Quantitative Marketing and Economics
https://doi.org/10.1007/s11129-019-09217-3

Search query formation by strategic consumers

Jia Liu[1] · Olivier Toubia[2]

Received: 20 September 2018 / Accepted: 11 July 2019 / Published online: 30 August 2019

Abstract
Submitting queries to search engines has become a major way for consumers to search for information and products. The massive amount of search query data available today has the potential to provide valuable information on consumer preferences. In order to unlock this potential, it is necessary to understand how consumers translate their preferences into search queries. Strategic consumers should attempt to maximize the information content of the search results, conditional on a set of beliefs on how the search engine operates. We show using field data that optimal queries may exclude some of the terms that are more relevant to the consumer, potentially at the expense of less relevant terms. In two incentive-aligned lab experiments, we find that consumers have some ability to strategically omit relevant terms when forming their search queries, but that their search queries tend to be suboptimal. In a third incentive-aligned experiment, we find that consumers' beliefs on how the search engine operates tend to be inaccurate. Overall, our results are consistent with consumers being strategic when formulating their queries, but acting on incorrect beliefs on how the search engine operates.

Keywords Search engines · Revealed preference · Experiments

JEL Classification M300

1 Introduction

Every minute, more than 3 million queries are submitted to Google (InternetLiveStats 2016). Each of these queries is some expression of one consumer's preferences

✉ Olivier Toubia
ot2107@columbia.edu

Jia Liu
jialiu@ust.hk

1 Marketing, Hong Kong University of Science and Technology, Hong Kong, Hong Kong

2 Graduate School of Business, Columbia University, New York, NY, USA

(Pirolli 2007), which is voluntary and incentive-compatible (to the extent that the consumer is motivated to find useful content). As such, online search query data present a massive opportunity for revealed preference measurement. Such data may not only be a source of consumer insights, they are also essential for search marketing, an industry expected to be worth \$80 billion by 2020 in the U.S. only (Borrell 2016). For example, the amount a marketer should be willing to bid on a particular keyword is a function of how well his/her content matches the preferences of consumers who usually type that keyword.

A long literature in marketing and economics has modeled search by strategic, utility-maximizing consumers (e.g., Stigler 1961; Weitzman 1979; Erdem et al. 2005; Gabaix et al. 2006; Seiler 2013; Honka and Chintagunta 2016). However, this literature has been primarily limited to situations where search is performed by a series of discrete choices (e.g., purchases, clicks). Therefore, it does not capture the most pervasive form of consumer search in the market today: text-based search. Text-based search is *not* a straightforward special case of discrete-choice search in which consumers would select from a very large universe of queries. In particular, search terms are semantically related to each other and to the search results, which creates a rich set of dependencies between queries and their results. Accordingly, extant economic search models are in most cases not suitable for modeling consumer query formation.

In order to develop a utility-based, revealed preference framework for analyzing online search queries, it is necessary to understand how consumers formulate queries given their preferences. That is, what is the link between the content a consumer wants to consume, and the queries he or she formulates on a search engine? Answering this question is a necessary first step to the development of tools and methods that enable academics and practitioners to fully leverage the textual data from online search, and eventually help advertisers improve their selection of keywords, bidding strategies, and design of ad copies and landing pages.

In this research, we define a consumer's "content preferences" as the textual content that this consumer desires to "consume." Depending on the context, content may consist of news stories, information about products and services, or information about any domain of knowledge. Consumers formulate queries in order retrieve certain content, based on their beliefs on how the search engine operates. A naive set of beliefs would be that the presence of a term in the search results is only driven by the presence of this term in the search query. That is, according to this naive view, the only way to ensure that a term appears in the results is to include this term in the search query. Such naive approach to query formation would lead consumers to simply formulate queries that contain the most relevant terms they are searching for. However, in reality there exist complex relationships between the terms in search queries and the content in the associated search results, which we describe using "activation probabilities" (the probability of seeing a term in the search results given the content of a search query). If consumers' beliefs on activation probabilities are *not* naive, then queries should *not* be direct and straightforward expressions of preferences, but rather the result of a strategic attempt to retrieve desired content. For example, consider a consumer typing the following query: "affordable sedan made in America." It is possible that the most important attributes for this consumer are in fact safety, comfort, and made in America, and that affordability is of lesser importance. This consumer might have decided to type the query "affordable sedan made in America" because he or she believes that cars made in America are generally safe and comfortable, but not necessarily affordable. In that case, the consumer anticipated that he or she would find results that match his or her preferences efficiently (i.e., with short queries) by only including "made in America" and "affordable" in his or her queries, but neither "safe" nor "comfortable," although these are important attributes. In other words, consumers may strategically leave out important terms from their search queries, and include less relevant ones.

In this paper, we illustrate using field data the potential benefits for consumers from strategically omitting some of the most relevant terms from their search queries. In addition, we explore directly and experimentally whether consumers have the ability to identify optimal queries, and whether the queries formulated by consumers are consistent with naive beliefs. We find that the queries formulated by consumers are typically *not* consistent with a naive benchmark according to which consumers would only include the most relevant terms in their queries. We also find that consumer's queries, while not being consistent with a naive view of query formation, are usually suboptimal, and that consumers' beliefs on activation probabilities tend to be incorrect. Together, these results paint a picture of consumer search query formation in which consumers attempt to strategically optimize search queries by leveraging activation probabilities, but act on incorrect beliefs on activation probabilities.

The rest of the paper is organized as follows. Section 2 reviews relevant research. Section 3 presents a simple utility framework which quantifies the performance of search queries, formally defines optimal queries, strategic query formation and naive query formation. In addition, we develop a set of testable predictions that would hold if consumers held naive beliefs on activation probabilities. Section 4 shows using field data that the best performing queries may omit some of the most relevant terms. Section 5 reports on two lab experiments that explore consumers' ability to formulate optimal queries, and tests the predictions from Section 3. Section 6 reports on a third experiment that directly elicits consumers' beliefs on activation probabilities (in an incentive-aligned manner), and compares them to the truth. Section 7 discusses implications of our results, and concludes with future research directions.

2 Relevant literature

Our paper is related to the large literature on search in marketing and economics (Erdem et al. 2005; Hui et al. 2009; Park and Chung 2009; Yang et al. 2015). Some of this literature has even studied search in the context of search engines (Kim et al. 2010; Shi and Trusov 2013; Jeziorski and Segal 2015). However, in this literature search is typically expressed by discrete choices among items such as products or links. Consequently, text-based online search queries have largely been ignored in marketing. Liu and Toubia (2018) take one step towards modeling text-based consumer search. These authors propose a topic model,[1] HDLDA (hierarchical dual latent Dirichlet allocation), which nests LDA (latent Dirichlet allocation), and which links the content in the search queries submitted to a search engine to the content in the top search results returned by the search engine. HDLDA focuses on modeling the way the search engine operates, and it is agnostic as to how consumers formulate search queries given their content preferences. Nevertheless, Liu and Toubia (2018) show that HDLDA can be leveraged to infer consumers' content preferences from their search queries, based on a set of assumptions on how consumers translate their content preferences into search queries. These authors test two specific assumptions: a naive assumption that search queries are a direct representation of content preferences, and a strategic assumption that consumers formulate search queries that will (in expectation) retrieve the content they would like to consume. These authors find that the content preferences estimated based on the strategic assumption are more accurate (i.e., closer to the true preferences) compared to the preferences estimated based on the naive assumption. In this paper, we explore consumers' ability to formulate optimal queries and test the naive assumption, directly and experimentally. In addition, we illustrate theoretically and empirically the potential benefits for consumers from strategically omitting relevant terms from their search queries, and we measure consumers' beliefs on activation probabilities directly.

A large body of research on online search comes from the Information Retrieval (IR) literature, which has focused primarily on the problem of finding the most relevant documents given a query (e.g., Salton and McGill 1986; Ruthven 2003; Manning et al. 2008; Li and Xu 2013; Santos et al. 2015). That is, this literature has typically focused more on optimizing information retrieval systems using a query as input, rather than on understanding the process by which consumers form their queries. As such, this literature does *not* provide answers to the question of how consumers formulate search queries given their content preferences. There is also a considerable body of descriptive research on online queries in the IR literature. It has been shown consistently that most queries are a list of one or more nouns; on average the length of a query is two to three terms; and at least 80% of queries contain three terms or fewer (Jansen et al. 2009, 2000a; Spink et al. 2001; Kamvar and Baluja 2006). It has been argued that consumers favor short queries because forming long queries is costly in terms of time, cognitive effort, physical typing, and so on (Ruthven 2003; Azzopardi et al. 2013). That is, if two queries are equally effective at retrieving relevant content, consumers should have a preference for the shorter of the two queries.

3 Definitions and theoretical analysis: utility framework, optimal queries, strategic and naive query formation

In this section we present a simple utility framework, which enables us to quantify the performance of search queries, and to define and characterize optimal queries, strategic query formation and naive query formation.

[1] A topic model is a statistical model that describes text using a set of topics rather than individual words, where topics are defined as probabilistic combinations of words.

3.1 Utility function

Consumers derive value from consuming the content of webpages, based on how well these pages satisfy their content preferences. Consistent with the existing models of user navigational behavior (Fu and Pirolli 2007; Wu et al. 2014), we assume that the utility of a webpage l for a user is a function of the words that are present in the page. Let $G = \{t_1, t_2, ..., t_W\}$ denote the set of relevant words, i.e., the set of words from which the user derives value. For the sake of illustration, we consider here the simple case of a linear additive utility function. We denote by β_j the value of word t_j for the user. We refer to terms for which $\beta_j > 0$ as being *relevant* to the consumer. For simplicity, we assume that the consumer derives utility from webpage l as long as the page contains word t_j. In this simple case, the utility of a webpage l may be written as follows:

$$U(l) = \sum_{t_j \in G} \beta_j \mathbf{1}(t_j \in l) \tag{1}$$

where $\mathbf{1}(t_j \in l)$ indicates whether webpage l contains word t_j. Alternative utility functions may be considered, for example, that replace this indicator function with a count of how many times the word appears on the page. It is straightforward to extend our theoretical analysis to this case, and we consider such utility function in Appendix A. Note that one could also include other webpage characteristics (e.g., type of webpage, type of content) in this utility function. Such characteristics should be particularly important in predicting which webpages consumers decide to visit, given a set of search results presented by the search engine. In this paper, we focus on parsimonious utility functions and do not consider such factors.

Here we take the view that the object being consumed by users is the content of webpages, and define utility accordingly based on content preferences, which are a function of the words present on a webpage. Content consumption is usually a means towards another consumption goal (e.g., purchasing a product, being informed on a news topic). Accordingly, content preferences may themselves be tied for example to product attributes (e.g., the user values content related to specific product attributes) or other consumption goals (e.g., a user who consumes news in order to stay informed on a topic values content that contains specific keywords).

3.2 Search cost

We allow for the possibility that formulating longer queries is more costly to the user, by introducing a cost function $c(q)$. This function denotes the cost of formulating query q, which for example may reflect the physical cost of typing or the cognitive cost of formulating this query. For simplicity, in this illustrative model we define $c(q)$ as a linear function of the number of terms in q: $c(q) = c|q|$, where $|q|$ denotes the number of words in query q and c is the unit cost of each term in query q.

