Site-Searching Strategies of Searchers Referred from Search Engines

Adan Ortiz-Cordova
College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802 Email: avo5018@psu.edu

Bernard J. Jansen
College of Information Sciences and Technology, The Pennsylvania State University, University Park PA 16802 Email: jjansen@acm.org

ABSTRACT
In this research, we analyze the referral queries and associated site-search queries at the session level from searchers coming from web search engines. Findings are based on a random sample of 10,000 from a total of 327,261 searching sessions of an online Spanish entertainment business collected over the course of a five month period from March 23, 2012 to August 26, 2012. We find six searching strategies that are correlated with the type of referral keywords (i.e., search terms) used at the major search engine. Of the six, the three major searching strategies are (1) the *explorers* who submit a broad query on the major search engine and then submit multiple broad queries on the site-search engine, (2) the *navigators* who submit a query to the major search engine that is part of a URL and then submit specific queries to the site-search engine, and (3) the *persisters* who submit the exact type of query on both the search engine and the site search. Implications for this research include developing better internal searching features, sponsored search keyword generation, and personalization of website content.

Keywords
Web queries, Web searching, Site-search, search strategies, query reformulation.

INTRODUCTION
The majority of online business websites rely on search engines (i.e., Google, Baidu , Bing, or Yandex) for most of their traffic. This traffic is critical for many online businesses as organic traffic from search engines is free, continual, and provides a direct funnel of potential customers, clients, users, etc. When a website has a high number of pages indexed by the search engines, any indexed page that appears on the search results has the potential to become a landing page for visitors. The page that the user lands on is the de-facto homepage. 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 go back to the search engine results page (SERP) to look elsewhere for the information they require. A user "bouncing" from a landing page is obviously not good for the online business because it results in the loss of a potential sale, registration, sign-up, or advertising revenue.

One way to combat a user bouncing from the landing page is to provide the user an internal site search engine so that users can quickly and efficiently find the information they are looking for while remaining on the site. Therefore, it is important to uncover different site-search strategies that manifest depending on the referral keyword used at the major search engines so that these may be leveraged by online businesses to develop better internal search capabilities, identify what sponsored keywords to bid on, and personalize website content.

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

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)

BACKGROUND
The theoretical basis for this research is human information processing (Wilson, 2000). Not all searchers search with the same intent in mind. Previous researchers (Broder, 2002) have classified web search queries into three categories *informational*, *navigational*, and *transactional*. Similarly, Bernard J. Jansen, Booth, and Spink (2008) classified

queries into three categories *informational*, *navigational*, and *transactional* determining that approximately 25% of queries have multiple intents. Bernard J. Jansen, Booth, and Spink (2009) also automatically classified 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* and *generalization*.

Concerning site search, Avrahami, Yau, Si, and Callan (2006) explore the use of federated verticals for site search. Ortiz-Cordova and Jansen (2012) used user behaviors on a site to identify potential high revenue customers for an online business. Kruschwitz, Lungley, Albakour, and Song (2013) use site query log to investigate query reformulation.

Despite the research that has been done in analyzing and classifying user intent of search queries, there has been no research that we could locate linking the referral keywords from a major search engine and the queries submitted to a site-search engine in order to identify possible multi-site searching strategies. Although Espadas, Calero, and Piattini (2008) note the importance of this referral traffic to a website. Additionally, Chau, Fang, and Sheng (2005) report that site searching behaviors are similar to that conducted on general web searching. Understanding the type of searching strategies based on referral and site-search queries could provide valuable intelligence to an online business identifying different searching strategies based on referral and site-search queries.

RESEARCH OBJECTIVE

Our research objective is *to identify session level site-searching strategies of searchers referred from a major search engine.*

The motivation for this research is to identify relationships between searches performed on a major web search engine and the internal site-search searches in order to highlight potential website/business intelligence. This would allow better support to engage users based on their searching strategy. We link these two searching sessions using the referral keyword from the search engine. For example, if the referral keyword is a transactional keyword, do the site search queries tend to be also transactional? If the referral keyword is broad, do the onsite search queries tend to shift to specific songs or artists? Are there trends that stand out during a sequence of onsite search queries for a given strategy? Or, is there no relation between the referral keyword and the internal site-search query?

To investigate this research objective, we use a transaction log from a popular online business website. The data in the transaction log was gathered using a custom server-side software tool. It is important to posit that when we talk about a "user" or "visitor" we are actually referring to the browsing session of a searcher.

