Where Can I Buy a Boulder? Searching for Offline Retail Locations

Sandro Bauer
University of Cambridge
Cambridge, UK
sandro.bauer@cl.cam.ac.uk

Filip Radlinski
Microsoft
Cambridge, UK
filiprad@microsoft.com

Ryen W. White
Microsoft
Redmond, WA, USA
ryenw@microsoft.com

ABSTRACT

People commonly need to purchase things in person, from large garden supplies to home decor. Although modern search systems are very effective at finding online products, little research attention has been paid to helping users find places that sell a specific product offline. For instance, users searching for *an apron* are not typically directed to a nearby kitchen store by a standard search engine.

In this paper, we investigate "*where can I buy*"-style queries related to in-person purchases of products and services. Answering these queries is challenging since little is known about the range of products sold in many stores, especially those which are smaller in size. To better understand this class of queries, we first present an in-depth analysis of typical offline purchase needs as observed by a major search engine, producing an ontology of such needs. We then propose ranking features for this new problem, and learn a ranking function that returns stores most likely to sell a queried item or service, even if there is very little information available online about some of the stores. Our final contribution is a new evaluation framework that combines distance with store relevance in measuring the effectiveness of such a search system. We evaluate our method using this approach and show that it outperforms a modern web search engine.

Keywords: Offline purchases; Location search; Map search

1. INTRODUCTION

Web search engines are often used as a starting point for product purchases, including the purchase of services [5, 34, 39]. This has inspired research in sponsored search [6, 13, 31, 37, 43], local search [27, 21, 22], and product search [25, 36, 38] directly. Product purchases can happen online (as in e-commerce) or offline (in person). The wide availability of recommendation systems is a key advantage of online purchases over traditional retail shopping. With price comparison web sites being widely used, it is straightforward to find out which store sells a particular item at the best price subject to constraints such as shipping or payment options. In traditional (offline) retail scenarios, on the other hand, customers often have to resort to merely satisfactory solutions based on criteria such as distance, prior experience with or loyalty to a particular store, a broad categorization of places (such as "bakery", "bank"), and a number of common-sense assumptions about the likelihood of an item being in stock [3, 10]. For example, most customers looking for a screw would pick the nearest large hardware store just because they can be relatively sure to find the right type there. The possible cost of driving to a closer but smaller store, only to discover that the item is unavailable, is commonly seen as prohibitive and leads customers to fail-safe solutions such as large shopping malls [10]. This, however, may result in a sub-optimal choice for a variety of reasons. For instance, there may be a specialist store much closer to the user's current location able to provide better advice and customer support than the well-known chain at the other end of town; or the range of products of a chain may not be as up-to-date as that in a store for enthusiasts. Finally, when searching for products where common intuition fails, many product search tasks end up time-consuming and frustrating.

We start from the observation that, while recommendation systems are becoming commonplace in online shopping (see [32] for an overview), they have not been commonly used to inform decisions about in-person purchases in the same way. There exist a number of commercial business directories (such as *yellowpages.com*), but these are often poor at answering product-specific queries. For instance, a user

Table 1: Example results for *where can i buy golf clubs in cleveland*. We consider both ranking by estimated relevance (top), and by estimated relevance per unit distance (bottom).

Sorted by relevance score:

Rank	Distance	Rel.	Store name
1	5.9 miles	N	Front Line Golf Cars
2	9.6 miles	Y	Canterbury Gold Pro Shop
3	12.7 miles	Y	Nieto Custom Golf Clubs
4	13.5 miles	N	M&M Golf Car and Mower Inc
5	21.5 miles	N	Golf Systems

Sorted by relevance score and distance:

Rank	Distance	Rel.	Store name
1	0.9 miles	Y	Caddy Shack Golf Pro Shop
2	1.1 miles	N	Ambassador Bowling Lanes
3	1.9 miles	N	Twin Lanes
4	0.8 miles	N	Rainbow Apparel Company
5	5.9 miles	N	Front Line Golf Cars

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2882998.

looking to buy wasabi in Phoenix, is presented a list of faraway Japanese restaurants that have wasabi in their name instead of the East Asian specialist supermarket that might be just round the corner. Our search system is designed to answer such product-specific queries; a user need not know beforehand in which type of store an item is likely to be sold.

In this work, we study how retrieval methods which have been successfully applied to e-commerce can be adapted to the traditional retail market, and build ranked lists of places for queries like "where can i buy golf clubs in cleveland" (see Table 1). The obvious practical problem with this is that many small and average-sized stores do not have professionally managed websites that comprehensively detail all of their products. Some websites within each category (such as "Korean restaurants", "bicycle dealers" or "hardware stores") are likely to, however, and therefore it should be possible to learn a good estimate of the kinds of products sold at a particular point of sale. One way we do this is by modeling what products stores in each category are likely to stock (even if a specific store does not have a website). This knowledge is used as the basis for a search system for places.

We note that offline purchases are different from e-commerce in many ways. For instance, risk management tends to play a role in offline shopping, where customers may rely on expert advice available at a specialist store. As an example, a customer may be hesitant to buy a fitted suit at a discounter (where no advice is available), while no expert advice is required for a standard food item purchase. This is different from an online setting, where different sizes and categories of stores are much more easily comparable, and hence users do not in general prefer a particular kind of store [20].

We make the following contributions with our research:

- Perform an analysis of the types of products and services commonly sought online yet purchased in person. This results in a general-purpose product ontology.
- Introduce an evaluation framework, and a novel distance-sensitive evaluation metric suitable for this task.
- Learn a model to rank stores for offline retail queries, including places without an online presence, and show that it substantially outperforms a modern baseline.

The paper is organized as follows: After reviewing related work, we characterize the information needs we are targeting in this paper, and develop an ontology of categories that these can be classified into. We next describe how we mine information about items typically sold by a given class of stores, as well as the ranking approach that we take. Following a presentation of suitable evaluation metrics, we discuss our evaluation experiments and their implications, and outline avenues for future work.

2. RELATED WORK

Understanding users' search intent has long been one of the core challenges in information retrieval. Seminal query analysis work by Broder [5] and Rose and Levinson [33] has provided a categorization of Web queries which is still valid today. At the top level, this consists of navigational queries, where the user already knows the web page they want to visit, but not the URL; informational queries, where a user wants to find information by reading one or more websites; transactional or resource queries, where a user wants to obtain some resource from the web (i.e. a file, a book) or interact with a web service (i.e. to book a flight ticket).

Finding Places

Rose and Levinson further subdivided informational and resource queries, providing statistics on the frequency of each of the classes in a set of AltaVista queries. We note that the sub-class of queries investigated in this paper was already found to be highly prominent in those early works. Queries where the search goal is to "find out whether/where some real world service or product can be obtained" were called "Locate" queries. In their analysis, about 25% of all queries were of this type. Although the quantitative query categories may now be somewhat outdated, "Locate" queries continue to be important. Other seminal analysis of query logs by Spink et al. [34] outlined the rise in interest in product searching (e-commerce in their study), which has continued to occupy a significant fraction of query volume [39].