3.3 Performance metrics and optimal queries

A search query q may be defined as any ordered subset of the words in G. Several metrics may be considered to quantify the performance of a query, combining and weighing the utility of each search result. In this paper, we consider two performance metrics. The first assumes that the consumer cares about the average utility of the top L results returned by the search engine. Taking search costs into account, this gives rise to the following performance metric:

$$Perf^{\text{average}}(q) = \frac{1}{L}\sum_{l} U(l) - c(q) = \sum_{t_j \in G} \beta_j \frac{\sum_l \mathbf{1}(t_j \in l)}{L} - c|q| \tag{2}$$

The second performance metric assumes that the consumer cares about the utility of the best search result, giving rise to the following:

$$Perf^{\max}(q) = \max_{l}\{U(l)\} - c|q| \tag{3}$$

We describe queries that maximize these performance metrics as *optimal* queries. Because of the complex relationship between search queries and search results, long queries that contain all relevant terms (i.e., terms for which $\beta_j > 0$), may be in fact less effective at retrieving all relevant terms, compared to shorter queries that omit some of the relevant terms. That is, optimal queries may omit some of the relevant terms, even when $c = 0$. We illustrate this phenomenon in Section 4.

3.4 Strategic query formation

We assume a consumer with a utility function given by Eq. 1. The search engine determines the relationship between search queries and search results. Conditional on how the search engine operates, the expected utility of a link l returned as a result of query q may be written as:

$$E[U(l)|q] = \sum_{t_j \in G} \beta_j Prob(t_j \in l|q) \tag{4}$$

where $Prob(t_j \in l|q)$ is the probability that word t_j appears in link l, given query q, which we refer to as an *activation probability*. This activation probability depends on the entire set of words present in the query (and their order).

Accordingly, if a consumer cares about the expected value of the average utility across the top L search results returned by the search engine, he or she should form queries that maximize the following objective function:

$$f^{\text{average}}(q) = E[U(l)|q] - c(q) = \sum_{t_j \in G} \beta_j Prob(t_j \in l|q) - c|q| \tag{5}$$

This objective function depends on the set of true activation probabilities, $\{Prob(t_j \in l|q\}$. These activation probabilities describe how the search engine operates, and their true values are typically unknown to the consumer. We denote as $\widetilde{Prob}(t_j \in l|q)$ the consumer's belief about the activation probability $Prob(t_j \in l|q)$.

We define strategic query formation as an attempt by the consumer to maximize the information content of the search results, conditional on his or her beliefs on activation probabilities. Hence, a strategic consumer who cares about the average expected utility of the top L search results maximizes the following objective function:

$$\tilde{f}^{\text{average}}(q) = \sum_{t_j \in G} \beta_j \widetilde{Prob}(t_j \in l|q) - c|q| \tag{6}$$

Similarly, if the consumer cares about the expected value of the maximum utility among search results, the objective function becomes:

$$\tilde{f}^{\max}(q) = E[\max_l \{U(l)\}|q] - c|q| \tag{7}$$

We note that query formation may also be thought of as a game played between consumers and the search engine, in which consumers formulate queries based on their belief on how the search engine operates, and the search engine retrieves results based on its belief on how consumers formulate queries as a function of their preferences. We leave the study of such game to future research. In this paper we consider the behavior of consumers, conditional on their beliefs on how the search engine operates. Accordingly, in our context, "strategic query formation" refers to the consumer strategically forming queries in order to maximize utility conditional on his or her beliefs on activation probabilities, but it does *not* imply that the consumer takes into account the search engine's reaction to his or her query formation approach. In other words, we treat consumers as atomistic agents (see for example Nair 2007, for other marketing papers that have assumed atomistic agents).

3.5 Naive beliefs

Strategic query formation requires a set of beliefs on a rich set of activation probabilities, $\{\widetilde{Prob}(t_j \in l|q)\}$. Such beliefs require some understanding of how the search engine operates. As a naive baseline, we consider the following set of beliefs:

$$\widetilde{Prob}(t_j \in l|q) = \begin{cases} \alpha^{high} & if\ t_j \in q \\ \alpha^{low} & if\ t_j \notin q \end{cases} \tag{8}$$

where $0 \leq \alpha^{low} < \alpha^{high} \leq 1$. That is, as a baseline we consider a naive consumer who believes that the probability that a term appears in the result is only influenced by whether the term appears in the query, but not by what other terms are included in the query. Intuitively, such naive beliefs should lead consumers to include the most relevant terms in their queries, in order to ensure these terms will be present in the search results. In addition, a consumer with naive beliefs would not include a less relevant term in his or her query, as he or she would not realize that including such a term might help retrieve other, more relevant terms.

Formally, when the consumer cares about the average utility across search results, this naive set of beliefs gives rise to a simple and intuitive query formation rule. The objective function in Eq. 6 becomes:

$$\begin{aligned}\tilde{f}^{\text{average}}_{\text{naive}}(q) &= \sum_{t_j \in G} \beta_j \left(\alpha^{high}\mathbf{1}(t_j \in q) + \alpha^{low}\mathbf{1}(t_j \notin q)\right) - c\sum_{t_j \in G}\mathbf{1}(t_j \in q)\\ &= \sum_{t_j \in G}\left[\beta_j \left(\alpha^{high}\mathbf{1}(t_j \in q) + \alpha^{low}\mathbf{1}(t_j \notin q)\right) - c\mathbf{1}(t_j \in q)\right] \qquad (9)\end{aligned}$$

where $\mathbf{1}(t_j \in q)$ is an indicator function for whether word t_j is contained in query q. The objective function becomes separable in the terms $\{t_j\}$, making it trivial to find the optimal set of terms to include in the query, by making a decision on each term separately. In particular, term t_j should be included in the query if and only if:

$$\beta_j\alpha^{high} - c \geq \beta_j\alpha^{low} \iff \beta_j \geq \frac{c}{\alpha^{high} - \alpha^{low}} \qquad (10)$$

That is, the query that maximizes the naive objective function the naive objective in Eq. 9 consists of the terms that have a marginal utility above a threshold. This threshold is increasing in the cost of including additional terms in the query c, and decreasing in the marginal impact of including a term in the query on the probability that this term appears in the search results, $\alpha^{high} - \alpha^{low}$. When $c = 0$, the threshold is 0 and hence all terms with positive marginal utility would be included in the query according to the naive approach. We show in Section 4 that this is suboptimal, and that consumers may be better off omitting relevant terms from their queries, and even sometimes including less relevant terms at the expense of more relevant ones.

When the consumer cares about the maximum utility across search results, the objective function in Eq. 7 is *not* separable in the terms $\{t_j\}$ even with naive beliefs, as the identity of the highest-utility search result depends on the *set* of terms present in each result, which in turn depends on the set of terms present in the query. Hence, it is not possible anymore to optimize the objective function by making decisions on each term separately. The case $c = 0$ still gives rise to the simple prediction that the consumer should include all relevant terms in the query, which we again show to be suboptimal in Section 4. When $c > 0$, there is no simple algebraic expression for the naive benchmark when the consumer cares about the maximum utility across results. In Appendix C, we solve the naive benchmark numerically given our experimental setup, i.e., with three words valued at \$2, and with two words are valued at \$2 and one word valued at \$1, assuming $c = 1$. We find that naive queries never include a low-value (\$1) term. Naive queries tend to include more terms when $\alpha^{high} - \alpha^{low}$ is larger, i.e., when including a term in the query has a higher impact on the occurrence of this term in the results. However, unlike in the case in which performance is measured by the average utility across search results, there exist pairs $\{\alpha^{low}, \alpha^{high}\}$ for which some but not all high-value terms are included in the query. In these situations, there are diminishing returns to including additional high-value terms in the query.

In this section, we have introduced three definitions: optimal queries, strategic query formation, and naive query formation. Optimal queries are those that actually maximize performance. In the following section, we illustrate that optimal queries may omit some of the more relevant terms, potentially at the expense of less relevant terms. In Section 5, we explore experimentally consumers' ability to identify optimal queries. Strategic query formation refers to the consumer strategically forming queries in order to maximize utility, conditional on his or her beliefs about activation probabilities. These beliefs may be correct or not. In Section 6, we measure consumers' beliefs on activation probabilities directly in an incentive-aligned manner, and compare these beliefs to the truth. Naive query formation refers to a special case of strategic query formation, in which the consumer naively believes that the probability that a term will be included in the results is only a function of whether this term is included in the query. In Section 5, we experimentally test the predictions derived in this section regarding naive query formation: when consumers care about the average utility across search results, naive queries include all terms whose relevance (i.e., marginal utility) is above a threshold; when consumers care about the maximum utility across search results, in our experimental setup naive queries would never include low-relevance terms.

4 Optimal queries may omit some of the more relevant terms: evidence from field data

Before exploring consumers' ability to identify optimal queries and test predictions from the naive benchmark in Section 5, in this section we explore characteristics of optimal queries, using field data. In particular, we show that shorter queries that omit some of the relevant terms may be more effective at retrieving desired content, compared to queries that include all relevant terms, even in the absence of search costs (i.e., when $c = 0$).

We start by assembling a set of popular short queries (e.g., "food network"). Next, we identify sets of terms related to these queries (e.g., "recipes"). We consider a consumer for whom the relevant terms consist of the union of the terms in one of the short queries and one related term (e.g., {food, network, recipes}). We compare the results this consumer would obtain by submitting the shorter query (that only contains a subset of the relevant terms, e.g., "food network"), to the results that would be retrieved by longer queries that combine the shorter query with the other relevant term (e.g., "food network recipes"). We now describe each step of this procedure in detail.

4.1 Data

Shorter queries We collect, in the food domain, the 100 most popular search queries among online users from a few major platforms (including Google, Bing, and Yahoo!) in March 2016. The ranking is obtained from the website http://tools.seobook.com. Each query contains on average 2.42 terms with a standard deviation of 0.70. We label these popular queries as "shorter queries," in contrast to longer queries that are formed by adding related terms. (Note however that these queries were selected based on their popularity, not their length.) We form a vocabulary consisting of 98 words for this domain, using all the unique words that appear in these queries.

Related terms For each shorter query q, we identify terms from the vocabulary that tend to be of interest to consumers who use query q, by leveraging data publicly available from Google Trends. For any target query, Google Trends reports a list of "related queries" that users who search the target query also tend to search for (typically 25 related queries are provided for each query). Accordingly, as a measure of whether word w is related to query q, we use whether word w is part of any "related queries" of query q on Google Trends. We find that 95 out of the 100 shorter queries have at least one related term from the vocabulary.[2] For interested readers, we have reported all shorter queries and their related words in Appendix D.

Relevant words We consider all possible sets of relevant words $\{G\}$ that are obtained by combining one of the shorter queries with one of its related words. For instance, if the shorter query is "food network" and the related terms are "recipes" and "tv," then there are two sets of relevant words corresponding to query q: {food, network, recipes} and {food, network, tv}. In total, we have 382 unique sets of relevant words, each containing 3.38 words on average.

Longer queries For each set of relevant words $G = \{q, w\}$, we compare the results from the shorter query q with the results from the queries that combine word w with query q. We consider all possible ways of combining word w with query q, e.g., in the first previous example we would consider the following queries: "recipes food network," "food recipes network," and "food network recipes."

Search results Consumers generally tend to focus on the top results when evaluating search results from a query (Narayanan and Kalyanam 2015). Hence, for each query in our data, we collect its top 10 search results, using the Google customer search API. We use a script to automatically download the actual webpage content of these links. We record which of the words in the corresponding set G are contained in each webpage. We record whether the word appears anywhere in the actual webpage associated with the result, the title or the snippet shown on Google. The analysis reported here is based on the utility function in Eq. 1, i.e., on *whether* the word appears on the webpage associated with a result, not how many times it appears. The analysis reported in Appendix A is based on an alternative utility function that depends on how many times each word appears on the page, title, and snippet all together.

