THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

COLLLEGE OF INFORMATION SCIENCES AND TECHNOLOGY

CLASSIFYING WEB SEARCH QUERIES IN ORDER TO IDENTIFY HIGH-REVENUE GENERATING CUSTOMERS

ADAN ORTIZ-CORDOVA
SPRING 2013

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Information Sciences and Technology
with honors in Information Sciences and Technology

Reviewed and approved* by the following:

Jim Jansen
Associate Professor in Information Sciences and Technology
Thesis Supervisor

Dr. Xiaolong (Luke) Zhang
Associate Professor in Information Sciences and Technology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Traffic from search engines is important for most online businesses, with the majority of visitors to many websites being referred by search engines. Therefore, an understanding of this search engine traffic is critical to the success of these websites. Understanding search engine traffic means understanding the underlying intent of the query terms and the corresponding user behaviors of searchers submitting keywords. In this research, using 712,643 query keywords from a popular Spanish music website relying on contextual advertising as its business model, we use a k-means clustering algorithm to categorize the referral keywords with similar characteristics of onsite customer behavior, including attributes such as click through rate and revenue. We identified 6 clusters of consumer keywords. Clusters range from a large number of users who are low impact to a small number of high impact users. We demonstrate how online businesses can leverage this segmentation clustering approach to provide a more tailored consumer experience. Implications are that businesses can effectively segment customers to develop better business models to increase advertising conversion rates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Words alone cannot express my gratitude to all the wonderful people who have helped me in one way or another throughout my college and graduate school career. First and foremost, I would like give thanks to God and Our Lady of Guadalupe for allowing me to complete this degree. I would like to express my deepest thankfulness to my parents for all their hard work and sacrifices. I very much appreciate the support and love of family. *¡Gracias, mamá y papá!*

In addition, I would like to thank Dr. Jansen for the guidance and mentorship while I conducted this research. I would also like to thank Dr. Zhang, Dr. Reddy, Dr. Lenze, Dr. Fonseca, The College of Information Sciences and Technology, and The Schreyer Honors College.

Furthermore, I would like to thank the College Assistance Migrant Program for the academic advising during my freshman and sophomore years. I would like to thank the Latino Caucus and the Noche Latina 2011 committee for working with me while I was a candidate competing for Mr. Latino. Being selected as Mr. Latino Penn State 2011-2012 was a wonderful experience and a nice memory that I will cherish for the rest my life! I want to thank my friends that I have met here at Penn State. I wish you all well and look forward to seeing your continued success! I would also like to say thanks to Denise Bartolome for proofreading this and several other research papers.

Lastly, I want to express my gratefulness to the businesses, organizations, and benefactors who awarded me scholarships. I believe that I have received an excellent education here at The Pennsylvania State University. *We are, Penn State!*

Introduction

Major search engines such as Google, Bing, and Yahoo! use complex algorithms to determine the relevance of a page (Brin & Page, 1998). Websites that appear on the first page of the search results are likely to get more traffic because most users click on first page results (Jansen & Spink, 2004). These search engines not only drive new visitors, but research has shown that repeat visitors use search engines as navigational tools (Jansen, Spink, & Pedersen, 2005). With search engines being the primary point of entry to the web for many people, the traffic from search engines is vitally important to websites. For online businesses, a visitor to their website could mean a sale, ad revenue, user registration, or exposure to branding.

In the context of web searching, the set of terms for which a user searches is called the query. If a user enters a query and then clicks on a result, these query terms are embedded within the URL that is passed from the search engine to the website. This URL is called the referral URL, and the query terms within the referral URL are called the referral keywords. The webpage pointed to by the link the user clicks is called the landing page. Both the referral URL and referral keywords provide important information to the website owner. Examples of such information include where traffic is coming from (i.e., which search engine, for example), what topics searchers are most interested in, and how a particular landing page is indexed by the search engines. Therefore, it is important to understand and study the search keywords and search phrases that are bringing people from the search engines to the websites (Hackett & Parmanto, 2009). When analyzed appropriately, these referral keywords can provide insightful information about user behavior and user intent, from which website owners can build better business models or provide more relevant content to visitors.

Many websites measure the success of a visit by conversion rate, which is the ratio of visits that result in users performing the end goal, as defined by the website owner, divided by the number total visits Booth and Jansen, 2008. The end goal of a conversion varies depending on the type of website. For websites that sell products, a conversion would be one where a shopper turns into an actual buyer. For a website that relies on contextual advertising, a conversion is one that results in a click on an advertisement. Websites and online businesses are continually looking for ways to improve conversion rate, as it's reported that only about three percent of visits result in a conversion (Betts, 2001).

Contextual advertising is a successful business model for many websites in which they generate revenue by displaying ads that closely match the content of the site's pages (Broder, Fontoura, Josifovski, & Riedel, 2007). If a website owner can determine which types of referral keywords bring in high performing or low performing customers, based on conversion rates for example, the website owner can then optimize the landing pages of the website to increase conversion rates for these consumers through personalized content. This is an example of behavioral targeted advertising, where the advertisements are personalized for users based on their individual web search and browsing behaviors (Yan et al., 2009).

What if, by using the referral keyword, webmasters could predict the onsite behavior of potential consumers sent to the site from search engines? What if the website owner could tell which referral keywords are more likely to generate contextual advertising revenue and how much? What if the website owner could know which referral keywords do not make any revenue and somehow move those users off the site as quickly as possible? What if the website owner could know which referral keywords produce the highest bounce rates and enhance the site in a way that retains the visitor longer than one page view? These are some of the questions that motivate our research. To address these questions, we develop clusters of users based on their referral keywords and the associated behavioral characteristics and attributes on the website.

In the following sections, we first present a review of the literature. We then discuss our research objectives for clustering based on referral keywords. In the methodology section, we review the k-means clustering algorithm, along with the website and data used in this research. We then discuss our results and implications, including how these findings could be used by an online business to improve the consumer experience and the conversion rate on these websites.

Literature Review

The theoretical basis for this research is human information processing, which is the methods that people use to acquire, interpret, manipulate, store, retrieve, and classify information (Wilson, 2000). Web analytics is typically the method used for understanding human information processing on the Internet, as there is much user, searcher, and consumer information collected by logs and other means. The Web Analytics Association defines web analytics as the process of measuring, collecting, analyzing and reporting website usage to understand and optimize web usage (Burby, Brown, & WAA Standards Committee, 2007), and the methodological approach has been used in information science, by marketers, and by other researchers to study and gain greater insight into user information behavior (Penniman, 2008; Peters, 1993).

For this research, we are interested in a subset of human information processing, namely, information searching (Jansen & Rieh, 2010; Marchionini, 1995). We are specifically interested in the use of query terms on search engines as indicators of intent, as our assumption is that these query terms could be the basis for segmenting visitors (i.e., potential customers) to a website. Prior work would indicate that this is a valid assumption. For example, Broder (2002) proposed three broad user intent classifications—*navigational, informational, and transactional*—based on query terms. Using survey results, Broder reports that nearly 26% are navigational, approximately 73% of queries are informational or transactional, with an estimated 36% are transactional. (Note: The researcher placed some queries into multiple categories.) Then, based solely on log analysis, Border reports that 48% of the queries were informational, 20% navigational and 30% transactional. (Note: We assume the missing 2% were unclassifiable or the result of rounding.)

In similar work, Rose and Levinson (2004) classified search engine queries using the categories of *informational, navigational, and resource*, along with hierarchical subcategories of

each. In determining the user intent, the researchers investigated using just the searcher's query, the results the searcher clicked on, and subsequent queries. Rose and Levinson (2004) reported that approximately 62% of the queries are informational, 13% navigational, and 24% resource. The researchers report only small differences in results when using the additional information beyond the query. However, like that of Broder (2002), this research was based on logs from the search engine and not the landing page website.

Researchers have also examined automatically classifying intent, which is related to the research that we propose here. For example, Lee, Liu, and Cho (2005) automatically classified informational and navigational queries using 50 queries collected from computer science students at a U.S. university. Kang and Kim (2003) classified queries as either topic or homepage using several iterations of classification. The researchers report a classification rate of 91% using selected TREC topics (50 topic and 150 homepage finding) and portions of the WT10g test collection. Dai, Nie, Wang, Zhao, Wen, and Li (2006) examined classifying whether or not a web query has commercial intent, noting that 38% of search queries have commercial intent. Baeza-Yates, Calderon-Benavides, and Gonzalez-Caro (2006) used supervised and unsupervised learning to classify 6,042 Web queries as either *informational, not informational*, or *ambiguous*. Jansen, Booth, and Spink (2008) provided a comprehensive automated multilevel analysis, reporting a 74% success rate in user intent classification using a decision tree approach. These approaches are similar to work by Özmutlu, Çavdur, and Özmutlu, 2006 that focused on topical classification.