RESEARCH DESIGN

We first present our data collection site and method.

About BuenaMusica.com

We collected data from www.BuenaMusica.com (BuenaMusica), a popular Spanish-based entertainment website. The website offers its visitors 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 even listen to streaming radio (interested readers may visit www.BuenaMusica.com to take a look at all of the site's features). The site displays advertising via Google AdSense, and the business is supported by the revenue it generates from these ads. At the time of the study (November 2012), the Google search engine 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 particularly popular in South America where it is in the top 6,000 sites in fourteen different South American countries.

BuenaMusica is particularly useful in this study because it already contains features that we used to gather data required for our research study. The majority of the users visiting the site were referred to by a search engine. Specifically, at the time of the study, 70.57% of all traffic is referred by a major search engine. 25.17% is direct traffic, and 4.25% is referred by other websites. This is typical of many online commercial sites. Because of these features, BuenaMusica served to be a good platform to collect our data and conduct our research study. BuenaMusica has a custom developed internal site search engine. This site search text box is located at the top right hand corner of every page.

Figure 1. BuenaMusica.com Homepage

DATA COLLECTION AND ANALYSIS

For data collection, we developed a transaction logging system that gathered key pieces of information and saved them in a relational database. The data was saved in a relational database with four tables: *referral_keywords*, *internal_keywords*, *sessions* and *URLs*. These four tables were linked via a unique *session ID*.

Classification	Content	Query (translated from Spanish)	Query Example
G	Genre	*rock, salsa, hiphop*	*rock, salsa, hiphop*
T	Transaction terms	*download, listen, watch, free, search*	*bajar, escuchar, gratis*
N	Navigation terms	*.com, http, www*	*.com, http, www*
B	Broad terms	*videos of, music of, lyrics, songs*	*videos, musica de*
I	Informational terms	*discography, biography, news, album*	*discografia, biografia*
S	Social terms	*chat, profile, friend*	*chat, perfil, amigo*
L	The term "lyric"	*lyric of*	*letra*
A	The term "artist" or an artist name	*shakira, pitbull*	*shakira, pitbull*
C	The term "song" or a song name	*confusing love*	*amor confuso*
V	The term "video" or a video name	*romeo video*	*video de romeo*
Q	Strings that contain an artist or song name and additional terms	*usher more*	*usher more*

Table 1. Coding scheme and characteristics used to classify session queries.

Development of Tracking System

We developed a tracking system using PHP in order to fully capture all of these user actions. This tracking system was specifically developed to identify if a visitor 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 together using a foreign key constraint. If an internal site search was performed, the on-site search query was linked with both the session ID and the respective referral keyword. A unique time stamp was included with each record at the time that it was saved to the database. The browsed URLs of the visitors were also recorded.

RESEARCH METHOD

In order to discard bot or crawler sessions, we performed an inner join on the *referral_keywords* and *internal_keywords* tables by the *session_id* attribute. This inner join also assured two things: 1) sessions that were referred by a search engine actually performed a site-search and 2) the site-search queries that were performed were from sessions that originated or referred by a major search engine. Joining these two tables in this manner allowed us to directly link the referral search engine queries to the site-search queries of the same session. This resulted in 327,261 sessions (i.e., a search from a major search engine followed by one or more searches on the site–search) composed of 896,410 internal site search queries.

Using search log analysis (B. J. Jansen, Taksa, & Spink, 2008), we then classified the queries of a random sample of 10,000 sessions (27,781 internal site search queries) into twelve different categories using the coding scheme and characteristics presented in Table 1. Similar to (Slone, 2002), we based on classification on searching behaviors. Wolfram, Wang, and Zhang (2009) used search patterns to cluster users. We wrote a PHP script that leveraged BuenaMusica's extensive database of artists, songs, lyrics, videos, and genres and performed an exact match query against the queries. If there was an exact match, the query was assigned to that particular classification. Although there is the probability that a query could be classified into more than one category, the possibility is small given our use of exact match criteria against the database and the defined jargon of the domain (Bernard J. Jansen, Spink, & Pederson, 2003). We understand that some music elements (i.e., 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 SQL query against BuenaMusica's database which, as previously described, contains thousands of music elements. The SQL 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, we believe that it is appropriate to label queries this way since the chance of mislabeling is minimal.