4.2 Results

We compare the shorter query to all longer queries, using both performance metrics, i.e., Eqs. 2 and 3. We assume that there is no search cost, i.e., $c = 0$, to provide a conservative test of the performance of the shorter query. Table 1 reports the proportion of times the shorter query performs at least equally well, or strictly better than all longer queries corresponding to the set G. We report the average across all sets, and across sets of sizes $|G| = 3, 4, 5$, for each scenario. Note that we make

[2]The five queries that do not have any related term in the vocabulary are: fast food com, food poisoning symptoms, boys food#, food lion weekly ad, pet food express.

Table 1 Benefits of omitting relevant terms

Performance metric	$\lvert G \rvert$	All words valued equally		One word in the shorter query valued at half of the others	
		Prob (shorter query $\geqslant$ all longer queries)	Prob (shorter query $>$ all longer queries)	Prob (shorter query $\geqslant$ all longer queries)	Prob (shorter query $>$ all longer queries)
Average utility of top search results	All	0.157	0.089	0.123	0.090
	3	0.113	0.054	0.067	0.042
	4	0.250	0.160	0.197	0.133
	5	0.296	0.185	0.232	0.204
Max utility among top search results	All	0.592	0.003	0.596	0.004
	3	0.529	0.000	0.523	0.000
	4	0.720	0.000	0.713	0.000
	5	0.593	0.037	0.556	0.037

Note: "Shorter query" refers to a query q that is among the 100 most popular queries in the food domain (according to seobook.com). "Longer queries" refer to queries that contain an additional word w which is related to query q (as indicated by Google Trends). Queries are evaluated based on their ability to retrieve the words in $G = \{q, w\}$. In the first panel, each word has the same value $\beta_j = 1$. In the second panel, one of the words in the shorter query is valued at $\beta_j = 1$, while all other words have $\beta_j = 2$. Average/Maximum utility of search results: the performance of a query is equal to the average/highest utility among all top search results no claim on whether the shorter query is the optimal query, as we do not test all possible queries. However, because we compare the shorter query against all longer queries that include all relevant words, when the shorter query performs better than all longer queries we can safely claim that the optimal query includes only a subset of the relevant words in G.[3]

We start by assuming that all relevant terms have the same utility, e.g., $\beta_j = 1$, in the first panel of Table 1. We see that when queries are evaluated based on the average utility across search results, it is possible to achieve weakly better performance with the shorter query than with any of the longer queries in 15.7% of the cases, and strictly better performance in 8.9% of the cases. When queries are evaluated based on the utility of the best search results, it is possible to achieve weakly better performance with the shorter query than with any of the longer queries in 59.2% of the cases, while it is very unusual for the shorter query to achieve strictly better performance (0.26%). Importantly, given the consensus in the extant literature that consumers tend to prefer shorter queries (Jansen et al. 2000b; Spink et al. 2001), a shorter query that achieves the same performance as a longer query should be more desirable for consumers. In our set up, such shorter query would strictly outperform all longer queries as soon as the search cost parameter c becomes positive.

Next, we take our analysis one step further, by investigating whether it may be optimal to omit some of the *most* relevant terms in the queries. For each set of relevant words $G = \{q, w\}$, we assume that all words have utility $\beta_j = 2$, with the exception of one of the words in the shorter query which has utility $\beta_j = 1$. That is, we consider cases in which the shorter query q actually omits a term that is among the most relevant ones for the consumer, and includes a less relevant term instead. For each set G, we consider all possible ways of choosing the lower-value term, e.g., if q contains three words, we consider the cases in which the first, second, or third terms are valued at 1, and all other terms are valued at 2. Results are presented in the right panel of Table 1. We see that there are still many cases in which the shorter query performs at least as well or strictly better than any longer query.

In order to provide more intuition for our results, Table 2 displays the candidate queries for two sets of relevant words as examples, along with the proportion of search results that contain each word and the performance of each query, when $\beta_j = 1$ for all words. In the first example, a consumer interested in recipes from the food network magazine would be weakly better off simply using the query "food network magazine" rather than a longer query that includes "recipes," even without taking into account the general preference for shorter queries. This is because 70% of the results from the shorter query already contain the term "recipes." Adding "recipes" into the query increases this proportion to 80%, at the expense of the other terms "network" and "magazine," i.e., the results may contain recipes that do not come from the food network magazine. All queries retrieve at least one page that contains all the relevant terms, i.e., they all perform the same on the "Max" metric. The shorter query performs slightly better on the metric based on the average

[3]The only reason this would not be true would be if the optimal query included all the words in G plus some additional non-relevant terms, which is highly unlikely. In this paper we focus on queries that only include relevant terms.

Table 2 Examples of results

	Prob(t_1)	Prob(t_2)	Prob(t_3)	Prob(t_4)	Avg.	Max.
{"food network magazine," "recipes"}	food	network	magazine	recipes		
Food network magazine	1.0	1.0	0.9	0.7	3.6	4
Food network magazine recipes	1.0	1.0	0.7	0.7	3.4	4
Food network recipes magazine	1.0	0.9	0.7	0.8	3.4	4
Food recipes network magazine	1.0	0.9	0.6	0.7	3.2	4
Recipes food network magazine	1.0	0.9	0.7	0.8	3.4	4
{"kitchenaid food processor," "cuisinart"}	kitchenaid	food	processor	cuisinart		
Kitchenaid food processor	0.8	1.0	0.9	0.1	2.8	4
Kitchenaid food cuisinart processor	0.5	0.8	0.7	0.4	2.4	4
Kitchenaid cuisinart food processor	0.5	0.9	0.8	0.5	2.7	4
Cuisinart kitchenaid food processor	0.5	0.9	0.8	0.5	2.7	4
Kitchenaid food processor cuisinart	0.5	0.9	0.8	0.5	2.7	4

Note: For each query, the first four columns report the proportion of top search results that contain each word, and the last two columns report the performance of the query under each performance metric. The value of each word β_j is assumed to be one. Avg./Max. denotes the average/largest utility across search results

across search results. In the other example, a consumer who is looking for information on Kitchenaid and/or Cuisinart food processors, may be better off using the query "Kitchenaid food processor" rather than a longer query that includes "Cuisinart." This is because Cuisinart food processors are often compared to Kitchenaid, and the shorter query will retrieve at least one search result that contains all relevant terms. Including "Cuisinart" into the query greatly increases the proportion of results that contain this term, but this comes at the expense of other terms, in particular "Kitchenaid."

In Appendix A, we consider an alternative utility function, that does not only take into account *whether* a word appears on a page, but also *how many times* it appears. Results tend to be stronger in favor of shorter queries under this alternative utility function. For example, when the consumer cares about the maximum utility among search results, the shorter query now performs *strictly* better than all longer queries in 20.94% of the cases.

In sum, our field data documents the existence of cases in which consumers stand to benefit from strategically omitting some of the most relevant terms from their search queries, as shorter queries may be at least as effective at retrieving desired content, compared to queries that contain all the terms the consumer is searching for. The fact that consumers tend to prefer shorter queries (Jansen et al. 2000b; Spink et al. 2001) makes these shorter queries even more attractive.

We close this section by noting that the existence of complex activation probabilities may not be the only factor that makes it optimal to sometimes omit relevant terms from search queries. For example, search engines typically do not allow users to assign different weights across terms in their queries. Consequently, it may be optimal to omit less relevant terms, in order to focus the search results on the most relevant terms. While this phenomenon may happen in practice, it does not explain the results in this section, as we focus on cases in which all relevant terms are equally valuable to the consumer, or in which it is optimal to omit one of the *most* relevant terms. Future research may further explore the various reasons why it may be optimal to omit relevant terms from search queries in practice.

5 Consumer query formation: experimental evidence

We developed an incentive-aligned experimental paradigm to directly test and measure consumers' ability to identify optimal queries, and to test predictions from the naive benchmark. Using this paradigm, we ran one experiment (Study 1) with the average-utility performance metric defined in Eq. 2, and another experiment (Study 2) with the maximum-utility performance metrics defined in Eq. 3.

5.1 Design

We built our experimental paradigm as a "search query game," with the following specifications: (i) the relevant words G should be set exogenously and provided to participants; (ii) the game should be incentive-aligned, i.e., participants' payment should be a function of the performance of their queries; (iii) the performance of a query should be independent of the particular computer on which the game is played; (iv) in order to focus exclusively on query formation, any other type of search behavior such as evaluating results and clicking on links should be excluded from the game; (v) the game should capture the essence of query formation on search engines; (vi) the game should be easy to explain to participants. Taking the above into consideration, our search query game asks each participant to form search queries on Google to win a cash bonus. We formed the sets of relevant words $\{G\}$ by selecting nouns in the car domain (for Study 1) and in the food domain (for Study 2), because both are very common domains in which we expect all participants to have at least some knowledge. In each domain, we formed 10 overlapping sets of 3 words corresponding to each search task. See Table 3.

Each participant completed 10 independent search tasks in a random order in each study. In each task, the participant was asked to form a query consisting of any non-empty subset of the words in G in any order. We randomized the order in which the three words were displayed to the participant in each task, in order to avoid any potential ordering effect. In both studies, we varied the values of the three words (β_1, β_2, β_3) by selecting randomly (with equal probabilities) one of the following four sets of values for each of the 10 tasks and each participant: (\$2, \$2, \$2), (\$1, \$2, \$2), (\$2, \$1, \$2), and (\$2, \$2, \$1). Like in the field data, we consider the pages associated with the top 10 results of each query. The utility of each page is given by Eq. 1.

In Study 1, the performance score associated with each query was the average utility of the top 10 results minus the cost of the query in dollars (Eq. 2). We simply set the cost of each query to \$1 times the number of words in the query, i.e., $c =$

Table 3 Search tasks

Task	Study 1			Study 2		
	t_1	t_2	t_3	t_1	t_2	t_3
1	minivan	safe	economy	candy	caffeine	sugar
2	sedan	comfortable	luxury	fish	tea	tomato
3	SUV	comfortable	luxury	milk	cheese	tea
4	minivan	comfortable	luxury	Easter	candy	egg
5	hatchback	efficient	economy	tomato	drink	pizza
6	sedan	efficient	Japanese	Easter	caffeine	ketchup
7	hatchback	comfortable	compact	sugar	cake	pizza
8	SUV	efficient	compact	egg	candy	drink
9	SUV	luxury	American	cake	cheese	Easter
10	sedan	economy	Japanese	ketchup	cake	tomato

1. Figure 1 shows an example from Study 1 in which G = {luxury, comfortable, minivan}, valued respectively at \$2, \$2, and \$2. In Fig. 1a, the participant is forming his or her query by deciding which words to use and in which order. After submitting the search query, on the next page the participant is shown the title and url of the top 10 links and the list of words that were found on each page, along with the participant's performance score for this task. For example, Fig. 1b shows the interface after submitting the query "luxury minivan" in which two words were picked.

In Study 2, the performance score associated with each query was equal to the utility of the best search result, minus the cost of the query in dollars (Eq. 3). Figure 2 shows an example from Study 2 in which G = {milk, cheese, tea}, valued respectively at \$1, \$2, and \$2. In Fig. 2a, the participant is forming his or her query by deciding which words to use and in which order. In this case, although the words "cheese" and "tea" are worth more than the word "milk," only "milk" has a strong association with both of the other two words. This implies that forming the one-word query "milk" may achieve the highest score. Figure 2b shows the result after submitting the query "tea cheese" in which the two words with the higher value were picked; Fig. 2c displays the result after submitting the query "milk."

