Linking External and Internal Search: Investigating the Site Searching Patterns of Referred Searchers

Adan Ortiz-Cordova
College of Information Sciences and Technology
The Pennsylvania State University
contact@adanortiz-cordova.com

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
jjansen@acm.org

Abstract
In this research, we investigate the relationship between external search on a major search engine and the subsequent internal search on an individual web site. Insights in the relationship can be a competitive advantage for websites. We use 295,271 searching sessions of an online Spanish entertainment business collected over a five month period. We develop a classification scheme for external and internal search queries using the referral query as the starting point. Using an n-gram approach, we identify query patterns for 295,271 searching episodes. We aggregate and identify six searching patterns. The three major searching strategies are *Explorers* (47%, a broad query for external search and then multiple broad queries during internal search), *Navigators* (16%, a navigational query for external search and then specific queries during internal search), and *Acquirers* (15%, transaction queries for both external and internal search). The remaining three patterns are *Shifters* (12%), *Persisters* (7%), and *Orienteers* (3%). Identification of searching patterns and related content can be a competitive advantage for websites dependent on providing relevant, fresh, and locatable information.

Author Keywords
Web queries, Web searching, site-search, search strategies, query reformulation

ACM Classification Keywords
H.3.3 Information Search and Retrieval – *Search process*

Figure 1. Screenshot of a page from BuenaMusica.com. Note the search box in the upper right of the screen. This search box appears on every page.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
ACM 978-1-4503-2474-8/14/04.
http://dx.doi.org/10.1145/2559206.2581199

Introduction
Many websites rely on search engines (e.g., Google, Yandex, or Baidu) for most of their traffic. With the majority of web users using a search engine as a starting point, this traffic is critical for many websites as traffic from search engines provides a direct funnel

Background Terminology

Site Search – the use of a search engine typically built by the web domain or web host that allows the user to search for content only to that particular website

Organic Traffic – visits referred by a major search engine based on relevance listings rather than ads

Landing Page – the page that a user is directed to after clicking on a listing on the search engine results page

Bounce Rate – the percentage of one page visits (i.e., the user left the site from the landing page)

Time on Site – the duration of a visit to the site

Referral Keyword (query) – the terms that the user typed in the search engine

Sponsored search – targeted, relevance-based advertisements that are displayed alongside major search engine results (e.g., Google AdWords)

of potential customers, clients, and users. However, when a website has a large number of pages indexed by search engines, any page has the potential to become a landing page for visitors. If a user determines that the landing page does not contain the information he/she is looking for, the user may leave the site to look elsewhere for the information sought. A user "bouncing" from a landing page is obviously not good for the online business. For example, a bounce for an online business means the potential loss of a sale, a registration, or of advertising revenue.

One approach to combating a user bouncing from a landing page is to provide the user an internal search capability for the site so that users can quickly and efficiently find the information they are looking for and thereby remain on the site. However, there has been little research into the linkage between external search (the searching conducted by a user on a major search engine that brings the user to a site) and internal search (the subsequent searching conducted by the searcher on the site).

Our research motivation is based on the importance of understanding the linkage between external and internal search in order for websites to develop better internal search capabilities, provide relevant content to searchers, and compete effectively for visitors.

Background

The theoretical basis for this research is human information processing [5], specifically information searching. Intent and patterns of searching can vary. Previous research [1] has proposed three broad classifications for web search intent, *informational, navigational* and *transactional*. Jansen, Booth, and Spink [3] automatically classified queries into three categories *informational*, *navigational*, and *transactional* finding that approximately 25% of queries have multiple intents. Prior research [4] has also automatically identified query patterns during web search. Much of the prior web research has focused on behaviors on the major search engine, with research examining what the searchers did once they arrived at specific websites.

In fact, we could locate no research investigating the linkage between external and internal search. Perhaps due to the challenges of data collection, the searching session are viewed as separate. Given the association of searching need, we view external search and subsequent internal search as comprising a complete search episode, which is consistent with theories of information searching [2]. Understanding the linkage could provide valuable competitive, market, or business intelligence to website owners.