However, these prior works all focused search engine data and not the corresponding user behavior on the landing page website, which could provide additional insights. For example, Nettleton, Calderon, and Baeza-Yates (2006) used 65,282 queries along with click stream data and clustered these queries based on various parameters to label queries as informational, navigational, or transactional. Fujii (2008) presented a method for identifying navigational

queries by comparing them to the anchor text in webpages, using 127 informational and 168 navigational queries. The researcher reported that anchor text can be used for query classification. Cao and colleagues (2009) used a set of previous queries from a user session as well as the webpages retrieved by these queries to topically classify queries based on taxonomy of web topics. Kathuria, Jansen, Hafernik, and Spink (2010) used k-means clustering to automatically cluster web queries into eight different clusters, six informational, one transactional, and one navigational.

However, these prior works have focused solely on identifying user intent of query terms. In this research, we extend the line of inquiry by examining (and predicting) actual user behavior on a website by clustering referral keywords based on similar onsite behaviors. So, our research provides a linkage between the user intent work focused on query terms and the consumer behaviors on the destination website. Based on the prior work showing that different query terms are implicit indicators of intent, it would seem reasonable that these query terms also act as gauges of different user behaviors because the underlying intent may be different. In fact, there is prior work that suggests this linkage. With general web searching, researchers have developed different classifications depending on the users' browsing behaviors or the queries entered (Caramel, Crawford, & Chen, 1992; Jansen et al., 2008; Marchionini, 1995; Rozanski, Bollman, & Lipman, 2001).

Given that visitors to a website (who are first searchers on the search engine and may be viewed later as potential customers by an online business) arrive via different keywords, it is reasonable to assume they might exhibit different behaviors on the website. If so, there would potential value in segmenting these visitors based on these keywords and behaviors. Such user generated data can endow businesses with valuable information about understanding users better and indicate needed modifications or enhancements in web systems (Jansen, 2009, p. 2). In fact, Carmel, Crawford, and Chen (1992) classified users into three different categories using a verbal

protocol analysis technique. Cheung, Kao, and Lee (1998) used web tools that analyzed web user data that allowed them to learn users' access patterns. Buchner, Mulvenna, Anand, and Hughes (1999) showed how mining web server traffic to discover patterns of user access could be used for the marketing and management of e-business and e-services. Rozanski, Bollman, and Lipman (2001) segmented Internet users into seven different categories by analyzing clickstream data and exploring session characteristics. However, this study only used four session variables (session length, time per page, category concentration, and site familiarity) to segment the users. Chen and Cooper (2001, 2002) used clustering and stochastic modeling to detect usage patterns in a web information system. Banerjee and Ghosh (2001) clustered users using a weblog on a website. The study found six different clusters using weighted longest common subsequences that took into account the trajectory and time spent at each page. Phippen, Sheppard, and Furnell (2004) state that companies can coordinate and audit website design by understanding user behavior using an array of web metrics.

By analyzing web usage patterns to segment users, one might be able to modify or enhance web systems in a way that effectively caters to segmented Internet traffic. Websites could then offer services in a more personalized way to their users. In addition, segmentation of online visitors allows advertising networks to behaviorally target advertisements.

Unlike other forms of Internet advertising, such as sponsored search advertising or content advertising, behavioral targeting is the practice of displaying advertisements based on past user behaviors. Advertising has the potential to be much more effective when using information science concepts such as relevance (Saracevic, 1975). As Yan and fellow researchers (2009) note, users who click on the same advertisement exhibit similar behaviors on the web. Therefore, the click through rate of an online advertisement can be significantly improved by segmenting users. Yan et al. (2009) also notes that segmenting using short-term behaviors is more effective for behavioral advertising than using long-term behaviors. So, by segmenting visitors,

not only can websites offer more personalized services, but also the click through rate of the advertisement on those websites can be improved, leading to more revenue being generated.

Despite the research on web searching, which drives much of the traffic to websites, there is little published research that attempts to find out how effective various segments of traffic from search engines really is or how users are behaving after arriving at the site from a search engine search. Addressing this issue has significant ramifications for areas from information science (e.g., information relevance and usefulness) to marketing (e.g., advertising and search engine optimization).

In this study, we investigate whether or not one can segment customers to a website based on user behavior. We leverage keyword referral data to a Spanish music website and user behaviors collected via an analytics program.

Research Objective

Our research objective is to automatically classify a large set of referral keywords into unique clusters that are meaningful for an online business. The motivation for this research objective is to demonstrate whether or not segmenting the market by referral keywords can provide actionable intelligence for online businesses. Market segmentation is the processes of dividing a market along some similarity where the market segments have something in common (Thomas, 2007), and it is considered important for tailoring aspects of a business, such as marketing and advertising, to particular customer groups.

To investigate this research objective, we use a referral keyword log from an online business website. In addition to the referral keywords, we also collect online consumer behaviors associated with these keywords, such as page views, time on site, and revenue generated. Based on the attributes, we employed a k-means clustering algorithm (Kanungo et al., 2002) to segment clusters of customers based on referral keywords and the onsite behaviors associated with those keywords. K-means clustering is a categorizing and labeling algorithm based on means of groups of similar data points.

Research Design

We first present our data collection site.

Data Collection from BuenaMusica.com

For this research study, we collected data from BuenaMusica.com, a Spanish-based entertainment business. The website offers customers the ability to play songs, watch music videos, look up song lyrics, read artist biographies, check the latest artist news, and communicate with other users in chat rooms, as well as listen to streaming radio. The site is financially supported by revenue generated through advertisements of Google AdSense, which is an advertising service where the website (i.e., publisher) allows Google to post advertisements on the site in exchange for a portion of the advertising revenue Google receives. This form of contextual advertising is a primary revenue source for many websites. Figure 1 shows a sample of BuenaMusica.com's homepage, illustrating the site's interface and features during the data collection period.

Figure 1. BuenaMusica.com homepage.

At the time of the study (June 1 to October 31, 2010), the Google search engine had indexed a total of 116,000 pages of the domain www.BuenaMusica.com. Alexa.com, a web traffic reporting company, had assigned BuenaMusica.com a worldwide traffic rank of 26,178. According to Alexa.com, the site is particularly popular in South America where it is in the top 1,000 visited sites in three countries. In Nicaragua, it is ranked 487, in Guatemala 919, and in Venezuela 929. In the United States, it has a 33,354 traffic rank, and in China, it has a 106,573 traffic rank. So, the site is a well-trafficked website from a multitude of countries and, therefore, it is a good candidate for our research. Further breakdown of worldwide traffic rank for the site is shown in Table 1.

TABLE 1. Worldwide traffic rank according to Alexa.com.

BuenaMusica.com worldwide traffic rank	
Country	Rank
Nicaragua	487
Guatemala	919
Venezuela	929
Honduras	1,038
Dominican Republic	1,567
Bolivia	1,570
Ecuador	2,619
Colombia	3,132
Peru	3,428
Mexico	4,278
Costa Rica	7,943
Argentina	16,013
Spain	29,098
United States	33,354
China	106,673

TABLE 2. Visitors by country breakdown according to Alexa.com.

Visitors by country	
Country	Percent of site traffic
Venezuela	22.1
Mexico	19.0
United States	13.7
Colombia	7.9
Peru	6.0
Dominican Republic	4.7
Guatemala	4.3
Spain	3.7
Nicaragua	3.1
China	2.7

The percentage of visitors by country according to Alexa.com can be seen in Table 2. According to Alexa.com, visitors from Venezuela, Mexico and the United States make up 54.8% of total site traffic.

Concerning demographics, Figure 2 provides a breakdown of the site's U.S. visitor demographics based on data from Quantcast.com, a widely used web traffic firm that provides web analytics data.