For each participant, we randomly chose at the end of the game the score from one of the 10 tasks and paid that amount as a bonus to the participant, in addition to a show-up fee.[4] To ensure that participants understood the instructions, they were given a short quiz after reading the instructions. Participants proceeded to the game

[4]Before running each study, we obtained all the activation probabilities using the same approach as with our field data (see "Search results" in Section 4.1). We ran all queries on a single computer to ensure that the results given to participants during the game would not be dependent on the computer on which the query was run. We used these results during the game, i.e., we did not actually run any query during the game. We also re-ran these queries using different computers, and the optimal queries and results were mostly consistent.

Fig. 1 Search query game interface in study 1. **a** shows the game interface when a participant forms a search query given the set of words, their values ($1 or $2 per word) and costs ($1 per word). The participant decides which word(s) to use and in which order. **b** shows the screen the participant would see after submitting the query "luxury minivan"

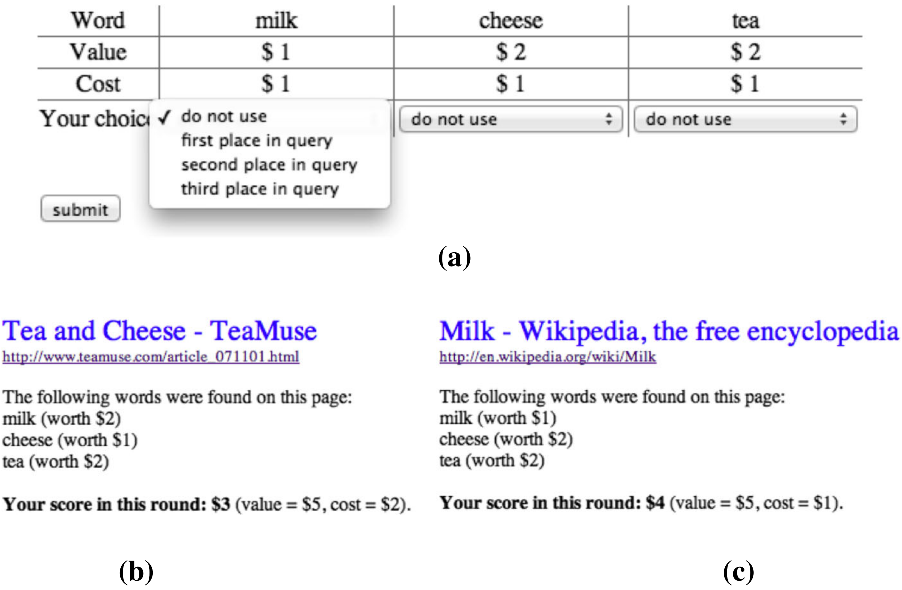


Fig. 2 Search query game interface in study 2. **a** shows the game interface when a participant forms a search query given the set of words, their values ($1 or $2 per word) and costs ($1 per word). The participant decides which word(s) to use and in which order. **b** and **c** show the screens the participant would see after submitting the queries “tea cheese” (**b**) and “milk” (**c**)

only after having answered all quiz questions correctly. While playing the game, participants were not allowed to use any other website. We enforced this by running the study in a lab in which we could observe and control the sites accessed by participants. The actual instructions of the game and the quizzes shown to participants are displayed in Appendix B.

5.2 Study 1

In Study 1, we collected data from N=199 participants recruited at a large university in the northeast of the United States.

When discussing each study, we first describe the optimal queries given the experimental setup. Next, we describe the queries that are consistent with naive beliefs. Finally, we describe the queries actually submitted by participants, report the proportion of queries that were optimal, and the extent to which the queries submitted by participants were consistent with the naive benchmark.

5.2.1 Optimal queries

We identify the query(ies) that actually maximized the payoff in each task played by each participant. In order to simplify the analysis, we classify queries into different types. Specifically, we have 502 observations in which all terms were valued at $2. In this case, there were four possible query types for each set of words: the empty query, and queries with one, two, or three terms. We find that payoffs were always maximized with two-term queries.

In the remaining 1,488 observations, one term was valued at \$1 and two terms at \$2. In this case, there were six possible types of queries for each set of words: the empty query, the query with the low-value term only, queries with one high-value term only, queries with one high-value term and the low-value term, queries with the two high-value terms, and queries with all the three terms. We find that the optimal query always had two terms: 38.78% of the optimal queries had two high-value terms, while the other 64.45% had one high-value term and the low-value term. If we exclude 46 observations in which the optimal payoff was achievable by more than one query types, then these percentages become 36.74% and 63.26% respectively.

5.2.2 Naive queries

According to Eq. 10, as long as $\frac{1}{2} \leq \alpha^{high} - \alpha^{low} < 1$, the threshold for including a term in the query is higher than 1 but lower than or equal to 2, and naive queries include all terms valued at \$2. The condition $\frac{1}{2} \leq \alpha_{high} - \alpha_{low}$ means that the marginal impact of including a term in a query is large enough and leads to inclusion of the \$2 terms in the query. The condition $\alpha^{high} - \alpha^{low} < 1$ means that a term may still appear in a search result even if it is not part of the search query ($\alpha^{low} > 0$), and/or there is no guarantee that all terms in the search query will appear in all search results ($\alpha^{high} < 1$), and leads to the omission of the \$1 term.

This means that in the range of $(\alpha^{high}, \alpha^{low})$ that covers all reasonable values, the query that maximizes Eq. 9 does not depend on α^{high} and α^{low}. When one term was valued at \$1, participants with naive beliefs should always choose a query consisting of two high-value terms. This query type was shown in the previous subsection to be optimal in only 38.78% of the observations. When all terms were valued at \$2, naive beliefs would lead participants to always choose a query with all three terms, which the previous subsection showed was never optimal. This confirms the sub-optimality of the naive queries. More importantly, this also provides a set of predictions that can be compared to the queries that were actually chosen by participants.

5.2.3 Queries chosen by participants

We now examine the queries actually submitted by participants. Table 4 reports the number of observations in which each query type was submitted by participants, crossed with whether the query achieved the maximum payoff for that task. We first discuss the results when one word was valued at \$1, reported in Table 4a. Naive queries, i.e., queries with two high-value terms, were chosen only in 42.20% of the observations. And in 33.26% of the observations, participants instead submitted a query of a type that includes the low-value term while excluding at least one of the high-value terms. Table 4b reports the results when all words were valued at \$2. Naive queries, i.e., queries with all three terms, were chosen in only 23.90% of the cases. These observations suggest that participants usually did not follow the naive benchmark when forming their queries.

Table 4 Participants' queries in study 1

Query type	Query did not achieve max. payoff	Query achieved max. payoff	%Obs.
(a) One term was valued at $1 (N=1,488)			
The low-value term	35	0	2.35%
One high-value term	135	0	9.07%
Two high-value terms*	509	119	42.20%
One high-value term + the low-value term	383	77	30.91%
All three terms	230	0	15.46%
Column Total	86.83%	13.17%	100%
(b) All terms were valued at $2 (N=502)			
One term	53	0	10.56%
Two terms	276	53	65.54%
All three terms*	120	0	23.90%
Column Total	89.44%	10.56%	100%

Note: Each table reports the number of times each query type was observed, crossed with whether the query achieved the highest payoff for that task. The last column reports the proportion of observations in which each query type was observed. *: naive query type

Nevertheless, we also see that participants were usually unable to identify the optimal query. Only 13.17% of the observed queries were optimal when one word was valued at $1, and 10.56% were optimal when all words were valued at $2. To illustrate this more clearly, Table 5 reports the proportion across all observations in which the observed query was / was not of the type consistent with naive beliefs, crossed with whether the query was optimal. We see that in the majority of cases (55.88 %), the query type was inconsistent with naive beliefs, but the query was not optimal.

These results bring up the question of whether participants simply submitted queries randomly, irrespective of the optimal query type. To shed light on this question, we investigate whether the type of query submitted by participants was different

Table 5 Participants' queries in study 1

	Query did not achieve max. payoff	Query achieved max. payoff
Query type consistent with naive beliefs	629 (31.61%)	119 (5.98%)
Query type inconsistent with naive beliefs	1,112 (55.88%)	130 (6.53%)

Note: Number of times (proportions in parentheses) query type was consistent with naive beliefs, crossed with whether the query achieved the highest payoff for that task

Table 6 Participants' queries in study 1 conditional on optimal query type - one term valued at \$1 (N=1,440)

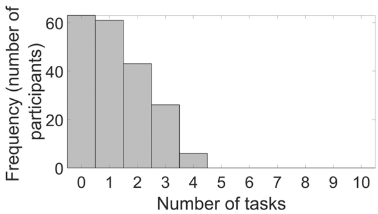
	Optimal query type	
Query type	One high-value term + the low-value term	Two high-value terms
The low-value term	2.96%	0.95%
One high-value term	8.01%	11.34%
Two high-value terms*	40.07%	49.15%
One high-value term + the low-value term	34.36%	23.25%
All three terms	14.60%	15.31%

Note: Each column shows the percentage of observations of each query type, in tasks where the optimal query type was the one indicated at the top. In order to split observations in a mutually exclusive manner, we focus on observations in which a single query type was optimal. There are no columns for the query types that never achieved the maximum payoff. *: naive query type

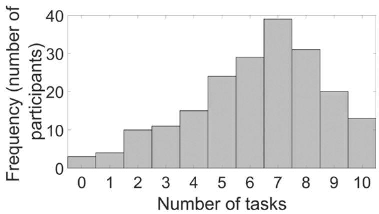
based on which query type was optimal. That is, we compute the probability of observing each query type, conditional on each query type being optimal. This conditional analysis can only be performed among the observations in which one term was valued at \$1, as the same query type was always optimal when all words were valued at \$2. We also limit this analysis to observations in which the optimal payoff was achieved by a single query type, in order to avoid counting the same observation multiple times. Table 6 reports the conditional probabilities. A χ^2 test of independence rejects the null hypothesis that the conditional probabilities are identical across columns ($p < 0.01$), i.e., we reject the hypothesis that participants did not adjust their queries based on the optimal query type. In addition, we see that participants were relatively more likely to submit a query with one high-value term and the low-value term when this query type was optimal vs. not (34.36% vs. 23.25%). Similarly, they were relatively more likely to submit a query with two high-value terms when this query type was optimal vs. not (49.15% vs. 40.07%). While such conditional analysis is not possible when all words were valued at \$2, we do observe that in these cases the most commonly observed query type was the optimal one (i.e., queries with two terms, see Table 4b).

Finally, Fig. 3 illustrates heterogeneity across participants, in their propensity to submit optimal queries (Fig. 3a) and to submit queries that deviate from the naive benchmark (Fig. 3b). Out of 10 tasks, participants submitted queries that were inconsistent with naive beliefs in 6.24 tasks on average (std.= 2.35), and they submitted queries that achieved maximum payoff in 1.25 tasks on average (std.= 1.13).

In conclusion, our results from Study 1 suggest that although participants were usually not able to identify the optimal query, they selected their queries neither in a completely random manner nor in a naive manner. In order to reconcile these results, it is important to remember that strategic query formation relies upon a set of beliefs on activation probabilities, $\{\widetilde{Prob}(t_j \in l|q)\}$. Consider a strategic consumer who would act upon beliefs on activation probabilities that are not naive, but that



(a) Distribution across participants of the number of tasks (out of 10) in which the submitted query achieved maximum payoff



(b) Distribution across participants of the number of tasks (out of 10) in which the submitted query was not consistent with naive beliefs

Fig. 3 Study 1 results at the participant level

are not completely correct either. Such consumer would form queries that are often suboptimal, but that do attempt to leverage activation probabilities, and that sometimes exclude some of the more relevant words / include some of the less relevant words. Therefore, our results are consistent with participants being strategic in forming their queries, but having incorrect beliefs on activation probabilities. In Study 3 (reported in Section 6), we directly measure (in an incentive-aligned manner) participants' beliefs on activation probabilities. We find that indeed, participants' beliefs tend to be inaccurate.

