we use SPSS to calculate the frequencies of the classification of the referral query from the ExS and the first InS search query (e.g., TA, TB, BA, NL, etc.). Based on the trend analysis, we identify six searching patterns as the searcher shifted from ExS to InS based on a frequency table.

Table 6

Internal search (InS) queries length with occurrences and percentage for no use and use of the query suggestion tool, (Average = 2.759 and 1.590, respectively).

Query length	Occurrence no assistance	Percentage no-assistance (%)	Occurrence query suggestion	Percentage query suggestion (%)
1	102,025	21.53	194,973	53.14
2	161,863	34.16	137,399	37.45
3	89,902	18.97	26,479	7.22
4	54,213	11.44	6415	1.75
5	32,379	6.83	1468	0.40
6	17,185	3.63	173	0.05
7	8535	1.80	10	~0.00
8	4205	0.89	3	~0.00
9	1757	0.37	1	~0.00
10	896	0.19	0	0.00
>10	930	0.20	0	0.00
	473,890	100.00	366,921	100.00

Table 7

Summary of external-internal searching (Ex2InS) patterns from search engine to site.

Searching pattern classifications	Search engine used ...	Internal site search used ...	Searching example (External → Internal)	Percentage of entire dataset (%)
Explorers	As an informational tool	As an information tool (discovery)	videos → pitbull	43.1
Navigators	As a navigation tool	As an information tool (specific)	*.com → daddy yankee	15.0
Acquirers	As a navigation tool	As a transactional tool	download music → daddy yankee	13.6
Shifters	Varied	Varied	pop music → mexican music	15.3
Persisters	For same purpose as site search	For same purpose as search engine	adele → romeo santos	7.4
Orienteers	As an information tool	As an information tool (information about something else)	ramon ayala discografia → adele	5.6

Explorers (43.1%) – This searching pattern consists of submitting a broad ExS query to a major search engine and then submitting multiple, different types of InS queries on the site search engine. For example, ExS queries might be *good music* or *music* and the InS queries entail a wide array of specific songs, artists, or genres. These searchers use the major search engine as a navigation tool to get to a content collection website versus a specific website. Once on the content website, their intent is generally exploratory.

Persisters (7.4%) – This searching pattern consists of submitting the same type of query during both ExS and InS. For example, an ExS query could be about a music genre, and the InS query could also be about the same genre. These same-value combinations, such as AA, GG, or CC each were less than 1% each, but combined, they represent 7.4% of the total traffic.

Navigators (15.0%) – This Ex2InS searching pattern consists of submitting an ExS query that includes URL navigation terms such as *.com* or *http*. These users use the site search engine as a direct navigation tool during InS. Their intent is navigational to get to a particular site, and they had varied content needs once on the web site.

Acquirers (13.6%) – This searching pattern consists of submitting a specific ExS query with a clear intent to perform a transaction such as *download music* or *listen to music*; these searchers then submit subsequent transactional queries during InS. For example, queries on the major search engine might entail *download free music*, and the site search engine keywords would entail specific artists or song names. These users use the major search engine as a navigation tool during ExS to locate a transactional service, and they use the site search engine as an information or transactional tool to locate the specific content in order to execute their goal during InS.

Shifters (15.3%) – This searching pattern consists of submitting one type of query to the major search engine during ExS and then submitting a query of a different, seemingly unrelated category, on the site search engine during InS. For example, ExS queries might entail a specific music genre or song lyrics, and the InS queries are about another music genre. The searching intent of these users varies, but it could be designated as fluctuating, as the content that they desire widely varies. Shifters pattern classification is composed of search combinations that are of less than 1% each. So, shifters do not adhere to specific patterns; however, in total, they comprise 15.3% of all sessions.

Orienteers (5.6%) – This searching pattern consists of submitting an ExS query that specifically looks for artist information, such as an artist biography, and then submitting InS queries that contain an artist or song name. For example, ExS queries entail an artist's biography or discography, such as *ramon ayala discography*; meanwhile, the InS queries entail queries of an artist name such *adele*. These users use the site as an information tool since they first explore the music site to determine what is there before performing queries for a particular artist. Their intent is informational, and the specific content they desire is information about specific artists.

The major searching patterns are presented in [Table 7](#).

The different columns of [Table 7](#) are discussed below.

