Marketing Science Institute Working Paper Series 2019
Report No. 19-113

Keyword Selection Strategies in Search Engine Optimization: How Relevant Is Relevance?

Mayank Nagpal and J. Andrew Petersen

Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance?

Mayank Nagpal
J. Andrew Petersen

October 2018

Mayank Nagpal (mayanknagpal@psu.edu) is a doctoral student in Marketing at the Smeal College of Business, The Pennsylvania State University, University Park, PA 16802, USA

J. Andrew Petersen (jap57@psu.edu) is an Associate Professor of Marketing at the Smeal College of Business, The Pennsylvania State University, University Park, PA 16802, USA

We would like to thank a digital media advertising firm for providing the data used in this study. We would also like to thank Arvind Rangaswamy, Gary Lilien, and participants of the 2018 Marketing Science Conference for providing feedback on an earlier version of this paper. Finally, we would like to thank the Marketing Science Institute (MSI) and its Young Scholars program for providing financial support for this research (MSI Grant #4-1921). This research is part of the first author's dissertation.

Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance?

Abstract

Search Engine Marketing (SEM) consists of efforts by a firm to increase clicks to their website through Sponsored Search Advertising (SSA) and Search Engine Optimization (SEO). To date, research in Marketing has focused more on SSA relative to SEO. This seems surprising given that organic search results are considered more trustworthy, account for most of the clicks, and firms spend significantly more on SEO than on SSA ($65B vs. $35B in the US in 2016). To create content for SEO to maximize organic clicks, managers need to identify appropriate keywords on which their content will be focused. Currently, most managers use simple heuristics to identify appropriate keywords and then write relevant content about those keywords. However, it is unclear whether these heuristics are effective techniques for SEO. In this paper we build a framework which provides model-based guidance to SEO practitioners for keyword selection and web content creation. It is often thought that content relevance is a key factor to improve the effectiveness of SEO. We find, however, that while content relevance is an important criterion in a consumer's organic click decision, improving content relevance regardless of the keyword selected may not always be effective. Specifically, we find that when the online authority of a website is more (less) than the average online authority of its competitors, creating relevant content for broader (more specific) keywords is more effective in improving organic rank and organic clicks.

Keywords: Search Engine Optimization, Tobit Model, Latent Semantic Analysis

Firms use Search Engine Marketing (SEM) techniques to promote websites by increasing their visibility on search engine results pages (SERPs). As a large majority of users begin their online browsing experience using search engines, SEM accounts for the largest share (47%) of digital marketing spend (Silverman 2016). SEM aims to increase the prominence of a link on the Search Engine Result Pages (SERP) by appearing higher on either the sponsored and/or the organic portions of the SERP (see Figure 1 for an example). SEM comprises of two parts: (1) search engine optimization (SEO) which aims at getting a higher rank and more clicks from the organic search results on the SERP and (2) sponsored search advertising (SSA) which aims at getting a higher rank and more clicks from the sponsored search results on the SERP.

While significant academic research has focused on SSA (e.g., Skiera and Nabout 2013; Wiesel, Pauwels, and Arts 2011; Li et al. 2016), research on SEO is relatively scarce. This seems surprising, given that organic links are considered to be more trustworthy by the users (Purcell 2012), account for the majority of clicks on a SERP (Baye, De los Santos, and Wildenbeest 2016; Jerath, Ma, and Park 2014), and also get a lion's share of SEM spending ($65 billion[1] against the $35 billion[2] spent on sponsored links in 2016).

Organic links on the SERPs are ranked by search engines using various criteria. These include factors such as authority of the website, quality of the incoming links, and relevance of webpage content to the keyword searched.[3] However, both SEO practitioners and academic researchers (Luh et al. 2015) have found content relevance to be among the top few factors affecting organic rank. According to Google's SEO guidelines, creating compelling and useful

[1] https://www.borrellassociates.com/industry-papers/papers/2016/trends-in-digital-marketing-services-april-16-detail
[2] https://www.statista.com/statistics/266627/projected-spending-on-search-marketing-in-the-us/
[3] https://moz.com/search-ranking-factors

content will likely influence website ranking more than any other factor.[4] Additionally, several market research surveys have found relevant content creation as the most effective SEO tactic.[5] While writing new content on their websites, managers need to select keywords on which they want to focus this content. Ideally, they would want to write content around those keywords which are most effective in getting them the highest potential for organic clicks. To better understand how firms can make better keyword selection decisions for web content creation, we focus this study on the impact of content relevance for a website on the search engine's ranking decision as well as the user's click decision. We try to understand if writing relevant content is effective regardless of the keyword on which the content is focused or whether its effectiveness varies based on the type of keyword and website. Based on the results of this study, SEO managers will be able to better select keywords which are most effective in getting clicks.

Keyword selection, often referred to as keyword research, is based on keyword characteristics (keyword traffic, keyword competition, keyword commercial value), and the keyword market segment to which a keyword relates. To date, it appears that many firms use simplifying heuristics such as setting a minimum traffic threshold and counting the number of main competitors in the market to select their desired keywords. Ideally, a firm would want to select keywords to create web content about with high search volumes and low levels of competition. This would allow the firm to create relevant content, get ranked highly on the SERP, and receive a large share of the large potential number of clicks for that keyword. However, it is uncommon to find such a keyword; keywords that receive higher search volume also tend to have a higher corresponding level of competition (see Table 1 for an example). A broad keyword (i.e., a short keyword that often represents a large topic area) often is searched by

[4] https://support.google.com/webmasters/answer/7451184?hl=en
[5] http://webpromo.expert/google-qa-march/

a larger number of users as compared to a more specific keyword (i.e., a longer keyword that often represents a narrow topic area), but often also has a larger number of websites creating landing pages relevant to it. Thus, firms often face a trade-off between creating pages relevant to broad keywords (potentially getting a small share of a large market) versus specific keywords (potentially getting a large share of a small market).

As an example, a website for a health care provider needs to decide whether it should create a landing page focused on a broader topic such as 'Malaria' or a more specific sub-topic such as 'Effects of Malaria on the Body'. To help the health care provider solve this problem, in this paper we analyze how focusing content on broad keywords compares to that on more specific keywords in terms of affecting organic rank and user click behavior. Accordingly, we propose a modeling framework to study how three key keyword characteristics (popularity, competition, and specificity) as well as two key website characteristics (content relevance and online authority) affect the organic clicks a website receives for a keyword. The framework provides model-based guidance to SEO practitioners in their keyword selection decisions by studying how effectiveness of writing relevant content can vary with keyword and website type.

We first present an overview of the existing literature on both SSA and SEO and build on this by presenting a conceptual model where we present the expected relationships between the keyword and website characteristics and the organic clicks a website receives for a given keyword. Second, we use data for 2,674 search queries relevant to three different firms from three different industries (online retailer, culinary school, and urgent health care provider) to empirically test our conceptual model.

After controlling for the endogeneity of competition and developing a new method for measuring text relevance between keywords and web content, we empirically show how selecting keywords with certain characteristics is related to the rank a website receives for a given keyword and that website's share of organic clicks. This allows us to rank keywords based on the estimated number of clicks a given website will receive. Further, we compare how this effect varies across keyword and website type.

We find that when website authority is more (less) than the average authority of the websites ranked on a keyword, improving relevance is more effective in improving rank when firms target broader (more specific) keywords. We also find that improving relevance affects the probability of a user clicking on the organic link both directly and indirectly through its effect on rank. The direct effect was found to be significant only if the website was ranked on the first page of the results, as getting placed on the first page is important to get noticed by the users.[6]

These findings have implications for both marketing theory and SEO practice. In addition to providing a better understanding of the organic rank generation process and the importance of content relevance in SEO to marketing researchers, the framework provides model-based guidance to SEO practitioners for their keyword selection decisions. The model shows that though content relevance is an important criterion in the consumer's click decision on the SERP, improving relevance may not always be effective in getting organic clicks for all keywords.

Literature Review

To date, extant work on SEM primarily focuses on SSA with little research dedicated so far to SEO. In this literature review we look at relevant research for both SSA and SEO. Table 2

[6] https://chitika.com/2013/06/07/the-value-of-google-result-positioning-2/

provides a summary of the related papers in these fields as well as the relevant contribution from each of the papers.

Research studying the relationship between SSA and SEO campaigns provide mixed evidence about the importance of sponsored results in SEO. On one hand, click-through rates, conversions rates, and revenues in the presence of both paid and organic search listings are significantly higher than those in the absence of paid search (Yang and Ghose 2010). On the other hand, a webpage with high attractiveness, which is likely to rank higher on the organic links, has a lower incentive to bid for sponsored links as consumer trust in sponsored links is lower (Katona and Sarvary 2010). Further, SEO campaigns have been shown to be more cost effective than SSA (Kritzinger and Weideman 2015) and increase consumer satisfaction (Berman and Katona 2013). However, given that the two sets of search results are interconnected, it is important to consider past literature from both these topics as important insights can be obtained from past studies on the sponsored results.

Sponsored Search Advertising (SSA)

The first set of papers in the field of SSA study how link allocation and auction strategies of search engines (Feng et al. 2007; Santos and Koulayev 2013) and bidding and attribution strategies of websites (Skeira and Nabout 2013; Li et al. 2016; Nabout 2015) affect financial performance of firms and consumer satisfaction. Researchers in this field study the importance of incorporating consumer choice (Santos and Koulayev 2013) and content relevance (Feng et al. 2007) in a search engine's ranking procedure along with the importance of website quality improvement (Nabout and Skeira 2012) and keyword heterogeneity (Rutz and Bucklin 2007; Rutz and Bucklin 2012; Kang and Kim 2004) in websites' SSA strategies. These findings

collectively provide insights on how firms could improve rank and get more clicks through SSA. Factors related to auction and bidding strategies are not directly applicable to improving the ranking on organic search results, as the ranking algorithm for organic results does not consider auction bids while ordering results. However, we expect that factors such as website quality, relevance of content and keyword heterogeneity, identified in the SSA literature, will apply to our context of SEO.

Another group of researchers in SSA study how consumer search patterns differ in terms of user expertise (White and Morris 2007; White, Dumais, and Teevan 2009), keyword type (Jerath, Ma, and Park 2004; Agarwal, Hosanagar, and Smith 2011; White, Dumais, and Teevan 2009) and search state, i.e. exploration state and evaluation state (Shi and Trusov 2013). Researchers in this area classify keywords based on the underlying need of the users (Broder 2002; Kim and Kang 2004; Rose and Levinson 2004), popularity (Jerath, Ma, and Park 2014), the stage of the purchase process (Li et al. 2016), user expertise (White and Morris 2007; White, Dumais, and Teevan 2009), and branded versus generic keywords (Rutz and Bucklin 2012). These classifications are useful for studying how different types of keywords can affect SEM strategies of firms. Rutz and Bucklin (2012) showed that incorporating keyword heterogeneity in SSA strategies can be profitable for firms as consumer behavior differs across keywords. Others have shown that even though the higher ranked links have higher CTRs, conversion rates increase with link positions (Agarwal, Hosanagar, and Smith 2011; Ghose and Yang 2009) as a larger proportion of users clicking the lower ranked links have high purchase intent. The increase in conversion rate is higher for more specific keywords (Agarwal, Hosanagar, and Smith 2011) as users tend to spend more effort in finding relevant links and thus click on results much lower in the sponsored results. Even while noting that findings from sponsored listing may not be directly

applicable to SEO (due to differences in competitive strategies of firms and user beliefs about sponsored versus organic links), we see that keyword types play a crucial role in SEO strategy.

Finally, the importance of market niches (Kotler 2003; Shani and Chalasani 1992; Toften and Hammervoll 2008; Dalgic and Leeuw 1994) is noted in SEM. Skeira et al. (2010) define long tail keywords as those which are searched for by fewer users but with a high probability of conversion. Accordingly, we define specific keywords based on whether they represent a more niche market or a broader market and study their importance in the context of SEO.

Search Engine Optimization (SEO)

Research on SEO is scarce likely due to the lack of publicly available data for important variables such as clicks on each link, the complexity of the ever-changing ranking algorithms, and the difficulty in measuring important variables such as the semantic relevance of website content. Extant research identifies the most important SEO strategies. For example, Baye, De los Santos, and Wildenbeest (2016) find that investments in quality and brand awareness increases organic traffic to a website both directly, by influencing consumer behavior on the SERP, and indirectly by improving rank or the prominence of a link on the SERP. In addition to website quality and brand related factors such as PageRank (Page and Brin 1998) and website authority, studies find content related factors such as the content relevance of the title and the snippet as the most important factors in determining organic ranks on Google SERPs (Luh et al. 2015). Additionally, other studies suggest improvements in SEO by incorporating semantic factors (Mavridis and Symeonidis 2015) and consumer information needs (Liu and Toubia 2015).

These papers study the effect of factors affecting organic rank and how SEO techniques such as investment in improving brand awareness (Baye, De los Santos, and Wildenbeest 2016) and content relevance (Luh et al. 2015) can help firms. In contrast, we study how the effect of these

content and quality improvement SEO techniques on organic clicks varies across the type of keyword. By doing so, we aim to provide model-based guidance to firms in making keyword selection choices based on keyword type. A key difficulty for any firm to determine the most important factors to be used while selecting keywords to get a high organic ranking is that not only do search engines use many factors while ranking the organic list, but also continuously keep updating their ranking algorithm (Evans 2007). We aim to do this by building a modeling framework which focuses on certain fundamental website and keyword characteristics, the importance of which is less likely to change over time.