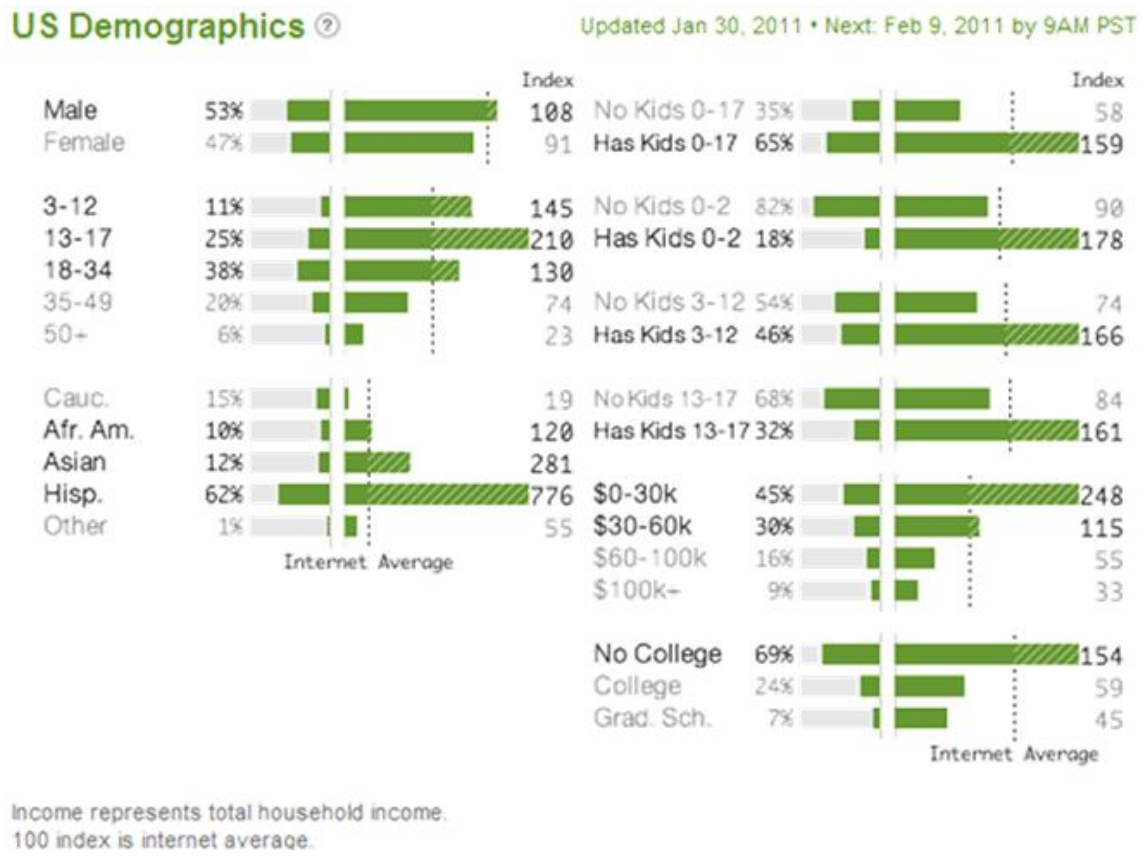

Figure 2. U.S. demographics breakdown according to Quantcast.com.

As seen in Figure 2, the majority of users in the United States who visit BuenaMusica.com are Hispanics between the ages of 13 and 34. In addition, 69% of U.S. users do not have a college education and almost half make less than $30,000 a year. However, this is certainly tied to some degree with the age of the website users, since 36% of users are under the age of 18.

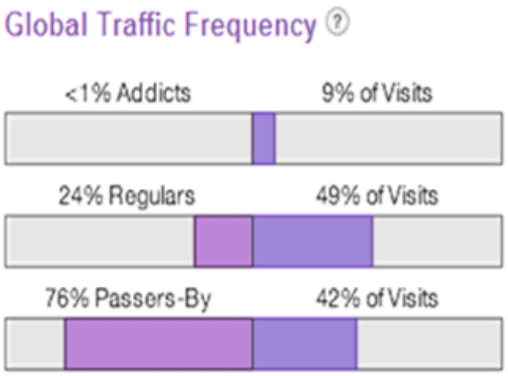

Figure 3. Global traffic frequency breakdown according to Quantcast.com.

From Figure 3, 76% of users are *passers-by*, meaning that they have a single visit over the course of a month. The second row of Figure 3 shows that 24% of users are *regulars* who visit the site more than once but less than 30 times a month. Less than 1% of users are *addicts* who visit the site more than 30 times per month. So, the traffic breakdown is fairly typical, in that there are a lot of occasional users, a fair portion of frequent users, and a small number of core users.

Data Collection and Preparation

The data used for this research study were collected using Google Analytics, an online website analytics tool, which is widely used in the industry. This web-based tool generates detailed statistics and reports about traffic and visitors to a website. Given the wide use of Google Analytics, we expect the procedures used in this research to be implementable by many online businesses and other websites. Google Analytics is integrated into a site by a page tag. A snippet of JavaScript code, known as the Google Analytics Tracking Code (GATC), is embedded on every page of the website. This code has a unique identification tag that identifies the website with the Google Analytics account holder. Whenever the page is loaded, the snippet of code runs, collects visitor data, and sends it to Google servers for processing and aggregation. The statistics collected range from the time a user spent on the site, to the number of page views, browser used, operating system of the computer, as well as screen resolution of the computer monitor. All of this information is available to the account holder via the Google Analytics interface.

Additionally, the analytics tool collects the referral information associated with the particular website. Referrals are page visits from different websites or search engines that direct traffic to a particular website via a hyperlink. A referral URL is the web address of the search engine or website that is directing traffic to another site. For example, say site A (e.g., http://www.twitter.com/buena_musica) has a link to site B (e.g., http://www.buenamusica.com). The traffic that site B receives from site A via the link is called referral traffic. The URL of site A where the link is placed is called the referral URL. In our example, the referral URL would be http://www.twitter.com/buena_musica. A similar process happens for traffic directed from web search engines.

Figure 4 features a screenshot of a typical Google search results page.

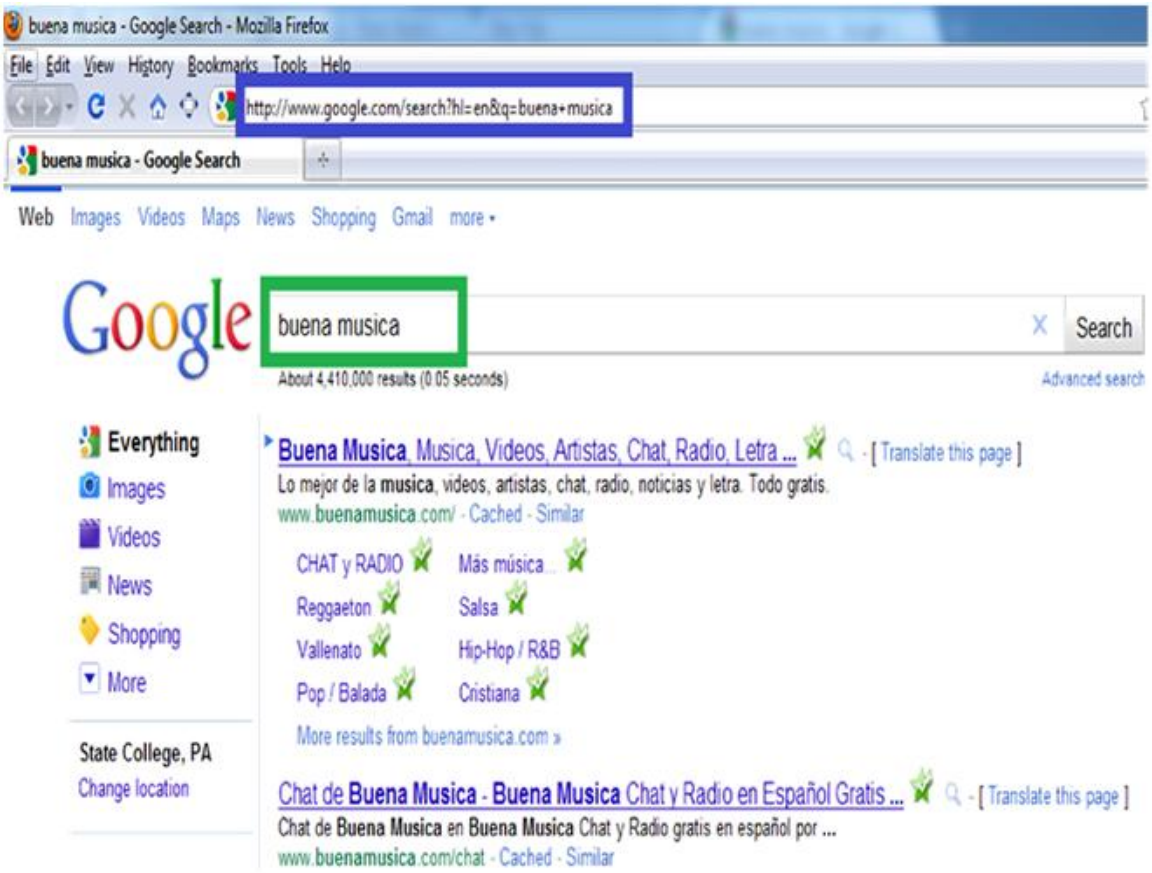

Figure 4. Google search results page example.

