Click Through Rate Prediction for Local Search Results

Fidel Cacheda
University of A Coruña
Facultad de Informática
15071, A Coruña, Spain
fidel.cacheda@udc.es

Nicola Barbieri
Tumblr
35 E 21st Street, Ground Floor
New York City, 10010, USA
barbieri@yahoo-inc.com

Roi Blanco
University of A Coruña
Facultad de Informática
15071, A Coruña, Spain
rblanco@udc.es

ABSTRACT

With the ubiquity of internet access and location services provided by smartphone devices, the volume of queries issued by users to find products and services that are located near them is rapidly increasing. Local search engines help users in this task by matching queries with a predefined geographical connotation ("local queries") against a database of local business listings.

Local search differs from traditional web-search because to correctly capture users' click behavior, the estimation of relevance between query and candidate results must be integrated with geographical signals, such as distance. The intuition is that users prefer businesses that are physically closer to them. However, this notion of closeness is likely to depend upon other factors, like the category of the business, the quality of the service provided, the density of businesses in the area of interest, etc.

In this paper we perform an extensive analysis of online users' behavior and investigate the problem of estimating the click-through rate on local search ($LCTR$) by exploiting the combination of standard retrieval methods with a rich collection of geo and business-dependent features. We validate our approach on a large log collected from a real-world local search service. Our evaluation shows that the non-linear combination of business information, geo-local and textual relevance features leads to a significant improvements over state of the art alternative approaches based on a combination of relevance, distance and business reputation.

Keywords

Information Retrieval; Distance; Model

1. INTRODUCTION

Every day people search for services, products, stores and events that are located near them. *Local search queries* encompass a class of search requests that include both a description of *what* the user is looking for and *where* (city name, street address, geographical coordinates, etc.). In this setting, relevant results for local search queries should address good relevance with respect to the user's intent and also meet a specified geographical constraint. Consumers search for local information online and as we tend to simply search for "restaurants" or "cinema show times", the local connotation of queries might often be implicit. Several online services, such as *Google Maps*, *Yahoo local* and *Bing Maps*, allow users to search for businesses or services within a geographical area and they integrate results with rich information, such as name and short description of the business, interactive maps, driving directions, reviews and ratings.

With the widespread availability of Internet connection and geo-localization services provided by smartphones, it is not surprising that a recent study[1] found that 56% of on-the-go of searches have a local intent. This huge and increasing volume of traffic driven by local search queries and the opportunities for the monetization through advertising (e.g. sponsored results) motivates a deep study of users' local search behavior and efforts in improving the quality of search results. To match queries with local businesses local search engines employ machine learning algorithms to learn a function that promotes, for each given user's query, businesses that are more likely to be clicked (in expectation). Since the number of candidate businesses matching a given query can be large, this scoring function decides which ones will be provided as result. Similarly to the web-search setting, complex interactions between properties of the queries and business listings are typically learned by leveraging the huge information available in historical logs that record anonymized users' activity on the service. The scoring function used to rank each pair query-business (q, b) depends mainly on three factors: *relevance*, *distance* and business *profile*. Relevance is a factor that estimates the degree of alignment between the intent of the query and the service/product provided by the business. Distance accounts for the geographical proximity between the location of the business and a) the location specified in the query (if provided) b) the location from which the query has been issued. Finally, the profile of the business includes information such as category, reviews or ratings.

In this paper, we address the problem of estimating the click-through rate on local search results ($LCTR$) as a proxy for deciding the best entries that should be displayed for a given user's query [5]. While in web-search (or traditional information retrieval) it is possible to rely upon human edito-

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018683

[1] https://think.storage.googleapis.com/docs/how-advertisers-can-extend-their-relevance-with-search_research-studies.pdf

rial judgments to determine the quality of a query-document match, this approach is not well-suited to our scenario for a number of reasons. First of all, it is not straightforward to assess whether the distance between the implicit location of the query and the location of the business is reasonable. In fact, this is likely to depend on other factors, like the category or the business, its rating, the density of business in the geographical area of interest, and user's attitude towards traveling short/long distances among others. The strength of the interactions between these variables is highly subjective and some of those factors might be not known to human labelers. Secondly, while in principle users tend to prefer businesses that are physically closer to them, there could be corner-case situations in which a business which is farther away is actually preferred over a similar one that is closer due to other reasons, e.g. this could happen if users consistently prefer a venue over other branches of the same commercial chain because of higher score reviews. Finally, due to the expected inter-dependency between distance, quality of the service and category of the business, an editorial evaluation approach would require the labeling of a large set of examples to cover all cases and deal with inconsistency in judgments.