Research Objective

Our research objective is to ***classify external – internal searching patterns*** for website visitors referred from a major search engine and who subsequently utilize a site's searching capability.

We link the external search with the internal search using the referral query from the search engine. When a searcher performs a search on a search engine and then clicks on a link on the SERP, the search terms are passed to that website. In cases where the referral keyword is masked, techniques have been developed to identify the keyword based on a given probability. Queries executed on the internal search service are readily available to the website owner.

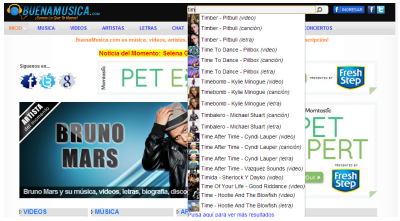

Figure 2. Screenshot of the Buenamuscia.com internal search with query suggestions.

More on BuenaMusica.com

BuenaMusica offers the ability to play songs on demand, watch music videos, view song lyrics, look up artist information such biographies, check the latest news, communicate in chat rooms, and streaming radio (www.BuenaMusica.com). The business is supported by the revenue from display advertising. At the time of the study, Google had indexed a total of 281,000 pages of the domain BuenaMusica.com. Alexa.com, a web site traffic reporting company, assigned BuenaMusica.com a worldwide traffic rank of 12,824. The site is popular in South America where it is a top music site in fourteen different countries.

To investigate the research objective, we use a search log from a popular online business website. The data in the search log was gathered using a custom server-side software tool.

Research Design

We first present our data collection site and method.

Data Collection Site

We collected data for this research from www.BuenaMusica.com (BuenaMusica), a popular Spanish-based entertainment website. BuenaMusica is particularly useful for this study because it has the characteristics common to many websites and where understanding the external to internal search would prove valuable. The majority of the users visiting the site are referred by a search engine (70.6%). This is typical of many online commercial sites. BuenaMusica also has a custom developed internal site search engine, which is used by about 10% of site visitors. This site search text box is located at the top right hand corner of every page (see Fig. 1 and 2). Both the percentage of the referral traffic and internal search usage are typical of many websites, so our research findings may be generalizable to other websites. Therefore, BuenaMusica served to be a good platform to collect data and conduct this research study.

Data Collection and Analysis

For data collection, we developed a search logging system that gathered key pieces of search data and saved them in a relational database. The data was saved in a relational database with four tables: *referral_keywords* (which stored the external search referral queries and user ids), *internal_keywords* (which stored the internal search queries and user id), *sessions* (which stored the session level information), and *URLs* (which stored the website pages viewed). These four tables were linked via a unique session ID.

Research Method

In order to discard bot sessions, which are a high percentage of traffic to most websites, we performed an inner join on the *referral_keywords* and *internal_keywords* tables by the *session_id* attribute. Joining these two tables in this manner allowed us to link the external search sessions to the subsequent internal search session. This inner join also assured that 1) users from external sessions performed an internal search and 2) the internal search sessions were from users who had conducted an external search. This operation resulted in 295,271 episodes (i.e., an external search followed by an internal search) composed of 295,271 external search queries and 896,410 internal search queries for a total of 1,191,681 queries.

We then classified the 1,191,681 queries into categories using the coding scheme and characteristics presented in Table 1, which we validated with music content professionals at BuenaMusica. We wrote a PHP script that leveraged BuenaMusica's extensive database of artists, songs, lyrics, videos, and genres and performed an exact match against the queries. If there was an exact match, the query was assigned to that particular classification from the BuenaMusica database. We then enumerated five arrays of terms that related to user intent, which we based on prior research. For example, navigation terms were *.com, http,* or *www*. Transaction terms such as *download, list, free,* or *watch* were in a different array.

Development of Search Logging System

We developed a tracking system using PHP in order to fully capture all of the user actions.

This tracking system was specifically developed to identify if a user was coming from a search engine. If that was the case, the script extracted the referral keyword, and stored it in the appropriate table of the database.

The PHP script also generated a unique session ID for each different session. The referral keyword and the session ID were linked using a foreign key constraint.