Looking at Figure 4, the phrase inside the green box is the query (“buena musica”) submitted by the searcher. The text inside the blue box is the URL of the results page (http://www.google.com/search?q=buena+musica). Embedded in this URL are the terms of the query (e.g., buena musica). When the user clicks on a result, the URL in the blue box gets passed to the website pointed to by link that the user has clicked on (i.e., the specific webpage is called the landing page). If the landing page has an analytics tracking tool, this tool can collect and analyze the information within the referral URL for aggregation and analysis. Referral keywords and the referral URL provide invaluable insight to webmasters and website owners. The referral

URL provides information such as where traffic is originating, and the referral keywords provide insights into what the users are searching for that ultimately leads them to the website. Google Analytics is particularly useful for the research presented here since it collects traffic source statistics including referral keywords along with user behaviors on the website associated with each referral keyword. Each referral keyword is collected with the following attributes:

- **Visits**: the number of visits to the site that the keyword generated in a given time period
- **Pages per visit**: the average number of pages viewed during a visit
- **Average time on site**: the average duration of a visit to the site
- **Percentage of new visits**: the percentage of visits by people who have never visited the site before (within a given period and based on IP address and cookie)
- **Bounce rate**: the percentage of single-page visits (i.e., visits where the user left the site from the landing page without browsing other pages)

Each of these attributes can be aligned to the referral keyword that brought the users to the website. Table 3 shows an example of a Google Analytics referral keyword log:

TABLE 3. Google Analytics referral keyword log with browsing behavior attributes.

Keyword	Visits	Pager per visit	Average time on site	New visits	Bounce rate
buena musica	773,030	5.81	685.18	0.35	0.29
buenamusica.com	688,533	6.26	750.57	0.32	0.26
Buenamusica	318,636	5.50	717.22	0.32	0.28
Musica	203,509	6.37	561.75	0.63	0.28

Content Advertising

BuenaMusica.com is supported by revenue generated by advertisements from Google AdSense, which is a service where website owners can offer advertisements based on the website's content. This service is implemented similar to Google Analytics where snippets of code are placed on web pages and when the pages are viewed, advertisements are displayed. Revenue is generated when a visitor clicks on an advertisement. The AdSense account publisher (e.g., BuenaMusica.com) makes a quantity of money based on the amount the advertiser is paying Google to service a particular advertisement. This quantity can range from a fraction of a penny per click to a couple dollars or more per click. The amount of money depends on the content of the page, the user's demographics, the past performance history of the site, and other factors.

Because both Analytics and AdSense are Google products, one can integrate them to share data. This integration is done in the Google Analytics web interface where the account holder can allow the Analytics account to receive AdSense data. When enabled, the Google Analytics program collects the ad revenue information of the site from the AdSense account, aligning these attributes with referral keywords.

For this research, we configured Google Analytics and AdSense applications to share data, allowing the referral keywords to have not only the website attributes but also the additional AdSense revenue data. The AdSense attributes are as follows:

- **Revenue**: the total amount of revenue generated for the website
- **Ads clicked**: the number of ads clicked
- **Page impressions**: the number of viewed pages where ads were displayed
- **CTR (click though rate)**: the ratio of the number of ads clicked on to the number of ads viewed

- **eCPM (cost per million impressions)**: the estimated revenue from AdSense per thousand ad page views

Table 4 is an example a referral keyword log with additional revenue attribute:

TABLE 4. Google Analytics referral keyword log with revenue and Google AdSense attributes.

Keyword	Revenue	Ads clicked	Page impressions	CTR	eCPM
buena musica	1,138.95	37,433	4,116,684	0.0091	0.27
buenamusica.com	883.52	33,518	3,974,888	0.0084	0.22
Buenamusica	336.61	10,146	1,611,535	0.0062	0.20
Musica	687.77	30,224	1,170,304	0.0246	0.55

Note. CTR = clickthough rate.

From Table 4, keyword is the referral keywords (i.e., the query that the user submitted to the search engine). Revenue is the amount of ad revenue generated by consumers who arrived at the site from a search engine using the referral keywords within a given period. Ads clicked is the number of times those consumers clicked on ads displayed on the site, which generates the revenue. Page impressions are the number of ads displayed to these consumers. CTR (click through rate) is ad clicked divided by page impressions. eCPM is the cost per 1,000 ads displays to these consumers.

Data Methodology

We now discuss our research data.

Data Collection

We collected data on referral keywords, visitor traffic, and advertising revenue data on BuenaMusica.com from June 1, 2010 through October 31, 2010. A total of 900,795 referral keyword records were collected during the 5-month data collection period. We extracted each month's data individually from Google Analytics web interface in a tab delimited file format with 20,000 keywords in each batch, since Google Analytics limited the size of each export. We then imported these batches into a relational database for data normalization and aggregation. Some referral keywords were repeated within each batch. After normalization and aggregation, the total number of keywords used for data analysis was 712,643. Of the 900,795 keywords, 188,152 were duplicates of some sort. We discuss data normalization and aggregation in more detail in the data preparation section.

Data Preparation

Our first step in analysis was to normalize all the attribute values since some were ratios (e.g., *bounce rate, CTR and eCPM*) and some were absolute numbers (e.g., *pages per visit, time on site, new visits, bounce rate, revenue, ads clicked*). Additionally, high traffic keywords such as "buena musica" (i.e., branded keywords in the search engine marketing realm) or "musica" bring

in thousands of visits to the site as opposed to other queries that only bring in a few visits. Because of the nature of the k-means clustering algorithm, clustering using ratios and absolute numbers would skew the results (i.e., comparing apples to oranges). To address this issue and get a more accurate representation of the keyword attributes, we clustered using attributes that were ratios or percentages. However, we wanted to use all of the attributes possible, so we created ratios using the attributes that were absolute numbers. The additional three ratios that we generated for this research are as follows:

- **Average revenue**: the average revenue the site makes from a visit based on a given referral keyword. We calculated this ratio using: total revenue / number of visits
- **Average ads clicked**: the average number of ads clicked based on a given referral keyword. We calculated this ratio using: ads clicked / number of visits
- **Average impressions**: the average number of ad impressions based on a given referral keyword. We calculated this ratio using: page impressions / number of visits

Using these three additional ratios, we were able to cluster using all attributes and at the same time without using raw attributes such as *visits or ads clicked* that would skew the results. At the same time, we did not lose any of the additional information provided by the data collection applications.

It is worth mentioning that some keywords in the logs showed that they bring in zero visits. This is because a visit is not recorded if the user has remained inactive for more than 30 minutes or has cleared the browser's cache within those 30 minutes. If a user visited the site within 30 minutes and then searched using a term that brought him/her back to the site within those 30 minutes, a visit would not be recorded under that or any subsequent keywords.

For referral keywords that had zero visits, we calculated the three additional ratios using the following formula:

IF (visits>0, value/visits, value)

The formula states that if the number of visits is greater than zero; divide the value (revenue, ads clicked, and impressions) by the number of visits. Otherwise, simply copy the value to the respective averages field.

Once we had prepared the data set, the final k-means clustering was done with the following nine attributes associated with each referral keyword: *pages per visit, average time on site, percentage of new visits, bounce rate, CTR, eCPM, average revenue, average ads clicked, average impressions*.

Once the formula was applied to all keywords, the spreadsheet was imported into SPSS, a statistical analysis computer program. After the keywords were aggregated, we implemented k-means clustering on entire data set.

Data Methodology and Analysis

Clustering of the data was done using the k-means clustering algorithm. This particular algorithm uses an unsupervised learning technique that makes it ideal for clustering big data sets. The objective of the algorithm is to segment n items (keyword attributes in our case) into k clusters where each item (i.e., keyword) belongs to the cluster that is of the nearest mean. Items in the same cluster are most similar to each and most dissimilar to those in other clusters.

The k-means clustering algorithm attempts to maximize the mean of each cluster while at the same time tries to minimize the standard deviation in these clusters. The algorithm uses a k

amount of centroids. Centroids can be defined as random points in the data that serve as the center of that cluster. The Euclidian distance is determined by:

$$\sqrt{D_{ij} = \sum_{k=1}^{n} (x_{ki} - x_{kj})^2}$$

where

D_{ij} distance between cases *i* and *j*
x_{ki} value of variable X_k for case *j*

Below is a step-by-step break down of the algorithm:

1. Randomly choose *k* centroids and use them as initial centroids (centers).
2. For each item, locate the closest center and assign the item to the cluster that the nearest centroids belongs to.
3. Update the centroids of each cluster derived from the times in that cluster.
    a. The new updated centroid will be the mean (average) of all items that belong to that cluster.
4. Until no item switches clusters, repeat steps 2 and 3.