Conceptual Model

When writing new web content, firms need to select keywords on which to focus content. If we consider each keyword as a separate "market" of consumers, then the decision of selecting a keyword is analogous to selecting a target market. Past research has shown that the most important factors which firms consider when selecting the target market are market size (Abratt 1993; Scaperlanda and Mauer 1969), level of competition in the market, and nature of customer needs (Abratt 1993).

In our conceptual model, we assume that firms who invest in SEO select keywords which are likely to get them the maximum number of clicks. The number of organic clicks attributable to a SERP is equal to the number of users who search for the keyword (market size) times the share of those users who click on the website link (market share). As the share of users who click on a link is largely dependent on the organic rank (Baye, De los Santos, and Wildenbeest 2016; Feng, Bhargava, and Pennock 2007; Shi and Trusov 2013), firms aim to improve their rank on the SERP to get a larger number of clicks. To capture these forces, our conceptual model (see Figure 2) includes a rank generation process (modeling factors determining the search engine's decision

to assign rank), the share – rank link (modeling a consumer's decision on which link to click from the organic results), and the clicks – share link (calculating the number of clicks).

Rank Generation Process

In our conceptual model, we first look at how website characteristics and keyword characteristics interact to affect rank. The interaction effects help us analyze whether the importance of writing relevant content varies with the type of keyword and website.

We consider two website characteristics in our conceptual model, online authority and content relevance. Online authority represents the overall quality and popularity of the website in its domain of expertise.[7] Content relevance represents the degree of overlap between the webpage content and the keyword searched. As both these characteristics make a website more attractive to the users searching for any keyword, we expect that an improvement in either of the two leads to a higher rank (Mavridis and Symeonidis 2015; Liu and Toubia 2015).

Turning to keywords characteristics, we focus on keyword competition, or the level of competition in the keyword market (market competition), keyword popularity, or the number of users who search for the keyword (market size), and keyword specificity, a measure of how broad or specific a keyword is (Li et al. 2016) within a product category (niche vs. broad market). We expect keyword specificity and keyword competition to directly affect rank. A firm selecting a keyword with high competition should expect to get a lower rank on SERP. For example, a firm should expect to get a higher rank on a less competitive keyword such as "symptoms of type 1 diabetes in a child" in comparison to a more competitive keyword such as "Type 1 diabetes", which has a higher number of firms competing for it (see Table 1 for details).

[7] https://www.searchenginejournal.com/seo-guide/search-authority/

Specific keywords are niche segments (Skiera et al. 2010), where customers have a distinct set of needs and pay premium to firms which best satisfy their needs (Kotler 2003). Researchers have explained how niches have greater growth and profit potential for firms due to economies of specialization (Kotler 2003; Shani and Chalasani 1992; Toften and Hammervoll 2008; Dalgic and Leeuw 1994). Thus, making content relevant to a more specific keyword should get a higher rank on the SERP.

In contrast to the other two keyword characteristics which have direct effects on rank, keyword popularity has an indirect effect on rank through its effect on keyword competition. Selecting a more popular keyword would indirectly lead to lower rank as more firms would compete for the large user base of these keywords. For example, in Table 1, a less popular keyword such as "symptoms of type 1 diabetes in a child" has a lower number of firms competing for it in comparison to a more popular keyword such as "Type 1 diabetes".

Moderating Role of Keyword Specificity and Online Authority

Past literature has shown that users searching for more specific keywords are more advanced in the search process (Jerath, Ma, and Park 2014; White, Dumais, and Teevan 2009), have high purchase intent (Moe 2003), and use more specific search queries such as "symptoms of type 1 diabetes in a child" rather than a general search of "Type 1 diabetes". As these users provide more details of their needs, search engines can provide more relevant content in their search results. However, when the search queries provided by the users are more general, we expect that search engines give greater importance to the quality of the site and the page as the user has not specified his exact needs. As users searching for specific keywords are more advanced and involved in their search, they tend to click on results much lower in the sponsored list, spending more effort in finding relevant links (Agarwal, Hosanagar, and Smith 2011). Thus, we expect

that content relevance is much more important to these users compared to users searching for broad keywords. As search engines try to mimic user preferences (Joachims 2002; Granka, Joachims, and Gay 2004; Broder 2002), it is expected that search engine would assign weights based on what the consumers are more attracted to, giving more weight to online authority for broad keywords as compared to content relevance for specific keywords. We provide some evidence for this in Table 3.

Table 3 shows that for a broader search query or keyword, the top ranked sites have greater page and domain authority, even if the semantic similarity between the webpage title and the query is not as high as the lower ranked pages. However, for more specific keywords or search queries, the semantic similarity between the title and the search query is high for higher ranked sites, even if the authority of the page is not very high.

This suggests that not only do keyword specificity and online authority affect the rank directly, but they interact together to moderate the effect of content relevance on rank. Thus, we expect that the importance of content relevance in determining rank is greater for more specific keywords as consumers searching for such queries are more informed and explicitly define their needs. Further, the importance of online authority in determining rank is greater for less specific keywords, as the consumers searching for such keywords being at an early stage of the search process are still trying to understand their exact needs. To analyze how these effects moderate the influence of content relevance on rank, we also explore the relationship between online authority, keyword specificity, and content relevance.

Share-Rank Link

The share-rank link represents the decision of the user to click on any link after the search results have been generated by the search engine. We expect users to click more on higher ranked links as they are visually more prominent on the SERP (Baye, De los Santos, and Wildenbeest 2016; Feng, Bhargava, and Pennock 2007; Shi and Trusov 2013). In addition to considering Rank as an indicator for a website's usefulness, users will also consider whether the content of the website seems relevant. Thus, content relevance of the website affects the user's click decision directly as well as indirectly through its effect on rank.

We also include an interaction effect of content relevance with a dummy variable, First Page. The dummy variable indicates whether the focal website is ranked on the first page of the results or not. A large majority of the users do not go beyond the first page of the results in their search process.[6] As content relevance will affect a user's click decision only if it seen by the user, it is expected that content relevance may have a lower effect in case the website is ranked lower than the first page.

Clicks – Share Link

The total number of organic clicks is mathematically equal to the share of users who click on the link multiplied by the total number of users who search for the keyword (i.e., keyword popularity). Thus, keyword popularity affects the number of organic clicks both directly and indirectly (see Figure 2). The positive direct effect is due to the larger market size (number of users) of popular keywords. A larger market size provides a greater number of potential clicks. A firm will get more clicks if its market share, or the share of users who click on the website, remains the same. This directly increases organic traffic on the website from the keyword. On the other hand, the negative indirect effect is due to its negative effect on rank caused by

increased competition in the keyword market. This, in turn, leads to a smaller market share from any given keyword, indirectly leading to a decrease in organic clicks on the website.

Model Development

In this section, we provide the methodology used for empirically validating the conceptual model presented in Figure 2. The full conceptual model translates to Equations 1 to 4. The first two equations (Eq. 1 and 2) associate with the two-stage rank generation process. Equation 1 looks at the effect of keyword popularity on keyword competition, whereas Equation 2 captures the effects of keyword and website characteristics on organic rank. Equation 3 captures the effect of organic rank and content relevance on share. Finally, Equation 4 captures the total number of organic clicks it gets from that SERP based on the share of clicks the website receives and the total search traffic for that search query.

Part 1: Rank Generation Models

$$Keyword\ Competition_k = f(Keyword\ Popularity_k, Keyword\ Specificity_k, Online\ Authority\ Diff_{ik}, Content\ Relevance\ Diff_{ik}, F) + \eta_k \quad (1)$$

$$Rank_{ik} = g(Keyword\ Competition_k, Keyword\ Specificity_k, Online\ Authority\ Diff_{ik}, Content\ Relevance\ Diff_{ik}, F) + \varepsilon_{ik} \quad (2)$$

Part 2: Share-Rank Model

$$Share_{ik} = h(Rank_{ik}, Content\ Relevance\ Diff_{ik}, First\ Page_{ik}, F) + \nu_{ik} \quad (3)$$

Part 3: Clicks-share Equation

$$Organic\ Clicks_{ik} = Share_{ik} * Keyword\ Popularity_k \quad (4)$$

where,

- $Rank_{ik}$ is the rank of website i on the organic list on SERP for search query k
- $Keyword\ Competition_k$ is the level of competitiveness for search query k
- $Keyword\ Specificity_k$ is the specificity of search query k
- $Keyword\ Popularity_k$ is the number of users who searched for search query k.
- $Online\ Authority\ Diff_{ik}$ is the relative online authority for website i on search query k
- $Content\ Relevance\ Diff_{ik}$ is the relative content relevance of website i to search query k
- $Share_{ik}$ is the share of organic clicks for website i for search query k
- $First\ Page_{ik}$ is 1 if website i appears on page 1 of the SERP for search query k; 0 otherwise

- *F* represents the firm fixed effects

Modeling Challenges

There are several modeling challenges that need to be addressed if we want to estimate the three models in Equations 1 to 3. These include the endogeneity of keyword competition, unobserved heterogeneity, limited dependent variables, and the correlation across equations.

Endogeneity

As stated earlier, keyword popularity affects rank indirectly through its effect on competition. This makes keyword competition endogenous in the model as it is determined by the popularity of the keyword. Moreover, firms which target competitive keywords are expected to spend more effort and more resources since ranking for such keywords requires more effort and resources. As this can directly affect the ranking for these firms, keyword competition is likely correlated with the error term and may lead to biased estimation of the parameters. To account for this endogeneity of keyword competition, we model Equations 1 and 2 using an instrumental variable approach by using keyword popularity as an instrument for keyword competition. We expect that keywords which are more popular will be targeted by a larger number of websites. This means that there should be a positive relationship with keyword competition. We also expect that the popularity of the keywords should not directly impact the rank of any webpage on the SERP. Thus, we believe keyword popularity is an appropriate instrument for this estimation.

Unobserved Heterogeneity

The data has unobserved heterogeneity among firms as well as among the keywords. There are unobserved differences among firms caused due to various factors such as varying profitability and competition in specific industries to which firms belong. To account for this unobserved heterogeneity of firms we use firm fixed effects in the models. Also, there are

inherent differences among keywords. We control for the observed differences by using the three keyword characteristics in our models. However, to account for the unobserved differences, we use relative measures of the two website characteristics, i.e. online authority and content relevance. For this, we subtract the average value of these two measures for all other websites ranked on the top three pages for a keyword from the value of these for the focal websites.

Limited dependent variables

The dependent variables in Equations 3 and 4 are limited dependent variables. Keyword competition is censored below at 0, whereas organic rank is censored above at the maximum rank that is available in the data (i.e., 30) and below at the minimum rank a webpage can achieve (i.e., 1). Thus, estimating the models using OLS may lead to biased estimates (Heckman 1976). To overcome the issue of censoring in the dependent variables, we use the Tobit Model for estimating Equations 1 and 2, which is an approach proposed by Tobin (1958) to model limited dependent variables.

Correlation Across Equations

Each of the three dependent variables (keyword competition, rank, and share) are part of the same overall process. As such, it is likely that Equations 1 to 3 are inherently related. As a result, we will estimate the equations jointly using a Conditional Mixed Process (CMP) in Stata (Roodman 2017).