By relying on *LCTR* as indicator of the quality of the match between query and business we can leverage upon a huge collection of online user preferences and aggregate over a very large number of queries. The main contributions of this work can be summarized as follows:

- We provide an extensive analysis of online user behavior on a real-word local search service (Section 3). This data-driven analysis extends previous studies on the characterization of geographical information needs [6]. More specifically, we investigate: (i) user intent on local search, (ii) the distribution of the distance from users to businesses, (iii) the relationship between such distance and other factors (e.g. category or rating). The results of this analysis provide interesting insights regarding the correlation between clicking behaviors, distance-to business and business reputation.
- We formalize the problem of *LCTR* prediction as a semi-supervised learning task and propose (Section 4) a large number of novel geo and business-dependent features to model patterns in click behavior on local search results. At a high level, geolocation features take into account distance, location of the query/business, characteristics of the neighborhood where the business is located, characterization of the query (explicit vs implicit); features related to the particular venue consider the business category and reputation (e.g. rating).
- Our evaluation (Section 5) shows that the combination of textual, business-dependent and geographical features proposed in this paper outperforms the current state-of-the-art geolocation-aware baseline that exploits textual information, rating and distance by 22% on μ-RMSE and 8% on μ-F1. We extensively discuss findings and analyze the importance of the different group of features.

2. RELATED WORK

Understanding users' behavior on local search has been an active research topic in the last decade, especially in the context of mobile where 35% of the information-needs are reported to be local [24]. Teevan et al. [25] present a study that determines the importance of location, time and social context on mobile local search. A survey on mobile information needs [6] shows that geographical intent accounts for over 30% of web queries issued from mobile devices and it introduces 3 subclasses for characterizing geographical queries: explicit, implicit and directions.

Bian and Chang [4] introduce a taxonomy of local search intents and an automatic local query classifier; local search queries are categorized into two types, "business category" and "business name" and the latter is refined further into "chain business" and "non-chain business". Henrich and Luedecke present a classification of the geographic aspects of information needs [11] by grouping different facets of geographic references in queries. West et al. [27] interpret queries that contain addresses and predict users' mobility patterns using the addresses retrieved with a search engine.

Distance and geolocation information have been exploited for improving the quality of recommendations in different scenarios. Saez-Trumper et al. [23] focus on the task of predicting event check-ins and propose a model that is based on a combination of user's frequency of check-ins, preferences, and closeness to the event. Lu and Caverlee [17] integrate distance as a geo-spatial feature for personalized expert recommendation, obtains up to 24% improvement on precision over several distance-agnostic baselines. Ye et al. [28] propose the concept of "geographical influence" for point-of-interests recommendation. Distance is modeled according to a power law distribution and this information is integrated into a collaborative filtering system that outperforms alternative approaches. Kumar et al. [15] focus on the dynamics of geographical choice of restaurants. In this case, the overall likelihood that the user will select a particular venue depends on its popularity, distance and on alternatives venues closer to the user and it is overall expressed by a combination of log-normals.

All the aforementioned works assume a typical recommendation set-up, whereas we are interested in a query-oriented search scenario. Agarwal et al. [1] use distance and geolocation as a feature for CTR prediction; the location of the user is used in to capture geographical and topical patterns for the consumption of online news. The proposed model is able to combine information from correlated locations which results in an improvement of the predictive performances over the ones achieved by a per-location model. A symmetric approach has been proposed by Balakrishnan et al. [2], who focus on predicting CTR on businesses exclusively taking into account their location and business category (thus not considering the user's query). The idea is that when predicting CTR for a specific business it is possible to use geographic neighborhood information.