5.3 Study 2

We obtained results from N=108 participants recruited at a large university in the northeast of the United States.

5.3.1 Optimal queries

We again start by identifying the queries that actually maximized the payoffs in each task played by each participant. Among the 273 observations in which all terms were valued at \$2, we find that it was always optimal to use queries with one (67.40%) or two (32.60%) terms. Among the remaining 807 observations in which one term was

Table 7 Optimal queries in study 2 – one term valued at \$1

Query type	All Obs. (N=807)	Single query type was optimal (N=709)
Empty query	0%	0%
The low-value term	23.67%	26.94%
One high-value term	56.51%	53.74%
Two high-value terms	14.87%	3.10%
One high-value term + the low-value term	26.39%	16.22%
All three terms	0%	0%

Note: This table reports the proportion of observations in which the optimal payoff was achieved by each query type valued at \$1, the optimal queries also contained either one or two terms. Details are provided in Table 7.

5.3.2 Naive queries

As mentioned in Section 3.5, when performance is measured based on the maximum utility across search results (3), the objective function is *not* separable in the terms $\{t_j\}$ under naive beliefs, it is not possible to optimize the objective function by making decisions on each term separately, and there is no simple algebraic expression that characterizes naive queries. We analyze the naive benchmark under this objective function in Appendix C. We consider situations in which all terms are valued at \$2, and in which two terms are valued at \$2 and one term is valued at \$1. We maximize the objective function numerically, when the two belief parameters α^{high} and α^{low} vary between 0 and 1. When all three words are valued at \$2, we find that queries with 0, 1, 2 or 3 terms may be selected, depending on the values of α^{high} and α^{low}. When two words are valued at \$2 and one word is valued at \$1, we find that different query types may be selected, but the naive objective function is never maximized by a query that includes the \$1 term. Hence, the naive benchmark makes the prediction that participants should never include \$1 terms in their queries. This further confirms the sub-optimality of naive queries, as according to Table 7, the query with the low-value term was in fact strictly optimal in 26.94% of the observations, and the query type with one high-value term and the low-value term was strictly optimal in 16.22% of the observations. More importantly, this again provides predictions which we compare against the actual queries submitted by participants. That is, in this study we use observations of queries that include the \$1 term as evidence against the naive benchmark.

5.3.3 Queries chosen by participants

Table 8 reports the number of observations in which each query type was observed, crossed with whether the query achieved the optimal payoff, when one word was valued at \$1 (Table 8a) and when all words were valued at \$2 (Table 8b). We see that a query including the low-value term was submitted in 34.94% of the observations in which one term was valued at \$1 (summation across the three types of queries that include the low-value term in Table 8a). All of these observations contradict the hypothesis that participants formulated their queries using the naive benchmark, as this benchmark predicts that the low-value term should never be used.[5]

Nevertheless, as in Study 1, participants were usually unable to maximize their payoffs. Only 24.04% of the queries were optimal when one word was valued at \$1, and 22.34% were optimal when all words were valued at \$2. This again brings up the question of whether participants simply submitted queries randomly, irrespective of the optimal query type. To shed light on this question, we report in Table 9 the

[5] As the specific query type selected under naive beliefs depends on the unobserved belief parameters α^{low} and α^{high} in this study, we do not create a table like Table 5 or a figure like Fig. 3b.

Table 8 Participants' queries in study 2

Query type	Query did not achieve max. payoff	Query achieved max. payoff	%Obs.
(a) One term was valued at \$1 (N=807)			
The low-value term	24	28	6.44%
One high-value term	110	80	23.54%
Two high-value terms	262	73	41.51%
One high-value term + the low-value term	144	13	19.45%
All three terms	73	0	9.05%
Column total	75.96%	24.04%	100%
(b) All terms valued at \$2 (N=273)			
One term	46	41	31.87%
Two terms	92	20	41.03%
All three terms	74	0	27.11%
Column total	77.66%	22.34%	100%

Each table reports the number of times each query type was observed, crossed with whether the query achieved the highest payoff for that task. The last column reports the proportion of observations in which each query type was observed. When one term is valued at \$1, naive queries should have one or two high-value terms (depending on the respondent's beliefs represented by α^{high} and α^{low}), but not include the low-value term. When all terms are valued at \$2, queries with 0, 1, 2 or 3 terms may be selected according to the naive benchmark, depending on the respondent's beliefs

probability of observing each query type, conditional on each query type being optimal, when one word was valued at \$1 (Table 9a) and when all words were valued at \$2 (Table 9b). We again limit this analysis to observations in which the optimal payoff was achieved by a single query type, in order to avoid counting the same observation in multiple columns. The null hypothesis that the conditional probabilities are identical across columns is rejected in both Tables 9a and b, based on χ^2 tests of independence ($p < 0.01$). That is, similar to Study 1, we reject the hypothesis that participants did not adjust their queries based on the optimal query type.

In Study 1, we found that each query type was most likely to be observed when it was optimal. As can be seen in Table 9, in this study, we find this is often, but not always the case. When one word was valued at \$1, three of the query types (low-value term, one high value term, two high-value terms) were each most likely to be observed when they were optimal. However, queries with one high-value term and the low-value term were most likely to be observed when the optimal query type was the one with the high-value term only. When all words were valued at \$2, queries with one term were most likely to be observed when this was the optimal query type, but queries with two terms were most likely to be observed when queries with one term were optimal.

Table 9 Participants' queries in study 2 conditional on optimal query type

(a) One term was valued at $1 (N=709)

	Optimal query type			
Query type	The low-value term	One high-value term	Two high-value terms	One high-value term + one low-value term
The low-value term	14.66%	4.99%	9.09%	0.87%
One high-value term	15.71%	31.23%	9.09%	22.61%
Two high-value terms	41.36%	32.28%	68.18%	52.17%
One high-value term + the low-value term	20.94%	24.93%	4.55%	10.43%
All three terms	7.33%	6.56%	9.09%	13.91%

(b) All terms valued at $2 (N=273)

	Optimal query type	
Query type	One term	Two terms
One term	35.87%	23.60%
Two terms	43.48%	35.96%
All three terms	20.65%	40.45%

Note: Each table shows the percentage of observed query type in tasks where the query type achieving the maximum payoff was the one indicated at the top. In order to split observations in a mutually exclusive manner, we focus on observations in which a single query type was optimal. There are no columns for the query types that never achieved the maximum payoff

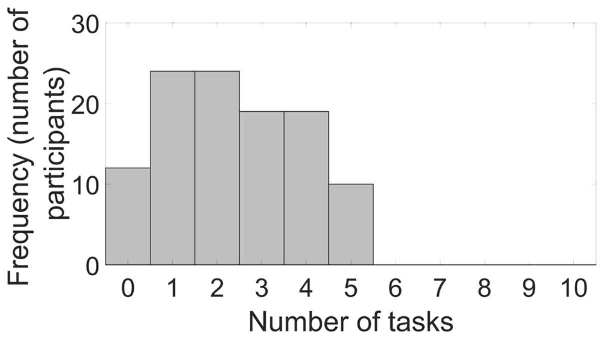


Fig. 4 Study 2 - Distribution across participants of the number of tasks (out of 10) in which the submitted query achieved maximum payoff

Finally, Fig. 4 illustrates heterogeneity across participants, in their propensity to submit optimal queries. Out of 10 tasks, participants submitted queries that achieved maximum payoff in 2.36 tasks on average (std.= 1.51).

To sum up, Studies 1 and 2 suggest that participants are able to strategically omit relevant words from search queries, and to include less relevant words at the expense of more relevant words. Nevertheless, we also find that participants were often unable to maximize their payoffs. As noted previously, this is consistent with participants being strategic in forming their queries, but having incorrect beliefs on activation probabilities.

Given that our study uses a somewhat artificial lab setting, we do not claim that the *extent* to which consumers are able to identify optimal queries / deviate from naive query formation in the real world is the same as in our study. Instead, we view our results as proof of existence that consumers have some ability to strategically formulate queries that contain only a subset of the terms they are interested in. An analogy may be made to the experimental economics literature. Games such as the dictator game are used in this literature to show that individuals have the potential to behave in ways that are inconsistent with maximizing their own financial well-being, although these games do not quantify the extent of such behavior in real life. We argue that our results are enough to warn researchers and practitioners against assuming that consumers formulate naive queries in which terms would be included only on the basis of their relevance. Such naive queries are not only suboptimal from a normative perspective, they also do not conform with our data.

6 Study 3: consumers' beliefs on activation probabilities

Both studies 1 and 2 are consistent with participants forming queries strategically, but having incorrect beliefs on activation probabilities. In order to further explore this possibility, Study 3 measures participants' beliefs on activation probabilities directly, in an incentive-aligned manner.

6.1 Design

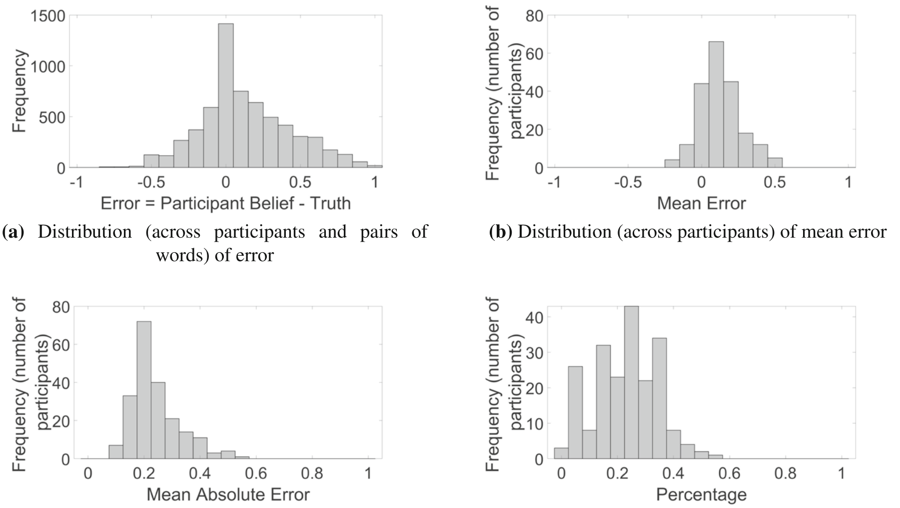
Activation probabilities may be defined at different levels. For example, they may be defined from queries to words, i.e., the probability that query q will activate a word t_j: $Prob(t_j \in l|q)$. For simplicity, here we focus on word-to-word activation probabilities, more precisely, the probability that a one-word query t_i will activate a certain word t_j: $Prob(t_j \in l|t_i)$. If consumers' beliefs on these simple activation probabilities are incorrect, their beliefs on more complex activation probabilities that involve longer queries, are likely to be even further from the truth.

We measured participants' beliefs on activation probabilities directly, by asking them for example to "estimate how many of the top 10 search results from the query *egg* on Google contain the word *fish*." Again, we told participants that a search result contains the word if it appears on the actual page, and not just the description provided by Google. Participants chose a number between 0 and 10 as their answers, i.e., they entered their best guess of the number of top results containing the target word. We formed 30 pairs of words using the same words as in Study 2, giving us 60 possible questions (recall that activation probabilities are directional). We chose the pairs of words to have a large range of true activation probabilities across pairs. Each participant answered 30 questions that were randomly selected from the pool. After participants answered these 30 questions, we presented them with the correct answers for all of them at once. This study was incentive-aligned. In addition to a \$3 show-up fee, each participant won a \$0.20 cash bonus for each correct answer.