Rank Generation Models

The mathematical representation of the two-step rank generation process is provided in Equations 5 and 6. Equation 5 represents the effect of the instrumental variable (keyword popularity) on competition and Equation 6 represents the direct and interaction effects of all

keyword and website characteristics on the rank of the website for a keyword. In each case we provide the actual model specification that is estimated.

$$log(Keyword\ Competition_k) = f1(\alpha_1 + \alpha_2 * log(Keyword\ Popularity_k) + \alpha_3 * Online\ Authority_k^* + \alpha_4 * Content\ Relevance_k^* + \alpha_5 * Keyword\ Specificity_k + F) + \eta_k \quad (5)$$

$$Rank_{ik} = f2(\beta_1 + \beta_2 * \log(Keyword\ Competition_k) + \beta_3 * Online\ Authority\ Diff_{ik} + \beta_4 * Content\ Relevance\ Diff_{ik} + \beta_5 * Keyword\ Specificity_k + \beta_6 * Keyword\ Specificity_k * Online\ Authority\ Diff_{ik} + \beta_7 * Keyword\ Specificity_k * Content\ Relevance\ Diff_{ik} + \beta_8 * Online\ Authority\ Diff_{ik} * Content\ Relevance\ Diff_{ik} + \beta_9 * Online\ Authority\ Diff_{ik} * Content\ Relevance\ Diff_{ik} * Keyword\ Specificity_k + F) + \xi_{1ik} \quad (6)$$

where,

- $Rank_{ik}$ is the rank of website i on the organic list on SERP for search query k
- $Keyword\ Competition_k$ is the level of competitiveness for search query k
- $Keyword\ Specificity_k$ is the specificity of search query k
- $Keyword\ Popularity_k$ is the number of users who searched for search query k.
- $Online\ Authority\ Diff_{ik}$ is the relative online authority for website i on search query k
- $Content\ Relevance\ Diff_{ik}$ is the relative content relevance of website i to search query k
- *F* represents the firm fixed effects
- *f1(.) (f2(.))* is the one-sided (two-sided) Tobit functional form

Share-Rank Model

As one of the major objectives of this paper is to understand how content relevance of the website affects the share of clicks that the website gets from a given keyword, we include the direct effect of content relevance on a customer's click behavior. We also expect the effect of content relevance on the click decision to vary based on where the website is located on the SERPs. Thus, we include an interaction effect of content relevance with a dummy variable, First Page. A large majority of the searchers do not go beyond the first page of the results in their search process.[6] As such, content relevance will likely affect a user's click decision more if it seen by the user. We provide the actual model specification that is estimated for the click share model in Equation 7.

$$Share_{ik} = \gamma_1 + \gamma_2 * Rank_{ik} + \gamma_3 * First\ Page_{ik} + \gamma_4 * Content\ Relevance\ Diff_{ik} + \gamma_5 * Content\ Relevance\ Diff_{ik} * First\ Page_{ik} + \nu_{ik} \quad (7)$$

where,

- $Share_{ik}$ is the share of organic clicks for website i for search query k
- $Rank_{ik}$ is the rank of website i on the organic list on SERP for search query k
- $Content\ Relevance\ Diff_{ik}$ is the relative content relevance of website i to search query k
- $First\ Page_{ik}$ is 1 if website i appears on page 1 of the SERP for search query k; 0 otherwise
- F represents the firm fixed effects

Calculating Expected Organic Clicks

The three equations (Equations 5 to 7) above are estimated to obtain the expected share of clicks for a website from any given search query based on website and keyword characteristics. The total number of expected clicks a website gets from a given search query can then be calculated as the product of this estimated share of clicks and keyword popularity.

$$E(Organic\ Clicks_{ik}) = \widehat{Share}_{ik} * Keyword\ Popularity_k \quad (8)$$

The estimated or the expected clicks can be used as a measure of how lucrative a keyword is to a firm and its website. The full model (Equations 5 to 8) can be used by firms in their keyword research campaigns to identify the most lucrative keywords based on the relationship between the keyword characteristics studied (keyword popularity, keyword competition and keyword specificity) and the expected organic clicks, given the characteristics of the website.

Empirical Application

Data Description

We empirically validate the relationships described in the conceptual model using data for three firms from three different industries. The dataset contains information on organic clicks on the three websites for search queries relevant to these three firms and their main competitors for a given month. The firms include an online retailer, a culinary school, and an urgent health care provider. We have data for the first 30 links of the SERP for 1475, 505, and 705 search queries respectively for the three firms. We use data from the first three pages as these pages typically account for more than 90% of the clicks (Moz Study 2015) from SERPs. The data contains

information on search query traffic and the number of organic clicks for the focal firm accumulated over a month, along with the Cost Per Click (CPC). It also contains information about the domain and page authority of the first 30 ranked links for each keyword. While the data contains actual search queries (and not keywords), these queries are user manifestations of underlying keywords which firms write content about. Thus, we expect the measures derived from keywords and search queries to be highly correlated. Table 4 provides a list of all variables used in the model along with their descriptions and sources and Table 5 provides some descriptive statistics and correlations of the variables.

Variable Operationalizations

Dependent Variables. Organic clicks are measured as the number of clicks received by a website from the organic list on the SERP. Click share is the total organic clicks on the website represented as a proportion of keyword popularity. Organic rank is the minimum rank of a website on the SERP of a search query.[8]

Online Authority (Diff). The online authority of a webpage is defined as the standing or the impact the page has in its field of expertise (Kleinberg 1999). We use two metrics, domain authority and page authority to derive our measure of online authority.[9] Moz's Page/Domain Authority is a metric on how high a given webpage/domain is likely to rank in search results regardless of its content. It is based on the Linkscape web index and includes link count,

[8] As the data contains instances where multiple webpages from the same domain are ranked for the same search query, we take the minimum rank obtained by any webpage from the firm as the rank of the page of that website. The minimum rank is used for the estimation as we assume that all the organic clicks received by the website from the SERP, come from the first instance the user sees a link from the domain or the website.

[9] The domain authority part of online authority represents the overall authority the domain is perceived to have across all its webpages. As a robustness check we found that the domain authority as computed by Moz creates the same rank order in terms of domain rank as we find from the website ranking service Alexa.com. We use Moz to create our measure of online authority as it includes a composite of both the domain and webpage authority whereas Alexa.com only provides a measure of the domain rank.

mozRank and several more metrics. The highest score is achieved for pages/domain that are heavily linked and for pages that are near to the top of SERPs. They are aggregates of several other metrics including MozRank, MozTrust, quality of the link profile, and other factors which are known to affect rank of the website. They are represented as integer values from 1 to 100 on a logarithmic scale and are calculated by combining more than 40 parameters into a single score. Given that the authority metrics comprise of several other important variables, they measure the overall quality of the site and the page. In a survey study conducted by SEOMoz surveying over 150 leading search marketers, it was found that Authority factors were considered most important among the 90 ranking factors surveyed. We create an online authority index from these two metrics by performing a Principal Component Analysis (PCA) on the two authority factors (Page Authority and Domain Authority) and taking the first component as the measure of online authority of a website. In our model, we consider online authority as a relative measure, i.e. we compare the online authority of a given website against the average online authority of the competitor websites on the keyword. We use the following formula to measure a given website's online authority relative to competing websites:

$$Online\ Authority\ Diff_{ik} = Online\ Authority_{ik} - \overline{Online\ Authority_{k}} \quad (9)$$

Here we subtract the average online authority for all websites for a given search query k which appear on the first three pages of a SERP from the online authority of the focal website i for search query k.

Content Relevance (Diff). The content relevance of a webpage to a keyword is the degree of similarity between the content of the webpage and the keyword or the degree of relevance of webpage content to keyword. There are a number of different textual relevance measures such as the General edit distance or the Levenshtein edit distance (Levenshtein 1966) which measure the

textual similarity of two phrases. However, these measures are not the most appropriate measures when measuring content relevance in the context of search engines as they do not consider the semantic relationship between phrases, as semantics play a very important role in SEM (Mavridis and Symeonidis 2012). Other researchers in the field of SEM have incorporated semantics in their ranking measures by using techniques such as Latent Semantic Analysis (Luh et al. 2015) and Latent Dirichlet Allocation (Mavridis and Symeonidis 2012; Mavridis and Symeonidis 2015; Liu and Toubia 2015). For calculating content relevance between a query or a keyword and the title of any document, we adapt the method used by Luh et al. (2015). Figure 3 presents the entire process we follow for calculating the content relevance of the webpage to the search query. A more detailed description of this process has been provided in ***Web Appendix A***.

The method for measuring content relevance involves partitioning the terms in the title of the webpage into two groups: query terms which include terms in the title which are present in the search query and non-query terms which include terms in the title not present in the query. The relevance score for the query and non-query terms is calculated separately and the overall content relevance is the sum of the non-query score and the query score.

The query score is calculated in three steps. First, we calculate the prominence score which measures how prominent the query terms are in the title. Second, we calculate the proximity score which measures the degree to which all terms of the query or sub-query occur together in the title. Third, we calculate the overall query score as the product of the prominence and proximity scores.

The second part of the relevance score, i.e. the score for the non-query terms is calculated based on the semantic relationship between the non-query and query terms derived using Latent

Semantic Analysis (LSA) which is a statistical technique for extracting and inferring relations of contextual usage of words in documents by using singular value decomposition. The overall non-query score between the title and a search query or a keyword is calculated as the average semantic relationship score (described in detail in ***Web Appendix A***) between each non-query term and each term in the keyword.

Like online authority, we operationalize content relevance as a relative measure, i.e. we compare the content relevance of the focal website against the average content relevance of the competitor websites for a given search query. Thus, we use the following formula to calculate content relevance:

$$Content\ Relevance\ Diff_{ik} = Content\ Relevance_{ik} - \overline{Content\ Relevance_{k}} \quad (10)$$

Here we subtract the average content relevance for all websites for a given search query k which appear on the first three pages of a SERP from the content relevance of the focal website i for search query k.

Keyword Specificity. Keyword specificity is the specificity or broadness level of the keyword. We use the number of terms in the keyword as a measure for specificity as a longer keyword is typically more specific (Ghose and Yang 2009;Rutz and Bucklin 2011; White, Dumais, and Teevan 2009). However, before calculating the number of terms in the keyword, we used a commonly used stop word list (Page Analyzer English Stop Words List) and remove any words in this list from our search query before computing keyword specificity.

Keyword Popularity. Keyword popularity represents the total number of users who search for a keyword. It is measured using the estimated number of searches for a given keyword by Google AdWords.

Keyword Competition. Keyword competition, which represents the level of competition in the keyword market, is measured by using the Cost per Click (CPC) or bidding price for each keyword. CPC is the average price in dollars that a website must pay for each click obtained from the sponsored links on the SERP for that given search query.

First Page. We measure First Page as a dummy variable which equals 1 if the website is ranked as 1 of the first 10 links on the SERP, i.e. shows up on the first page, or 0 if the website does not appear in the first 10 links on the SERP.

Results

We provide the results for the joint estimation in Table 6. The table provides coefficient estimates and standard errors for coefficients in all the three stages of the model. As expected, we find that the coefficient for keyword popularity in the first stage of the IV regression is positive ($\alpha_2 = .095$; s.e. = .009) and statistically significant. Also, in the second stage we find that keyword competition affects the rank positively ($\beta_2 = 6.361$; s.e. = .861), where a more positive rank number means a website is farther down the SERP (i.e., 1 is the highest rank). To understand the effect of the rest of the variables (keyword specificity, online authority and content relevance) on rank is more complex as in addition to the main effects of these variables, the model also includes the two-way and the three-way interactions among these variables. Thus, to understand the effect of the variable of our interest, i.e. content relevance, from our estimated model, we calculated the marginal effects of content relevance at different levels of online authority and keyword specificity.

Figure 4 provides the marginal effect of content relevance across different keyword lengths for websites having high (+1 s.d. from the mean) and low (-1 s.d. from the mean) online

authority. The figure shows that for low online authority websites, the marginal impact of writing relevant content on improving rank is higher for more specific keywords. On the other hand, for high authority websites, the marginal impact of writing relevant content on improving rank is higher for broader keywords. This means that high authority websites see a greater improvement in rank (i.e., getting a lower rank) by writing content focused on broader keywords (μ = -9.39; s.e. = 1.62) as compared to more specific keywords (μ = 3.49; s.e. = 2.86) whereas low authority websites see a greater improvement in rank (i.e., getting a lower rank) for more specific keywords (μ = -4.01; s.e. = 3.70) compared to more broad keywords (μ = 2.66; s.e. = 1.88). This suggests that in terms of getting a better rank, a high (low) online authority website would be better off by writing content for broader (more specific) keywords.

The third part of the model looks at the click decision of the users on the SERP. The results show that content relevance of a website affects the probability of the user clicking on the link of that website both directly as well as indirectly. The indirect effect is through the effect of content relevance on rank (from the rank model) as a user is more likely to click on a website which is ranked higher as shown by the significant effect of rank on share (γ_2 = -0.038; s.e. = 0.006). The results also show evidence of the direct effect of content relevance on the user's click decision. We see that content relevance has a direct and positive impact on click share (γ_3 = 0.237; s.e. = 0.089). Additionally, this impact is higher when the website is ranked on the first page of the search engine results shown by the positive and significant coefficient of the interaction term of content relevance and first page (γ_4 = 0.187, s.e. = 0.097). This is likely because users are more likely to click on a link on the first page rather than navigate to the second page of the SERP in search of a link to click. Thus, if the website is not ranked high enough to be placed on the first

page of the SERP, writing very relevant content may not be as effective of a strategy for getting clicks from the SERP.

Discussion

To provide some additional insights from our findings, we used the estimates of our model to compare the keyword selection decisions of a high online authority website and a low online authority website. For this we consider two websites: a well-known and reputed website in the health care category (cdc.gov) with Page and Domain authority of 78 and 98 respectively, and a relatively less known website in the health care category (malaria.com) with Page and Domain authority of 13 and 26 respectively. Using the estimates of our model we calculated the expected rank, share, and organic clicks for these two websites if they were to write content relevant for a specific and broad keyword. As an example, we take two keywords relevant to these two websites, the broad keyword "Malaria" and the specific keyword "Effects of Malaria Parasite on the Body". We compare the expected values of the three outcome variables to see whether these websites are better off writing content on the broad or the specific keyword. The three Figures (5, 6, and 7) compare these values for these two keywords for the high and low online authority websites.

We see in Figure 5 that the low online authority website (malaria.com) gets a better expected rank by writing content about the specific keyword "Effects of Malaria Parasite on the Body" (8 versus 19), whereas the high online authority website (cdc.gov) gets a better expected rank by writing content about the broader keyword "Malaria" (4 versus 6). We see in Figure 6 that the low online authority website (malaria.com) gets a higher expected share by writing content about the specific keyword "Effects of Malaria Parasite on the Body" (26% versus 2%), whereas the

high online authority website (cdc.gov) gets a higher expected share by writing content about the broader keyword "Malaria" (63% versus 47%). Finally, we see in Figure 7 that the low online authority website (malaria.com) gets a higher number of expected organic clicks by writing content about the specific keyword "Effects of Malaria Parasite on the Body" (1,300 versus 1,000 or 30% more clicks), whereas the high online authority website (cdc.gov) gets a higher number of expected organic clicks by writing content about the broader keyword "Malaria" (31,500 versus 2,350 or 1,240% more clicks). This provides some further evidence that websites with higher (lower) online authority benefit significantly more when they target broader (more specific) keywords.