For data analysis, we experimented with a minimum of two cluster groupings and a maximum of 10 cluster groupings. Early analysis and comparison of frequency numbers for each cluster showed that in all of those cluster groupings, there was always one cluster that only had two keywords. Upon examination, these two keywords were determined to be outliers because they had a very large average time on site (about 35,000 seconds). Considering that we were clustering over 700,000 keywords, we believe that the removal of these two outlier keywords was not going to affect our end results, so those two keywords were removed from our data to prevent skewing the clusters.

We re-ran the k-means clustering groupings again from 2 to 10 clusters (iterations up to 20 for each clustering attempt), and saw that the frequency distribution for each cluster was more realistic based on our known traffic patterns reported above. After analysis and comparison between these nine cluster groupings, the group of six clusters for the 712,643 referral keyword most accurately described the customer segmentation while showing the maximum differences between each cluster based on the elbow method (Sugar & James, 2003), which is based on the amount of variance in the data that is explained by adding an additional cluster.

To check for effect of data order on clustering results, we did a three-fold cross validation, ending with six clusters on each validation. Given this result, we believe our methodological approach to be valid.

Each of the six clusters represents a grouping of keywords or search key phrases that share commonality among the attributes specified. Each cluster is also dissimilar with the other clusters.

Results

Table 5 shows the frequency of each cluster. Each row is a different cluster and the two columns are the frequency and percent of that cluster relative to the entire data set.

Table 5 shows an uneven distribution of the cluster frequency. Cluster 1 is the biggest cluster with 83.3% followed by cluster 2 with 11.3%. The next biggest cluster is cluster 5 with 3.9% followed by cluster 4 with 1.2%. Cluster 3 and cluster 6 each had 0.3% and less than 0.0%, respectively.

TABLE 5. Frequency and percentages of six clusters.

Cluster	Frequency	Percent
1	593,532	83.3
2	80,583	11.3
3	2,069	0.3
4	8,399	1.2
5	27,715	3.9
6	343	~0
Total	712,641	100.0

Table 6 shows the final cluster centers (i.e., means) and standard deviation *(SD)* for each cluster. Each column is a different cluster and each row is an attribute of that cluster. The final cluster table provides insightful information about user behavior based on web factors, giving us a consolidated snapshot of the cluster groupings.

We now discuss each cluster and the implications for online businesses by categorizing each cluster along two axes, *onsite behavior* and *revenue generation*. Onsite behavior addresses the engagement of the visitor while on the site (e.g., percentage of new visits, pages per visit, average time on site, bounce rate, CTR, average ads clicked, and average impressions), and the revenue generation addresses the business concern from the perspective of the website owner

(e.g., eCPM, average revenue). For both onsite behavior and revenue generation, we classify each cluster as high, medium, or low relatively.

TABLE 6. Final cluster centers of each cluster.

	Cluster											
	1		2		3		4		5		6	
Attributes	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Pages per visit	1.59	1.42	5.61	4.4	34.59	28.5	18.32	16.7	9.32	8.9	71.74	58.1
Average time on site (sec.)	28.74	56.20	474.15	168.7	4,780.92	899.1	2,508.16	475.7	1,252.47	280.3	9,462.36	2463.6
Percentage of new visits	0.81	0.37	0.74	0.4	0.64	0.4	0.66	0.4	0.69	0.4	0.66	0.5
Bounce rate	0.67	0.44	0.11	0.2	0.01	0.1	0.03	0.1	0.05	0.1	0.004	0.04
CTR	0.02	0.13	0.02	0.1	0.009	0.04	0.015	0.07	0.022	0.09	0.006	0.03
eCPM	0.72	10.26	0.95	9.0	0.35	1.7	0.63	6.0	0.84	6.4	0.25	1.0
Average revenue ($)	0.001	0.04	0.004	0.02	0.009	0.04	0.007	0.04	0.005	0.04	0.014	0.05
Average ads clicked	0.035	0.19	0.11	0.4	0.18	0.5	0.17	0.5	0.14	0.4	0.26	0.6
Average impressions	1.6	1.80	5.11	4.3	31.74	28.1	16.65	16.5	8.4	8.6	63.96	59.2
Classification engagement	Low		Low		High		Medium		Medium		High	
Classification revenue	Low		Medium		High		Medium		Low		High	

Note. SD = standard deviation. CTR = clickthough rate

Cluster 1—Low Engagement, Low Revenue

This cluster was both low engagement and low revenue. With 593,532 keywords, cluster 1 accounted for 83.3% of the dataset. Of the visits in this cluster, 81.21% were new visits (i.e., first time visitors within the preceding 30 days) and about 67% of visits were bounces (i.e., the visitor viewed one page view only). For cluster 1, users viewed 1.59 pages per visit, by far the lowest of any cluster. Users in this cluster generally spent 28 seconds on the site per visit. The number of ad page impressions per visit was 1.60 and the ad click through rate was 2.2%. In terms of revenue per visit, with users in this cluster clicking on the least number of ads. One thousand ad page views (eCPM) in this cluster generated $0.25. So, for engagement, those visiting the website in this cluster are typically new visitors, who spend very little time on the site, and visit very little content. In terms of revenue, visitors in the cluster generated by far the lowest revenue.

Search phrases in cluster 1 typically comprised natural queries that are expressed in the form of a question usually asking for some specific information. The information that queries looked for ranged from wanting to know an artist's nationality, names of members of bands, and who wrote specific songs. For example, there were queries that looked for specific information about an artist such as "what is the real name of {artist name}," "who is {artist name}," "where is {artist name} from." The queries asking for an artist nationality were composed of "nationality {artist name}." Queries asking for names of members of a specific band were composed of "names of the members of {band name}." Cluster 1 also had queries that were in the form of requests such as "I want to know" followed by any of the previously aforementioned examples. In addition, there were queries that wanted to know who wrote specific songs such as "who wrote {song name}." Also, there were search phrases that looked for lyrics of songs and the query was

composed of “lyrics of the song {song name}.” In addition, there were queries that actually seemed to be malformed URLs. These queries begin with “www.” followed by any combination of genre, song name, or video name with a space or typo somewhere in the middle of the text between the “www” and the “.com.” (Note: We provide the English translation as most of the referral keywords were in Spanish).

For the online business, these customers are the ones that generate most of the traffic but generate the lowest average revenue per visit. In some aspects, they are primarily cost, as they use server cycles, access information, and generate little revenue. However, they represent a significant portion of the traffic to the website, which aids in website ranking. So, they do provide some indirect benefit.

Cluster 2—Low Engagement, Medium Revenue

Cluster 2 was low engagement and medium revenue users. There were 80,583 keywords in cluster 2 that accounted for 11.3% of the dataset. For cluster 2, users viewed 5.61 pages per visit, higher than cluster 1 but lower than that of the other clusters. Users in this cluster generally spent 474 seconds on the site per visit and 84 seconds per page view. Of the total, 74% of visits in this cluster were new visits, and about 11% of visits were bounces, which is a very low bounce rate. The number of ad page impressions per visit was 5.11 and the ad click through rate was 2.76%, resulting in $0.004 in revenue per visit. One thousand ad page views (eCPM) in this cluster generated $0.95. In terms of engagement, visitors in this cluster were most like cluster 1. In terms of revenue, these visitors were most like cluster 4.

Search phrases in cluster 2 also consisted of queries that looked for information using broader terms than those in cluster 1. The queries looked for information on artists about their music productions such as discographies and less about their personal lives as in cluster 1.

Queries in this cluster had a variety of combinations and were ordered in a variety of ways that included an *artist name* and the term *biography* or *discography*. For example, some of the queries were composed of "biography {artist name}" or "discography {artist name}." In addition to wanting to know artists' information, there were queries that looked for information about specific albums. Some of these queries comprised "in what album" followed by terms related to that album. Search phrases in cluster 2 also comprised terms related to the domain name such as "buena musica." Queries also looked for chat rooms that were either "free" or had the domain of the site. Such search phrases were composed of the word "chat" with a derivative of the term "buena musica" or "free." This cluster also comprised some search phrases that were composed the term "listen to music" and a combination of "free" and a genre or artist name.

For the business, these customers arrive at the site after searching for terms that are similar to the domain name. Many of these customers, although new, have heard of the website as evidenced by the branded referral keyword. These customers stay on the site for a reasonable amount of time and generate a fair amount of revenue per visit. As such, searchers with in this cluster are a potential source of new repeat customers for the online business.

Cluster 3—High Engagement, High Revenue

Although small with 2,069 keywords accounting for 0.3% of the dataset, cluster 3 users are high engagement and high revenue. Visitors in this cluster viewed 34.59 pages per visit and spent 9,462 seconds on the site per visit, the most pages and time of any cluster. With 64% of visits in this cluster being new visits (the fewest of any cluster) and a bounce rate of only 1%, these users are highly engaged with the site. The number of ad page impressions per visit is 31.74 and the ad click through rate is 0.97%, which resulted in $0.009 in revenue per visit. One thousand ad page views (eCPM) in this cluster generate $0.35, the second highest of any cluster.