If an internal search was performed, the internal search query was linked with both the session ID and the respective referral keyword.

A unique time stamp was also included with each record at the time that it was saved to the database.

Lastly, the browsed URLs of the users were also recorded.

Code	Content	Query (translated from Spanish)	Actual Query
A	"artist" or Artist Name	*shakira, pitbull*	*shakira, pitbull*
B	Broad Terms	*videos of, music of, lyrics, songs*	*Videos, musica de*
C	"song" or a Song Name	*confusing love*	*amor confuso*
G	Genre	*rock, salsa, hiphop*	*rock, salsa, hiphop*
I	Informational Terms	*Discography, biography, news, album*	*Discografia, biografia*
L	"lyric"	*{lyric of}*	*{letra}*
N	Navigation Terms	*.com, http, www, buenamusica*	*.com, http, www,*
P	Artist Name and additional terms	*usher more*	*usher more*
Q	Song Name and additional terms	*we found love in a hopeless place*	*we found love in a hopeless place*
S	Social Terms	*chat, profile, friend*	*Chat, perfil, amigo*
T	Transaction Terms	*download, listen, watch, free, search*	*Bajar, escuchar, gratis*
V	"video" or a Video Name	*romeo video*	*video de romeo*

Table 1. Coding scheme and characteristics used to classify external and internal searching queries.

We iterated through the lists in the following order: transaction, navigation, broad, informational, and social. For example, a query such as *download music*

would be classified as a transaction (T). A term such as *music of shakira* would be classified as broad (B). A term such as *biography of shakira* would be classified as informational (I).

Lastly, for any queries that were not yet classified we performed a reverse match against BuenaMusica's database of artists and songs. We iterated through all artists' names in the database and if there was a wildcard match (i.e., an artist name was anywhere inside the query string) with a query then that query was classified as *P* (Artist Name and additional terms). For example, a query such as *shakira addicted to you* would be classified as *P* since the string contains an artist name in it. We then repeated this process for songs and classified any queries that matched a song name as *Q* (Song Name and additional terms).

Examples of queries from internal search are:

- *sexy back*
- *El buen ejemplo calibre50*
- *diomedito que pasa contigo*
- *descargar solamente tu pablo alboran*
- *guns n roses*
- *los temerarios*

The classification scheme resulted in twelve categories (see Table 1) and provided a systemic way of identifying the relationship between the external search queries to the internal search queries that belong to the same session using n-grams. For example, in a given episode, a user might use a broad term (B) in external search and then internal search to search for a specific artist (A) and then a specific song (C). That search

Table 2 Attribute Definitions

Strategies – the names assigned to each different set of search patterns

Action Plan – how the user employs the major search engine in external search and internal search on the site

Content Goal –inference on what the user's searching intent is for interacting with BuenaMusica's collections of data (i.e., discover new music, download or listen to music, look up artist information)

N-gram External – the classification code assigned to the referral query that belongs to that pattern

N-gram Internal – the classification code assigned to the site search queries that belongs to that pattern

Query Example: an example for the pattern

% - percentage that the pattern represents of the entire data set

Strategies	Action Plan Intent	Content Goals	N-gram combinations External | Internal		Query Example	%
Explorers	External as informational tool Internal varies	Discover music	B	All possible combinations	buena musica-> pitbull	47.0%
Navigators	External as navigation tool Interval varies	Look up music	N	Artist or song	*.com -> daddy yankee	15.8%
Acquirers	External as navigation tool to transactional site Internal as navigation tool	Download or acquire music	T	Artist or song	download music -> daddy yankee	15.0%
Shifters	Varies for both External and Internal	Varies	Varies	Different from external	pop music -> mexican music	12.5%
Persisters	Varies but same for External and Internal	Varies	Varies	Same as external	adele -> romeo santos	6.5%
Orienteers	External as informational tool Internal as an information tool	Look up information about an artist	I	Artist	ramon ayala discografia -> adele	3.1%

Table 2. Summary of External and Internal Searching Episodes

interaction would have a search episode n-gram of BAC (broad → artist → song). N-grams is a method for probabilistic modeling and widely used for predicting the next item in a sequence using a (n - 1) order

Markov model. N-grams have many inherent advantages in pattern processing. Aside from the model's simplicity, one can scale n-grams efficiently by simply increasing the order of n. One can use n-grams for both descriptive and predictive analysis. As such, n-grams was an appropriate methodological approach for our research.