6.2 Results

We conducted this study in a lab with N=206 participants. Among the 60 questions we selected, 0 is the correct answer for about one third. Hence, choosing 0 could be an attractive strategy for participants in this study. However, we found that all participants gave at least three different answers across questions. Therefore, no participant blindly selected 0 as their answers to all questions.

We compare participants' beliefs to the true activation probabilities in each question. Overall, participants provided correct answers (i.e., correctly guessed the true activation probability) in 22.88% of the cases. The distribution (across participants and pairs of words) of the error (i.e., participant's answer minus the truth) is displayed in Fig. 5a. The distribution has a mean of 0.130 and a standard deviation of 0.301. i.e., participants' beliefs are biased upward. The mean *absolute* error across participants and pairs is 0.239, with a standard deviation of 0.224. We report additional figures that explore heterogeneity across participants. Figure 5b reports the distribution across participants of the average error across pairs of words. The average error is positive (i.e., the participant's beliefs are biased upward) for 84.47% of the participants, and the standard deviation of the average error is 0.138. Figure 5c plots the distribution across participants of the mean absolute error across pairs. The mean absolute error has a standard deviation across participants of 0.086. Figure 5d shows the distribution across participants of the proportion of correct answers.



(a) Distribution (across participants and pairs of words) of error

(b) Distribution (across participants) of mean error

(c) Distribution (across participants) of mean absolute error

(d) Distribution (across participants) of percentage of correct responses

Fig. 5 Study 3 results

The proportion of correct answers has a standard deviation across participants of 0.113.

In sum, this analysis suggests that consumers' beliefs on activation probabilities are only approximate. In addition, we find that beliefs on activation probabilities tend to be biased upwards.

7 Discussion, implications, and future research

Our research findings may be summarized as follows. First, consumers stand to benefit from being strategic when formulating their search queries: a short query that does not contain all the relevant terms a consumer is searching for is often as effective or more effective at retrieving these terms, compared to longer queries that contain all relevant terms. The fact that consumers tend to prefer shorter queries make these queries even more attractive. Second, the queries formulated by consumers are typically *not* consistent with a naive benchmark according to which consumers would only include the most relevant terms in their queries. Instead, consumers have the ability to formulate queries that omit some of the more relevant terms. That is, queries formed by consumers are not necessarily straightforward expressions of their content preferences, but rather the outcome of a strategic attempt to leverage activation probabilities in order to retrieve relevant content. In particular, our results are consistent with the argument that consumers attempt to formulate queries that are more likely to *retrieve* the content they are searching for, rather than merely being *similar* to that content. Third, consumer's queries, while not being consistent with a naive view of query formation, are usually suboptimal. Moreover, consumers beliefs on activation probabilities tend to be incorrect. Together, these results paint a picture of consumer search query formation in which consumers attempt to strategically optimize search queries by leveraging activation probabilities, but acting on incorrect beliefs on activation probabilities.

Our research provides implications for both practitioners and researchers. For researchers, our findings pave the way for the development of utility-based models that link consumers' content preferences to their textual queries. In today's environment, search is primarily text-based, and marketing models of search may be adapted to capture this reality. Our research suggests that utility-based models capturing text-based search should allow consumers to be strategic when formulating search queries.

In terms of managerial implications, our results are relevant for practitioners interested in leveraging online search queries as a source of consumer insights. More specifically, our findings have clear implications for practitioners in the SEO/SEM industries. SEO/SEM deal prominently with determining relevant search keywords (i.e., keywords on which to bid in the case of SEM, or on which to improve organic search rankings in the case of SEO), and optimizing the content delivered to consumers (link description or ad copy, and landing page). Despite the importance of the SEO/SEM industries, many marketers still struggle with these decisions (Wiltshire 2015). Marketers should focus their efforts on keywords and queries that reflect content preferences that are well aligned with the content they are trying to promote. If consumers formulated queries naively, a search query would simply provide the list of the most relevant terms for the consumer. In contrast, our research suggests that the content that is of interest to consumers is not simply the content in their search queries, but also the content retrieved by their search queries. Therefore, our results suggest that the set of keywords/queries on which a piece of content should be promoted should not be limited to keywords/queries that have more words in common with the target content, but also include queries that are more likely to retrieve the target content. In addition, ad copies and page descriptions provided to consumers should highlight the relation between the content being promoted and the content preferences revealed by the query.

Our findings also challenge the common way in which SEO/SEM campaigns are currently structured and tracked. Marketers typically rely on data from Google Analytics and manual research, and maintain spreadsheets that track performance at the keyword or ad group level (Wiltshire 2015).[6] That is, the raw search queries

[6]For more examples, one can refer to these articles/blogs: https://neilpatel.com/blog/seo-excel-hacks/; https://blog.hubspot.com/marketing/on-page-seo-template; https://www.distilled.net/excel-for-seo/; https://medium.com/@jacobjs/beginners-guide-to-seotools-for-excel-part-1-db84ed54daff; https://moz.com/blog/one-formula-seo-data-analysis-made-easy-excel; https://mainpath.com/using-excel-as-an-seo-tool/; https://cleverclicks.com.au/blog/seo-excel-formula-toolkit/.

are typically ignored. We argue that focusing on keywords or ad groups leads to a loss of valuable information about consumers, as our results suggest that a search keyword is best studied in the context of the original search query in which it appears. Indeed, if consumers are strategic in query formation, the information contained in a keyword about consumer preferences should vary based on the context in which the keyword is used. Our research suggests that taking a consumer-centric approach to SEO/SEM may involve focusing on consumer content preferences (or needs) as the basis of analysis, rather than keywords or ad groups. These content preferences should be mapped onto search queries, search keywords, and search results. Such mapping, which may be constructed by leveraging publicly available data from search engines (e.g., Google API, Google Trends, Google Correlate), may help marketers develop a better understanding of the variations in performance across keywords. It may also help marketers identify a broader set of keywords that are relevant to the content they would like to promote, as well as content that users would be interested in seeing in an ad copy or landing page.

We close by highlighting additional areas for future research. First, while in this paper we focus on the consumers' decision of *which* words to include in their queries, future research may study how consumers decide to *order* the words in their queries. Second, future research may study the impact of query auto-completion (i.e., the search engines auto-completes the user's queries) on the link between content preferences and query formation. Auto-complete suggestions reflect co-occurrence of words in queries across all users. In contrast, leveraging activation probabilities between queries and results allows consumers to improve the quality of the search results given their *own* particular content preferences. Hence, from the perspective of consumers, auto-completion should not be a substitute for leveraging activation probabilities between queries and search results. Third, we hope that future research will integrate query formation with downstream behavior such as clicking, into a comprehensive revealed preference framework. Finally, in this paper we focused on the context of online search engines because it is prominent and lends itself to precise measurement. Future research may study similar strategic behavior in offline search contexts, where consumers interact with real estate agents, guidance counselors, car dealers, etc. (see for example Dzyabura and Hauser 2018).

Appendix A: Alternative utility function

We reproduce Table 1 under an alternative utility function, which captures the number of times each word appears on a page, not just whether it appears. That is, the utility function in Eq. 1 is replaced with:

$$U(l) = \sum_{t_j \in G} \beta_j \, n_{t_j,l} \tag{11}$$

where, $n_{t_j,l}$ is the number of times word t_j appears on webpage l. The results are summarized in Table 10.

Table 10 Benefits of omitting relevant terms

Performance metric	$\lvert G\rvert$	All words valued equally		One word in the shorter query valued at half of the others	
		Prob (shorter query $\geqslant$ all longer queries)	Prob (shorter query $>$ all longer queries)	Prob (shorter query $\geqslant$ all longer queries)	Prob (shorter query $>$ all longer queries)
Average utility of top search results	All	0.312	0.312	0.299	0.296
	3	0.313	0.313	0.283	0.281
	4	0.340	0.340	0.337	0.337
	5	0.333	0.333	0.315	0.306
Max utility among top search results	All	0.414	0.209	0.421	0.197
	3	0.338	0.213	0.317	0.192
	4	0.590	0.170	0.580	0.190
	5	0.444	0.259	0.444	0.259

Note: "Shorter query" refers to a query q that is among the 100 most popular queries in the food domain (according to seobook.com). "Longer queries" refer to queries that contain an additional word w which is related to query q (as indicated by Google Trends). Queries are evaluated based on their ability to retrieve the words in $G = \{q, w\}$. In the first panel, each word has the same value $\beta_j = 1$. In the second panel, one of the words in the shorter query is valued at $\beta_j = 1$, while all other words have $\beta_j = 2$. Average/Maximum utility of search results: the performance of a query is equal to the average/highest utility among all top search results

Appendix B: Instruction page for search query game

Overview
You will be playing a web **search query game**. You will submit **search queries** and you will score **points** based on the **search results** from these queries. The whole study will take you about 15 minutes.

Rules of the Game
You will play **10 rounds** of the game.

In each round:

- You will be given **3 words** to search for.
- Each word has a **value ($1 or $2)** associated with it, and a **cost (always $1)**.
- You will form a **search query** that contains any of these 3 words in any order. That is, your query may contain one, two or three of the words, in the order of your choice.
- After you submit your search query, in the background we will automatically run this query on Google and scan the webpages associated with the top 10 search results.
- Each search result has a different value based on which of the 3 words it contains. The value of each search result is obtained by checking which of the words are contained in the search result, and adding up the values of these words. That is, the more words are contained in the search result, the higher its value.
- But each query also has a cost. The cost of the query is equal to the number of words contained in the query. That is, you are penalized for including too many words in your query.

Your score in each round is the average value of the top 10 search results, minus the cost of your search query. That is, your score is equal to:

The proportion of the top 10 search results that contain the first word × the value of the first word
+ The proportion of the top 10 search results that contain the second word × the value of the second word
+ The proportion of the top 10 search results that contain the third word × the value of the third word
– The number of words in your query.

In other words, **the best queries are those that allow you to find many of three words in the set, while being short.**

Example
For example, suppose the following 3 words are given to you in a round:

Word	fruit	salad	chicken
Value	$2	$1	$2
Cost	$1	$1	$1

Suppose that you choose the query **‘chicken fruit.’**. That is, you select the words 'chicken' and 'fruit' out of the three words, in that order.

Suppose that:

- 8 of the top 10 search results from the query ‘chicken fruit’ contain the words ‘chicken'
- 9 out of the top 10 search results from the query ‘chicken fruit’ contain the word ‘fruit’
- 4 of the top 10 search results from the query ‘chicken fruit’ contain the word 'salad'

Then your score would be:

The proportion of the top 10 search results that contain the first word (8/10) × the value of the first word ($2)
+ The proportion of the top 10 search results that contain the second word (9/10) × the value of the second word ($1)
+ The proportion of the top 10 search results that contain the third word (4/10) × the value of the third word ($2)
– The number of words in your query ($2)
=$1.8

A different query may give you different proportions of top 10 search results that contain each word, and a different cost.

Your Final Bonus
At the end of the study, one round will be selected randomly. As a bonus, you will receive your score in this round in cash. For example, if your score is $4, you will receive a $4 bonus. This bonus will be in addition to your show up fee.

Note:

- What matters is whether a word appears anywhere on the actual webpage associated with a search result, not just the url or the blurb given by Google.
- It doesn’t matter how many times each word appears on the webpage associated with a search result, as long as it appears at least once.
- Your score cannot be negative, that is, you cannot loose money.
- You are not allowed to use any other web site while playing the game. Doing so will exclude you from the study.