- *Searching Classification* – the names assigned to each different set of search tactics (i.e., deduced based on the change in focus between the referral query and the first onsite search query)

- *Search Engine used* ... – how the user employs the major search engine at the start of the searching episode (i.e., based on the referral query)
- *Internal search used* ... – how the user employs the site search at the start of the searching episode (i.e., the first site search query)
- *Searching Example (External → Internal)* – example of an Ex2InS sequence
- *Percentage Overall* – the percentage that the strategy comprises of the entire dataset (295,271 classified sessions)

Table 7 was formulated in the following manner. First, we group the external and internal search category patterns that had combinations greater than 1%. This comprises 77.3% of the dataset and yielded the *Explorers*, *Navigators*, *Acquirers* and *Orienteers* searching patterns. Then, we took the combinations that were of the same values (i.e., TT, CC, BB, etc.) regardless of percentage amount and allocate them as the *Persisters* pattern. Lastly, the *Shifters* pattern consists of the remaining pattern combinations that were less than 1%. The *Shifters* pattern has no individual search pattern that was greater than 1%; however, the total set of *Shifters* patterns represented 15.3% of the searches.

6.3. State transition matrix

Concerning research objective 2, we formulate a state transition matrix using subsequent site searching. Instead of just looking at the first InS state, using this matrix, we map out the InS queries to the 2nd, 3rd, and 4th order by searching pattern classification.

Table 8 shows the top four strategies and their corresponding top search episodes up to the 4th state (i.e., InS query classification). The first column is the ExS pattern classification. The second column, State 1, is the first InS query classification corresponding to the specified searching pattern. The third column, State 2, is the InS classification given State 1, and so forth for State 3 and State 4.

For Table 8, each table cell the percentage value is the frequency of the InS query classification for that state within the pattern type. From this, we develop an n-gram of Ex2InS patterns. For example, using Table 8, if a searcher is classified as an

Table 8

Top search episode conditional probability table for each strategy.

State 0 (external query type) (% in pattern & data set at state 0)	Site State 1 (internal query type) (% in pattern at state 1)	Site State 2 (internal query type) (% in pattern at state 2)	Site State 3 (internal query type) (% in pattern at state 3)	Site State 4 (internal query type) (% in pattern at state 4)
B (Explorers) (43.1%)	A (40.67%)	A (38.61%) Q (11.56%) C (11.56%)	A (40.71%) Q (30.66%) A (24.58%)	A (47.11%) Q (35.33%) A (35.09%)
	Q (36.29%)	Q (35.09%) A (13.04%) C (3.87%)	Q (36.81%) A (36.78%) C (26.05%)	Q (45.26%) A (40.28%) C (32.07%)
	C (9.80%)	C (29.44%) Q (15.13%) A (11.22%)	C (36.08%) Q (31.74%) A (32.64%)	C (45.20%) Q (34.36%) A (44.01%)
I (Orienteers) (5.6%)	Q (40.26%)	Q (32.42%) A (8.89%) C (1.68%)	Q (36.40%) A (33.72%) Q (29.16%)	Q (46.17%) A (35.93%) Q (23.80%)
	A (37.68%)	A (34.38%) Q (8.52%) I (2.27%)	A (38.68%) A (35.38%) A (35.71%)	A (46.70%) A (35.26%) A (16.92%)
N (Navigators) (15.0%)	A (42.73%)	A (38.50%) Q (11.92%) C (3.24%)	A (41.66%) A (30.65%) C (23.33%)	A (48.37%) A (39.76%) C (35.44%)
	Q (36.04%)	Q (34.96%) A (14.92%) C (4.78%)	Q (37.98%) A (34.62%) C (26.60%)	Q (44.13%) A (43.23%) C (33.48%)
	C (12.01%)	C (31.87%) Q (14.65%) A (12.33%)	C (36.40%) Q (32.01%) A (33.88%)	C (44.41%) Q (33.81%) A (33.87%)
T (Acquirers) (13.6%)	Q (41.09%)	Q (38.11%) A (8.02%) C (2.23%)	Q (37.13%) A (34.75%) C (28.76%)	Q (47.86%) A (38.08%) Q (35.15%)
	A (34.78%)	A (39.81%) Q (9.67%) C (1.59%)	A (38.25%) Q (33.49%) A (24.53%)	A (46.23%) Q (38.16%) A (40.90%)
	C (7.07%)	C (32.80%) Q (13.83%) A (7.95%)	C (35.82%) Q (37.63%) A (33.08%)	C (48.50%) Q (28.65%) A (43.33%)

Note: Shifters not listed as each pattern was 1% or less.