More recently, a number of authors have investigated queries issued from mobile devices. Hinze et al. [12] categorized such mobile information needs via a user study, finding that between 15% and 20% of these queries start with "where". They also conducted further analysis of the direction-related queries that they identified; 64% of those referred to places already known to the user, while 36% simply asked for information about nearby venues. A similar diary study by Church and Smyth [8] led to comparable numbers: 31.1% of all information needs were geographical in nature, and the vast majority of them were submitted when the user was not at home.

Taken together, this shows that queries for places are common, and often fit clear templates. In this work we study similar types of information needs. However, the subset of queries that we address does not correspond exactly to mobile information needs. While we expect that a large number of queries related to offline purchases are indeed sent from mobile devices (particularly those describing food and similar short-term needs), the remainder, which might include queries for expensive products or products for an expert audience, will often be issued from a desktop computer. Our work targets users planning an in-person purchase in a nearby store, no matter whether the planning occurs at home, in transit, or elsewhere.

Local Search

More broadly, it has also been observed that certain information needs are more "local", i.e. dependent on the user's current location, than others. A number of authors have attempted to infer how location-specific a given query is. For instance, Jones et al. [16] evaluated geographic features from queries and the documents retrieved; other works harvest geographic hot-spots from background information on the Web or through crowdsourcing [29, 28]. Lagun et al. [21] conducted a user study eliciting explicit relevance feedback to understand whether offering location-sensitive search results might improve the user's search experience; they report that users frequently make use of this possibility. Wu et al. [42] investigated differences in the information needs of locals and non-locals, reflecting their different motivations for being in the city. White and Buscher [40] found that locals tend to search for information related to everyday activities, while visitors are more interested in tourist activities. Finally, Tang et al. [35] concluded that depending on the nature of an expressed local information need, further entities might be of relevance to the user: Buying fish or meat is often associated with buying spices and fur-

ther ingredients, and hence we might want a search engine to recommend a nearby specialist in addition to the actual fishmonger's.

Interpreting Queries

There has also been considerable work on the analysis of search query logs more broadly; Jiang et al. [15] recently presented an overview. For example, Lee et al. [23] built on the work of [5] by automatically identifying user goals behind queries. Dai et al. [9] detected "on-line commercial intention" in web queries, i.e. how likely a user is to make a decision to buy a product soon. Guo and Agichtein [11] classify commercial intents into "research" (where the goal of a user is to find out more about a product) and "purchase" (where the user wants to buy the product). Ashkan et al. [2, 1] used clickthrough logs to classify queries into commercial or non-commercial, as well as into informational and navigational. Our work addresses a particular sub-class of these queries which are related to offline purchases, and most importantly, our focus is on answering these queries rather than detecting them. Weber and Jaimes [39] analyzed a large query log according to several dimensions such as the users' demographics, query topics and search behavior. They provide an estimate of the frequency of shopping-related queries, but their analysis does not investigate the distribution across product categories that we present in this article, nor do they focus on the important challenge of identifying locations from which the products could be purchased offline. Given our focus on products, prior work in areas such as sponsored search [6, 13, 31, 37], local recommendations [27, 21, 22], and product search (primarily e-commerce related) (e.g., [25, 36, 38]) is also relevant to our research.

Classes of Products

There is a clear case for investigating offline shopping decisions, since there are many classes of items that most users rarely order online (even where this is possible). The literature typically subdivides retail goods into search, experience and credence products [20]. Typical *search* products (such as books) can be evaluated without prior inspection; we also refer to them as *standardised* products. *Experience* products, on the other hand, are analyzed in person by the customer before purchasing. This may be because important features of a product can only be analyzed in person, or because the cost of getting personal product experience is perceived to be lower than that of searching for information [19]. Further advantages of offline shopping from the customers' perspective include the ability to select an individual instance of a product, no-hassle exchange, and immediate availability [24]. Finally, credence products are those where even personal experience is not enough to judge the quality of a product (e.g., home maintenance services). A number of user studies [20, 26, 17, 24] confirm the intuitive assumption that users are more inclined to buy search products online than experience products. With experience products, customers are willing to accept a decrease in convenience (driving time, longer check-out lines etc.) for the sake of finding the right product and minimizing the risk of having to request a refund [7].

Importance of Distance

As with queries issued from mobile devices, the location of the user plays an important role in answering product search queries: A user is relatively unlikely to be willing to travel to a store that is very far from their current whereabouts. Lee et al. [22] provide empirical evidence for this, although preference for the closer locations varies a lot between categories. They report, for example, that for movie theaters users are much more likely on average to choose the closest location than for restaurants. Lv et al. [27] provide a similar study which suggests that distance (and the difficulties that a greater distance entails) must not be ignored by a search engine servicing mobile information needs.

Importantly, we also note that researchers involved in studying traditional retail environments have developed empirical models of a customer's decision process before choosing a store. For instance, Bell et al. [3] describe the total cost of an offline retail act as consisting of a fixed cost (independent of the items purchased) and a variable cost (the cost of the items purchased). The fixed cost is determined by the travel cost from the user's current location to a store, as well as a household's loyalty towards a store. Grewal et al. [10] extended this by highlighting the role of uncertainty; if a retailer can ensure that items are in stock most of the time, a customer will be more willing to travel to this place than if that is not the case. We draw on these results in our work, balancing the two criteria: availability and distance. The inclusion of further criteria, such as prices, place recommendations or previous purchases, is left to future work.

To summarize, previous work has largely focused on online (e-commerce) purchases, while offline purchasing needs are a lot less well understood. In this work we attempt to address this shortcoming. Second, previous work has not even implicitly modeled the products that are stored in particular types of stores. In this paper, we show how to map from a product query to a set of places, taking into account the category of the store (grocery store, Chinese restaurant, dental surgery) as well. Finally, distance to such physical locations has been shown to be important when evaluating user needs, and our work shows how to combine this with relevance information for ranking places.

3. OFFLINE PURCHASE NEEDS

In this section, we study what types of products and services users typically search for online with the intention to purchase in person.

3.1 Methodology

To develop an understanding of typical information needs related to offline purchases, we studied a sample of query logs from the Bing commercial search engine from July 2015. It is hard to reliably identify offline purchase intent in arbitrary queries: We need to identify "purchase" rather than "research" queries (according to Guo and Agichtein's classification), and secondly, the user must intend to purchase offline rather than online. While identifying such (offline) purchase intent is a research question in its own right, for this work we adopted a precision-oriented approach: We exploit the fact that a fraction of users tend to submit full-sentence queries such as "where can i buy flowers in san jose" to search engines; it is those queries that we analyze here.