A local search engine that integrates relevance with contextual parameters, such as distance, weather and personal preferences, is presented by Lane et al. [16]. Following up on this work, there have been different efforts in improving the ranking on local search results by exploiting relevance labels assigned by human judges. Berberich et al. [3] and Kang et al. [14] propose to improve local search results by combining textual relevance with geolocation of the user, distance and reputation of the venue. Kang et al. [14] recognize the difficulty of the editorial evaluation task and propose some aggregation guidelines to combine the different quality of matching, distance and reputation into a unique local rel-

	Top Query Terms
ALL	restaurant, sports, store, discount, bar, comedy
Events&Performances	clubs, jazz, comedy, museum, view, bar
Landmarks	churches, monument,lane, beach, park, cathedral
Automotive	car, rental, motorcycles, dealer, honda, nissan
Education	school, colleges, bartending, university, classes
Entertainment&Arts	clubs, showtimes, bar, theater, comedy, movie
Computers&Electronics	apple, store, micro, att, computer
Gov&Community	housing, hospital, care, library, museums
Food&Dining	cipriani, eataly, bob's, restaurants
Home&Garden	furniture, discount, home, ikea, barrel
Legal&Financial	chase, health, exchange, bank, insurance
Retail shopping	store, zara, nike, toys, bakery
Professional services	babies, b&h, puppies, agencies, photo
Business to business	leather, spa, moving, radio, limo
Travel&Lodging	hyatt, hotel, marriott, sheraton, cruise
Recreation&Sports	sports, clubs, tours, yoga, bike
Real estate	rooftop, apartments, rent, atrium, real estate
Health&Beauty	bath, massage, cycle, spa, beauty
Other	service, babies, church, exchange, carmel, money

Table 1: *Sample of the top query-terms by category.*

evance label. Our work is inspired by such contributions, but we focus on an alternative task by following a full data-driven approach.

Finally, Lymberopoulos et al. [18] tackle the task of predicting click behavior on mobile local search results by casting it as a click prediction problem. Although this setting is close to ours, the two contributions ultimately differ substantially. In fact, while we explicitly introduce features that measure the textual relevance between the query and the business, this aspect is not considered in [18]. Remarkably, their proposed approach is mainly based on learning the interactions between the rank of the business in the search result ("position") and click behavior. However, especially in a mobile setting, local search engines tend to provide directly search results on a map; in this scenario, any approach based on display position will not be suitable.

3. MODELING LOCAL SEARCH DATA

We are given as input a log $\mathcal{D}$ which records past users' interactions with the results provided by the local search engine. This log consists of a collection of tuples of the form $\langle id, q, q_{loc}, b, \mathsf{click}\rangle$, where each entry denotes the impression of a business among the results for a particular user's query. More specifically, *id* is an anonymized identifier associated with the user's session on the search engine, q denotes the user's query (free-form text), q_{loc} represents information regarding where the query was issued (typically inferred from the user's IP address), b denotes an identifier associated with the business, and click is a binary indicator denoting whether on the considered session the user clicked on the business (click=true) or not (click=false).

We also assume that the following information regarding each business is available:

- the *title* of the business;
- a *description* of the service/product provided by the business, which could be either an editorial summary or automatically extracted from the web-page associated to the business;
- the *display url* (human friendly version of the URL without parameters) of the corresponding landing page;
- the most relevant *business category*;
- the average user's *rating*, which is assumed to reflect the quality of service/product provided by the business;
- the *location* (GPS coordinates) associated with the business.

Given this data, we focus on the problem of estimating the likelihood that particular business will be clicked given a query and its location. This likelihood is computed by aggregating information from multiple user's sessions, which leads to the definition of *local search click-through-rate* ($LCTR$):

$$LCTR(b; q, q_{loc}) = \frac{|\langle \_, q, q_{loc}, b, \mathsf{true}\rangle \in \mathcal{D}|}{|\langle \_, q, q_{loc}, b, \_\rangle \in \mathcal{D}|}.$$

This measure is a special instance of the classical CTR, one of the most widely used metrics in online advertising, which is defined as the ratio between the number of clicks and the number of impressions of an ad for a query.

3.1 User behavior on local search

To characterize user's behavior on local search we rely on the analysis of a large sample of user interactions with the Yahoo search engine in a time window of several months. For the sake of the following analysis, we consider only events that lead to a click on businesses located in New York City and for which the information about the location of the user is available, which led to a volume of nearly 600k events. We borrow the categorization of geographical intent introduced in [6] but we deem as implicit even those queries where the *explicit location* is broader than user's true location, e.g. *pizza in new york*, while the user-zip is a postcode in NYC. According to this categorization, we find that the location can be considered as explicit for roughly 28% of the queries.

Each business in this data is associated with the most relevant out of a list of 18 categories. To preserve users' anonymity, the query log simply stores the zip code where the user is located; hence, we associate each click event with the coordinates (longitude and latitude) of the geographical center of the users' zip code. In order to compute the distance between each user and the location of the business, we make use of the *Haversine*[2] formula, which computes the *great circle distance* between two points on the Earth.