Bold indicates most common occurrence.

Table 9

Top search episode maximum likelihood probability table for each strategy.

ExS query type State 0 (%data set at state 0)	InS query type State 1 (% in data set at state 1)	InS query type State 2 (% in data set at state 2)	InS query type State 3 (% in data set at state 3)	InS query type State 4 (% in data set at state 4)	
B (Explorers) (43.1%)	A (18.86%)	A (7.28%)	A (2.96%)	A (1.39%)	
		Q (2.18%)	Q (0.66%)	Q (0.23%)	
		C (2.18%)	A (0.12%)	A (0.04%)	
	Q (16.83%)	Q (5.90%)	Q (2.17%)	Q (0.98%)	
		A (2.19%)	A (0.80%)	A (0.32%)	
		C (0.65%)	C (0.16%)	C (0.05%)	
	C (4.54%)	C (1.33%)	C (0.48%)	C (0.21%)	
		Q (0.68%)	Q (0.21%)	Q (0.07%)	
		A (0.51%)	A (0.16%)	A (0.03%)	
I (Orienteers) (5.6%)	Q (2.88%)	Q (0.93%)	Q (0.34%)	Q (0.15%)	
		A (0.25%)	A (0.08%)	A (0.03%)	
		C (0.04%)	Q (0.01%)	Q (<0.01%)	
	A (2.7%)	A (0.92%)	A (0.35%)	A (0.16%)	
		Q (0.23%)	A (0.08%)	A (0.02%)	
		I (0.06%)	A (0.02%)	A (<0.01%)	
N (Navigators) (15.0%)	A (7.05%)	A (2.71%)	A (1.13%)	A (0.54%)	
		Q (0.84%)	A (0.25%)	A (0.10%)	
		C (0.22%)	C (0.05%)	C (0.01%)	
	Q (5.95%)	Q (2.08%)	Q (0.79%)	Q (0.34%)	
		A (0.88%)	A (0.30%)	A (0.13%)	
		C (0.28%)	C (0.13%)	C (0.02%)	
	C (1.98%)	C (0.63%)	C (0.23%)	C (0.10%)	
		Q (0.29%)	Q (0.09%)	Q (0.03%)	
		A (0.244%)	A (0.08%)	A (0.02%)	
	T (Acquirers) (13.6%)	Q (6.72%)	Q (2.56%)	Q (0.95%)	Q (0.45%)
			A (0.53%)	A (0.18%)	A (0.07%)
			C (0.15%)	C (0.04%)	Q (0.01%)
A (5.68%)		A (2.26%)	A (0.86%)	A (0.40%)	
		Q (0.55%)	Q (0.18%)	Q (0.07%)	
		C (0.09%)	A (0.02%)	A (<0.01%)	
C (1.15%)		C (0.37%)	C (0.13%)	C (0.06%)	
		Q (0.16%)	Q (0.06%)	Q (0.01%)	
		A (0.09%)	A (0.03%)	A (0.01%)	

Note: Shifters not listed as each pattern was 1% or less.

Explorer currently at state 0, then there is a 40.67% probability that the first InS query that searcher submits (i.e., state 1) will be A (an artist query on the site search engine), with an n-gram pattern of BA. If the searcher transitions to State 2, then there is 38.61% probability that it will be also be an A query with an n-gram pattern of BAA. If the searcher moves to State 3, there is a 30.66% probability that the query is of type Q), with an n-gram pattern of BAAQ.

Table 9 presents the same state transition information, but the percentage in each cell is for the overall 295,571 searching sessions. For example, a searching episode of BA represents 18.86% of the entire data set. In the interest of space, only the top three states for each pattern are listed.

An interesting trend observed from Tables 8 and 9 is that as the searcher engages on the 4th InS state, the query is nearly always directed toward finding a particular artist or song. We can see on the State 4 column, most of the values are A (artist), Q (artist or song with additional terms), or C (song); however, this is only if the searcher actually performs four site search queries, which occurred for 19% of all Ex2InS episodes. States 1 through 3 inclusive are extremely varied, as shown in Tables 8 and 9.

Interestingly, InS sessions of four or more queries happens quite frequently (56,299 sessions, 19.0% percentage of all sessions). For three InS queries, 10.8% percentage of sessions (31,830 sessions), and 22.9%percentage were two InS queries sessions. The remaining 47.3% (139,688 sessions) were one InS query sessions.