We started by manually compiling a set of query patterns likely indicating purchase intent such as *where can I <verb>* and *in what store can I <verb>*. The <verb> slots were then filled with all synonyms of the verb *buy* according to

the Collins English Thesaurus; these include *buy*, *purchase*, *get*, *score*, *secure*, *pay for*, *obtain*, *acquire*, *invest in*, *shop for* and *procure*. We further added a number of patterns containing the verb *sell*. This resulted in a total of 176 query templates. For each, we obtained matching queries from a uniform random sample of all recent search queries submitted to the search engine in the United States. Templates that did not have any matching queries were excluded. This resulted in 53 purchase query templates. We manually inspected the matching queries and found almost of all of them to represent purchase intents.

We then filtered the queries matching the templates using an existing proprietary query classifier, such that only queries classified as indicating a geographically constrained intent were considered. This classifier selects for queries that explicitly include the name of a known geographic entity (usually a city name or state). Our assumption is that such a filtering indicates that the user intends to purchase the product in person, since for online purchases the exact location of the store is arguably irrelevant and would therefore be omitted. Further, the existence of a specific named place means that the intended location of the purchase is known, which is critical for evaluation. While future work may provide a more accurate way of identifying offline purchase intent, we are confident that the process applied here yielded a representative sample of common offline purchase intents.

3.2 Special Properties

Queries related to offline purchase needs are different from those related to online purchases. An initial inspection of the queries collected suggests that the following aspects are of particular interest for ranking offline retail locations:

Query specificity. Offline retail queries range from queries for broad classes of items (such as "food") to an exact description of the form of a screw. It is therefore important to detect how specific a query is. For instance, a user searching for an exact model name (such as "where to buy laptop lenovo x1 carbon in new york"[1]) may not be keen to buy a different model since they will likely have done all the research already. Availability of the chosen item in store is paramount for such users. For queries such as "where can i buy a refrigerator in philadelphia", other aspects (such as price, quality, distance and the range of products offered) can be expected to be more important. An ideal ranking algorithm should take into consideration such differences.

Different types of stores. Stores are different in size and nature. A smaller store might be providing more up-to-date products and better service, while malls often offer lower prices. Queries may express a preference for one or the other type of store, which should ideally be taken into account when recommending places. For instance, a user looking to go to a Turkish restaurant on a date will not be satisfied with a kebab stand only because Turkish food is sold there.

Urgency. Some requests are for short-lived needs, while for others opening hours and similar constraints are less important. A search engine answering offline retail queries should ideally analyze whether a user needs a product immediately, and, if so, exclude closed stores. Sometimes users express their immediate need explicitly by adding cues such as "now" or "tonight", but most of the time this will not be obvious.

[1]Note: For user privacy reasons, all complete queries shown in this paper were manually modified, although are completely representative of actual queries.

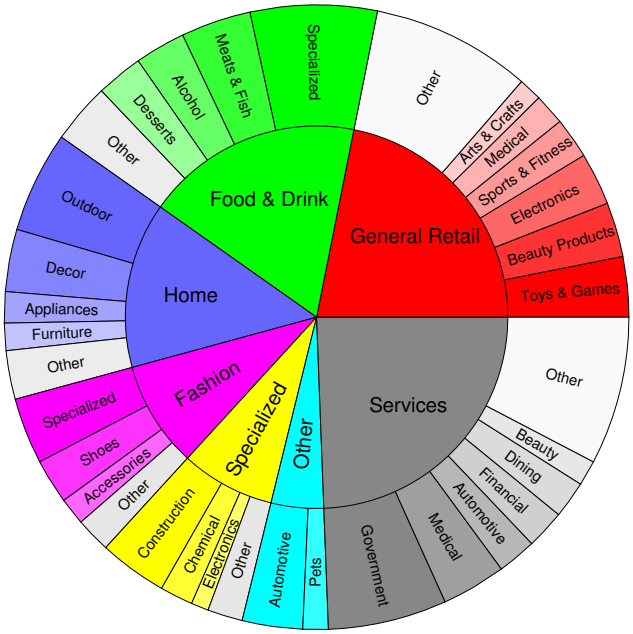

Figure 1: Ontology of offline purchase needs

Travel cost. Unlike visiting websites to find a good online store, traveling to a place comes at a cost. The willingness of users to accept longer travel times likely depends on the item queried; an algorithm should take this into account.

In this paper, we focus on the likely availability of an item in store and the travel cost incurred; other factors (such as different types of stores and urgency) are left for future work.

3.3 Offline Purchase Ontology

In our query sample, we manually labeled a random subsample of 1,000 queries from the data described earlier into an ontology of product purchase intents. Of these 1,000 queries, under 1% of queries matching our templates were not purchase queries. We exclude those in what follows.

We iteratively developed a two-level ontology of offline purchase intents. The aim was to capture most products in a balanced ontology that groups offline purchase intents by similarity. Our final ontology has 8 top-level categories with 54 subcategories. The ontology is presented in Figure 1, where the size of each slice corresponds to the number of queries in the corresponding (sub-)category. The complete ontology appears in Table 8.

Note that this ontology is different from existing directories such as *YP.com* and *superpages.com*: Our ontology is about categories of products, not stores. Second, it provides a quantitative overview of the important categories of products searched for by real users, which is useful from a research perspective as well as for search engine providers.

Overall, we found that three quarters of offline purchase queries are for products, and one quarter are for services. In the 24% of the queries that are for services, the most

common services are administrative/government needs such as where to get a particular type of license, a birth certificate, etc. (6% overall). The next biggest class is medical, comprising vaccines, tests, treatments and so on.

We found that 19% of queries are for general retail items, defined as products that can usually be purchased in many large and general purpose stores. These include toys and games (e.g. *where to buy balloons near temecula, ca*), beauty products (*where to buy molton brown hand wash in seattle*), and so forth. A further 18% are for food and drink, e.g., *where to buy lobsters in tacoma wa*, *where to buy free range eggs in troy mi* and *where to get beer keg in juno ak*.

Beyond general retail and food, the next most common categories are Home (14%) and Fashion (9%). Home is dominated by the outdoors (from stone boulders to tulips), while fashion requests tend to ask for specific brands or types of fashion (male winter jackets or platform shoes). We were surprised that fashion queries are not more common. Our hypothesis is that people already know the fashion retailers nearby, or search for stores from the brand home pages.

Finally, specialized queries account for 18% of volume. The 8% Specialized top-level category includes requests for items not usually found at a mall such as construction materials, chemical supplies, or specialized electronics. Queries in specialized sub-categories of the Fashion (mostly specific brands) and Food & Drink (less common foods, often ethnic or satisfying specific dietary constraints) categories constitute the rest.

3.4 Importance of Travel Distance

To better understand the importance of distance, it is important to know how far users would typically travel to make offline purchases. For instance, for meat or fish we expect relevance to degrade more quickly as distance increases than for pianos or furniture.