Local search intent. We start our analysis by characterizing user's intent on local search, mainly investigating on which subjects users are interested in when they interact with the local search engine. We do so by analyzing the distribution of business categories on clicked entries, which is reported in Figure 1. The top categories of interests are "Food and dining" (38%, really close to value reported by [18]) and "Entertaining and arts" (16%). Furthermore, we provide in Table 1 a sample of the most-frequent query-terms for each business category.

Analysis of the distance. Next, we focus on distribution of the distance from the user to the clicked business, given in Figure 2. This data clearly follows a heavy tailed distribution; this holds especially true on shorter distances (the frequency decreases as the distance increases), while for larger distances we observe a noisier relationship. Interestingly, the first dot in the plot shows that the frequency for business closest to the user (less than 1km) has not the highest value. In this case, probably the user's knowledge of her own neighborhood reduces the need to for local search.

To better characterize this data we evaluate the goodness-of-fit between the data and the Power Law and compare this fit with an alternative hypothesis, namely Log-Normal, via

[2]https://en.wikipedia.org/wiki/Haversine_formula

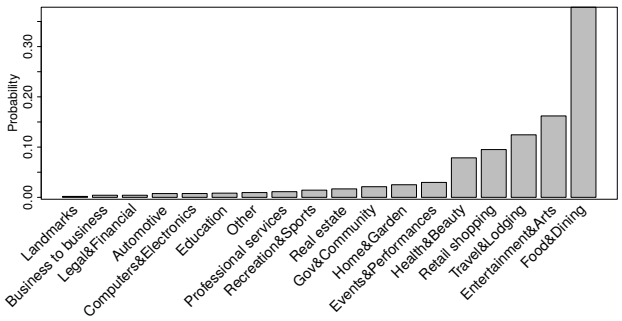

Figure 1: *Distribution of categories on local search results.*

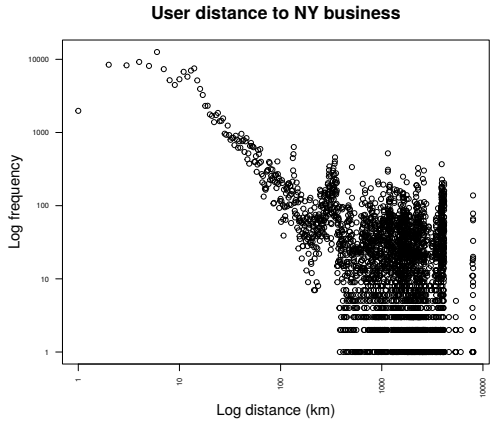

Figure 2: *Frequency plot for distance from user to business.*

a likelihood-ratio test. This statistical analysis is based on the fitting procedure for quantifying power-law behavior in empirical data described by Clauset et al. [7]. Table 2 summarizes the best setting of parameters for both hypotheses on the overall data and considering different business categories. This analysis suggests that the distance between user and business is usually larger for entries related to "Travel and Lodging". According to the log-likelihood ratio test, the distribution that best fits these values is always the Log-Normal. Figure 3 shows the cumulative distribution function on the distance values recorded on entries of two categories and the fit of the two hypotheses.

Distance vs rating. The relationship between the distance and the reputation of each business can be assessed by measuring the median and average distance per rating, as reported in Table 3. The great majority of the business receives ratings between 3 and 4 (ratings are given in a decimal scale). As the rating increases, so does the distance from the user to the business (both median and mean), with the exception of the top rating, most likely due to the location of 5-star rated business in lower business density areas, reducing the user mobility as there are less alternatives (mean distances are statistically significantly different among all ratings). These values suggest that users are willing to travel more distance for better (i.e. higher rated) businesses.