These findings indicate that we can deduce different InS patterns based on ExS (0th state) and then the 1st through the 4th InS queries (1st, 2nd, 3rd, and 4th states); however, for subsequent InS queries in the same Ex2InS episode, there is a higher likelihood that the searcher will search for an artist or song no matter how they arrive at the site. Of course, queries in the State 4 column also do occur in other classifications, but it appears that A, Q, or C are the ones with the highest frequencies. This shows a common goal for these searchers, perhaps constrained by the music collection; although, the technique for locating said information varies by searcher.

7. Discussion and implications

As one of the first studies to investigate ExS on a web search engine and InS on a site as related searching sessions of the same searching Ex2InS episode, this research has several important implications.

7.1. Practical implications

In terms of practical implications, identifying the relationship between ExS and InS sessions and viewing these sessions as a complete Ex2InS episode can provide great insight into online searcher behavior and highlight potential business intelligence, as there is limited InS research that focuses on ecommerce sites, as this study does. For example, if an online business can describe and predict what a site visitor is looking using the ExS referral keyword, the business can provide better support to engage users in several ways.

First, online businesses can leverage the linkage between the ExS queries with InS queries that indicates the user intent of the site visitor. Based on this correlation, Ex2InS patterns can be detected and characterized. These research findings can have fruitful implications for online businesses. Using a state transition matrix, online businesses could identify the next state that the searcher is likely to transition to after performing an ExS or InS query and display an appropriate advertisement. By knowing the probability that a searcher is going to enter a higher order state beforehand, the business could personalize the site's content by changing aspects as the searcher enters the next state. For example, if the user is searching for songs or artists, the site could display a widget or menu that links to the artist's landing page or the songs sections page in order to show the searcher that the site indeed has that type of content. The business could also display a recommended search query (Teevan, Dumais, & Horvitz, 2010) beneath the search bar that says "users like you most commonly search for (artist name)" or "(song name)".

Second, the business could identify what topics and queries searchers execute during InS sessions and use this information to generate sponsored search keywords. The business could bid on these lower priced keywords as opposed to more expensive keywords. For example, an online business can bid on the InS keywords that belong to the *Explorers* or *Shifter* strategies; these InS keywords are more specific and have a lower price than the ExS keywords used to find the website. By doing this, the business could potentially decrease its advertising costs. According to the Google AdWords Traffic Estimator, a query such as *bajar musica* [translates to *download music*, classified as (T)] had an average cost-per-click of \$4.93 at the time of the study. In contrast, a query of an artist name such as *shakira* [classified as (A)] had an average cost-per-click of \$0.82. In the case of BuenaMusica, the users who execute a *Shifter* strategy tend to search for an artist or song during InS after searching using a transactional query during ExS. So, if BuenaMusica knows consumers will most likely search for a song or an artist, then the business can directly target potential customers who are searching for these terms at the major search engine, before they even come to the site.

Lastly, the business can aggregate queries beyond State 4 and compare these terms to its content collection in order to proactively refresh its collection ahead of popularized media stories (i.e., use the InS as a crowdsourcing business intelligence). If the searcher is still looking for an artist at State 4, it could be that the artist or music element is not yet posted on the site. The online business could leverage this information and add the artist and its corresponding songs or videos to increase its overall collection. Over time, adding new content could improve search engine ranking, provide better engagement, and increase site traffic.

To demonstrate the real potential of the approach presented in this research, BuenaMusica used these lengthy sessions following the findings of this research. BuenaMusica employees now monitor the InS queries log and compare these queries to the content within business's database. If that content (i.e., artist, song, or video) does not exist in the database, it is procured and subsequently uploaded/posted. So, the InS has become an indicator of shifting interest for the website customers, which translates into a competitive advantage for an online business.

This advantage is important because large music sites such as BuenaMusica have difficulty identifying and ranking which music elements to purchase and upload and how to allocate limited resources for collection procurement. By using the InS queries as indications, it can give priority to the type of content that it purchases. This makes financial sense for the business because now BuenaMusica can pay royalties to music labels for songs that users are actually expressing an interest. It can also get a jump on its competition by purchasing and posting desired content earlier.

These implications are potentially transferable to other types of online businesses that are oriented toward entertainment and content on demand. In addition, an online ecommerce business can also employ the methods described in this research to identify the path that customers are likely to take to locate a product. Online business can identify what kind of products or services the searchers are likely to search for and use this information for possible future growth and/or expansion. Additionally, a pattern analysis, as outlined in this research, could provide insights for other, similar, websites.