Given a product request, we asked annotators to provide an estimate of the maximum allowable travel distance in miles (between 0 and 50). For each request, judgments were obtained from 10 participants using a popular crowdsourcing platform; no expert knowledge was assumed on their part. Crowdworkers were paid 20 US cents for each block of 10 judgments. We then used Bayesian Classifier Combination [18] to combine multiple ratings, as it was shown to allow noisy crowd judgments to be reliably combined. Finally, we computed distance statistics for each subcategory of needs. The average distances for the top five and bottom five subcategories (in terms of distance traveled) for which we have at least 10 queries are shown in Table 2. Due to the nature of the subject matter, we did not collect estimates for products in "Services : Adult" and "General Retail : Illegal" (which combine to form 1% of our query set).

There are significant differences between the subcategories (analysis of variance (ANOVA): $F(51, 9628) = 209.95$, $p < 0.001$) and qualitatively there are clear differences in nature between the products at these two extremes. Products for which people would expect to travel short distances are typically essentials (e.g., food, clothing), whereas those that are further afield are largely non-essential (toys and games, entertainment). Interestingly, the Automotive subcategory appears in both groups. Expected travel distance was low for services such as oil changes and emission tests, but much higher for automobile purchases and automobile parts, including tires. For garments, people expected to travel shorter distances for basic clothing needs, and further for specialized items (e.g., sports team apparel, workwear).

Table 2: Top 5 and bottom 5 subcategories by expected distance. Average and standard deviation distances (in miles) are reported. N is the number of judgments over all queries in each category.

Subcategory	Avg	StDev	N
Food & Drink : Groceries	7.54	10.90	110
Fashion : Garments	8.83	10.15	120
Services : Restaurants & Cafes	9.19	11.42	190
Services : Automotive	10.72	11.79	180
Home : Other	11.13	14.65	120
...	...	...	...
General Retail : Medical	15.12	16.49	110
Other : Automotive	15.33	16.35	310
General Retail : Toys & Games	15.74	16.44	300
Fashion : Specialized	15.89	14.78	340
Other : Pets	16.19	17.04	120

Table 3: Expected distances traveled depending on the starting location. Average and standard deviation distances are reported. N is the total number of judgments over all queries in each location type.

Location type	Avg	StDev	N
Rural	24.65	17.40	1488
Suburban	11.65	12.82	5230
Urban	11.30	13.47	2962

Since the journey is affected by the nature of the starting location, we also asked judges to specify whether the point where they expected to start their journey to purchase the product was urban, suburban, or rural. Table 3 reveals the average expected travel distance for each location type. While many more factors may have an impact on how far a user is willing to travel, we leave a more comprehensive modelling of such user preferences to future work.

An ANOVA revealed significant differences between the urban, suburban and rural starting locations ($F(2, 9677) = 568.68$, $p < 0.001$). Post-hoc analysis using Tukey tests revealed significant differences between rural and the other location types (both $p < 0.001$), but not between suburban and urban ($p = 0.508$). It seems reasonable that people would expect to travel further from rural locations. The distinction between suburban and urban may have been less clear to judges. Also, metropolitan areas typically have many retail options to serve the large populations in these areas; meaning that there may be just as many options in suburban areas as are found in urban environments. A two-way ANOVA (with *location type* and *subcategory* as factors) revealed no significant interaction between location type and the product subcategory ($F(102, 9524) = 1.157$, $p = 0.134$).

4. RANKING PURCHASE DESTINATIONS

The main goal of this work is to recommend places given a query such as *where can i buy a vw polo in chicago*. In this section we first give an overview of our approach, then describe the algorithmic details.

4.1 Overview

The main difficulty of our task is that the information required – whether a specific item is sold at a particular

retail location – is not generally available in the form of a document in the classic IR sense that can be mapped to the query directly. The majority of stores (especially smaller boutiques) do not have a website, and even if they do, rarely list all the items sold. We therefore must take a different approach and develop techniques to generalize across stores to obtain an estimate of what each store is likely to sell.

For this to work, we need to use suitable external knowledge transforming both the user query and a relevant place entry to allow them to be matched. One problem is that item queries are often single terms, and hence an expanded form is required to match items precisely. Users may also enter simply a brand name (*where in nyc can i buy burberry clothing*) or the name of a city directly after the name of the product (*where to buy prime beef portland or*). On the document side, we have to identify suitable data sources that can serve as place representations, even in the absence of a website. Here, we supplement the data of each place with plausible information harvested from places that are similar.

Using the resources found, we build a term vector for each place and each query, and compute a match. We do this in two phases: First, we build an expanded model of the user's need by only matching locations for which richer data is available. Then, in a second pass, we compute more sophisticated features that can be combined to score any place, whether or not it has a homepage.

In particular, place categories (such as "hardware store") are commonly available even for very small stores, which allows us to link them to similar stores for which more data is available. For this, we use a category taxonomy constructed by a major search engine. The hierarchy starts with generic classes (i.e. "restaurants"), which are divided into subclasses of specialist stores ("Bengali restaurants"). Since places are typically annotated with more than one category, our algorithm considers all categories available for a given place.

This section only deals with the problem of calculating the relevance of a place with respect to a query. For evaluation, we take into account distance of a place from the user's location as an additional criterion (cf. Section 5.3).

4.2 Terminology

We represent a place p (which can be a store, practice, etc.) as always having a name $name_p$, geographic location loc_p, and belonging to a set of categories cat_p.

Let P denote a set of all places. Further, let $P(c)$ be the set of all places in category c. On the other hand, let q be a query that is known to have been issued from some location loc_q. Although our sample queries are long form, for the remainder of this work we assume q to simply consist of the required product[2].

4.3 Content-based Place Ranking

To rank locations, we start by constructing a vector representation for the query $\vec{v}_q$. We follow a standard pseudo-relevance feedback query expansion procedure using results returned by Bing. The details are given in Algorithm 1.

Suppose that place p also has some online content (say, a homepage) associated with it. Each such place can be represented by several component place vectors $\vec{v}_{p_i}$, which are term vectors constructed using the data sources listed in Table 4 (homepage, homepage title, queries, link content and ODP category information). Specifically, for data source i, $\vec{v}_{p_i}$ is a normalized term vector consisting of all unigrams and bigrams from that data source.

Places for which we have such content can then be ranked with a simple model that combines the similarity of each component to the expanded query vector:

$$score(q, p|\vec{w})^{minimal} = \sum_i w_i \cdot \text{ cosine}(\vec{v}_q, \vec{v}_{p_i}) \qquad (1)$$

where $cosine()$ is a simple cosine similarity function:

$$cosine(\vec{v_p}, \vec{v_q}) = \frac{\vec{v}_p \cdot \vec{v_q}}{||\vec{v}_p|| \cdot ||\vec{v}_q||} \qquad (2)$$

Given a query q, let $initial(q)$ be the ranking obtained by sorting places according to the score computed using Equation 1 with a uniform weight vector $\vec{w}$.