Implications to Marketing Theory and Practice

Keyword research forms an integral part of both SSA and SEO. For SSA, firms select keywords for auction bids, whereas in SEO, website content is built around the selected keywords. Though significant research has been done about keyword research in SSA, such literature in SEO is scarce. This has led many SEO strategies to rely on sets of common heuristics to select keywords for SEO. In this paper, we build a modeling framework to study the effect of keyword and website characteristics on rank, share, and organic clicks to understand how SEO strategies can vary with keyword type. Specifically, we establish the link between two keyword characteristics (keyword popularity and keyword specificity) and two website characteristics (content relevance and online authority) with the rank, share, and organic clicks obtained by a website from firm for a given search query. The findings of this paper have implications for both marketing researchers and practitioners in the field of SEO.

Implications to Theory

This paper enhances the research available in the field of keyword research for SEO as it investigates the relationships between keyword and website characteristics and the expected rank, share, and organic clicks. The modeling framework studies two important parts of the organic click generation process. It studies how the type of selected keyword influences organic rank and then further studies how the rank a website receives translates into organic clicks. It provides researchers with a broad outlook on how these relationships work and how understanding these relationships can help in selecting appropriate keywords for content optimization.

We show that the type of keyword the firm selects to write content about not only influences the rank directly, but it also influences the ranking process which search engines use to order organic links. Results from our model show that online authority (content relevance) is more important for getting a higher rank for broader (more specific) keywords. This is likely because users searching for more specific keywords are typically more involved and advanced in the purchase process (Jerath, Ma, and Park 2014; White, Dumais, and Teevan 2009). Such users which have a high purchase intent tend to be more focused in their search (Moe 2003) and are known to submit longer and specific queries (White, Dumais, and Teevan 2009; White and Morris 2007). As these users are more advanced in the purchase process, they are likely more well-informed about products, and thus provide more specific information that is more relevant to their needs. Thus, search engines find it easier to provide the most useful suggestions by matching the website content to the specified queries. However, when the search query is broad, users are typically in the early stages of the purchase process and unsure of their exact needs. Such users likely give less importance to content relevance as the keyword searched does not

represent their exact needs. In such a case, search engines try to help the users advance in their search process by suggesting them the most well-known websites in their area of interest.

Further, we show that relevant content influences a user's decision to click on a link directly as well as indirectly through its effect on organic rank. Further, we provide evidence for the importance of getting ranked highly – as even the most relevant website will not receive many clicks if it is not found on the first page of the SERP.

Implications to SEO Practice

In this paper, we study how the effectiveness of SEO strategies involving improvement of online authority and content relevance can vary across broad and specific keywords. While improving online authority is akin to brand-building and could be viewed as long-term investment, increasing content relevance to selected keywords can improve rank and organic clicks immediately. The managerial importance of the paper stems from the fact that keyword research is one of the primary methods used by SEO marketers to enhance search traffic. The modeling framework in this paper can be used by SEO practitioners to understand how the characteristics of the selected keywords will affect the expected clicks the firm gets. This would provide guidance to identify the type of keyword characteristics which would maximize the expected number of organic clicks to their websites and select appropriate keywords to target accordingly.

The paper also sheds light on the moderating influence of keyword specificity and online authority on the effectiveness of writing relevant content. We find that for low authority websites, improving content relevance is more effective in getting a higher rank for specific keywords, whereas, for high authority websites, improving content relevance is more effective for broader keywords. The finding suggests that a website having higher (lower) online authority

would be better off targeting broader (more specific) keywords to increase their expected organic clicks. This finding provides some evidence to help solve the common tradeoff firms face between selecting market strategies which focus on trying to get higher market share from smaller markets (specific keywords) or a smaller market share from larger markets (broad keywords). This is in accordance with previous niche marketing research which discussed how smaller firms are better off targeting more niche markets as they have the ability to capture a larger share of the smaller market (Dalgic 1993; Ferguson and Morris 2002).

Limitations and Future Research

As with any empirical analysis there are several limitations of our study. Our analysis is based on data from a single snapshot of search and click behavior where we focus on explaining variation across keywords, websites, and domains. It would be useful for future studies to examine these relationships using data across multiple time periods to see whether changes in firm strategies to target different types of keywords led to differences in the impact of content relevance on organic click.

Additionally, we use organic clicks as our outcome variable. It is not always the case that the first click on a SERP leads the searcher to the content that they want to consume or the retailer they want to purchase from. Future research could use data related to conversion rates to analyze the effect of content relevance on the conversion rate or the likelihood of purchase on the website for different types of websites and keywords. Users are expected to complete a purchase only if the relevance of the link to their requirements is high. Our finding suggests that the top ranked links for broad keywords may not be very relevant to the user's requirements as search engines give less importance to keyword relevance for these keywords. Thus, it would be interesting to

see if the conversion rate for broader keywords is higher for lower ranked websites compared to the higher ranked websites even though they get a smaller proportion of clicks.

Future research may also want to explore if the findings from this paper can be applied in the offline market as well. Research may be conducted to test whether the brand value of a product is a more important determinant of market share in a broader market as compared to a smaller market. As a larger market consists of a more diverse set of consumers who are not very aware of all the functionalities of a product, they will likely be more concerned about brand equity (like online firms with higher online authority). A more niche market is expected to have a smaller, more homogeneous set of consumers who are aware of the functionalities they need in the product and care less about the brand equity if they can get their required functionalities from the product. Thus, we would expect that a brand with higher (lower) brand equity is likely better off targeting a larger (smaller), more (less) diverse market to get the best sales results.

References

Abou Nabout, Nadia (2015), "A novel approach for bidding on keywords in newly set-up search advertising campaigns," *European Journal of Marketing*, 49 (5/6), 668-91.

Abratt, Russell (1993), "Market segmentation practices of industrial marketers," *Industrial marketing management*, 22 (2), 79-84.

Agarwal, Ashish, Kartik Hosanagar, and Michael D Smith (2011), "Location, location, location: An analysis of profitability of position in online advertising markets," *Journal of marketing research*, 48 (6), 1057-73.

Bar-Isaac, Heski, Guillermo Caruana, and Vicente Cuñat (2009), "Costly search and design."

Baye, Michael R, Babur De los Santos, and Matthijs R Wildenbeest (2016), "Search engine optimization: what drives organic traffic to retail sites?," *Journal of Economics & Management Strategy*, 25 (1), 6-31.

Berndt, ET (1991), "The Practice of Econometrics: Classic and Contemporary. Massachusetts Insitute of Technology and the National Bureau of Economic Research," Addison-Wesley Publishing Company, Inc.

Berman, Ron and Zsolt Katona (2013), "The role of search engine optimization in search marketing," *Marketing Science*, 32 (4), 644-51.

Broder, Andrei (2002), "A taxonomy of web search," in ACM Sigir forum Vol. 36: ACM.

Dalgic, Tevfik (1998), "Dissemination of market orientation in Europe: a conceptual and historical evaluation," *International Marketing Review*, 15 (1), 45-60.

Dalgic, Tevfik and Maarten Leeuw (1994), "Niche marketing revisited: concept, applications and some European cases," *European Journal of Marketing*, 28 (4), 39-55.

De los Santos, Babur and Sergei Koulayev (2013), "Optimizing click-through in online rankings for partially anonymous consumers," *Unpublished manuscript*.

Evans, Michael P (2007), "Analysing Google rankings through search engine optimization data," *Internet research*, 17 (1), 21-37.

Feng, Juan, Hemant K Bhargava, and David M Pennock (2007), "Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms," *INFORMS Journal on Computing*, 19 (1), 137-48.

Ferguson, Charles H and Charles R Morris (2002), Computer Wars: The Post-IBM World: Beard Books.

Ghose, Anindya and Sha Yang (2009), "An empirical analysis of search engine advertising: Sponsored search in electronic markets," *Management Science*, 55 (10), 1605-22.

Granka, Laura A, Thorsten Joachims, and Geri Gay (2004), "Eye-tracking analysis of user behavior in WWW search," in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval: ACM.

Heckman, James J (1976), "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," in *Annals of Economic and Social Measurement, Volume 5, number 4*: NBER.

Jerath, Kinshuk, Liye Ma, and Young-Hoon Park (2014), "Consumer click behavior at a search engine: The role of keyword popularity," *Journal of Marketing Research*, 51 (4), 480-86.

Joachims, Thorsten (2002), "Optimizing search engines using clickthrough data," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining: ACM.

Kang, In-Ho and Gil Chang Kim (2004), "Integration of multiple evidences based on a query type for web search," *Information processing & management*, 40 (3), 459-78.

Katona, Zsolt and Miklos Sarvary (2010), "The race for sponsored links: Bidding patterns for search advertising," *Marketing Science*, 29 (2), 199-215.

Kleinberg, Jon M (1999), "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, 46 (5), 604-32.

Kotler, P. (2003), Marketing Management, 11th ed., Prentice-Hall, Upper Saddle River, NJ

Kritzinger, Wouter T and Melius Weideman (2015), "Comparative case study on website traffic generated by search engine optimisation and a pay-per-click campaign, versus marketing expenditure," *South African Journal of Information Management*, 17 (1), 1-12.

Levenshtein, Vladimir I (1966), "Binary codes capable of correcting deletions, insertions, and reversals," in Soviet physics doklady Vol. 10.

Li, Hongshuang, PK Kannan, Siva Viswanathan, and Abhishek Pani (2016), "Attribution strategies and return on keyword investment in paid search advertising," *Marketing Science*, 35 (6), 831-48.

Liu, Jia and Olivier Toubia (2015), "A framework for modeling how consumers form online search queries."

Luh, Cheng-Jye, Sheng-An Yang, and Ting-Li Dean Huang (2016), "Estimating Google's search engine ranking function from a search engine optimization perspective," *Online Information Review*, 40 (2), 239-55.

Mavridis, Themistoklis and Andreas L Symeonidis (2015), "Identifying valid search engine ranking factors in a Web 2.0 and Web 3.0 context for building efficient SEO mechanisms," *Engineering Applications of Artificial Intelligence*, 41, 75-91.

Moe, Wendy W (2003), "Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream," *Journal of consumer psychology*, 13 (1-2), 29-39.

Nabout, Nadia Abou and Bernd Skiera (2012), "Return on quality improvements in search engine marketing," *Journal of Interactive Marketing*, 26 (3), 141-54.

Page, Lawrence and Sergey Brin (1998), "Pagerank, an eigenvector based ranking approach for hypertext," in 21st Annual ACM/SIGIR International Conference on Research and Development in Information Retrieval.

Purcell, K., Brenner, J., & Rainie, L. (2012). Search Engine Use 2012,‖ PEW Research Center.

Roodman, David (2017), "CMP: Stata module to implement conditional (recursive) mixed process estimator," *Statistical software components*.

---- (2009), "Estimating fully observed recursive mixed-process models with cmp."

Rose, Daniel E and Danny Levinson (2004), "Understanding user goals in web search," in Proceedings of the 13th international conference on World Wide Web: ACM.

Rutz, Oliver J and Randolph E Bucklin (2011), "From generic to branded: A model of spillover in paid search advertising," *Journal of Marketing Research*, 48 (1), 87-102.

---- (2007), "A model of individual keyword performance in paid search advertising."

Rutz, Oliver J, Randolph E Bucklin, and Garrett P Sonnier (2012), "A latent instrumental variables approach to modeling keyword conversion in paid search advertising," *Journal of Marketing Research*, 49 (3), 306-19.

Scaperlanda, Anthony E and Laurence J Mauer (1969), "The determinants of US direct investment in the EEC," *The American Economic Review*, 59 (4), 558-68.

Shani, David and Sujana Chalasani (1992), "Exploiting niches using relationship marketing," *Journal of consumer marketing*, 9 (3), 33-42.

Shi, Savannah Wei and Michael Trusov (2013), "The path to click: Are you on it," working paper.

Silverman, D. (2010). IAB internet advertising revenue report. Interactive Advertising Bureau. New York, 26.

Skiera, Bernd and Nadia Abou Nabout (2013), "Practice prize paper—prosad: a bidding decision support system for profit optimizing search engine advertising," *Marketing Science*, 32 (2), 213-20.

Skiera, Bernd, Jochen Eckert, and Oliver Hinz (2010), "An analysis of the importance of the long tail in search engine marketing," *Electronic Commerce Research and Applications*, 9 (6), 488-94.

Tobin, James (1958), "Estimation of relationships for limited dependent variables," *Econometrica: journal of the Econometric Society*, 24-36.

Toften, Kjell and Trond Hammervoll (2010), "Niche marketing and strategic capabilities: an exploratory study of specialised firms," *Marketing Intelligence & Planning*, 28 (6), 736-53.

White, Ryen W, Susan T Dumais, and Jaime Teevan (2009), "Characterizing the influence of domain expertise on web search behavior," in Proceedings of the second ACM international conference on web search and data mining: ACM.

White, Ryen W and Dan Morris (2007), "Investigating the querying and browsing behavior of advanced search engine users," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval: ACM.