Results

We were able to classify 91% of the sessions using our classification method. Based on the frequency table, we see that several patterns (summarized in Table 2) emerge that we now discuss.

Explorers **–** (47.0%) This pattern starts with a broad external search query and then multiple different queries during internal search. For example, queries on the major search engine were *good music* or *music* and the search queries entail a wide array of specific song, or artist names. These users use the major search engine as an informational tool. Their intent is exploratory, and the content that they desire is to listen to music, although the internal search intent is extremely varied.

Navigators – (15.8%) This pattern begins with an external search query that includes URL navigation terms such as *.com* or *http*. This pattern used the major search engine as a navigation tool. Their intent is navigational. Once at the site, the internal search intent was varied, but the content was to typically look up an artist or a song.

Top Combinations

Combination of External and 1st Internal	Percentage Of External Search Pattern
Explorers	
BA	40.7%
BC	36.3%
Navigators	
NA	42.7%
NQ	36.0%
Acquirers	
TQ	41.1%
TA	34.8%
Orienteers	
IQ	40.1%
IA	37.7%

The *Shifters* are composed of n-grams that are less than 1 percent values and different query types.

The *Persisters* are composed of combinations that have the same values (i.e., TT, CC, BB) grams that are less than 1 percent values and different query types.

Acquirers – (15.0%) This pattern begins with a specific external search with a clear transactional intent such as *download music* or *listen to music* and is followed by specific internal search queries for songs. For example, queries on the major search engine entail *download free music* and the internal search queries were specific artist or song names. These users use the major search engine as a navigation tool to locate a transactional service and use the internal search as a navigational service to get to the specific content. Their intent is transactional, and the content that they desire is to download music of a specific artist or acquire free content.

Shifters – (12.5%) This pattern begins with an external search and followed by an internal search of a completely different type. The intent varies and the content that they desire widely varies also. There were no Shifter n-grams that were greater than 1%.

Persisters – (6.5%) This pattern has the same type of external and internal search query. For example, n-grams for this pattern would be combinations such as AA, CC, BB, or TT. Again, each individual n-gram was general less than 1%.

Orienteers – (3.1%) This pattern entails an external search query specifically looking for artist information, such as an artist biography and then internal search queries containing an artist or song name. For example, external queries on the major search engine entail an artist biography or discography such as *ramon ayala discography*, and the internal search queries entail queries of an artist name such *adele*. These users use the internal search as an information tool, since they first explore the music site and then perform queries looking for a particular artist. Their intent is informational, and the content they desire is information about a specific artist.

Discussion and Implications

As one of the first studies to explore the linkage between external and internal searching episodes, our results highlight several important implications. First, we can tell the necessity for some type of internal search. The entire dataset taken over the five month period averaged about 2,000+ daily internal search sessions that originated from an external search. It is apparent that these are continuations of the same searching episode. Based on this analysis, for future work we will investigate the motivations of why searchers utilized an internal search service, including those that come from direct traffic.

References

[1] Broder, A. A Taxonomy of Web Search. *SIGIR Forum,* 36, 2 (2002), 3-10.

[2] Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J., "Using Information Scent to Model User Information Needs and Actions on the Web," in ACM CHI 2001 Conference on Human Factors in Computing Systems, Seattle, WA, 2001, pp. 490-497.

[3] Jansen, B. J., Booth, D., and Spink, A. (2008) *Determining the informational, navigational, and transactional intent of Web queries,* Information Processing & Management. 44(3), 1251-1266.

[4] Jansen, B. J., Booth, D. L., & Spink, A. (2009). *Patterns of query modification during Web searching*. Journal of the American Society for Information Science and Technology. 60(7), 1358-1371.

[5] Wilson, T.D. Human information behavior. *Informing Science,* 3, 2 (2000), 49-55.