Algorithm 1 Constructing the query vectors

1: $q \leftarrow$ noun-phrase($query$), the noun phrase named in the query, eg. "flip flops" in "where can i buy flip flops".
2: Submit q to a web search engine, setting location to loc_q.
3: $snippets(q) \leftarrow$ snippets for the top 50 search results.
4: Tokenise all snippet text in $snippets(q)$.
5: Query term vector $\vec{v}_q \leftarrow ngrams(snippets(q))$, where $ngrams$ counts all unigrams and/or bigrams.

Table 4: Content for ranking places.

Data source	Description
Place homepage	Text content of the place homepage, parsed using the *Html Agility Pack*[3].
Homepage title	The content of the <title> tag on the place homepage.
Anchor text	The text of all links pointing to the place's website (which is known to the search engine).
ODP categories	The place's homepage classified into a class in the ODP category hierarchy using [4]. The set of these classes is used as a feature.
In-queries	All queries issued to Bing during a 30-day period in August and September 2015 after which users immediately click on the place's homepage.

4.4 Ranking Places with Minimal Metadata

We note that while there are many databases of places (including *YP.com* and *superpages.com* mentioned earlier), smaller stores are usually only annotated with basic metadata (name, location and categories), lacking their own homepage. It is hence not possible to compute the features listed in Table 4. We now consider how such places can be ranked.

Let $P^+ \in P$ be those places for which we have content as described in Section 4.3. Similarly, let $P^+(c) \in P(c)$ be the places with content in a given category c.

The first feature we present to score places without content does so by estimating the content that could be associated with a given place based on other nearby places in

[2]Any standard NLP tool can be used to extract this from a long-form query. We used MSR SPLAT [30].
[3]http://htmlagilitypack.codeplex.com

Algorithm 2 Calculating top categories from a ranking $r(q)$

```
score(c) ← 0 {initialize score for all categories}
for all i ← 1 ... k do
    p_i ← the i^th place in r(q)
    for all category c in categories(p_i) do
        score(c) ← score(c) + 1.0/i
    end for
end for
TopCategories(q) ← Highest-scoring 10% of categories
    in r(q) (no fewer than 5 if there are too few)
```

the same category for which online content is available. In particular, given a query q:

1. We identify the most likely relevant categories for q.
2. We build a location-specific average vector for each of these categories using all nearby places with metadata.
3. We use these average category descriptors to rank places that do not have metadata.

Taking $initial(q)$, the ranking of places with content, we pass this ranking to Algorithm 2 to produce $TopCategories(q)$, the set of categories of places that are most likely to sell the product or service required. For every category in this set, we then compute the mean vector per data source:

$$\vec{v}_{\overline{p}i}(c) = \frac{1}{|P^+(c,q)|} \sum_{p \in P^+(c,q)} \vec{v}_{p_i}, \tag{3}$$

where $P^+(c,q)$ are all places in $P^+(c)$ that are within a threshold distance of loc_q (we use $\Delta = 50$ miles).

Given a query q, each place p can then be represented by category-based content features:

$$\vec{v}_{p_i^c} = \sum_{c \in cat_p \bigcap TopCategories(q)} \vec{v}_{\overline{p}i}(c) \tag{4}$$

Now we are able to construct a more powerful scoring function for any given place, taking into account both raw content features and category-based features:

$$\begin{aligned} score(q,p|\vec{w^1},\vec{w^2})^{cross-cat} &= \sum_i w_i^1 \ \text{cosine}(\vec{v}_q, \vec{v}_{p_i}) \\ &+ \sum_i w_i^2 \ \text{cosine}(\vec{v}_q, \vec{v}_{p_i^c}) \end{aligned} \tag{5}$$

4.5 Additional Ranking Features

Beyond the basic content available for a place, and category-average content information, we propose a number of additional features that can be used to score how well a given place matches a query.

Category Overlap

A simple additional feature for scoring the match of a place to a query is how well the categories of the place overlap with the top categories of the query:

$$CatOverlap(p,q) = \frac{|TopCategories(q) \bigcap cat_p|}{|cat_p|} \tag{6}$$

Name Match

People are able to infer what sorts of products are sold by some stores based simply on the name: For instance, it is clear what stores named *Needle and Thread*, *Raw Box Organics* and *Tom's Garden Supplies* would likely sell. Thus, we also create a simple name-based ranking feature:

$$NameMatch(p,q) = cosine(name_p, NamesModel(initial(q))) \tag{7}$$

where $NamesModel(initial(q))$ is a term vector constructed from the concatenation of the names of all places in $initial(q)$.

Web Name Match

Finally, we propose to also use the expanded query as a name model. In contrast to *NameMatch*, this is a global model, based on general web content. We hypothesize that where the name of a store matches web content, it may provide information about the items and services the store offers.

$$WebNameMatch(p,q) = cosine(name_p, \vec{v}_q) \tag{8}$$

4.6 Learning to Rank

In producing the initial ranking, all component feature vectors are assigned the same weight. To obtain an improved ranker that correctly combines the features proposed, we use a standard learning-to-rank approach.

While any learning-to-rank algorithm would be suitable, we choose to use the LambdaMART boosted decision trees [41], as such models have recently been shown to be state-of-the-art for a number of learning-to-rank challenges. We provide all the features described in this section to LambaMART, along with the labels discussed in the next section, to produce our final ranking function. For evaluation, we use a 20-fold cross validation setup: In each round, 95 % of queries are used to train the learning-to-rank model, and the remaining 5 % are used for testing. Each query is in the test set exactly once.

5. EVALUATION

In evaluating the utility of a ranking of places that may sell a given product or service, literature in marketing (e.g., [10]) has noted at least three major aspects that must be considered: (1) How confident am I that the store sells the product and has it in stock? This is closest to the traditional concept of relevance in information retrieval, hence we also refer to it as *relevance*; (2) How difficult is it to reach the store? This measures the cost of getting to the store, for which we use straight line *distance* from the user to the store as a proxy, and; (3) How *expensive* are products in this store? While this is often a key attribute for satisfaction when searching for products to be purchased online, we leave it as future work for the purposes of offline purchase ranking.

5.1 Relevance Data

There is no established test collection for our task; hence we obtain suitable relevance judgments ourselves. We selected 200 queries uniformly at random from 10 of the most popular subcategories described in Section 3. We extracted the noun phrase (required item) and location from each. As the location was usually a geographic region such as a city, we sampled a random location within the boundary defined by this name (for example, "in brooklyn" became (latitude = 40.696°, longitude = -73.945°)). For each (query, location) pair, we then ran 12 configurations of the basic ranking algorithm (representing places with and without different features) and pooled the top five places returned in each configuration. We also ran each query against a baseline system

(described below) as part of the pooling. All systems considered, including the baseline system, operate on the full set of places contained in Bing Maps. This procedure resulted in a dataset of about 5,000 query/place pairs.

The obvious way of evaluating our system would have been to call all 5,000 stores and enquire whether the relevant items are sold there. Since this was prohibitively expensive for such a large number of query/place pairs, we opted for a simpler approach and instead asked human judges to assess, using world knowledge and any information available on the Web, how likely a store is to sell a particular item.