Wiesel, Thorsten, Koen Pauwels, and Joep Arts (2011), "Practice prize paper—Marketing's profit impact: quantifying online and off-line funnel progression," *Marketing Science*, 30 (4), 604-11.

Yang, Sha and Anindya Ghose (2010), "Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence?," *Marketing Science*, 29 (4), 602-23.

Figure 1: Search Engine Results Page

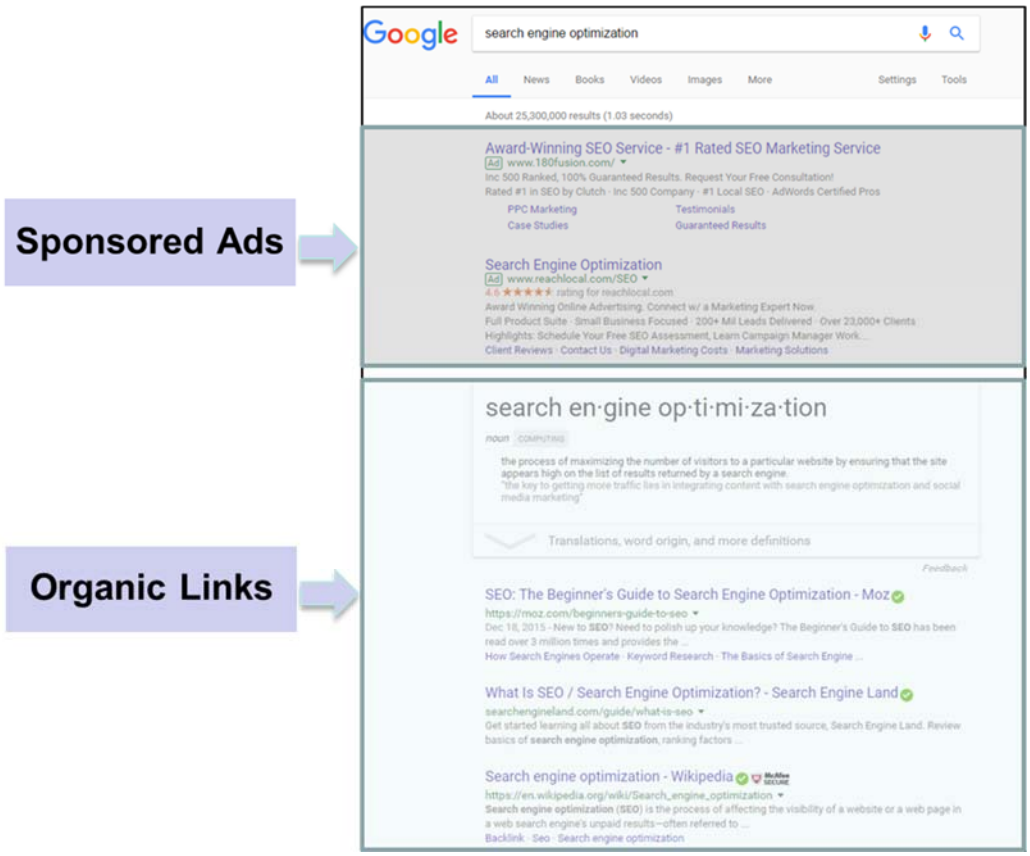


Figure 2: Conceptual Model

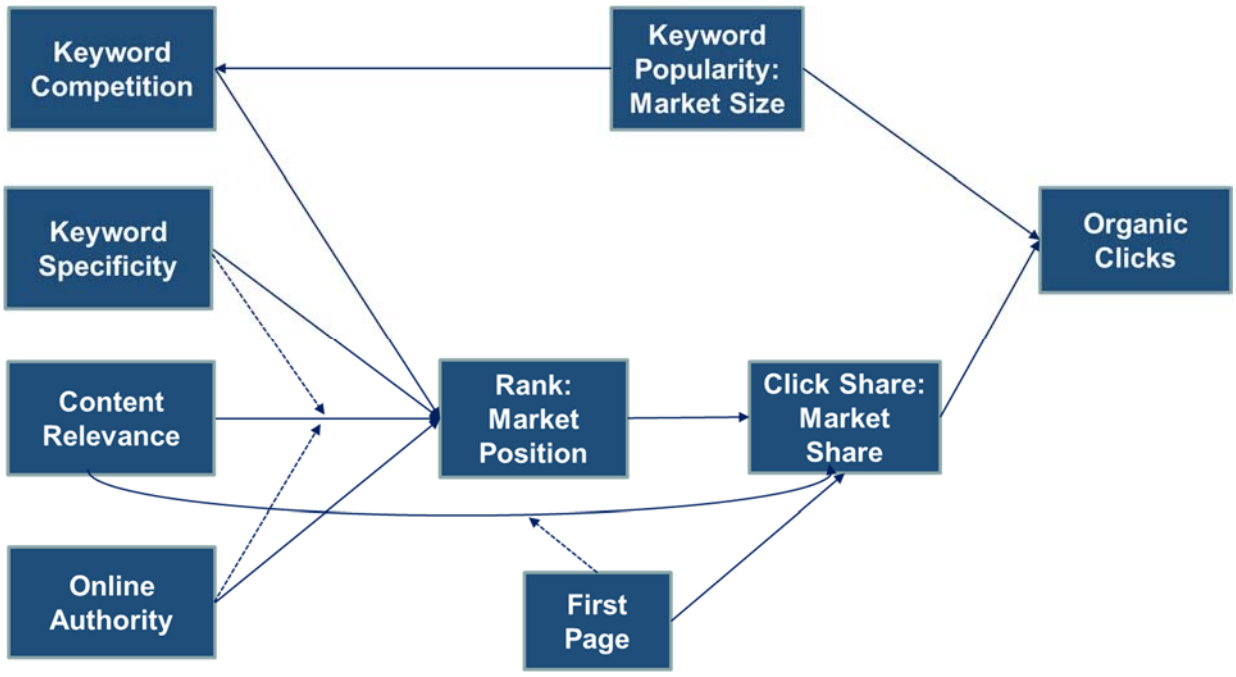


Figure 3: Calculating Keyword Relevance

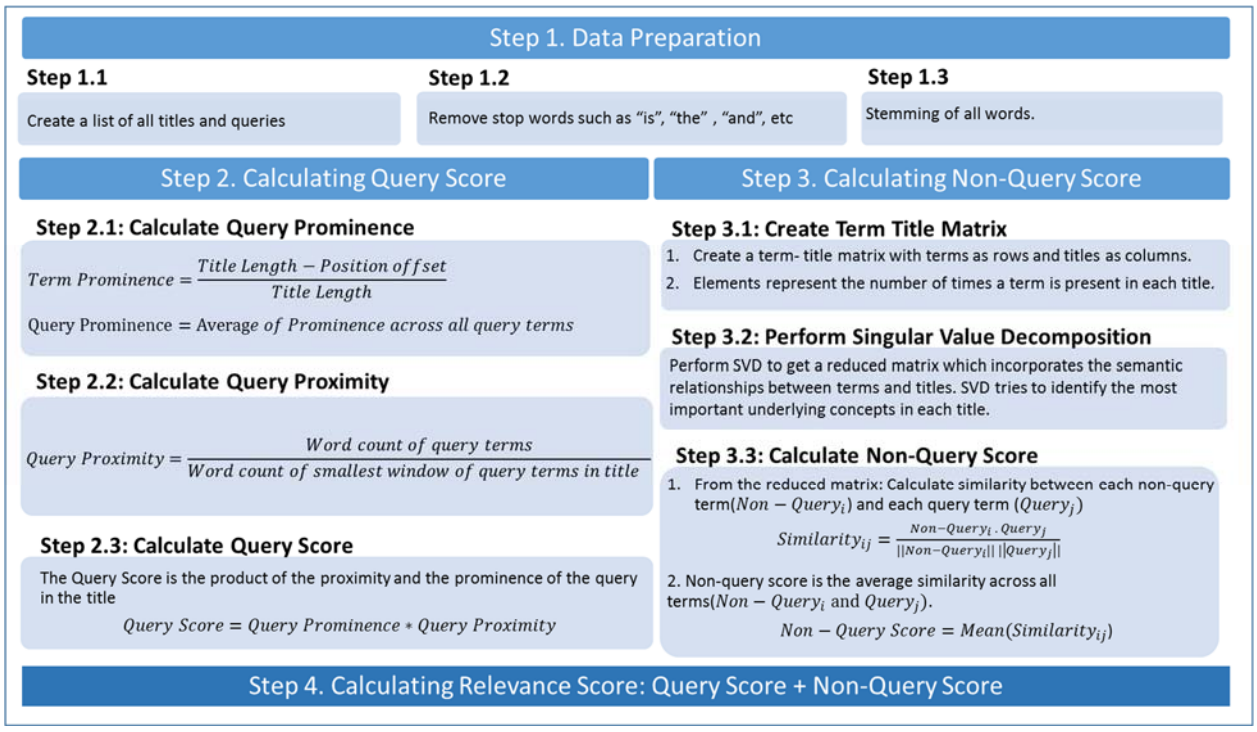


Figure 4: Marginal Effect of Keyword Relevance on Rank by Keyword Length: High vs. Low Authority