7.2. Theoretical implications

The research results also shed light on a more complex and nuanced aspect of web searching involving multiple content platforms relative to prior work that studied web searching (i.e., ExS) and site searching (i.e., InS) as separate sessions. It has been documented through observation and surveys that users many times visit multiple sources in order to address their searching need (Choo et al., 1998). The conceptual linking of ExS and InS, as presented in this research, acknowledges that people use multiple sources when attempting to address a task that requires searching. Therefore, there are several theoretical implications of this research.

First, there is an obvious linkage, as demonstrated in this research, between an ExS session on a web search engine and a subsequent InS session on a site. Exploration of this linkage has been limited in previous research. Based on our n-gram analysis, these ExS and InS sessions are part of the same Ex2InS episode, and these Ex2InS episodes occur quite frequently. The

five-month dataset used in this research averaged 2000 + daily Ex2InS episodes or about 10% of daily site traffic. Prior research has not examined the ExS session and the InS session as a part of a coherent search episode. Research examining the integration of ExS and InS sessions as a combined Ex2InS episode can lead to better insights into overall searching behavior on the web. As such, Ex2InS linkage is an important and novel research area, deserving of further study.

Second, there are apparent approaches to how users conduct InS that can be identified from referral queries from ExS. In this research, we identify six overall strategies (*Explorers, Navigators, Acquirers, Shifters, Persisters, and Orienteers*). Although it is yet to be determined if these patterns hold for other Ex2InS episodes on other sites and domains, the indications based on outline prior research are that they will hold. The informational, navigational, and transactional classifications that appear in this research have also appeared in prior work that focused on external web search (Broder, 2002; Jansen et al., 2008). Additionally, other search patterns similar to what we found have appeared in prior work (Jansen et al., 2009; Kellar, Watters, & Shepherd, 2007; Teevan et al., 2004). For example, Teevan et al. (2004) report that searchers take incremental steps in searching, which fits well with our research on the linkage of ExS and InS. Kellar et al. (2007) investigate web searching from Fact Finding, Information Gathering, Browsing, and Transactions paradigms, reporting that each have unique searching characteristics, supporting our approaching of grouping searching patterns. Jansen et al. (2009) report that there are predictable patterns in how searchers re-formulate queries. Based on similarities with these and other prior work, we believe that our general findings concerning Ex2InS patterns will be transferable to other domains.

Third, predicting the next searcher action in a sequence can be beneficial for web search systems attempting to provide useful content to users' requests. Predicting a user's action cannot only aid in the website's long term planning and strategy formulation, but it can assist in real-time website personalization. Using search data and implementing the queries as states approach, as used in this research, can model the search using an n-gram pattern as the searchers engage in InS after an ExS, effectively linking ExS sessions with InS sessions into a combined Ex2InS episode. The searcher's Ex2InS sequences can be used to create a state transition matrix and, using the resulting matrix, can predict the next possible query state. As we see in the state transition matrix presented from this research (Tables 8 and 9), the user's intent typically changes as the user transitions from ExS to InS. Being able to predict the searcher's intent as the searcher engages in InS and moves into higher InS states can be beneficial for the online business in order to select real-time targeted advertisements or to display InS query recommendations. Naturally, the specific query classifications may change from site-to-site depending on the searching environment; however, with the more limited content of an InS, our research shows that the internal site search codes are manageable.

8. Limitations and strengths

This research has limitations and strengths. Concerning limitations, our classification taxonomy may not directly extend to other non-music online businesses due to the specific lexicon of the music industry. Nevertheless, a particular online business can classify queries according to their industry lexicon using the process used in this research. So, the methodology is transferrable. Also, based on prior work mentioned above and the research presented here, we believe the general searching trends will be similar across websites. Another possible limitation is that the data we collected belongs to only one business, so the patterns we uncovered might not be directly generalizable to other businesses. However, based on the overlap between our research findings and prior work mentioned, we believe many of the patterns will be similar. Finally, there is the issue that we addressed only one possible Ex-InS pattern type, Ex2InS; however, as we discussed earlier, there may be different possible patterns as the searcher transitions among content sites. However, our analysis of the Ex2InS pattern contributes to the body of work in the Ex-InS literature.