Each (query, place) pair was presented to 3 crowdsourced judges on the platform described earlier, tasked with determining whether a given item or service could be purchased at a given place. Again, crowdworkers were not expected to have any expert knowledge. Due to the increased complexity of the task, we paid a higher per-judgment compensation (4 US cents) than for the annotation task described in Section 3.4. We asked annotators to consider all available information (including the Web and their own experiences) to improve judgment quality. For each place, we provided a link to an information page describing what was known about the place by Bing, including the place name and address, as well as the main website (if any). We required the annotator to visit this website before providing a judgment.

A graded relevance scheme was used, also capturing uncertainty in whether a given place sells a particular item or service. For instance, a government office which is not responsible for the specific service required may still be able to direct the user to the correct office. Similarly, it can sometimes be said with certainty that a store sells some item, while in other cases this depends on details that are difficult to answer without contacting the store. The four alternatives we presented to judges are shown in Table 5. Note that both labels 2 and 3 refer to relevant stores, as with rare products it is often hard to say whether a store sells a particular model and has it in stock. We also had error labels for when store information was insufficient, or where judges could not estimate intent from the query.

Once more, we combined the individual labels for each place into a single label using a standard Bayesian Classifier Combination [18]. We verified that our dataset captures a variety of relevant and non-relevant locations for each user need, with 189 of the 200 queries having at least one location labeled 2 or 3, and all queries having a number of less relevant locations. The overall distribution of labels is 3,059 (query, place) pairs with score 0, 73 pairs with score 1, 486 pairs with score 2 and 1,111 pairs with score 3.

Confirming the importance of ranking locations with minimal metadata (i.e. no known website, and hence no click or anchor data), we note that of the 3.6 million locations in our dataset, 1.5 million (42%) do not have a known website. Once we apply the pooling to build our evaluation set, we find that 1,277 (24%) of the 5,392 judged locations do not have a website. While on average such locations were found to be somewhat less relevant to the sampled queries, we note that 18% of the locations labelled with 2 or 3 (i.e. relevant) do not have a website, and thus retrieving these has a substantial impact on result quality.

5.2 Evaluation Metrics

To the best of our knowledge, there are also no standard metrics that combine distance and relevance in measuring utility of a ranking system. Therefore, we perform our evaluation with two complementary approaches.

Table 5: Graded scores and descriptions for measuring location relevance (as shown to judges).

Score	Description
3	It is highly likely the store would offer the item / service.
2	It's a reasonable place to look, although whether they offer the product or service depends on attributes such as the store's size, the brands stocked etc. I might call the store to confirm they offer it.
1	The store is unlikely to offer this item/service but is sufficiently related that they could advise me on where to obtain the item/service.
0	The store is unrelated and it's highly unlikely that they would offer the item/service.

Approach 1: Maximum Travel Radius

If we cap the maximum distance a person is willing to travel to satisfy a product or service need, we can simply measure the relevance of the top items within this distance using a standard IR metric. We take a radius of $\Delta = 50$ miles, and use DCG@k [14] to measure the quality of a ranked list for $k \in \{1, 3, 5\}$, hypothesizing that users are unlikely to want to visit more than a handful of places for a given product.

Approach 2: Success within a Range

Our second approach assumes that each user may have a different maximum distance they are willing to travel to find a specific item. After they have traveled this distance, they are either successful or not. We therefore start by assuming that given a ranking of places, the user would travel to the places in order, and stop once they have found the item needed or once they have traveled their maximum distance. For simplicity, the total distance traveled is taken as twice the sum of the distances from the user to each place in order (to the place and back). For a given threshold distance Δ we report the fraction of queries where the user was successful (i.e. visited a place with label 2 or 3), as well as the actual average distance traveled before being successful (in those cases where the user is successful).

Naturally, both approaches have benefits and drawbacks. DCG@k does not account for travel distance below the threshold, while the success model does not take into account that users could also choose to make a phone call before traveling, or be more efficient by choosing to visit sets of nearby places even if they are not ranked in that order. Although a number of variants were considered, due to space constraints we chose simpler metrics for our setting and leave the design of improved evaluation metrics as future work.

5.3 Experimental Results

We evaluate our place ranking algorithms using both approaches. The results are in Table 6. The left-hand side of the table shows results as measured using DCG@k, only requiring the places ranked to be within a radius $\Delta = 50$ miles around the query location. Here, the retrieval system simply ranks places by relevance score. The learned model produces rankings where on average the top location has a

Table 6: Ranking performance for location ranking algorithms. DCG results are shown on the left. The right-hand side of the table shows success analysis. Specifically, capping the total travel distance at 100 miles, % Success shows the fraction of queries for which the user would be successful, while E[dist] shows the average total distance they would have traveled to be successful. Values marked ▼ are statistically significantly worse than the Learned ranking, while those marked ▲ are significantly better (t-test; $p < 0.05$).

	DCG@k (50 mile cap)			No Distance Discount		Inverse Discount	
Method	k=1	k=3	k=5	% Success	E[Dist]	% Success	E[Dist]
Map Search (baseline)	1.54▼	2.28▼	2.62▼	11.5 %▼	9.9 miles▲	11.5 %▼	9.9 miles▲
Original Ranking	2.78▼	5.65▼	7.51▼	68.5 %▼	16.5 miles▼	67.0 %▼	7.2 miles▲
Learned	4.39	8.50	11.08	84.5 %	14.8 miles	79.0 %	13.0 miles
Learned model *excluding* the following source content / features:							
Homepage content	3.86▼	7.68▼	10.08▼	81.5 %	15.2 miles	77.5 %	14.2 miles
In queries	4.52	8.48	11.03	82.5 %	14.6 miles	79.0 %	12.4 miles
Anchor text	4.15	8.21	10.68	83.5 %	14.2 miles	78.5 %	12.7 miles
ODP categories	4.43	8.34	10.87	80.5 %	14.4 miles	76.5 %	13.2 miles
"Additional" features	4.09	7.81▼	10.11▼	79.5 %	13.5 miles	79.5 %	14.1 miles

rating of 2 or 3 (defined as in Table 5). This substantially and significantly outperforms the baseline, which consists of searching for the noun phrase using the Bing Maps API (that also allows constraints on the location of the user and required target locations to be specified). We tried a number of commercial search providers for baseline estimates, and achieved comparable (low) success rates for all.

The right-hand side of the table compares ranking places by relevance (*No Distance Discount*), against reranking them by the relevance score divided by the distance to the place (*Inverse Discount*). The latter greedily maximizes expected relevance per unit distance[4]. We see that the learned model is more successful than the non-learned model, and that taking distance into account reduces expected travel distance at the cost of likelihood of success.

At the bottom of the table we also evaluate the importance of the different data sources for computing place features, finding that removing individual information sources does not significantly hurt the performance of the model. The exception is the homepage content, which clearly provides useful content. However, it is interesting to note that without the homepage, the additional features based on search engine usage still allow the model to perform better than a uniformly weighted combination of all features including the content (the "original ranking"). The "additional" features refer to those introduced in Section 4.5.