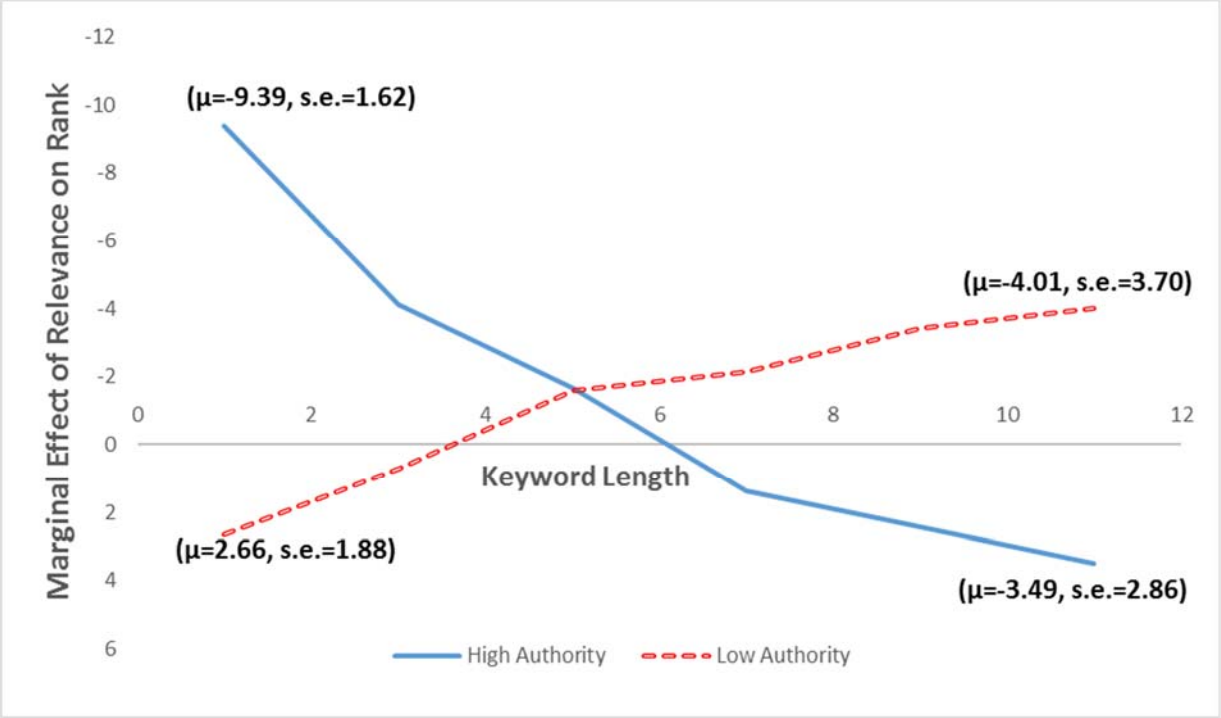


Figure 5: Expected Rank: Low vs. High Online Authority

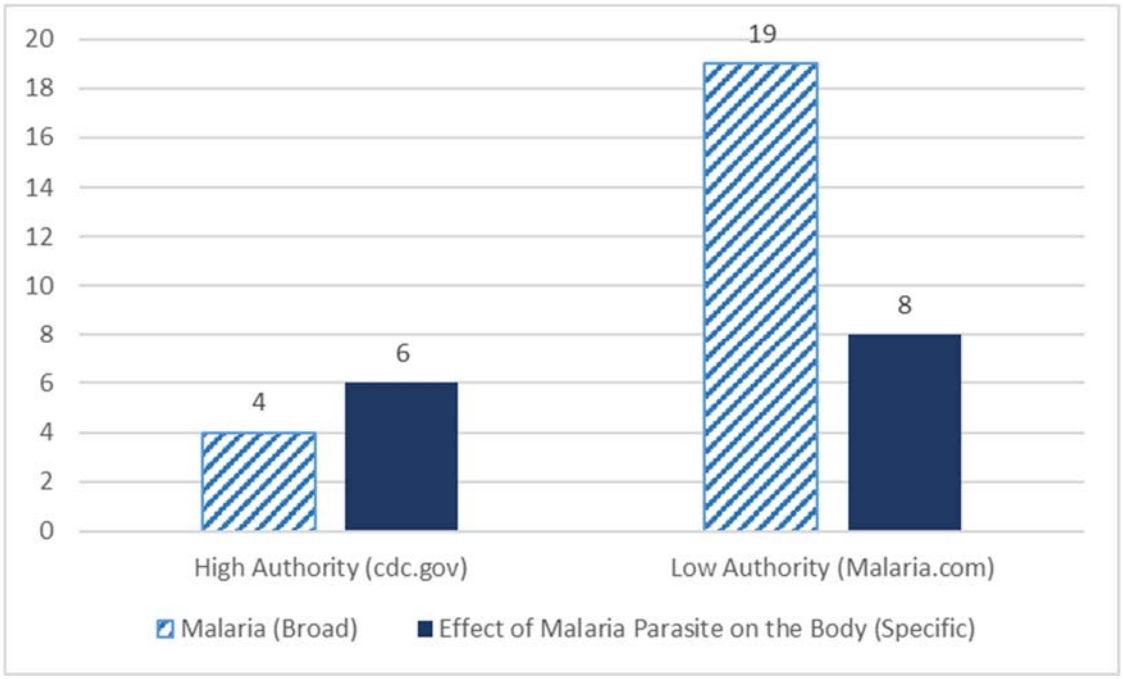


Figure 6: Expected Share: Low vs. High Online Authority

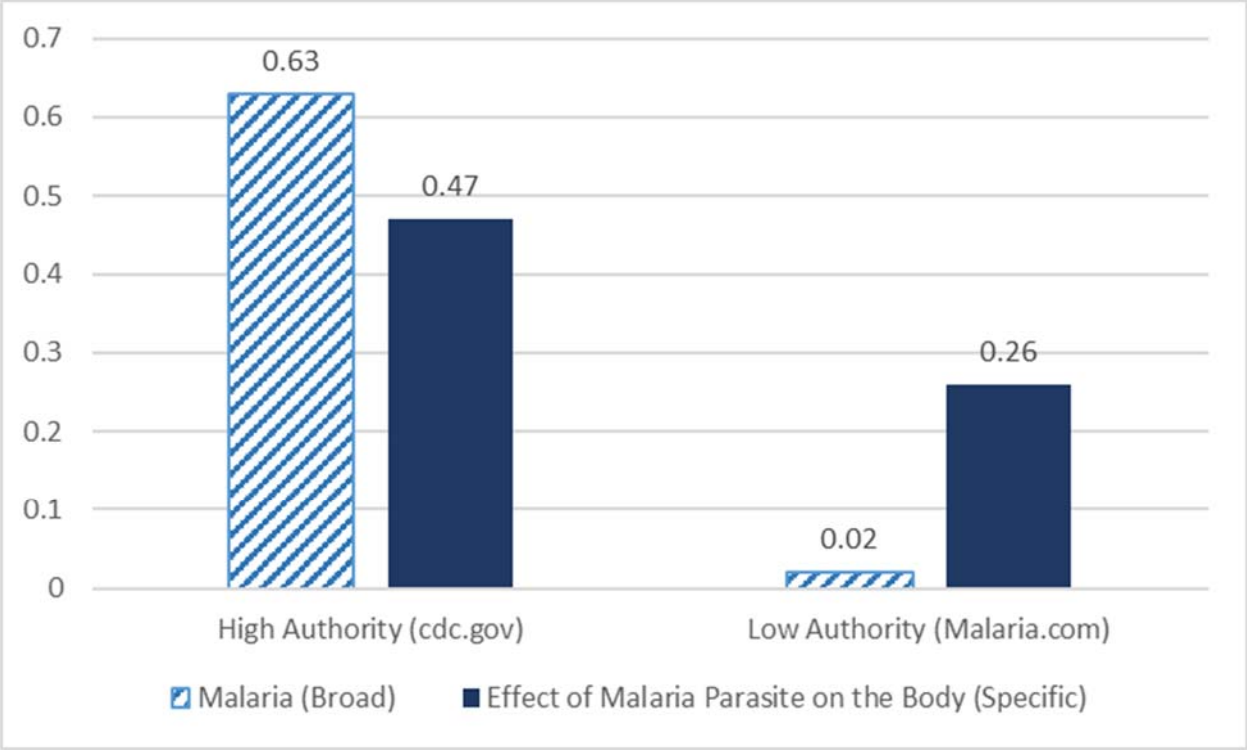


Figure 7: Expected Clicks: Low vs. High Online Authority

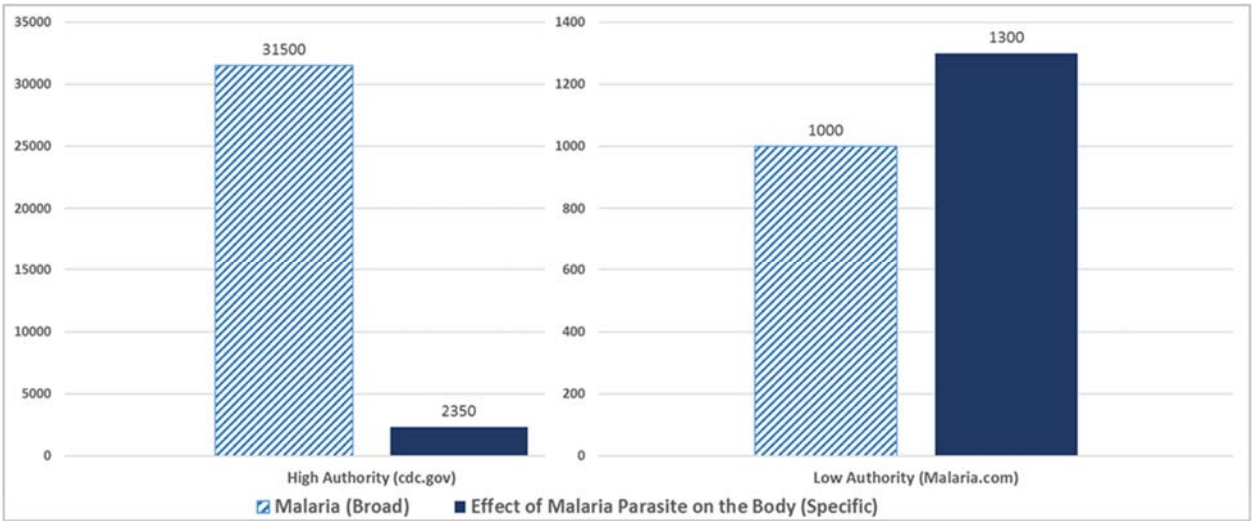


Table 1: Example of Keywords Commonly Used by Health Care Firms

	Keywords	Traffic	Relative # of Competitors
1	malaria parasite life cycle	36	0.15
2	symptoms of type 1 diabetes in a child	451	0.21
3	sore throat medicine for toddlers	1701	0.54
4	type 1 vs type 2 diabetes symptoms	1701	0.07
5	malaria symptoms	26651	0.13
6	sore throat medicine	26651	0.97
7	type1 diabetes symptoms	26651	0.80
8	type2 diabetes symptoms	65701	0.83

* Data calculated based on information from Google AdWords

Table 1 Notes: Table 1 provides information on the traffic each keyword receives and the number of sites bidding on each keyword relative to all keywords across Google. The first four keywords in Table 1 are specific keywords. A specific or long-tail keyword is often longer, in this case each keyword is at least 4 words in length, and more descriptive as the searcher has already defined a narrower topic to search. This leads to smaller search volumes on that keyword and as a result, fewer firms wanting to compete on that keyword. The second four keywords in Table 1 are broad keywords. A broad or generic keyword is often shorter, in this case each keyword is 3 or fewer words in length, and less descriptive as the searcher is often trying to define a narrower topic to search. This leads to larger search volumes on that keyword and as a result, more firms wanting to compete on that keyword. They need to select between getting a larger part of the small user base of specific keywords and getting a smaller part of the larger user base of broad keywords.

For instance, even though a specific keyword such as “malaria parasite life cycle” has low relative number of competitors (0.15), meaning that it will be easier for a firm to rank highly with relevant content, it also has a relatively small volume of search traffic (36). Additionally, we can see that a broad keyword such as “sore throat medicine” has a relatively high volume of search traffic (26,651), meaning that there are a lot of potential clicks available, but it also has a high relative number of competitors (0.97). Thus, a firm must decide whether it is better to create web content for keywords with relatively higher traffic and competition (i.e., potentially get a smaller share of a bigger market) or for keywords with relatively lower traffic and competition (i.e., potentially get a larger share of a smaller market).

Table 2: Literature Review of SSA and SEO Research

Authors	Field of Study	Relevant Contribution
Nabout and Skeira (2012)	SSA	Return on investment in quality improvement not always positive due to the negative effect on CPC due to the increased rank. Disentangled the negative direct price effect from the positive indirect price effect of quality improvement.
Skiera and Nabout (2013)	SSA	Developed and implemented the PROSAD (Profit Optimizing Search Engine Advertising) bidding decision support system to automatically determine optimized bids that maximize the advertiser's profit.
Li, Kannan, Vishwanathan, and Pani (2016)	SSA	Studied the impact of attribution strategies on the realized ROI of keywords in search campaigns. They find that first-click attribution leads to lower revenue returns and a more pronounced decrease in CTR for more specific keywords.
Baye, De los Santos, and Wildenbeest (2016)	SEO	A retailer's investments in factors such as the quality and brand awareness of its site increases organic clicks both directly by making the site more attractive to consumers and indirectly by improving its rank on the SERP.
Jerath, Ma, and Park (2014)	SEO/SSA	Consumers who search for less popular keywords expend more effort in their search for information and are closer to a purchase. This makes them more targetable for sponsored search advertising
Kritzinger and Weideman (2015)	SEO/SSA	After a certain period of time, an investment in search engine optimization rather than a pay-per-click campaign appears to produce better results at lower cost.
Berman and Katona (2013)	SEO/SSA	SEO improves search engine's ranking quality and thus customer satisfaction. This increases consumer's trust in organic links lowering SE's revenue from sponsored links. They find an inverse U-shaped relationship between the minimum bid and search engine profits.
Yang and Ghose (2010)	SEO/SSA	There is a positive interdependence between the click through rate on organic and paid listings. The positive impact of organic clicks on paid clicks is 3.5 times stronger than the opposite impact.
Taylor (2013)	SEO/SSA	As high-quality organic links cannibalize sponsored clicks SE have an incentive for quality degradation of the organic results to increase revenues.
White (2013)	SEO/SSA	When improvements in search quality benefit all users equally, advertisers will charge a higher price. However, when improvements in search quality provide a greater benefit to novice searchers, advertisers will charge a lower price.
Rutz and Bucklin (2011)	SSA	There is a significant spillover from generic to branded search as generic search causes an awareness of relevance of the brand. Incorporating this spillover considerably improves the financial performance of generic keywords for any firm.
Rutz and Bucklin (2007)	SSA	Developed a model for studying individual keyword performance using hierarchical Baye's model demonstrating the importance of keyword-level covariates and heterogeneity in conversion estimates.
Nabout (2015)	SSA	Compared multiple algorithms for finding the optimal profit maximizing bids.
Yao and Mela (2011)	SSA	A dynamic structural model of the sponsored search advertising market finds the following 1. Enabling firms to vary bids by consumer segment causes revenue gains for both firms and SE along with improving consumer welfare 2. Second price auctions increase firm's bids 3. Consumer search tools increase consumer welfare and SE revenues but reduce advertiser profits due to reduced exposure.

Chen, Liu, and Whinston (2009)	SSA	Find the optimal share of exposure allocated to each bidder by SEs and how this changes with the price elasticity of advertisers.
Feng, Bhargav, and Pennock (2007)	SSA	Propose a rank-revision strategy weights clicks on lower ranked items more than clicks on higher ranked items. This method converges to optimal ordering faster and more consistently.
Santos and Koulayev (2013)	SEO/SSA	Propose an optimal ranking strategy of search results that maximizes consumers' click-through rates (CTR) based on their preferences. This ranking system also increases consumer welfare.
Rutz, Bucklin, and Sonnier (2012)	SSA	Propose a modeling approach for assessing keyword performance in a sparse data environment. They find that higher positions have higher click-through and conversion rates.
Kang and Kim (2004)	Information Processing	Compared the performance of multiple scoring algorithms for different types of user queries, classified based on user intent.
Agarwal, Hosanagar, and Smith (2011)	SSA	Evaluate the impact of ad placement on revenues and profits generated from sponsored search 1. CTR decreases with position 2.Conversion rate increases and then decreases for long keywords
Ghose and Yang (2009)	SSA	Analyzed the relationship between keyword covariates and SSA performance: 1. CTR and conversion rate decreases with rank. 2. CTR is less for more specific keywords 3. Top ranked position not the most profitable due to the difference in CPC.
White and Morris (2007)	SEO/SSA	There are marked differences in the queries, result clicks, post-query browsing, and search success of advanced and novice users.
White, Dumais, and Teevan (2009)	SEO/SSA	Develop a model to predict expertise based on search behavior and describe how knowledge about domain expertise can be used to improve search results help increase user expertise
Shi and Trusov (2013)	SEO/SSA	Content of listings, textual information of previously viewed links and search intent influence the scanning behavior of users on SERP.
Broder (2002)	SEO/SSA	Classified information needs or search queries into informational, navigational and transactional
Rose and Levinson (2004)	SEO/SSA	Propose a framework for understanding the underlying goals of user searches.
Brynjolfssen, Hu, and Smith (2003)	Digital Marketing	Increased product variety leads to a larger increase in consumer surplus in the online market (Long tail phenomenon)
Brynjolfssen, Hu, and Smith (2006)	Digital Marketing	Identified supply side and demand side drivers of the long tail phenomenon along with its effects on consumers as well as producers.
Brynjolfssen, Hu, and Simester (2011)	Digital Marketing	Internet search and discovery tools, such as recommendation engines, are associated with the increase in share of niche products.
Shani and Chalasani (1992)	Niche Marketing	Provides a framework for implementation of relationship marketing for niche markets in the packaged goods industry.
Skiera, Eckert, and Hinz (2012)	SSA	Top 20% of all keywords attract on average 98.16% of all searches and generate 97.21% of all clicks. Hence, advertisers do not need to bother too much about the performance of keywords in the long tail.
Page and Brin (1998)	SEO/SSA	This paper describes PageRank, a method for rating web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

Luh, Yang and Huang (2015)	SEO	PageRank (PR) is the most dominant factor in Google ranking function. The title follows as the second most important, and the snippet and the URL have roughly equal importance with variations among queries.
Mavridis and Symeonidis (2015)	Information Technology	Developed a benchmark crawler called LHS Rank which incorporates semantics and compares its performance against established metrics.
Liu and Toubia (2015)	SEO/SSA	Develops a topic model Hierarchically Dual Latent Dirichlet Allocation (HDLDA), to find a relationship between main topics in search queries and search results. This helps understand consumer's content preferences using the semantic mapping between search queries and results.
Evans (2007)	SEO	Studied the SEO techniques used by top practitioners to find that: 1. Multiple pages were generated to influence ranking with limited success 2. PageRank very important in SEO. 3. Firms use older domains for higher rankings.