So, in addition to being highly engaged, visitors in cluster 3 are also high revenue generators. Visitors in this cluster were most like visitors in cluster 6 in terms of both engagement and revenue.

Search phrases in cluster 3 comprised broad terms that included verb format queries such as “listen.” Users of cluster 3 clearly wanted to listen to music and the search phrases were composed of “listen to” followed by a combination of an artist name, genre, or song name. Another derivative of these types of queries included the term music, and the phrases comprised “listen to music of” and a particular genre or artist name. Search phrases in cluster 3 also comprised the word “music” without the term listen followed by an artist name or song name. In addition, search phrases that looked for song collections of albums were present in this cluster and comprised “songs of” followed by a particular artist name.

For the business, these customers come to the site with the clear intent of listening to music and browsing lots of pages. These customers can be considered the ones that are the most expensive to have because they are most likely using a lot of bandwidth and server resources. However, as they also generate a considerable amount of revenue, as this traffic segment does contribute to cash flow.

Cluster 4—Medium Engagement, Medium Revenue

There were 8,399 keywords in cluster 4 (1.2% of the dataset), and these users were medium engagement with site and revenue generators. For cluster 4, users viewed an average 18.32 pages per visit, substantially more than the low engagement clusters, such as clusters 1 and 2, but substantially less than the high engagement clusters 3 and 6. Users in cluster 4 spent 2,508 seconds on the site per visit. Of the visits, 66.58% were new visits in cluster 4, with a 3.29% bounce rate. The number of ad page impressions per visit was 16.65, and the ad click through rate

was 1.58%. This resulted in $0.007 cents in revenue per visit. One thousand ad page views (eCPM) in this cluster generate $0.63 dollars. So, visitors in this cluster had good engagement and revenue generation. These visitors were most like those in cluster 5 for engagement and most like those in cluster 2 in terms of revenue.

Search phrases in cluster 4 had a combination of verb format queries as well as general queries. The verb format queries comprised “watch videos of” followed by an artist name or genre. Cluster 4 also had queries that were composed of “biography” followed by a particular artist name. There were some queries in the cluster that comprised “music” or “listen” with a combination of genre or artist name. Also, there were queries that comprised a particular genre and an artist associated with that genre such as “salsa {artist name}.” In addition, there were search phrases that simply had the word “video” followed by an artist name or genre.

For the business, these are average customers that yield a decent amount of revenue per visit, and their information interests are fairly focused. As such, given the relatively low time on site versus the revenue generated, they are a small but worthwhile customer group.

Cluster 5—Medium Engagement, Low Revenue

Cluster 5 visitors are medium engagement and low revenue. There were 27,715 keywords in cluster 5 that accounted for 3.9% of the dataset. For cluster 5, users viewed 9.32 pages per visit. Users in this cluster generally spent 1,252 seconds on the site per. There were 69.56% of visits in this cluster that were new visits, and about 5.55% of visits were bounces. The number of ad page impressions per visit was 8.40, and the ad click through rate was 2.20%, which resulted in $0.0005 in revenue per visit. One thousand ad page views (eCPM) in this cluster generated $0.84. So, visitors in cluster 5 are most like those in cluster 4 and most like cluster 1 in revenue generation.

Search phrases in cluster 5 comprised phrases such that include the domain name “buena musica” and a combination or an artist name or genre. These queries were composed of “buena musica {artist name}” or “buena musica {genre}.” In addition, the term “free” is also present in queries of cluster five. Users in cluster 5 were looking for free music as shown by queries such as “free good music.” This cluster also had some verb format queries that were looking to listen to specific songs such as “listen to {song name}.” Users in this cluster not only were looking for music but were also looking at lyrics as well. These queries included terms such as “lyrics and music of” followed by a combination of an artist name or song name.

For the business, these customers are similar to customers in cluster 2 in the sense that the search phrases in both clusters are somewhat alike, although they are more engaged, using more site resources. However, cluster 5 customers are low revenue generators, so these are the least attractive customers—they spend a considerable time on the site and don’t generate much revenue.

Cluster 6—High Engagement, High Revenue

Visitors in cluster 6 are few, but they are highly engaged and generate a high amount of revenue. There were 343 keywords in cluster 6 that accounted for 0.01% of the dataset. For cluster 6, users viewed 71.74 pages per visit. Users in this cluster generally spent 4,780 seconds on the site per visit and 131 seconds per page view, with 66.55% of visits in this cluster being new visits and about 0.45% of visits being bounces, a near zero bounce rate.

The number of ad page impressions per visit was 63.96 and the ad click through rate was 0.66%, resulting in $0.14 in revenue per visit. One thousand ad page views (eCPM) in this cluster generated $0.25. Users in cluster 6 are most like those in cluster 3 in terms of both engagement and revenue. However, relative to cluster 3, they are less engaged and generate more revenue.

Search phrases in cluster 6 were very broad in nature. The queries were comprised of general terms looking for music from a particular genre such as “music {genre}.” Other search phrases in this cluster included a genre name followed by the term music. Also, search queries in this cluster comprised verb format queries that explicitly stated what the searcher wanted to do such as “listen to music” followed by a specific genre. In addition, this cluster also had queries that looked for general music sites such as “listen to music page.” For the business, these customers are the music addicts that browse a large number of page views and stay on the site for a long amount of time. These are the ideal customers even though they might use a lot of server resources, as these users yield the highest ad click through rate and generate the highest revenue per visit.

Discussion and Implications

The implications of this research is that using referral keywords, combined with online user behaviors, to segment consumer traffic can provide critical customer insight to online businesses.

We demonstrate this by analyzing search engine traffic to the BuenaMusica website. Specifically in our analysis, we identified six clusters:

- Cluster 1 (low engagement, low revenue) keywords brought in the most number of visits to the site as well as the most number of new visits. We believe that user behavior (i.e., the high bounce rate) in this cluster was mainly because users arrived at a landing page that did not contain the information they were looking for there. The intent of users in this cluster was mostly for information seeking purposes as opposed to listening to music. Even if the landing page did contain the information they were looking for, they still left the site almost immediately without showing much interest in browsing other pages. We can see how this would be the case because users in cluster 1 are generally information seekers looking for artist-specific information and using natural language queries in the form of questions. Users in cluster 1 are not interested in listening to music. Even if they do start to listen to a song or watch a video, they do not stay on the site for the entire length of the song or video, which on average is about 3 minutes.

- Cluster 2 (low engagement, medium revenue) keywords brought in the second most number of visits to the site as well as the second highest number of new visits. We believe that users in this cluster were mainly on the site to listen to a couple of songs or videos since the average time on the site was almost 8 minutes. So, these are casual users.

- Cluster 3 (high engagement, high revenue) keywords brought in the second lowest number of visits to the site and the lowest number of new visits. Users in cluster 3 seem to be interested in taking their time to read the content of the page, listening to almost the whole song, or watching the entire video, because they spend a little over 2 minutes on each page.

- Cluster 4 (medium engagement, medium revenue) keywords brought in the third lowest number of visits to the site and the third lowest number of new visits. Users in cluster 4 seemed to listen to a whole song or watch an entire video since they spent over 2 minutes on each page. Unlike cluster 3, users in cluster 4 spent less than 20 minutes per visit (about one third of cluster 3) and browsed half of the page views relative to cluster 3.

- Cluster 5 (medium engagement, low revenue) keywords brought in the third highest number of visits to the site, along with a low number of new visits. In comparison to other clusters, users in cluster 5 spent the third lowest amount of time on the site. The average revenue per visit was the third lowest of all clusters. The average number of ads clicked per visit was also the third lowest as was the average number of page impressions. Users in cluster 5 spent almost the same amount of time per page view as users in cluster 4 and cluster 3. However, the pages per visit and average time on the site were half those in cluster 4.

- Cluster 6 (high engagement, high revenue) keywords brought in the lowest number of visits to the site and the second lowest number of new visits. In comparison to other clusters, users in cluster 6 spent the highest amount of time on the site. The number of pages per visit was also the highest and the bounce rate was the lowest. The CTR was the lowest and the eCPM is also the lowest. The average revenue per visit was the highest of all clusters. The average number of ads clicked per visit and average number of page impressions were also the

highest. Users in cluster 6 can be classified as outliers because they spent double the time on the site and viewed twice as many pages per visit.

TABLE 7. Final cluster centers with descriptions.