In terms of strengths, this research uses real-world data from a site that is popular among several different countries, gets the majority of its traffic referred by major search engines, and has a site search engine box located on every page. The data was collected over an extended period of time using a server-side script and did not require a browser plug-in or toolbars tools, so our data collection method functions regardless of users' device, browser, or operating system. Furthermore, our classification method is thorough and leveraged BuenaMusica's wide-ranging database of music elements to cross-check the classified search queries. We expect our findings to be directly transferable to other music websites and many of our findings to be transferrable to other online businesses. The query state prediction matrix has a high degree of naturalistic applicability and generalizability, since it uses search episodes of users searching for information in a real-world, lab-free environment. Also, our practical implications of linking ExS queries to InS queries are generalizable to many types of online businesses regardless of what industry the website might belong.

9. Conclusion

The objective of this research was to classify ExS session queries and InS session queries within the same Ex2InS episode to deduce possible overall searching strategies and to investigate the linkage between these searching sessions. We classify 295,571 Ex2InS episodes from an online music business collected over a five-month period. Queries from these Ex2InS sessions were classified into 12 different states to develop n-grams of Ex2InS patterns. By analyzing the n-grams, we deduce Ex2InS patterns. These Ex2InS patterns range from *Explorers* who submit a broad ExS query on the major search engine and then submit different types of InS queries on the site search engine to the *Persisters* who submit the same type of query

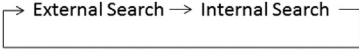
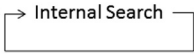
External - Internal Searching Patterns		
Code	Pattern of External - Internal Search	Description
Ex2InS	External Search → Internal Search	Linear external to internal search
(Ex2InS) _n		Repeated external to internal searches
Ex2In _n S	External Search → 	External to multiple internal searches
Ex2In2ExS	External Search → Internal Search → External Search	Linear external to internal to external search
In2Ex2InS	Internal Search → External Search → Internal Search	Linear internal to external to internal search

Fig. 5. Examples of major external – internal searching patterns with code and description of each.

during both ExS and InS. We then develop predictive patterns for Ex2InS episodes. Implications for this research include gaining better insight on creating enhanced InS capabilities, sponsored search keyword generation, and target advertisement selection, as well as the identification of possible new content.

For future research, similar studies can be conducted on a different type of websites to see if there are similarities among patterns and to confirm if these Ex2InS patterns apply to a wide range of online business websites. It would also be interesting to compare the InS patterns of users who visit the website directly. It would likewise be valuable to implement some of the recommendations in this paper, based on the discussion above, and analyze the effect of adding new content to the site based on InS queries to gauge the effect it has on future searchers and traffic ranking. Finally, there is the concept of investigating other external to internal searching patterns, as highlighted in Fig. 5.

The analysis presented in this research focuses on the Ex2InS pattern. However, as shown in Fig. 5, there may be other possible patterns. The searcher may start with an InS first and then move to an ExS. There may be multiple integrations of the ExS and InS, possibly involving multiple external sites. Investigations into these varied searching patterns would all be beneficial avenues for future research.

Acknowledgements

We would like to thank www.BuenaMusica.com for allowing us to implement our tracker and to have access to their database. We appreciate their willingness to collaborate on academic research. We also thank the two anonymous reviewers for many, lengthy, and detailed suggestions, they the recommendations substantially improved both the content and presentation of the research presented in this manuscript.

References

- Andreas, J., & Pascal, J. (2013). Forecasting the pulse: How deviations from regular patterns in online data can identify offline phenomena. *Internet Research*, 23(5), 589–607.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Indexing and searching. In *Modern information retrieval*, pp. 191–228. Addison Wesley.
- Bainbridge, D., Dewsnip, M., & Witten, I. H. (2005). Searching digital music libraries. *Information Processing & Management*, 41(1), 41–56.
- Bainbridge, D., Cunningham, S. J., & Downie, S. J. (2003). In how people describe their music information needs: A grounded theory analysis of music queries. In *Paper presented at the 4th international conference on Music Information Retrieval (ISMIR 2003)*, Baltimore, Maryland.
- Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. In Sanderson, M., Järvelin, K., Allan, J., & Bruza, P. (Eds.), *Paper presented at the 27th annual international conference on research and development in information retrieval*, Sheffield, U.K (pp. 321–328).
- Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3), 379–395.
- Blair, D. C. (2002). The challenge of commercial document retrieval. Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. *Information Processing & Management*, 38(2), 273–291.
- Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: Model and applications. In *Paper presented at the 17th ACM conference on information and knowledge management*. Napa Valley, California, USA: ACM.
- Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2), 3–10.
- Bucklin, R. E., & Sismeiro, C. (2003). A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, 40(2), 249–267.
- Butlion, J. (2013). *8 Important stats gathered from analyzing over 18,000 small to medium ecommerce sites (Vol. 2015)*. KISSmetrics.

- Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6), 1065–1073.
- Chapman, J. (1981). A state transition analysis of online information seeking behavior. *Journal of the American Society for Information Science*, 32(5), 325–333.
- 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.
- Chen, H. M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11), 888–904.
- Chen, H. M., & Cooper, M. D. (2002). Stochastic modeling of usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*, 53(7), 536–548.
- Chen, M., LaPaugh, A. S., & Singh, J. P. (2002). Predicting category accesses for a user in a structured information space. In *Paper presented at the 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'02), Tampere, Finland* (pp. 65–72).
- Chi, E. H., Pirolli, P., & Pitkow, J. (2001). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a Web site. In T. Turner & G. Szwillus (Eds.), *Paper presented at the SIGCHI conference on human factors in computing systems* (pp. 161–168). The Hague, Netherlands: ACM.
- Choo, C., Detlor, B., & Turnbull, D. (1998). A behavioral model of information seeking on the Web: Preliminary results of a study of how managers and IT specialists use the Web. In *Paper presented at the 61st annual meeting of the American Society for Information Science* (pp. 290–302). Pittsburgh, PA: ASIS.
- Clark, M., Kim, Y., Kruschwitz, U., Song, D., Albakour, D., Dignum, S., et al. (2012). Automatically structuring domain knowledge from text: An overview of current research. *Information Processing & Management*, 48(3), 552–568.
- Deng, H., King, I., & Lyu, M. R. (2009). Entropy-biased models for query representation on the click graph. In *Paper presented at the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 339–346). Boston, MA, USA: ACM.
- Downey, D., Dumais, S., Liebling, D., & Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. In *Paper presented at the 17th ACM conference on information and knowledge management* (pp. 449–458). Napa Valley, California, USA: ACM.
- Downie, J. S. & Cunningham, S. J. (2002). Toward a theory of music information retrieval queries: System design implications. In *Paper presented at the third international conference on Music Information Retrieval (ISMIR 2002)* (pp. 13–17), Paris, France.
- Dua, N., Gupta, K., Choudhury, M., & Bali, K. (2011). Query completion without query logs for song search. In *Paper presented at the 20th international conference companion on World Wide Web (WWW2011)* (pp. 31–32). Hyderabad, India: ACM.
- Hert, C. A., & Marchionini, G. (1998). Information seeking behavior on statistical websites: Theoretical and design implications. In C. M. Preston (Ed.), *Paper presented at the 61st American Society for Information Science annual meeting* (Vol. 35, pp. 303–314). Medford, NJ: Information Today.
- Hoa, C.-I., Lin, M.-H., & Chen, H.-M. (2012). Web users' behavioural patterns of tourism information search: From online to offline. *Tourism Management*, 33(6), 1468–1482.
- Huntington, P., Nicholas, D., & Jamali, H. R. (2007). Employing log metrics to evaluate search behaviour and success: Case study BBC search engine. *Journal of Information Science*, 33(5), 584–597.
- Inskip, C., MacFarlane, A., & Rafferty, P. (2010). Organizing musics for movies. *Aslib Proceedings*, 62(4/5), 489–501.
- Itoh, M. (2000). Subject search for music: Quantitative analysis of access point selection. In *1st international symposium on music information retrieval, Amherst, MA*.
- 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. (2009). *Understanding user – Web interactions via Web analytics*. San Rafael, CA: Morgan-Claypool.
- Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management*, 44(3), 1251–1266.
- Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during Web searching. *Journal of the American Society for Information Science and Technology*, 60(7), 1358–1371.
- Jansen, B. J., & Spink, A. (2005). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & 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.
- Katz, M. A., & Byrne, M. D. (2003). Effects of scent and breadth on use of site-specific search on e-commerce Web sites. *ACM Transactions on Computer-Human Interaction*, 10(3), 198–220.
- Kaushik, A. (2007). *Kick butt with internal site search analytics*. Occam's Razor. Vol. 2014.