Fig. 6 Study 1 – Instructions

Suppose that you are in a round in which the words and their values are given by:

Word	fruit	salad	chicken
Value	$2	$2	$1
Cost	$1	$1	$1

Suppose your query is 'fruit salad chicken.' Suppose that 8 of the top 10 results contain 'fruit,' 9 of the results contain 'salad,' and 7 of the results contain 'chicken.' Then your score would be:

○ 8 X $2 + 9 X $2 + 7 X $1 – $3=$38

○ 0.8 X $2 + 0.9 X $2 + 0.7 X $1 – $3=$1.1

○ 8 X $2 + 9 X $2 + 7 X $1=$41

○ 0.8 X $2 + 0.9 X $2 + 0.7 X $1=$4.1

Suppose your query is 'fruit' and 10 of the top 10 best result contains 'fruit,' 6 contain 'salad,' and 5 contain 'chicken.' Then your score would be:

○ 1 X $2 + 0.6 X $2 + 0.5 X $1 – $3=$0.7

○ 1 X $2 + 0.6 X $2 + 0.5 X $1 – $2=$1.7

○ 1 X $2 + 0.6 X $2 + 0.5 X $1 – $1=$2.7

○ 1 X $2 + 0.6 X $2 + 0.5 X $1=$3.7

Fig. 7 Study 1 – Quiz

Welcome to our Search Query Game!

Overview
You will be playing a web **search query game**. You will submit **search queries** and you will score **points** based on the **search results** from these queries. After you finish the game, you will fill out a short survey. The whole study will take you about 20 minutes.

Rules of the Game
You will play the game for **10 rounds**. In each round, you will be given **3 words** to search for. Each word has some **value** ($1 or $2) if you find it. In order to find these words you will form a **search query** that contains any of these 3 words in any order. Each word also **costs $1** if you use it in your query. That is, the fewer words you use in your query, the lower its cost.

Your task is to decide which words to use in your query and in which order.

For example, suppose the following 3 words are given to you in a round:

Word	fruit	salad	chicken
Value	$2	$1	$2
Cost	$1	$1	$1

You may decide to only use 'fruit', i.e., your query is 'fruit' and it costs $1; or you may decide to use 'fruit' first followed by 'salad,' i.e., your query is 'fruit salad' and it costs $2.

After you submit your search query, in the background we will automatically **run this query on Google and scan the webpages associated with the top 10 links**. Each link has a different value **based on which of the 3 words it contains.**

Your score in each round is the value of the best link minus the cost of your search query:
Score = Value of best link – Cost of the query

In other words, **the best queries are those that are short but allow you to find many of the words you are looking for.**

In the above example, suppose your query is 'chicken fruit' and we find that the best link contains the words 'chicken' and 'fruit' but not 'salad.' Then your score is $2 ($2 for 'chicken' plus $2 for 'fruit' minus $2 because there are 2 words in the query). Suppose now that your query is 'chicken salad' and we find that the best link contains the words 'chicken' and 'fruit' and 'salad.' Then your score is $3 ($2 for 'chicken' plus $1 for 'salad' plus $2 for 'fruit' minus $2 because there are 2 words in the query).

Your Final Bonus
At the end of the study, one round will be selected randomly. As a bonus, you will receive your score in this round in cash. For example, if your score is $4, you will receive a $4 bonus. This bonus will be in addition to your show up fee.

Note:

- You score points as long as the word appears anywhere on the actual webpage associated with the link, not just the url or the blurb given by Google.
- It doesn't matter how many times each word appears on the webpage associated with a link, as long as it appears at least once.
- Your score cannot be negative, that is, you cannot loose money.
- You are not allowed to use any other web site while playing the game. Doing so will exclude you from the study.

Fig. 8 Study 2 – Instructions

Suppose that you are in a round in which the words and their values are given by:

Word	fruit	salad	chicken
Value	\$2	\$2	\$1
Cost	\$1	\$1	\$1

Suppose your query is 'fruit salad chicken' and the best result contains 'fruit,' 'salad,' and 'chicken.' Then your score would be:

- ○ \$1
- ○ \$2
- ○ \$3
- ○ \$4
- ○ \$5

Suppose your query is 'fruit' and the best result contains 'fruit' and 'salad' but NOT 'chicken.' Then your score would be:

- ○ \$1
- ○ \$2
- ○ \$3
- ○ \$4
- ○ \$5

Suppose your query is 'chicken' and the best result contains 'salad' and 'chicken' but NOT 'fruit.' Then your score would be:

- ○ \$1
- ○ \$2
- ○ \$3
- ○ \$4
- ○ \$5

Fig. 9 Study 2 – Quiz

Appendix C: Naive queries in study 2

C.1 Two words valued at \$2, one word valued at \$1

Without loss of generality, we assume: $\beta_1 = \beta_2 = 2, \beta_3 = 1$. The support of the utility of webpages is {0,1,2,3,4,5}. We compute the probability distribution of the maximum utility across L pages, under the naive beliefs represented in Eq. 8. We simplify notations by setting $\widetilde{Prob}(t_i \in l|q) = p_i$ for $i \in \{1, 2, 3\}$. We start by computing the cumulative density function of the maximum utility, i.e, $\phi_j = Prob(\max_l\{U(l)\} \leq j)$ for $j \in \{0, 1, 2, 3, 4, 5\}$, as follows:

$$\begin{aligned}
\phi_5 &= 1 \\
\phi_4 &= (1 - p_1 p_2 p_3)^L \\
\phi_3 &= (1 - p_1 p_2)^L \\
\phi_2 &= \left[\sum_{i=1}^{3} \left(p_i \Pi_{j \neq i}(1 - p_j)\right) + \Pi_{i=1}^{3}(1 - p_i)\right]^L \\
\phi_1 &= (1 - p_1)^L (1 - p_2)^L \\
\phi_0 &= (1 - p_1)^L (1 - p_2)^L (1 - p_3)^L
\end{aligned}$$

Given this, we can express the objective function as (recall that $c = 1$ in the experiment):

$$\begin{aligned}
\tilde{f}^{\text{naive}}(q) &= E[\max_l\{U(l)\}|q] - |q| \\
&= 5(\phi_5 - \phi_4) + 4(\phi_4 - \phi_2) + 3(\phi_3 - \phi_2) + 2(\phi_2 - \phi_1) + (\phi_1 - \phi_0) - |q| \\
&= 5 - (\phi_4 + \phi_3 + \phi_2 + \phi_1 + \phi_0) - |q| \qquad (12)
\end{aligned}$$

There are six query types such that all queries from the same type achieve the same value of the objective function in Eq. 12:

1. Empty queries ($|q| = 0$): $p_1 = p_2 = p_3 = \alpha^{low}$
2. Queries with only the low value term ($|q| = 1$): $p_1 = p_2 = \alpha^{low}$, $p_3 = \alpha^{high}$
3. Queries with only one high value term ($|q| = 1$): we assume that $p_1 = \alpha^{high}$ and $p_2 = p_3 = \alpha^{low}$ without loss of generality
4. Queries with only one high value term and the low value term ($|q| = 2$): we assume that $p_1 = \alpha^{high}$ and $p_2 = \alpha^{low}$, $p_3 = \alpha^{high}$ without loss of generality

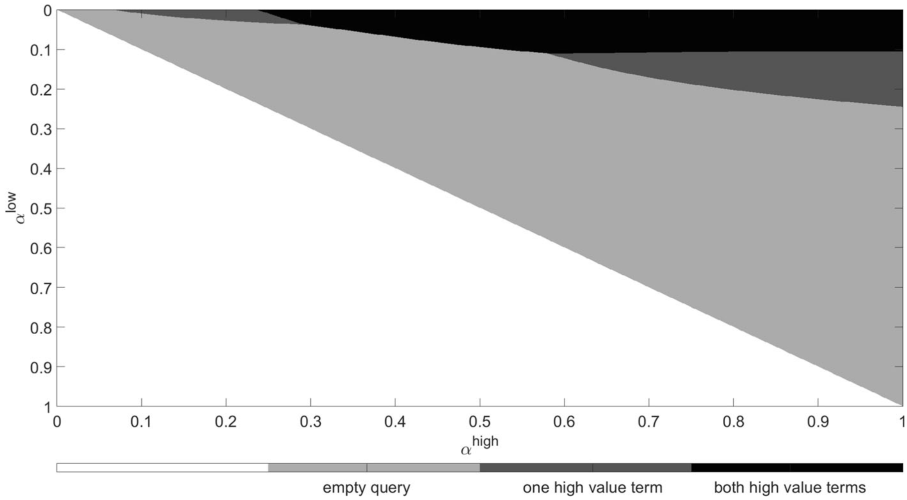


Fig. 10 Naive query type – one term is valued at $1

5. Queries with only the two high value terms ($|q| = 2$): $p_1 = p_2 = \alpha^{high}$, $p_3 = \alpha^{low}$
6. Queries with all three terms ($|q| = 3$): $p_1 = p_2 = p_3 = \alpha^{high}$

We compute the naive objective function, i.e., Eq. 12, for each query type, when both α^{low} and α^{high} vary between 0 and 1 under the constraint that $\alpha^{low} \leq \alpha^{high}$, for $L = 10$. Figure 10 displays the query type that maximizes the naive objective function, as a function of α^{high} and α^{low}. We see that the query types containing the low-value term (type 2, 4 and 6) never maximize the naive objective function. The other three query types may maximize the objective function, depending on the values of the parameters α^{high} and α^{low}.

C.2 All words valued at $2

The support of the webpage utility is {0,2,4,6}. We compute the probability distribution of the maximum utility across L pages, under the naive beliefs defined in Eq. 8. We also simplify notations by setting $\widetilde{Prob}(t_i \in l|q) = p_i$ for $i \in \{1, 2, 3\}$. We first compute the cumulative density function of the maximum utility, i.e, $\phi_j = Prob(\max_l\{U(l)\} \leq j)$ $j \in \{0, 2, 4, 6\}$, as follows:

$$\begin{aligned}
\phi_6 &= 1 \\
\phi_4 &= (1 - p_1 p_2 p_3)^L \\
\phi_2 &= \left[\sum_{i=1}^{3} \left(p_i \Pi_{j \neq i}(1 - p_j)\right) + \Pi_{i=1}^{3}(1 - p_i)\right]^L \\
\phi_0 &= (1 - p_1)^L (1 - p_2)^L (1 - p_3)^L
\end{aligned}$$

Given this, we can express the objective function as (recall that $c = 1$ in the experiment):

$$\begin{aligned}
\tilde{f}^{\text{naive}}(q) &= E[max_l\{U(l)\}|q] - |q| \\
&= 6(\phi_6 - \phi_4) + 4(\phi_4 - \phi_2) + 2(\phi_2 - \phi_0) - |q| \\
&= 6 - 2(\phi_4 + \phi_2 + \phi_0) - |q| \qquad (13)
\end{aligned}$$

In this case, there are four query types such that all queries from the same type achieve the same value of the objective function in Eq. 13:

1. Empty queries ($|q| = 0$): $p_1 = p_2 = p_3 = \alpha^{low}$
2. Queries with only one term ($|q| = 1$): we assume that $p_1 = \alpha^{high}$ and $p_2 = p_3 = \alpha^{low}$ without loss of generality
3. Queries with two terms ($|q| = 2$): we assume that $p_1 = p_2 = \alpha^{high}$ and $p_3 = \alpha^{low}$ without loss of generality
4. Queries with all three terms ($|q| = 3$): $p_1 = p_2 = p_3 = \alpha^{high}$

We compute the naive objective function, i.e., Eq. 13, for each type of query, when α^{low} and α^{high} vary between 0 and 1 under the constraint that $\alpha^{low} \leq \alpha^{high}$, for $L = 10$. Figure 11 shows that under naive beliefs, all four query types may maximize the objective function, depending on the values of the parameters α^{high} and α^{low}.