Table 3: SERP for Broad and Specific Query*

Search Query (Broad)	Search Query: Type 1 diabetes		
Rank	**Title**	**PA**	**DA**
1	Type 1 Diabetes: Causes and Symptoms	81	88
2	Type 1 Diabetes (Juvenile Diabetes) Causes, Symptoms, Treatments	68	91
3	Type 1 diabetes - Mayo Clinic	76	94
4	Type 1 Diabetes Facts - JDRF	66	77
5	What is Type 1 Diabetes? - Diabetes Research Institute	56	57
Search Query (Specific)	**Search Query: Type 1 diabetes symptoms in children**		
Rank	**Title**	**PA**	**DA**
1	Type 1 diabetes in children Symptoms and causes - Mayo Clinic	1	94
2	Diabetes Symptoms: Early Warning Symptoms & Signs of Diabetes	37	77
3	Type 1 Diabetes Symptoms - How to tell if your child has type 1 ...	38	54
4	Type 1 Diabetes (Juvenile Diabetes) Causes, Symptoms, Treatments	68	91
5	Type 1 Diabetes: What Is It? - KidsHealth	46	86

* PA = Page Authority; DA = Domain Authority; Data calculated based on information from Google AdWords

Table 4: Variable Descriptions and Data Sources

Variable Name	Description	Source
Outcome Variables		
Rank$_{ik}$	Minimum Rank of any webpage associated with website i on a SERP for search query k	Focal Firm
Click Share$_{ik}$	Share of the clicks for website i of the total traffic for search query k	Focal Firm
Organic Clicks$_{ik}$	Number of organic clicks on website i from search query k	Focal Firm
Relevant Drivers of Rank, Share, and Organic Clicks		
Online Authority (Diff)$_{ik}$	Online Authority is measured as the principal component for Domain and Page Authority for a given website i. Thus, Online Authority Diff$_{ik}$ = Online Authority of focal website i for search query k – average Online Authority of all other websites ranked on the first three pages of keyword k	Moz
Content Relevance (Diff)$_{ik}$	Content Relevance is measured as the relevance of the content of focal website i to search query k calculated using our proposed content relevance scoring algorithm (see ***Web Appendix A*** for full details). Content Relevance Diff$_{ik}$ = Content Relevance of focal website i for search query k – average Content Relevance of all other websites ranked on the first three pages of keyword k	Focal Firm (Computed)
Keyword Specificity$_{k}$	Length of search query k after removing stop words	Focal Firm (Computed)
Keyword Popularity$_{k}$	Number of users who searched for search query k	Google AdWords
Keyword Competition$_{k}$	Average Cost Per Click (CPC) for getting placed in top 3 sponsored search results for search query k	Google AdWords
First Page$_{ik}$	Equal to 1 if the focal website i is ranked in the top 10 on the SERP for search query k; 0 otherwise	Focal Firm (Computed)

Table 5: Descriptive Statistics and Correlations

	μ	s.d.	Share	Rank	Keyword Competition	Keyword Popularity	Keyword Specificity	Content Relevance (Mean)	Online Authority (Mean)	Content Relevance (Diff)	Online Authority (Diff)	First Page
Share	0.209	0.372	1.000									
Rank	8.324	7.464	-0.170	1.000								
Keyword Competition	0.603	0.826	-0.297	0.011	1.000							
Keyword Popularity	3.574	2.660	-0.529	0.264	0.439	1.000						
Keyword Specificity	2.576	1.820	0.374	-0.233	-0.318	-0.496	1.000					
Content Relevance (Mean)	0.558	0.142	-0.314	0.199	0.222	0.482	-0.385	1.000				
Online Authority (Mean)	-0.145	0.407	-0.065	0.283	0.141	0.222	-0.056	0.003	1.000			
Content Relevance (Diff)	-0.021	0.142	-0.314	0.199	0.222	0.482	-0.385	1.000	0.003	1.000		
Online Authority (Diff)	-0.145	0.407	-0.065	0.283	0.141	0.222	-0.056	0.003	1.000	0.003	1.000	
First Page	0.688	0.463	0.132	-0.867	0.005	-0.226	0.185	-0.157	-0.259	-0.157	-0.259	1.000

Table 6: Estimation Results

Variables	Keyword Competition coeff. (s.e.)	Rank coeff. (s.e.)	Share coeff. (s.e.)
Intercept	0.192** (0.097)	10.421*** (1.197)	0.776*** (0.09)
Keyword Popularity$_k$	0.095*** (0.009)		
Keyword Specificity$_k$	0.009 (0.021)	-1.147*** (0.257)	
Content Relevance (Mean)$_k$	0.196* (0.114)		
Online Authority (Mean)$_k$	0.136*** (0.042)		
Keyword Competition$_k$		6.361*** (0.861)	
Content Relevance (Diff)$_{ik}$		-5.469*** (1.647)	0.237*** (0.089)
Online Authority (Diff)$_{ik}$		-2.580*** (0.692)	
Content Relevance (Diff)$_{ik}$ * Keyword Specificity$_k$		0.914** (0.452)	
Online Authority (Diff)$_{ik}$ * Keyword Specificity$_k$		0.074 (0.214)	
Online Authority (Diff)$_{ik}$ * Content Relevance (Diff)$_{ik}$		-5.627*** (1.993)	
Content Relevance (Diff)$_{ik}$ * Online Authority (Diff)$_{ik}$ * Keyword Specificity$_k$		0.969* (0.533)	
Rank$_{ik}$			-0.038*** (0.006)
First Page$_{ik}$			-0.099* (0.055)
Content Relevance (Diff)$_{ik}$ * First Page$_{ik}$			0.187* (0.097)
Firm Fixed Effects	Included	Included	Included
Log Likelihood	-9080.91		

* 0.05<P value <0.1; ** 0.01<P value <0.05; *** P value < 0.01

Keyword Selection Strategies in Search Engine Optimization: How Relevant is Relevance?

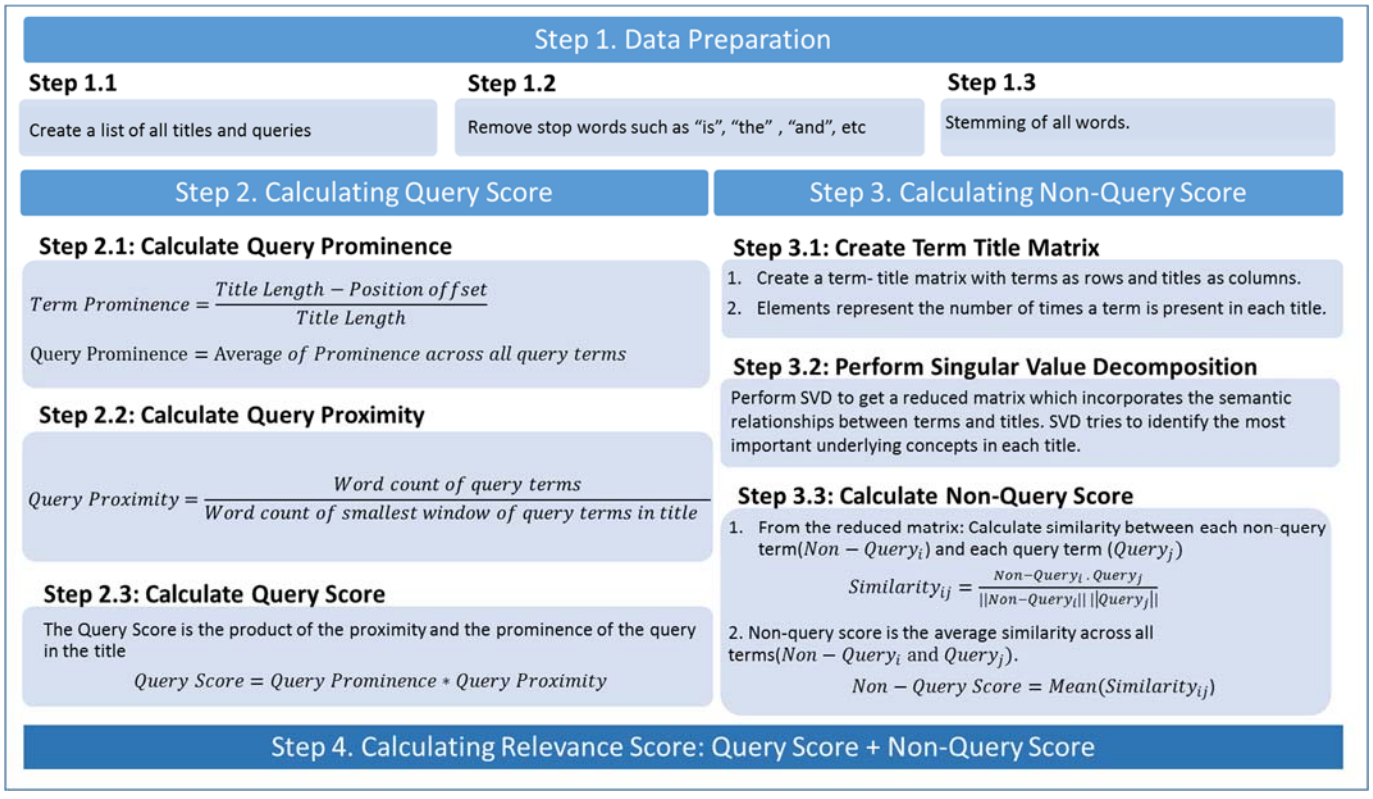
Web Appendix A: Content Relevance Score Algorithm

In this section we present the methodology for calculating the content relevance score between a search query or a keyword and the title of the webpage. To start, we partition the terms in the website title into two groups:

1. Query terms: Terms in the title which are present in the query or keyword.
2. Non-query terms: Terms in the title not present in the query or keyword.

The content relevance score for the query and non-query terms is calculated separately. The final content relevance score is calculated as the sum of the query score and non-query score. The score for the query terms is calculated based on the prominence and the proximity of the query terms in the title. Whereas, the score for the non-query terms is calculated based on the semantic relationship between the non-query and query terms derived using the Latent Semantic Analysis (LSA). The below figure provides the steps followed in calculating the relevance score between the titles and the queries.

Figure A1: Summary of Relevance Score Calculation Steps



Step 1: Data Preparation

The original data needs to be modified by removing the stop words from the titles as well as the

queries and stemming all the words. Data cleaning is important to identify the most important words and concepts in the text. Also, as some similar words relate to the same concept, it is important to group them together.

- Step 1.1: Create a List of Titles and Queries

  We create a list of titles and queries from which we try to determine the semantic relationship between words using LSA.

- Step 1.2: Stop Words

  Document content is often dominated by stop words. These are words such as "the", "it", "and", etc. which occur in most documents. Calculating similarity without removing these will be less precise as a large portion of the similarity between documents and phrases will be due to the stop words (Blei, Ng, and Jordan 2003). We used a commonly used stop word list (Page Analyzer English Stop Words List) and removed any words in this list from our data.

- Step 1.3: Stemming Words

  Stemming involves putting words like "Dog" and "Dogs" in the same basket, making the similarity calculation more meaningful. We use Porter's Stemming algorithm (Porter 1980) for stemming the words in the dataset. It is one of the most widely used stemming algorithms. We use the modified version (Snowball) of the algorithm developed by Porter (2001). Porter (2001) uses suffix stripping based on a set of rules, or transformations, applied in a succession of steps to identify the stems or roots of words.