- Kellar, M., Watters, C., & Shepherd, M. (2007). A field study characterizing Web-based information seeking tasks. *Journal of the American Society for Information Science and Technology*, 58(7), 999–1018.
- Kruschwitz, U., Lungley, D., Albakour, M.-D., & Song, D. (2013). Deriving query suggestions for site search. *Journal of the American Society for Information Science and Technology*, 64(10), 1975–1994.
- Kumar, R., & Tomkins, A. (2010). A characterization of online browsing behavior. In M. Rappa & P. Jones (Eds.), *Paper presented at the 9th international conference on World Wide Web* (pp. 561–570). Raleigh, NC, USA: ACM.
- Lee, J. H., & Downie, J. S. (2004). Survey of music information needs, uses, and seeking behaviours: Preliminary findings. In *Paper presented at the fifth international conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain* (pp. 441–444).
- Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, 40(1), 54–66.
- Obendorf, H., Weinreich, H., Herder, E., & Mayer, M. (2007). Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Paper presented at the 25th ACM conference on human factors in computing systems (CHI2007)* (pp. 597–606). San Jose, CA, USA: ACM.
- Ortiz-Cordova, A., & Jansen, B. J. (2013). Site-searching strategies of searchers referred from search engines. In *Paper presented at the 76th annual meeting of the American Society for Information Science and Technology (ASIST 2013), Montreal, Canada*.
- Ortiz-Cordova, A., & Jansen, B. J. (2012). Classifying web search queries to identify high revenue generating customers. *Journal of the American Society for Information Science and Technology*, 63(7), 1426–1441.
- Ortiz-Cordova, A., & Jansen, B. J. (2014). Linking external and internal search: Investigating the site searching patterns of referred searchers. In A. Schmidt & T. Grossman (Eds.), *Paper presented at the CHI conference on human factors in computing systems (CHI 2014)* (pp. 1345–1350). Toronto, Canada: ACM.
- Pachet, F., & Cazaly, D. (2000). A taxonomy of musical genres. In *Paper presented at the 1st conference of content-based multimedia information access conference (RIA0)* (pp. 1238–1245). Paris, France: ACM.
- Penniman, W. D. (1975). A Stochastic process analysis of online user behavior. In C. W. Husbands & R. L. Tighe (Eds.), *Paper presented at the annual meeting of the American Society for Information Science* (pp. 147–148). Washington, DC: ASIS.
- Shen, X., Dumais, S., & Horvitz, E. (2005). analysis of topic dynamics in Web search. In *Paper presented at the fourteenth international World Wide Web conference (WWW2005), Chiba, Japan* (pp. 1102–1103).
- Shenoi, A. (2013). An essential guide to organising your music library. <<http://www.digitaldjtips.com/2013/05/MUSIC-LIBRARY-ORGANISATION-PART-4/>>. Retrieved 10.03.15.
- Stolz, C., Barth, M., Viermetz, M., & Wilde, K. D. (2006). Searchstrings revealing user intent: A better understanding of user perception. In *Paper presented at the 6th International Conference on Web Engineering (ICWE '06), Palo Alto, California, USA* (pp. 225–232).
- Su, Z., Yang, Q., & Zhang, H. -J. (2000). A Prediction system for multimedia pre-fetching in internet. In *Paper presented at the eighth ACM international conference on Multimedia 2000, Marina del Rey, California, USA* (pp. 3–11).
- Sunghun, C. (2014). An empirical analysis of usage dynamics in a mobile music app: Evidence from large-scale data. *Internet Research*, 24(4), 436–456.
- Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Paper presented at the IGCHI conference on human factors in computing systems (CHI2004)*. Vienna, Austria: ACM.

- Teevan, J., Dumais, S. T., & Horvitz, E. (2010). Potential for personalization. *ACM Transactions on Computer–Human Interaction*, 17(1), Article 4.
- Wang, P., Berry, M. W., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743–758.
- White, R. W., & Drucker, S. M. (2007). Investigating behavioral variability in web search. In *Paper presented at the 16th international conference on World Wide Web (WWW2007)* (pp. 21–30). Banff, Alberta, Canada: ACM.
- Ylikoski, T. (2005). A sequence analysis of consumers' online searches. *Internet Research*, 15(2), 181–194.
- Zhou, K., Cummins, R., Halvey, M., Lalmas, M., & Jose, J. M. (2012). Assessing and predicting vertical intent for web queries. In *Paper presented at the 34th European conference on advances in information retrieval* (pp. 499–502). Barcelona, Spain: Springer-Verlag.