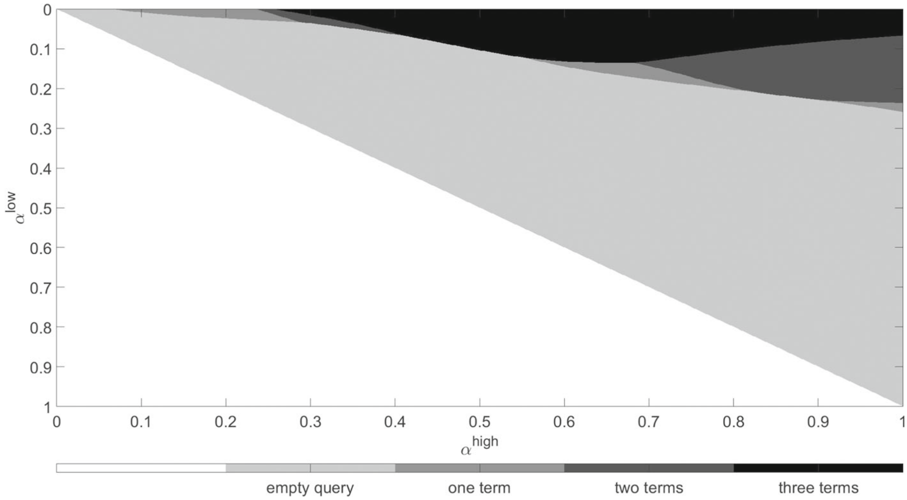


Fig. 11 Naive query type – all terms are valued at $2

Appendix D: Field data: all shorter queries and related words

Table 11 Shorter queries and related words (part 1)

Shorter query	Related words
Food	fast, chinese, network, dog, delivery, recipes, baby, restaurants, stamps, city, healthy, indian, mexican, poisoning
Food city	chinese, restaurants, delivery, ad, fast, mexican, indian, lion, network, stamps
Food delivery	chinese, restaurants, ad, fast, mexican, city, indian, lion, network, stamps
Food porn	fast, chinese, indian, network, recipes, delivery, japanese, healthy, restaurants
Boys food	porn, fast, bank, chinese, delivery, ideas, indian, network
Food depot	ad, chinese, city, fast, giant, lion, network, pet
Food com	fast, recipes, network, chinese, stamps, healthy, service
Food lion	ad, weekly, store, application, city, stores, giant
Food network	recipes, star, channel, magazine, restaurants, shows, tv
Halal food	restaurants, chinese, delivery, fast, places, indian, korean
Key food	fast, network, chain, chinese, bank, stores, web
Organic food	baby, health, stores, store, delivery, dog, healthy
Blue buffalo dog food	menu, recipes, ideas, shower, cake, baby

Table 11 (continued)

Shorter query	Related words
Fast food	restaurants, menu, places, healthy, delivery, nation
Food near me	chinese, delivery, restaurants, places, fast, mexican
French food	recipes, network, fast, italian, menu, restaurants
Indian food	recipes, restaurants, delivery, menu, baby, chinese
Jamaican food	recipes, restaurants, menu, chinese, delivery, indian
Japanese food	chinese, menu, restaurants, recipes, delivery, korean
Mr food	chinese, recipes, delivery, fast, greek, tv
Baby food recipes	homemade, indian, healthy, ideas, shower
Chinese food near me	delivery, restaurants, menu, places, express
Dog food	pet, blue, buffalo, homemade, iams
Food emporium	health, delivery, pet, key, network
Food network shows	tv, recipes, channel, star, cake
Food tv	network, shows, channel, recipes, fast
Frys food	store, ad, city, weekly, stores
Giant food	stores, store, ad, lion, weekly
Greek food	recipes, restaurants, menu, delivery, mediterranean
Homemade dog food	recipes, healthy, pet, blue, organic
Mediterranean food	restaurants, recipes, greek, menu, indian
Mexican food	restaurants, menu, recipes, delivery, chinese
Cuban food	restaurants, recipes, mexican, network
Food delivery near me	chinese, restaurants, places, fast
Healthy fast food	restaurants, healthiest, places, recipes

Table 12 Shorter queries and sample related words (part 2)

Shorter Query	Related words	Shorter query	Related words
Chinese food	delivery, menu, restaurants, express	food for less	fast, ad, weekly
Food for thought	recipes, network, restaurants	food 4 less	ad, store
Delivery food	chinese, restaurants, service, fast	food matters	health, recipes, tv
Ethiopian food	recipes, restaurants, indian, menu	hummingbird food	homemade, cake, recipes
Food bank	fast, network, stamps, chinese	iams dog food	pet, blue, buffalo
Food blogs	healthy, recipes, indian, health	food pyramid	healthy, chain, web
Food channel	network, recipes, tv, shows	fromm dog food	pet, blue, buffalo

Table 12 (continued)

Shorter Query	Related words	Shorter query	Related words
German food	dog, recipes, french	chinese food delivery	menu, restaurants
Food network star	recipes, tv, shows, channel	giant food store	stores, ad, lion
Food places near me	fast, restaurants, delivery, chinese	health food stores	store, organic, healthy
Food stamps	apply, florida, stamp, application	healthiest fast food	restaurants, healthy, menu
Giant food stores	store, ad, lion, weekly	healthy food	recipes, fast, health
Gordon food service	store, city, ad, depot	angel food cake	recipes
Food processor	cuisinart, kitchenaid, recipes	blue dog food	buffalo, diamond, pet
Italian food	restaurants, recipes, menu, network	junk food	healthy, fast, health
Korean food	chinese, menu, recipes, japanese	food web	chain, pyramid
Best dog food	blue, pet, buffalo	food inc	city, fast
Genetically modified food	organic	food chain	web, fast, pyramid
Chinese food menu	delivery, restaurants, express	fast food nation	restaurants
Diamond dog food	blue, pet, buffalo	food dehydrator	recipes
Fast food restaurants	places, healthy, delivery	food lion ad	weekly
Filipino food	recipes, menu, restaurants	food network com	recipes
Florida food stamps	apply, stamp, application	food poisoning	symptoms
Food and wine	network, city, tv	food trucks	sale
Food bazaar	restaurants, fast, nation	food trucks for sale	restaurants
Food network recipes	cake, healthy	halloween food ideas	recipes
Diatomaceous earth food grade	depot	food stamp application	stamps, florida
How long does food poisoning last	symptoms	kitchenaid food processor	cuisinart, recipes
Cuisinart food processor	kitchenaid, recipes	food network magazine	recipes, wine, tv
Baby shower food ideas	menu, recipes, cake	apply for food stamps	application, florida, stamp

References

Azzopardi, L., Kelly, D., Brennan, K. (2013). How query cost affects search behavior. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 23–32).

Borrell (2016). Forecast says seo-related spending will be worth $80 billion by 2020. Available online from http://searchengineland.com/forecast-says-seo-related-spending-will-worth-80-billion-2020-247712.

Dzyabura, D., & Hauser, J.R. (2018). The role of preference discovery in product search. *Marketing Science*. forthcoming.

Erdem, T., Keane, M.P., Öncü, T.S., Strebel, J. (2005). Learning about computers: an analysis of information search and technology choice. *Quantitative Marketing and Economics*, *3*(3), 207–247.

Fu, W.-T., & Pirolli, P. (2007). Snif-act: a cognitive model of user navigation on the world wide web. *Human–Computer Interaction*, *22*(4), 355–412.

Gabaix, X., Laibson, D., Moloche, G., Weinberg, S. (2006). Costly information acquisition: experimental analysis of a boundedly rational model. *American Economic Review*, *96*(4), 1043–1068.

Honka, E., & Chintagunta, P. (2016). Simultaneous or sequential? Search strategies in the us auto insurance industry. *Marketing Science*, *36*(1), 21–42.

Hui, S.K., Bradlow, E.T., Fader, P.S. (2009). Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior. *Journal of Consumer Research*, *36*(3), 478–493.

InternetLiveStats (2016). Google search statistics. Available online from http://www.internetlivestats.com/google-search-statistics/.

Jansen, B.J., Booth, D., Smith, B. (2009). Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, *45*(6), 643–663.

Jansen, B.J., Spink, A., Pfaff, A., Goodrum, A. (2000a). Web query structure: implications for ir system design. In *Proceedings of the 4th World multiconference on systemics, cybernetics and informatics (SCI 2000)* (pp. 169–176).

Jansen, B.J., Spink, A., Saracevic, T. (2000b). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*, *36*(2), 207–227.

Jeziorski, P., & Segal, I. (2015). What makes them click: empirical analysis of consumer demand for search advertising. *American Economic Journal: Microeconomics*, *7*(3), 24–53.

Kamvar, M., & Baluja, S. (2006). A large scale study of wireless search behavior: Google mobile search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 701–709): ACM.

Kim, J.B., Albuquerque, P., Bronnenberg, B.J. (2010). Online demand under limited consumer search. *Marketing Science*, *29*(6), 1001–1023.

Li, H., & Xu, J. (2013). Semantic matching in search. *Foundation and Trends in Informational Retrieval*, *7*(5), 343–469.

Liu, J., & Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, *37*, 6.

Manning, C.D., Raghavan, P., Schütze, H. (2008). *Introduction to information retrieval* Vol. 1. Cambridge: Cambridge University Press.

Nair, H. (2007). Intertemporal price discrimination with forward-looking consumers: application to the us market for console video-games. *Quantitative Marketing and Economics*, *5*(3), 239–292.

Narayanan, S., & Kalyanam, K. (2015). Position effects in search advertising and their moderators: a regression discontinuity approach. *Marketing Science*, *34*(3), 388–407.

Park, J., & Chung, H. (2009). Consumers travel website transferring behaviour: analysis using clickstream data-time, frequency, and spending. *The Service Industries Journal*, *29*(10), 1451–1463.

Pirolli, P.L. (2007). *Information foraging theory: adaptive interaction with information*. Oxford University Press.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 213–220): ACM.

Salton, G., & McGill, M.J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.

Santos, R.L., Macdonald, C., Ounis, I. (2015). Search result diversification. *Foundations and Trends in Information Retrieval*, *9*(1), 1–90.

Seiler, S. (2013). The impact of search costs on consumer behavior: a dynamic approach. *Quantitative Marketing and Economics*, *11*(2), 155–203.

Shi, S.W., & Trusov, M. (2013). The path to click: are you on it? Working paper.

Spink, A., Wolfram, D., Jansen, M.B., Saracevic, T. (2001). Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 226–234.

Stigler, G.J. (1961). The economics of information. *Journal of Political Economy*, *69*(3), 213–225.

Weitzman, M.L. (1979). Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, 641–654.

Wiltshire, C. (2015). Are you still using google analytics and excel to manage your seo strategy? Available online from https://www.gshiftlabs.com/seo-blog/are-you-still-using-google-analytics-and-excel-to-manage-your-seo-strategy/.

Wu, W.-C., Kelly, D., Sud, A. (2014). Using information scent and need for cognition to understand online search behavior. In *Proceedings of the 37th International ACM SIGIR conference on Research & Development in Information Retrieval* (pp. 557–566): ACM.

Yang, L., Toubia, O., De Jong, M.G. (2015). A bounded rationality model of information search and choice in preference measurement. *Journal of Marketing Research*, *52*(2), 166–183.