We then enumerated five array-lists of words that were closely related based on content. For example, navigation terms such as *.com, http, or www*, were in an array and transaction terms such as *download, list, free,* or *watch* were in a different array.

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) since the user is explicitly stating their intended action. A term such as *music of shakira* would be classified as broad (B). Likewise, a term such as *biography of shakira* would be classified as informational (I) since it can be deduced that the user is looking for information about that artist.

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 with a query then that query was classified as (Q). For example, a query such as *shakira addicted to you* would be classified as (Q) since the string contains an artist name in it. We repeated this process for songs and classified any queries that matched a song name as (Q). Essentially, (Q) means that the query contained a long-tailed string that matched an artist or song name anywhere in the string.

The purpose of classifying the queries was so that we could have a systemic way of identifying the relationship between the referral queries to the site search queries that belong to the same session using n-grams. For example, in a given session a user might use for a broad term (B) in a major search engine and then use the site-search engine to search for a specific artist (A) and then a specific song (C). That search interaction would have a session search episode n-gram of BAC (broad → artist → song).

RESULTS

We were able to classify 91% of the sessions using our classification method using SPSS in order to get the frequencies of the classification of the referral query and first site-search query (e.g., TA, TB, BA, NL). Based on the frequency table, we see that several patterns (summarized in Table 2) emerge that we now discuss. We formulated Table 2 by grouping the referral code categories that had combinations greater than 1%. The shifters are made up of combinations that are less than 1%. The persisters are made up of combinations that have the same values (i.e., TT, CC, BB).

Explorers - Submit a broad query to the major search engine and then submit multiple different queries on the site-search engine. For example, queries on the major search engine were good music or music and the site-search queries entail a wide array of specific song, or artist names. These users use the major search engine as a navigation tool. Their intent is exploratory, and the content that they desire is to listen to music. However, 1.9% perform transactional queries on the site-search engine.

Orienteers - Submit a query to the major search engine that specifically looks for artist information such as an artist biography and then submit queries that contain an artist or song name to the site-search engine. For example, queries on the major search engine entail an artist biography or discography such as *ramon ayala discography*, and the site-search 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 and then perform queries looking for a particular artist. Their intent is informational, and the content they desire is information about a specific artist.

Navigators - These users arrive at the site after submitting a term that included URL navigation terms such as *.com* or *http*. These users use the search engine as a navigation tool. Their intent is navigational, and the content they desire is to typically look up an artist or a song.

Shifters - Submit a query of a particular type to the major search engine and then submit a completely different type of query on the site-search engine. For example, queries on the major search engine entail a specific music genre or song lyrics, and the queries on the site-search engine are about another music genre. Their intent varies but could be designated as exploratory, and the content that they desire widely varies also. Shifters are made up of search combinations are of less than 1% each. However, in total, they make up 12.5% of all classified sessions. So, shifters do not adhere to specific patterns.

Goal-Oriented - These users submit a specific query to the major search engine with a clear intent such as *download music* or *listen to music* and then submit specific queries on the site-search engine. For example, queries on the major search engine entail *download free music* and the site-

Strategies	Action Plan	Intent	Content	Referral Code	Site-Search Code	Query Example	Percentage
Explorers	MSE as navigation tool	Exploratory	Discover music	B	A, C, G, Q, T	buena musica → pitbull	47.0%
Orienteers	Site as an information tool	Informational	Look up information about a song or artist	I	A, Q	ramon ayala discografia → adele	3.1%
Navigators	MSE as navigation tool	Navigational	Look up music	N	A, C, Q	*.com → daddy yankee	15.8%
Shifters	MSE as navigation tool	Varies	Varies	Varies	Varies	pop music → mexican music	12.5%
Goal-Oriented	MSE as navigation tool Site as information tool	Transactional	Download or acquire music	T	A, C, Q	download music → daddy yankee	15.0%
Persisters	Varies	Varies	Varies	Same value combinations		adele → romeo santos	6.5%

Table 2. Summary of Site-Searching strategies. MSE stands for Major Search Engine (i.e., Google or Bing).

search engine keywords entail specific artist or song names. These users use the major search engine as a navigation tool to locate a transactional service and use the site search engine as an information tool. Their intent is transactional, and the content that they desire is to download music of a specific artist or acquire free content.

Persisters - These users submit the same type of query on both the major search engine and the site search engine. These same value combinations such as AA or CC each was less than 1% with the exception of BB and TT which had 1.9% and 1.1%, respectively.