	Cluster					
Attributes	1	2	3	4	5	6
Pages per visit	Low	Medium low	High	Medium high	Medium	Very high
Average time on site	Low	Medium low	High	Medium high	Medium	Very high
Percentage of new visits	Very high	High	Medium low	Medium	Medium high	Medium
Bounce rate	Very high	High	Medium low	Medium	Medium high	Low
CTR	High	High	Medium	Medium high	Very high	Medium Low
eCPM	High	Very high	Medium low	Medium	Medium high	Low
Average revenue	Low	Medium low	High	Medium high	Medium	Very high
Average ads clicked	Low	Medium low	High	Medium high	Medium	Very high
Average impressions	Low	Medium low	High	Medium high	Medium	Very high

Note. CTR = clickthough rate.
Scale is low, medium low, medium, medium high, high, and very high.

TABLE 8. Final cluster centers in ascending order.

	Cluster					
Attributes	1	2	5	4	3	6
Pages per visit	Low	Medium low	Medium	Medium high	High	Very high
Average time on site	Low	Medium low	Medium	Medium high	High	Very high
Percentage of new visits	Very high	High	Medium high	Medium	Medium low	Medium
Bounce rate	Very high	High	Medium high	Medium	Medium low	Low
CTR	High	High	Very high	Medium high	Medium	Medium Low
eCPM	High	Very high	Medium high	Medium	Medium low	Low
Average revenue	Low	Medium low	Medium	Medium high	High	Very high
Average ads clicked	Low	Medium low	Medium	Medium high	High	Very high
Average impressions	Low	Medium low	Medium	Medium high	High	Very high
Percent of dataset (%)	83.3	11.3	3.9	1.2	0.3	~0

Note. CTR = clickthough rate.
Scale is low, medium low, medium, medium high, high, and very high.

In Table 7, we describe the attribute using several different adjectives using a 6-point Likert scale from (1-low) to (6-very high). The adjectives are based on how the attributes of each cluster relatively compare with the same attribute of another cluster. The attributes correlations for pages per visit, average time on site, average revenue, and average ads clicked as well as average impressions display a clear and identical pattern.

In Table 8, the clusters are arranged in ascending order based on the number of pages per visit. Several inferences can be made from the table and Figure 5, which show the trend lines for the attributes by cluster. We can see that there is a correlation between average time on site (for the more engaged clusters), average impressions, and pages per visit. We would expect these attributes to be somewhat correlated. Interestingly, metrics typically used as a surrogate for ecommerce (e.g., CTR), are not correlated with revenue.

Theoretical Implications

These research results extend prior work on user intent (e.g., Broder, 2002; Jansen et al., 2008; Rose & Levinson, 2004) by relating query terms with not only intent but also

different user behaviors on the landing websites. Much of the prior work in user intent has been in classifying and understanding the motivations behind the use of certain query terms during information searching (Jansen & Rieh, 2010; Marchionini, 1995). What has been lacking in this body of work is the downstream implications of the effect on user behavior. Understanding these behavioral effects can lead to improved processes and technology to enhance the searcher's overall experience.

The research presented here, combined with prior research, shows the beginning of a possible framework for extending understanding of web information searching, especially in the commercial domain. An end-to-end understanding has profound implications for ecommerce-related searching, especially. This research shows that query terms can be predictors of user behaviors. User behavioral tendencies associated with these terms appear to indicate that initial intent are important factors in indicating online commercial searching and evaluation of search results. These user behavioral tendencies and concepts provide the linkage between information searching to related, but usually separately researched, ecommerce processes such as consumer searching (Johnson, Moe, Fader, Bellman, & Lohse, 2004).

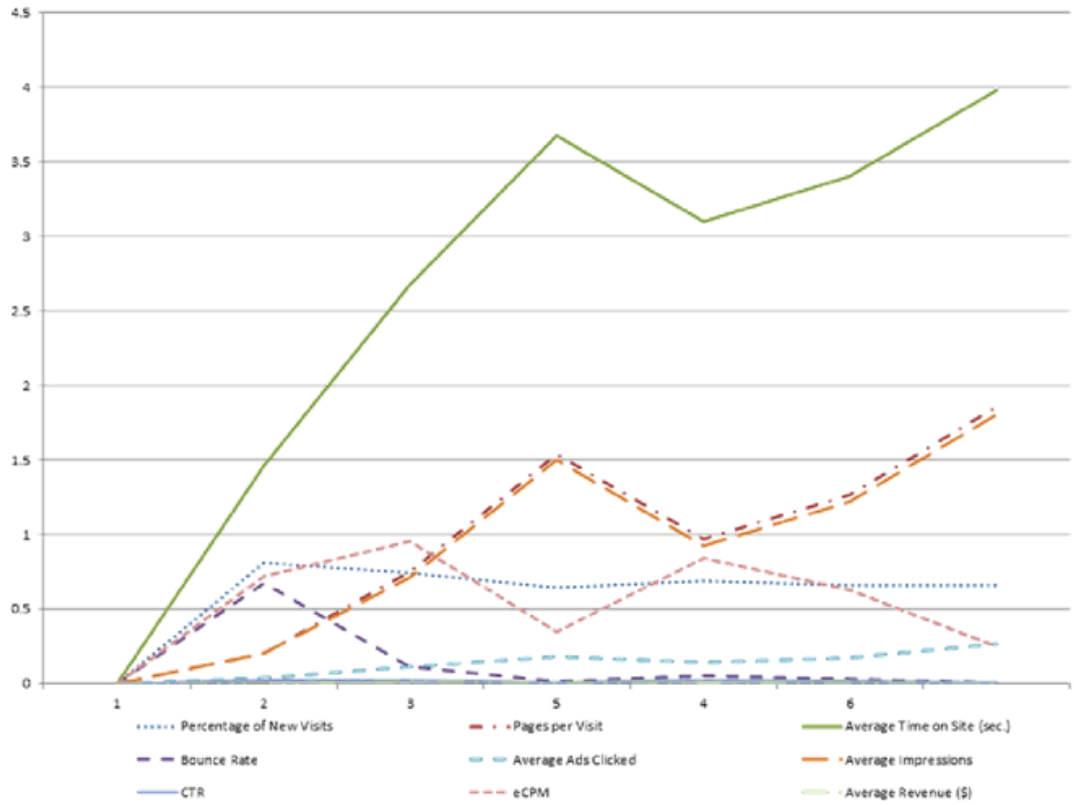

Figure5. Trend lines of cluster attributes.

Practical Implications

Understanding user intent through web search keywords and ecommerce factors can result in profitable market segmentation for online businesses. It is possible to segment website traffic by categorizing those customers being referred by the search engines. By segmenting customers to an online business, one is able to find the best and worst performing customers in terms of revenue generation. By segmenting these customers in terms of revenue generation, one can optimize the site. For example, one could attempt to optimize the site for bad performing customers in a way that attempts to increase customer conversion rates, page views, or time on site.

As a possible example to leverage this research, one way would be to optimize the site for best and worst performing customers. First, one could create a database with the referral keywords and their cluster groupings. Second, develop a predefined set of rules on how the landing page should display according to the cluster to which the referral keyword belongs. Third, create a server-side script that will extract the referral keyword of a user coming from a search engine and compare to the database. If there is a match, then follow the predefined rule for how to display the landing page according to that cluster.

Limitations and Strengths

One of this study's limitations is that the data belong to only one business that relied on contextual advertising. Users of the website mostly are Spanish speakers and are based in Latin America. This means that our results might not be applicable to other websites. However, all business and websites can collect the same data using Google Analytics. Therefore, it would be possible to use our research methodology and procedure to segment other customers. By segmenting their own customers into different clusters, website owners will be able to employ better advertising or other business models and make adjustments to their websites accordingly.

Concerning strengths, this research study used real-world searching, user behavior, and operating business data. The data were collected over an extended period of 20 weeks and represents actual searching behaviors of users on both search engines and the website. In addition, we provide a framework so that any website owner who collects the same type of data with Google Analytics will be able to cluster their own customers.

Conclusion

In this research, we categorized 712,643 referral keywords from major search engines, such as Google, Yahoo!, and Bing over a 20-week period. These keywords were categorized into clusters based on searcher behaviors and ecommerce-related attributes such as sales and revenue. Using a k-means clustering algorithm, we categorized search keywords that shared similar consumer behavior attributes and revenue characteristics. We developed a methodology for determining the user intent for web queries to identify profitable market segments using search engine referral keyword logs. We also outlined recommendations that webmasters can leverage this approach to increase engagement or revenue. Results from our data show that there are six clusters of users based on the engagement and revenue attributes that we investigated. Along with specific findings, results show that users who search using natural language queries spend a very short amount of time on the site and generate low income. Users who search for broader terms spend more time on the site and generate more revenue. The most significant implication is that website owners can leverage the k-means clustering for customer segmentation to build better business models and better website design. For website owners who want to increase their website conversion rates, they can cluster their customers based on browsing behaviors and ecommerce factors to engage the user at a more personalized level upon arriving at a landing page.