Step 2: Calculating Query Score

The query score calculates the importance of the query terms to the website title. If the terms in the query are also present in the title, their importance can be measured by how visually prominent they are in the title and how spread out they are in the title. If the terms occur towards the beginning of the title, they are more prominent. If the terms occur together in the title, they are expected to have a similar meaning to what they have in the title. Thus, the query score is comprised of the prominence score and proximity score of the query terms.

- Step 2.1: Calculate Prominence Score

  Prominence Score measures how prominent the query terms are in the title. A term closer to the beginning of the title is more prominent. We calculate the prominence of any given word in the title as below:

  Term prominence =
  (Title length - Offset to the term's position in query) / Title Length (A1)

  The average offset is calculated as the average of the difference in the position of the

term in the title and in the query.

Offset= Max (Term Position in Title – Term Position in Query, 0) (A2)

We floor the value of the offset at 0 to ensure that the prominence score remains between 0 and 1. If a particular term is present multiple times in the title, then the offset is taken as the average offset.

The overall query prominence score of the title is the average prominence score of all terms.

Example: Consider the Title, "Culinary and arts professional institute" and a query "Culinary Institute". The offset for "Culinary" is (1-1) =0 as "Culinary" occurs in the first position in the title as well as the query. Similarly, the average offset for "Institute" is (4-2) = 2 as it occurs in the fourth position in the title and the second position in the query. Thus, the Term prominence for "Culinary" is (4-0)/4 =1 and for "Institute" is "(4-2)/4=0.5. The prominence score is the average prominence of each term in the query. In the above example the average prominence is 0.75.

Thus, for the query, "Culinary academy", the prominence score of the title, "Culinary Academy in Pennsylvania state" would have a larger prominence score (=1) compare to the title "Art Institute and Culinary academy" (Prominence Score=0.5).

- Step 2.2: Calculate Proximity Score

  Proximity Score measures the degree to which all terms of the query or sub-query occur together in the title. If the terms occur together, they are expected to have a similar meaning in the title to what they have in the query. On the other hand, if the terms are spread out they may not have the same meaning as in title as in this case the terms occur in the title individually rather than as a phrase.

  Proximity Score = (Word count of the query) / (Word count of the smallest window in title containing all terms query)[1] (A3)

  Thus for the query "Culinary Institute", which has a word count of 2, the proximity score of the title, "Culinary Institute and arts academy" would be 2/2= 1 as the entire phrase occurs together in the title. On the other hand, the proximity score for the title "Culinary and arts professional institute" would be much smaller (2/4=0.5) as the two terms are spread out.

- Step 2.3: Calculate Query Score

  We use prominence and proximity scores to obtain an overall query score of the title as the product of the two scores:

[1] If all terms in the query or keyword are not present in the title, then we multiply the proximity score with a penalty factor (penalty factor = number of query terms in the title/total number of query terms).

Score for query terms in the title = Query Prominence Query proximity* (A4)

Step 3: Calculating Non-Query Score

To calculate the non-query score we use Latent Semantic Analysis (LSA) to find the semantic relationship between the non-query terms and the query terms. LSA is a statistical technique for extracting and inferring relations of expected contextual usage of words in documents. The LSA measures the relatedness between a term and a document. The process for calculating the Non-Query Score is as follows:

- Step 3.1: Create Term-Title Matrix

  The input data for LSA is an initial term-title matrix (A) of order m x n, where the m rows represent m terms obtained from the n titles represented by n columns. We use both titles and keywords for creating the matrix as this helps getting a larger collection of terms and thus better identifying the semantic relationships.

  Each entry in the matrix is an initial approximation of the term frequency–inverse document frequency (tf-idf), which is a numerical statistic that is intended to reflect the importance of a word to a document in a collection documents or a corpus. Thus, an entry in the matrix, aij represents the significance of term i to title j.

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \text{(A5)}$$

  We initialize matrix A, such that a_{ij} is initialized as the number of times the term i occurs in title j. For example, if we have a list of titles as given below:

Table A1: Titles Example

T1	a culinary program is what
T2	about culinary school
T3	academy of culinary arts
T4	accredited culinary school
T5	advanced culinary techniques and management
T6	bobby flay education
T7	art of baking course
T8	baking school
T9	baker pastry chef schools
T10	bakery and pastry arts
T11	bakery chef college
T12	bakery classes nyc
T13	bakery program

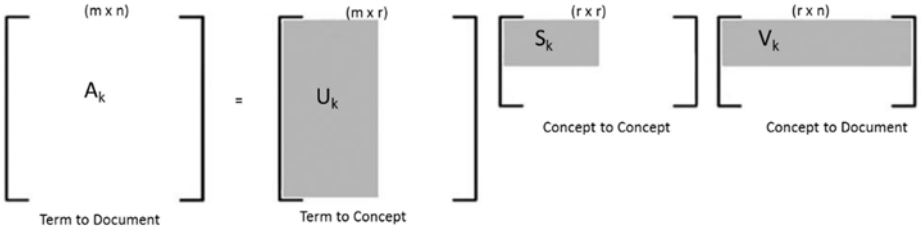
The term to title matrix is given as:

Table A2: Term-Title Matrix Example

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
academy	0	0	1	0	0	0	0	0	0	0	0	0	0
art	0	0	1	0	0	0	1	0	0	1	0	0	0
bakery	0	0	0	0	0	0	1	1	1	1	1	1	1
chef	0	0	0	0	0	0	0	0	1	0	1	0	0
class	0	0	0	0	0	0	0	0	0	0	0	1	0
college	0	0	0	0	0	0	0	0	0	0	1	0	0
course	0	0	0	0	0	0	1	0	0	0	0	0	0
culinary	1	1	1	1	1	0	0	0	0	0	0	0	0
education	0	0	0	0	0	1	0	0	0	0	0	0	0
flay	0	0	0	0	0	1	0	0	0	0	0	0	0
management	0	0	0	0	1	0	0	0	0	0	0	0	0
nyc	0	0	0	0	0	0	0	0	0	0	0	1	0
pastry	0	0	0	0	0	0	0	0	1	1	0	0	0
program	1	0	0	0	0	0	0	0	0	0	0	0	1
school	0	1	0	1	0	0	0	1	1	0	0	0	0
technique	0	0	0	0	1	0	0	0	0	0	0	0	0

- Step 3.2: Perform Singular Value Decomposition (SVD)

The initial term-title matrix (A) is subjected to a Singular Value Decomposition (SVD). A factor analysis is performed to decompose the matrix into a product of three matrices (U_k, S_k, and V_k). The most significant k singular factors indicate the most important hidden concepts or dimensions in the matrix.



(A6)

The elements of A_k represent the semantic relationship between the terms and each document. The new matrix contains the adjusted tf-idf values incorporating the semantic relationships. The algorithm considers two words to be semantically related if they co-occur in some documents. Thus, a word will have a significant relationship with a document if a related word is present in the document, even though the word itself is not present in the document. The degree of this relationship depends on the number of times the word occurs together with related words in the set of documents.

For example: After applying the Singular Value Decomposition to the above term to document matrix, we obtain a modified matrix given below:

Table A3: Term-Title Matrix (tf-idf) Example

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
academy	4.2E-16	2.5E-16	1.0E+00	7.7E-16	2.7E-16	-1.1E-16	-1.4E-16	1.9E-16	-3.6E-16	-8.3E-16	2.5E-16	5.7E-17	-3.6E-16
art	2.8E-16	3.5E-17	1.0E+00	7.8E-16	4.0E-16	3.3E-17	1.0E+00	2.8E-16	5.1E-16	1.0E+00	1.1E-15	7.3E-16	1.3E-16
bakery	-5.3E-16	-1.6E-15	-1.7E-16	-6.2E-16	-8.7E-16	-3.0E-17	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00
chef	-2.4E-16	-3.2E-16	-9.7E-17	-2.0E-16	-8.4E-16	2.6E-16	5.8E-16	-1.3E-16	1.0E+00	1.1E-15	1.0E+00	3.7E-16	-1.7E-16
class	1.1E-17	-8.4E-17	4.0E-17	-1.6E-16	-7.0E-16	-2.3E-16	3.7E-16	9.8E-16	9.0E-17	1.1E-16	3.2E-16	1.0E+00	1.7E-16
college	-1.1E-16	-4.9E-16	8.2E-16	-6.2E-16	-5.6E-16	1.6E-16	2.5E-16	8.1E-17	2.8E-16	6.9E-16	1.0E+00	1.9E-16	4.7E-17
course	-1.8E-16	-2.3E-17	-6.8E-16	-1.5E-17	-8.8E-17	-1.8E-17	1.0E+00	2.3E-16	8.6E-16	2.8E-16	2.5E-16	5.5E-16	-1.5E-16
culinary	1.0E+00	1.0E+00	1.0E+00	1.0E+00	1.0E+00	-3.5E-17	-5.6E-16	3.0E-17	4.4E-16	-4.6E-16	-1.7E-16	-4.6E-16	-4.8E-16
education	-5.5E-18	2.2E-17	2.2E-16	-4.4E-17	-2.5E-17	1.0E+00	-1.4E-16	-2.1E-16	1.5E-17	-1.9E-16	5.9E-16	1.7E-17	-1.7E-16
flay	1.7E-18	-1.1E-17	-3.5E-17	-3.0E-16	3.5E-16	1.0E+00	9.6E-18	-1.9E-16	2.8E-16	-1.6E-16	2.1E-16	-1.9E-16	-4.9E-18
management	6.4E-16	4.8E-16	1.9E-16	4.6E-16	1.0E+00	4.7E-16	-1.1E-16	2.8E-16	-1.1E-16	2.1E-16	-8.7E-16	-6.3E-16	1.2E-16
nyc	-1.6E-16	-1.4E-16	1.3E-16	-1.2E-16	-6.4E-16	-3.0E-16	4.0E-16	6.5E-16	2.5E-16	1.2E-16	3.5E-16	1.0E+00	-1.3E-16
pastry	-2.1E-16	-2.0E-16	-1.0E-15	-7.9E-17	-2.1E-16	2.4E-16	8.3E-16	3.4E-16	1.0E+00	1.0E+00	7.4E-16	2.3E-17	2.1E-16
program	1.0E+00	-3.0E-16	1.5E-16	-3.0E-16	1.5E-16	7.4E-17	7.5E-16	7.3E-16	1.5E-16	6.7E-16	3.4E-16	6.2E-16	1.0E+00
school	1.2E-15	1.0E+00	9.6E-16	1.0E+00	3.0E-16	-2.9E-16	1.7E-16	1.0E+00	1.0E+00	5.8E-16	-5.5E-16	2.3E-16	6.4E-16
technique	6.4E-16	4.7E-16	1.9E-16	4.7E-16	1.0E+00	3.9E-16	-1.1E-16	2.8E-16	-1.1E-16	2.1E-16	-8.7E-16	-6.3E-16	1.2E-16

We see that there is now a positive relationship between "School" and the first title even though it is not present in the title. This happens because the term "Culinary" and "School" are present together in two titles.

- Step 3.3: Calculate Non-Query Score

To find the semantic relevance score among terms or words, we calculate the cosine similarity among the rows of the reduced matrix, A_k. The similarity among words is calculated using the formula below:

$$Similarity_{ij} = \frac{Non{-}Query_k \,.\, Query_l}{||Non{-}Query_k||\ ||Query_l||} \quad \text{(A7)}$$

where, $similarity_{ij}$ represents the similarity between the kth non-query term and the lth query term.

The overall score for non-query terms in the title is calculated as the mean of the semantic relevance score between the non-query terms and the terms in the search query:

$$Non\text{-}Query\ Score = Mean(Similarity_{kl}) \quad \text{(A8)}$$

where, the mean is calculated across all k and l. In other words, we take the average of the similarity between each term in the query with each non-query term in the title.

Continuing with the earlier example of the search query, "Culinary institute" and the title "Culinary and arts professional institute". The score of non-query terms would be the

mean of relevance scores between the only non-query terms “professional” & “arts “and the two terms in the query, i.e. “Culinary” and “institute”. The non-query score thus depends on how many times the query terms co-occur with the non-query terms in the set of documents.

Step 4: Calculating Relevance Score

The overall relevance score between the title and the query score is the sum of the query score based on the prominence and proximity, and the non-query score which is based on the semantic similarity among the terms in the query and the non-query terms in the title. The relevance score is thus calculated as:

$$\textit{Relevance Score} = \textit{Query Score} + \textit{Non-query Score} \quad \text{(A9)}$$

Provided below are the descriptive statistics for the relevance score calculated for each dataset:

Table A4: Descriptive Statistics: Content Relevance Scores

		Mean	Std. Dev.	Min	Max
Urgent Care Data	**Query Score**	0.434	0.269	0	1
	Non- Query Score	0.073	0.057	0	0.447
	Relevance Score	0.507	0.301	0	1.32
Culinary Data	**Query Score**	0.456	0.236	0	1
	Non- Query Score	0.078	0.037	0	0.283
	Relevance Score	0.534	0.249	0	1.13
Retail Data	**Query Score**	0.552	0.001	0	1
	Non- Query Score	0.072	0.001	0	0.43084357
	Relevance Score	0.624	0.003	0	1.41

Web Appendix References:

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3 (Jan), 993-1022.

Porter, Martin F. (1980), "An Algorithm for Suffix Stripping," *Program*, 14 (3), 130-37.

Porter, Martin F. (2001), "Snowball: A language for stemming algorithms."