DISCUSSION AND IMPLICATIONS

As one of the first studies to link referral queries and site-search queries from the same session, our results highlight several important implications. First and foremost, we can tell the necessity for some type of site-search engine. The entire dataset taken over the five month period averaged about 2,000+ daily sessions that originated from search engines that subsequently performed a site-search. Second, it is important for online businesses to set-up some sort of linkage between the referral queries and site-search queries since there is certainly a correlation between these two. Based on this correlation several site-searching strategies can be deduced and characterized.

In addition, the site can aggregate the site search queries and identify the current state that the searcher is on. If the searcher is searching for songs or artists, the business could immediately personalize the site and display a widget or menu that links to the artists or songs sections page in order to show the searcher that the site indeed has that type of content. Lastly, the business could identify what artists or songs are being searched for on the site search by users that belong to *explorers* or *goal-oriented* strategies. The business could use this information in order to generated sponsored search keywords and bid on these lower cost-per-click keywords as opposed to more expensive transactional keywords such as *download music*. Thereby, the business could potentially decrease its search engine marketing costs.

LIMITATIONS AND STRENGTHS

As it is the case with all research studies, this study has limitations and strengths. Our classification taxonomy may not extend to other non-music online businesses. However, all businesses are able to collected the same type of data and classify queries accordingly to their own industry lexicon.

In terms of strengths, this research study uses data form a real-world operating business. In addition, the data was collected over an extended period of time. Also, our classification method was manually assembled, thorough, and used BuenaMusica's wide-ranging database of music elements to cross-check the classified search queries. Lastly, our implications are applicable to many different types of online businesses regardless of what industry vertical they belong to.

For future research, a similar study could be conducted on all 327,261 searching sessions in order to identify if the same patterns have similar frequency distribution ratios. In addition, a possible research study that utilizes n-grams and attempts to map out and predict patterns as the users engage in site search could yield fruitful theoretical implications.

CONCLUSION

The objective of this research was to identify session level site-searching strategies of searchers referred from a major search engine. We classified a random sample of 10,000 sessions from a total of 327,261 searching sessions of an online entertainment business that were collected over a five month period. Queries were classified into 11 different categories. From an analysis of the frequencies we find six searching strategies that are correlated with the type of referral keywords (i.e., search terms) used at the major search engine. The six searching strategies range from the *explorers* who submit a broad query on the major search engine and then submit multiple broad queries on the site-search engine, to the *persisters* who submit the same type of query on both search engines. Implications for this research include developing better internal searching capabilities, identification of sponsored search keywords, and personalization of website content.

ACKNOWLEDGMENTS

We would like to thank www.BuenaMusica.com for allowing us to implement our data collection tool and have access to their database. We strongly persuade other high-ranking and high-traffic websites to actively look for ways to collaborate with the academic research community.

REFERENCES

Avrahami, T. T., Yau, L., Si, L., & Callan, J. (2006). The FedLemur project: Federated search in the real world. *Journal of the American Society for Information Science and Technology, 57*(3), 347–358.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3-10. doi: 10.1145/792550.792552

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.

Espadas, J., Calero, C., & Piattini, M. (2008). Web site visibility evaluation. *Journal of the American Society for Information Science and Technology, 59*(11), 1727–1742.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Inf. Process. Manage., 44*(3), 1251-1266. doi: 10.1016/j.ipm.2007.07.015

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. doi: 10.1002/asi.21071

Jansen, B. J., Spink, A., & Pederson, J. (2003). *An Analysis of Multimedia Searching on AltaVista.* Paper presented at the the 5th ACM SIG Multimedia International Workshop on Multimedia Information Retrieval, Berkeley, California.

Jansen, B. J., Taksa, I., & Spink, A. (2008). Research and methodological foundations of transaction log analysis. In B. J. Jansen, A. Spink & I. Taksa (Eds.), *Handbook of Research on Web Log Analysis* (pp. 1–17). Hershey, PA: IGI.

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, Early View*.

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.

Slone, D. J. (2002). The Influence of Mental Models and Goals on Search Patterns During Web Interaction. *Journal of the American Society for Information Science and Technology, 53*(13), 1152–1169.

Wilson, T. D. (2000). Human information behavior. *Informing Science, 3*(2), 49-56.

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology, 60*(5), 896–910.