For future research, a similar study utilizing Google Analytics data from several different websites that uses the same research method can be conducted to see if there are similarities between the cluster groupings of each site. It would also be interesting to correlate these findings with attributes from the actual landing pages or with user profile information. Finally, it would

also be worthwhile to implement some of the website recommendations, based on the analysis presented here, and analyze the effect on customer behavior to determine if more targeted website experiences would actually increase revenue streams.

References

Baeza-Yates, R., Calder´on-Benavides, L., & Gonz´alez-Caro, C. (2006). The Intention Behind Web Queries. String Processing and Information Retrieval. In F. Crestani, P. Ferragina, & M. Sanderson (Eds.), (Vol. 4209, pp. 98–109): Springer Berlin/Heidelberg.

Banerjee, A., & Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. In R. Grossman, J. Jan, & V. Kumar, Proceedings of the Web Mining Workshop, First SIAM Conference on Data Mining (pp. 33–40). Philadelphia: SIAM.

Betts, M. (2001). Turning browsers into buyers. MIT Sloan Management Review, 42(2), 8–9.

Booth, D.L., & Jansen, B.J. (2008). A review of methodologies for analyz-ing websites. In B.J. Jansen, A. Spink & I. Taksa (Eds.), Handbook of research on web log analysis (pp. 1–17). Hershey, PA.: IGI.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1), 107– 117.

Broder, A. (2002). A taxonomy of web search. SIGIR Forum, 36(2), 3–10. Broder, A., Fontoura, M., Josifovski, V., & Riedel, L. (2007). In W. Kraaij & A.P.d. Vries (Eds.), A semantic approach to contextual advertising. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'07) (pp. 559–566). New York: ACM.

Buchner, A.G., Mulvenna, M.D., Anand, S.S., & Hughes, J.G. (1999). An Internet-enabled Knowledge Discovery Process. In G. Quirchmayr & E. Schweighofer, Proceedings of The Ninth International Database Conference (pp. 13–27). Springer: London.

Burby, J., Brown, A., & WAA Standards Committee. (2007). Web analytics definitions. Washington DC: Web Analytics Association.

Cao, H., Hu, D.H., Shen, D., Jiang, D., Sun, J.-T., Chen, E., et al. (2009). In J. Allan & J. Aslam (Eds.), Context-aware query classification. Proceed-ings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '09) (pp. 3–10). New York: ACM.

Caramel, E., Crawford, S., & Chen, H. (1992). Browsing in hypertext: A cognitive study. IEEE Transactions on Systems, Man and Cybernetics, 22(5), 865–883.

Carmel, E., Crawford, S., & Chen, H. (1992). Browsing in hypertext: A cognitive study. IEEE Transactions on Systems, Man, and Cybernetics, 22(5), 865–884.

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.

Cheung, D.W., Kao, B., & Lee, J. (1998). Discovering user access patterns on the World Wide Web. Knowledge-Based Systems, 10(7), 463–470.

Dai, H.K., Nie, Z., Wang, L., Zhao, L., Wen, J.-R., & Li, Y. (2006). Detecting Online Commercial Intention (OCI). In L. Carr, D. De Roure, & A. Iyengar, Proceedings of the 2006 World Wide Web Conference (WWW '06) (pp. 829–837). New York: ACM.

Fujii, A. (2008). Modeling anchor text and classifying queries to enhance web document retrieval. In J. Huai, R. Chen, & H.-W. Hon (Eds.), Proceedings of the 17th International Conference on World Wide Web (WWW '08) (pp. 337–346)

Hackett, S., & Parmanto, B. (2009). Homepage not enough when evaluating web site accessibility. Internet Research, 19(1), 78–87.

Jansen, B.J. (2009). Understanding user—Web interactions via web analytics. San Rafael, CA: Morgan-Claypool.

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

Jansen, B.J., & Rieh, S. (2010). The seventeen theoretical constructs of information searching and information retrieval. Journal of the American Society for Information Sciences and Technology, 61(8), 1517–1534.

Jansen, B.J., & Spink, A. (2004). An analysis of documents viewing patterns of web search engine users. In A. Scime (Ed.), Web mining: Applications and techniques (pp. 339–354). Hershey, PA: Idea.

Jansen, B.J., Spink, A., & Pedersen, J. (2005). Trend analysis of AltaVista web searching. Journal of the American Society for Information Science and Technology, 56(6), 559–570.

Johnson, E.J., Moe, W.W., Fader, P.S., Bellman, S., & Lohse, G.L. (2004). Depth and dynamics of online search behavior. Management Science, 50(3), 299–308.

Kang, I., & Kim, G. (2003). Query type classification for Web document retrieval. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development In Information Retrieval (ACM SIGIR ’03) (pp. 64–71). New York: ACM.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., & Wu, A.Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), Article 2.

Kathuria, A., Jansen, B.J., Hafernik, C., & Spink, A. (2010). Classifying the user intent of web queries using k-means clustering. Internet Research, 20(5), 563–581.

Lee, U., Liu, Z., & Cho, J. (2005). Automatic Identification of user goals in web search. Paper presented at the World Wide Web Conference (WWW ’05) (pp. 391–401). New York: ACM.

Marchionini, G. (1995). Information seeking in electronic environments. Cambridge, UK: Cambridge Unviersity Press.

Nettleton, D.F., Calderon, L., & Baeza-Yates, R. (2006). Analysis of web search engine query and click data from two perspectives: Query session and document. In T. Eliassi-Rad, L. Ungar, M. Craven, & D. Gunopulos

(Eds.), Proceedings of the12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’06). (pp. 207–226). New York: ACM.

Özmutlu, H.C., Çavdur, F., & Özmutlu, S. (2006). Automatic new topic identification in search engine transaction logs. Internet Research, 16(3), 323–338.

Penniman, W.D. (2008). Historic perspective of log analysis. In B.J. Jansen, A. Spink & I. Taksa (Eds.), Handbook of research on web log analysis (pp. 18–38). Hershey, Pennsylvania: IGI.

Peters, T. (1993). The history and development of transaction log analysis. Library Hi Tech, 42(11), 41–66.

Phippen, A., Sheppard, L., & Furnell, S. (2004). A practical evaluation of Web analytics. Internet Research: Electronic Networking Applications and Policy, 14(4), 284–293.

Rose, D.E., & Levinson, D. (2004). Understanding user goals in web search. In S. Feldman, M. Uretsky, M. Najork, & C. Wills (Eds.). Pro-ceedings of the World Wide Web Conference (WWW ’04) (pp. 13–19),. New York: ACM.

Rozanski, H.D., Bollman, G., & Lipman, M. (2001). Seize the occasion! The seven-segment system for on-line marketing. Strategy and Business, 24(3), 42–51.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society of Information Science, 26(6), 321–343.

Sugar, C., & James, G. (2003). Finding the number of clusters in a data set: an information theoretic approach. Journal of the American Statistical Association, 98(January), 750–763.

Thomas, J.W. (2007). Market segmentation. Retrieved from http:// www.decisionanalyst.com/Downloads/MarketSegm.pdf

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

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How much can behavioral targeting help online advertising? In J. Quemada & G. León (Eds.), Proceedings of the Eighth International Conference on World Wide Web (WWW '09) (pp. 261–270). New York: ACM.

ACADEMIC VITA

Adan Ortiz-Cordova

Education

B.S., Information Sciences and Technology, 2013, The Pennsylvania State University, University Park, PA

Honors in Information Sciences and Technology

M.S., Information Sciences and Technology, 2013, The Pennsylvania State University, University Park, PA

Honors and Awards

Mr. Latino Penn State 2011-2012

Google Scholar 2010 – Hispanic College Fund

Centre County Hispanic Heritage Most Outstanding Student Award

Princeton University Race Relations Prize

Bunton-Waller Scholarship

Renaissance Scholarship

Mary Lou Shuman Trustee Scholarship

Professional Experience

Research Assistant
The Pennsylvania State University
August 2010 – May 2013

Teaching Assistant
The Pennsylvania State University

August 2011 – May 2013

Web Application Developer
BuenaMusica.com
May 2012 – August 2012

Software Engineer – Intern
Cisco Systems – San Jose, CA
May 2010 – August 2010

Hardware Test Engineer – Intern
Cisco Systems – San Jose, CA
May 2009 – August 2009

Systems Administrator – Intern
Chatham Financial – Kennett Square, PA
May 2008 – August 2008

Desktop Support – Intern
Chatham Financial – Kennett Square, PA
April 2006 – August 2007