We further investigate the contrast between success and distance traveled. Figure 2 considers the case where we have different thresholds on how far the user may be willing to travel to satisfy their purchase need. We see that if the user is only willing to travel 10 miles (5 miles to a store, then 5 miles home), they would be successful in finding their item less than 30% of the time. Users willing to travel further would be much more successful.

Noting in passing that all our approaches clearly outperform the baseline commercial system, we focus on (1) the effect of distance discounting, and (2) the effect of ranking with a learned system versus a simpler ranking model. Considering the solid lines (for the learned ranking), we find that when the distance threshold is low, ranking by relevance per unit distance (*Inverse Discount*) is substantially more effective than ranking simply by relevance (*No Distance Disc.*), increasing success from 30% to 40% at 20 miles for example. However, if the user is willing to travel further, distance discounting decreases the probability of success relative to pure relevance ranking. This is because the top positions are too heavily biased towards nearby places when distance discounting is used, thereby making it less likely that a far away but highly relevant place is ever reached by the user. It is interesting to see that this effect is even stronger for the non-learned ranking model, which performs best at low distances. In fact, the learned distance-discounted model is never optimal according to this metric, despite it clearly outperforming the Unigram ranking model according to DCG.

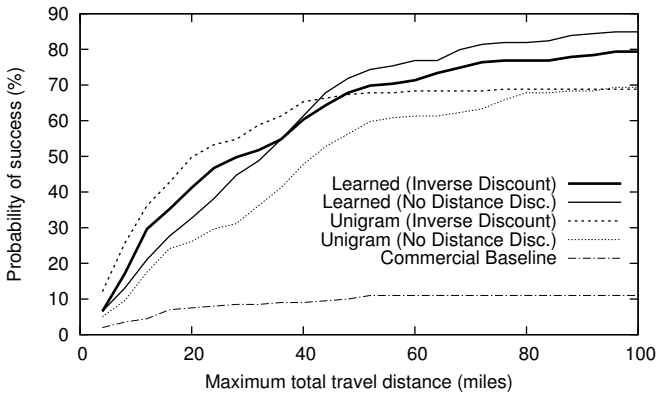

Figure 2: Cap on distance traveled versus success probability for different metrics, with and without distance-based reranking.

Other than mean scores across all 200 queries, we also investigated whether particular classes of queries (corresponding to the subcategories introduced earlier) are substantially easier than others. Table 7 lists the DCG@k scores for the subcategories in our sample of queries. We see that the government service category (where often only one place is correct) is significantly harder for the algorithm than classes with many matching places (such as Meats & Fish, or Shoes). This also suggests that for some categories of products or services other approaches may be more beneficial (for instance,

[4]The exception is the baseline commercial system (Bing), which always ranks in terms of decreasing distance.

Table 7: Ranking performance of ten of the most frequent categories of products, as compared to the aggregate performance. We see that the different categories vary significantly in difficulty.

Category	DCG@1	DCG@3	DCG@5	% Success	E[dist]
Aggregate	4.39	8.50	11.08	84.5 %	14.8 miles
Fashion : Shoes	5.75▲	11.16▲	14.58▲	95 %▲	12.6 miles
Home : Outdoor	5.75▲	10.59	13.46	85 %	11.5 miles
Services : Medical	5.40	9.26	11.45	80 %	16.4 miles
Food & Drink : Specialized	4.45	8.45	10.56	100 %▲	16.0 miles
General Retail : Electronics	4.32	9.54	12.86	95 %▲	12.5 miles
Home : Decor & Accessories	4.30	8.79	11.14	90 %▲	14.1 miles
Food & Drink : Meats & Fish	4.15	8.70	12.03	80 %	13.3 miles
General Retail : Toys & Games	3.85	7.13	8.93	70 %▼	15.0 miles
Fashion : Specialized	3.35	6.91	10.03	95 %▲	21.1 miles▼
Services : Government	2.75▼	4.99▼	6.35▼	55 %▼	16.0 miles

Table 8: Complete query ontology. Starred categories were studied in depth in this paper. Expected distances are the averages from the labeling study in Section 3.4, reported in miles.

Category	Example	Volume	Expected Distance (miles)
General Retail		**21.7 %**	
Toys & Games*	skylanders	3.1 %	15.74
Beauty Products	beard oil	2.8 %	13.52
Electronics*	a computer	2.7 %	13.36
Sports & Fitness	skateboards	2.1 %	13.60
Arts & Crafts	fabric	1.2 %	14.58
Health supplements	vitamin c	1.2 %	13.35
Books & Magazines	textbooks	1.1 %	14.43
Medical	test kits	1.6 %	15.12
Guns	guns	0.7 %	13.39
Illegal	heroin	0.7 %	-
Adult	sex toys	0.6 %	13.70
Jewellery	bracelets	0.6 %	18.10
Other	postcards	3.3 %	14.81
Home		**13.9 %**	
Outdoor*	a boulder	5.2 %	13.00
Decor & Accessories*	lava lamp	3.2 %	13.10
Appliances	refrigerator	1.6 %	11.73
Furniture	a bed	1.4 %	14.80
Home Improvement	window tint	1.3 %	13.84
Other	carpet sweepers	1.2 %	11.13
Fashion		**8.9 %**	
Specialized*	a kilt	4.0 %	15.89
Shoes*	timberlands	2.3 %	11.21
Accessories	hats	1.4 %	13.56
Garments	shapewear	1.2 %	8.83
Other		**4.4 %**	
Automotive	vw engines	3.1 %	15.33
Pets	turtles	1.3 %	16.19
Not a purchase		**0.8 %**	-

Category	Example	Volume	Expected Distance (miles)
Services		**24.2 %**	
Government*	passport	6.1 %	14.90
Medical*	measles shot	3.3 %	12.33
Automotive	oil change	2.0 %	10.72
Financial	title loan	2.0 %	11.22
Restaurants/Cafes	best burger	1.9 %	9.19
Health & Beauty	nose pierced	1.3 %	12.05
Repairs	glass fixed	1.0 %	11.69
Entertainment	broadway tickets	0.9 %	17.46
Photography	glamour pictures	0.8 %	11.96
Training	ged classes	0.8 %	15.38
Transport	bus pass	0.8 %	22.35
Real estate	properties	0.3 %	9.40
Pets & Animals	cat washed	0.3 %	15.80
Adult	-	0.3 %	-
Other	guitar restringed	2.4 %	12.16
Food & Drink		**18.2 %**	
Specialized*	raw honey	6.5 %	12.67
Meat & Fish*	lobsters	3.6 %	14.51
Alcohol	beer	2.6 %	12.57
Desserts	candied fruit	2.4 %	11.80
Fruit & Veg	plouts	1.2 %	11.68
Groceries	eggs	1.1 %	7.54
Tobacco	e-cigarettes	0.8 %	12.95
Specialized Retail		**7.9 %**	
Construction	knox box	3.5 %	13.47
Chemical supplies	cobalt chloride	1.7 %	13.74
Electronics	capacitors	0.9 %	18.68
Antiques/Collectables	silver coins	0.8 %	18.73
Engineering supplies	sealed bearings	0.7 %	16.87
Farming supplies	calves	0.3 %	17.84

there may be suitable fixed databases of the locations where the most common government-related services, such as birth certificates or passports, can be obtained).

6. CONCLUSION

This paper introduced the new problem of retrieving a list of nearby physical stores in response to a query for a product. We presented the first analysis of queries of this sort in real search behavior, and introduced a novel evaluation approach for this task. The features proposed for our system proved useful for creating an effective retrieval system, although substantial avenues for improvements are available. For instance, the diversity in smaller stores could potentially be modeled more accurately by taking into account additional information sources, such as online question and answer forums, crowdsourcing platforms and social network data. Clustering may also be useful for suggesting areas with a high density of likely matches, instead of presenting a greedy list of locations which potentially leads to long travel times if a matching place turns out not to have the item in stock. A more fundamental challenge is how to estimate the utility of visiting a particular place, combining expected travel time and the probability of items being available with other features such as a store's typical prices and user-specific attributes. This will likely lead to improved evaluation metrics for this particular task.

7. REFERENCES

[1] A. Ashkan and C. L. Clarke. Characterizing commercial intent. In *Proc. CIKM*, 2009.

[2] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. In *Proc. ECIR*, 2009.

[3] D. R. Bell, T. H. Ho, and C. S. Tang. Determining Where to Shop: Fixed and Variable Costs of Shopping. *J. Marketing Research*, 35(3), 1998.

[4] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *Proc. WWW*, 2010.

[5] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Sept. 2002.

[6] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using web relevance feedback. In *Proc. CIKM*, 2008.

[7] K.-P. Chiang and R. R. Dholakia. Factors driving consumer intention to shop online: An empirical investigation. *J. Consumer Psychology*, 13(1-2):177 – 183, 2003. Consumers in Cyberspace.

[8] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proc. Intelligent User Interfaces (IUI*, 2009.

[9] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (OCI). In *Proc. WWW*, 2006.

[10] D. Grewal, P. Kopalle, H. Marmorstein, and A. L. Roggeveen. Does travel time to stores matter? the role of merchandise availability. *J. Retailing*, 88(3):437 – 444, 2012.

[11] Q. Guo and E. Agichtein. Exploring searcher interactions for distinguishing types of commercial intent. In *Proc. WWW*, 2010.

[12] A. M. Hinze, C. Chang, and D. M. Nichols. Contextual queries express mobile information needs. In *Proc. Human Computer Interaction with Mobile Devices and Services*, 2010.

[13] B. J. Jansen and T. Mullen. Sponsored search: an overview of the concept, history, and technology. *IJEB*, 6(2):114–131, 2008.

[14] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, 2002.

[15] D. Jiang, J. Pei, and H. Li. Enhancing web search by mining search and browse logs. In *Proc. SIGIR*, 2011.

[16] R. Jones, A. Hassan, and F. Diaz. Geographic features in web search retrieval. In *Proc. International Workshop on Geographic Information Retrieval*, 2008.

[17] J. J. Kacen, J. D. Hess, and W.-Y. K. Chiang. Bricks or clicks? consumer attitudes toward traditional stores and online stores. *Global Economics and Management Review*, 18(1):12 – 21, 2013.

[18] H. Kim and Z. Ghahramani. Bayesian classifier combination. In *Proc. AISTATS*, 2012.

[19] L. R. Klein. Evaluating the potential of interactive media through a new lens: Search versus experience goods. *J. Business Research*, 41(3):195 – 203, 1998.

[20] P. Korgaonkar, R. Silverblatt, and T. Girard. Online retailing, product classifications, and consumer preferences. *Internet Research*, 16(3):267–288, 2006.

[21] D. Lagun, A. Sud, R. W. White, P. Bailey, and G. Buscher. Explicit feedback in local search tasks. In *Proc. SIGIR*, 2013.

[22] C.-J. Lee, N. Craswell, and V. Murdock. Inter-category variation in location search. In *Proc. SIGIR*, 2015.

[23] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. WWW*, 2005.

[24] A. M. Levin, I. R. Levin, and C. E. Heath. Product category dependent consumer preferences for online and offline shopping features and their influence on multi-channel retail alliances. *J. Electron. Commerce Res.*, 4(3):85–93, 2003.

[25] B. Li, A. Ghose, and P. G. Ipeirotis. Towards a theory model for product search. In *Proc. WWW*, 2011.

[26] T.-P. Liang and J.-S. Huang. An empirical study on consumer acceptance of products in electronic markets: a transaction cost model. *Decision Support Systems*, 24(1):29 – 43, 1998.

[27] Y. Lv, D. Lymberopoulos, and Q. Wu. An exploration of ranking heuristics in mobile local search. In *Proc. SIGIR*, 2012.

[28] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765–772, Sept. 2006.

[29] L. Mummidi and J. Krumm. Discovering points of interest from users' map annotations. *GeoJournal*, 72(3-4):215–227, 2008.

[30] C. Quirk, P. Choudhury, J. Gao, H. Suzuki, K. Toutanova, M. Gamon, W.-t. Yih, L. Vanderwende, and C. Cherry. Msr splat, a language analysis toolkit. In *Proc. NAACL HLT*, pages 21–24, 2012.

[31] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: A query substitution approach. In *Proc. SIGIR*, 2008.

[32] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, 2011.

[33] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proc. WWW*, 2004.

[34] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, Mar. 2002.

[35] J. Tang, R. W. White, and P. Bailey. Recommending interesting activity-related local entities. In *Proc. SIGIR*, 2011.

[36] D. Vandic, F. Frasincar, and U. Kaymak. Facet selection algorithms for web product search. In *Proc. CIKM*, 2013.

[37] B. C. Vattikonda, S. Kodipaka, H. Zhou, V. Dave, S. Guha, and A. C. Snoeren. Interpreting advertiser intent in sponsored search. In *Proc. SIGKDD*, 2015.

[38] J. Wang and Y. Zhang. Opportunity model for e-commerce recommendation: Right product; right time. In *Proc. SIGIR*, 2013.

[39] I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proc. WSDM*, 2011.

[40] R. White and G. Buscher. Characterizing local interests and local knowledge. In *Proc. SIGCHI*, 2012.

[41] Q. Wu, C. J. C. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270, 2010.

[42] S. Wu, S. Liu, D. Cosley, and M. Macy. Mining collective local knowledge from google mymaps. In *Proc. WWW*, 2011.

[43] C. Xiong, T. Wang, W. Ding, Y. Shen, and T.-Y. Liu. Relational click prediction for sponsored search. In *Proc. WSDM*, 2012.