Distance vs. category. Another interesting interplay to analyze is the relationship between the distance and the

	PL α	LN μ	LN σ	Best fit
ALL	1.432	3.612	2.154	LN(0.006)
Events&Performances	1.421	3.671	2.086	LN(0.138)
Landmarks	1.391	3.671	2.086	LN(0.057)
Automotive	1.500	3.671	2.086	LN(0.190)
Education	1.546	3.319	1.978	LN(0.001)
Entertainment&Arts	1.463	3.530	2.066	LN(0.047)
Computers&Electronics	1.562	3.530	2.066	LN(0.001)
Government&Community	1.475	3.388	2.050	LN(0.046)
Food&Dining	1.410	3.498	2.168	LN(0.083)
Home&Garden	1.474	3.153	1.931	LN(0.610)
Legal&Financial	1.437	3.153	1.931	LN(0.005)
Retail shopping	1.338	3.662	2.234	LN(0.001)
Professional services	1.428	3.662	2.234	LN(0.050)
Business to business	1.433	3.662	2.234	LN(0.007)
Travel&Lodging	2.321	4.979	2.284	LN(0.001)
Recreation&Sports	1.529	3.370	1.910	LN(0.017)
Real estate	1.317	3.370	1.910	LN(0.001)
Health&Beauty	1.519	3.110	1.928	LN(0.391)
Other	1.448	3.591	2.123	LN(0.054)

Table 2: *Comparing PowerLaw vs LogNormal fit on clicked entries. The best fit according to Vuong's test [26] is always the LogNormal Distribution (two-sided P values are given in the last column).*

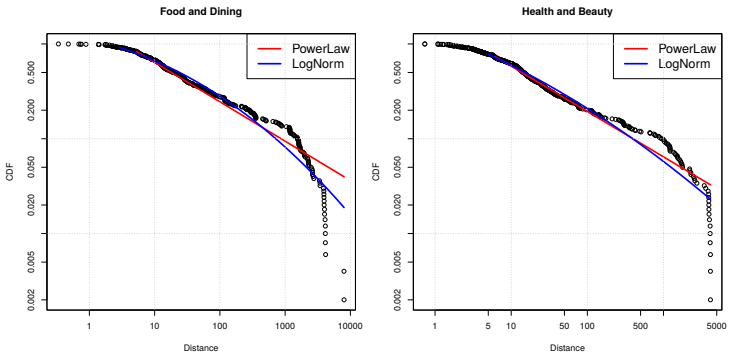

Figure 3: *Cumulative distribution functions on distance data for two business categories. Lines represent the best fit for the Power Law and Log Normal distributions. The best fit according to Vuong's test [26] is always the LogNormal.*

business category. Simply put, we are interested in determining how the notion of distance and relevance varies across different local intents. Figure 4 provides a box-plot for the logarithm of the distance per each category and it shows that some categories tend to have longer distances between the user and the business. More specifically, there are three categories that have a clear higher median (presented in order, from higher to lower): "Travel&Lodging" (median 476.7km), "Retail shopping" (median 147.2km) and "Real estate" (median 92.1km). This analysis shows different behaviors depending on the business category. For example, users searching for "restaurants" are mainly close to the business, while the location of those searching for "hotels" spans up to thousands of kilometers.

Distance vs. *LCTR*. Finally, we analyze the correlation between *LCTR* and (log) distance which is summarized in Table 4. The correlation values are mostly negative which suggests that users prefer businesses that are physically closer to them. The overall correlation is mild and

Rating	Percentage	Median	Mean
1	1%	13.42	381.20
2	5.2%	22.74	526.90
3	40.4%	26.62	556.30
4	49.7%	33.26	633.70
5	3.7%	17.60	486.70

Table 3: *Distance statistics (in kilometers) per rating.*

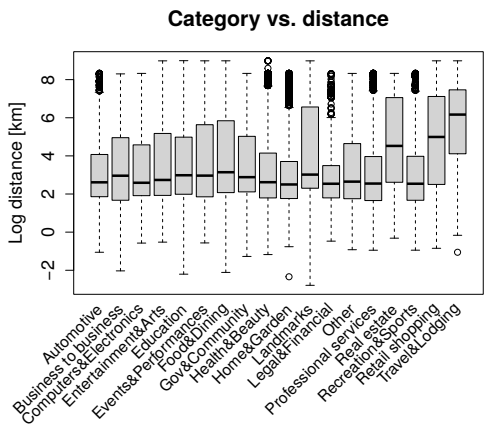

Figure 4: *Box-plot for log(distance) per category.*

interestingly it tends to weaken when considering businesses with higher ratings, suggesting that users are prone to discount the distance factor when considering venues with high reputation.

		Rating	
	Overall	<3	≥3
ALL	-0.02	-0.03	-0.02
Events&Performances	-0.06	-0.09	-0.06
Landmarks	0.00	0.06	0.00
Automotive	-0.06	-0.07	-0.05
Education	-0.05	-0.13	-0.04
Entertainment&Arts	-0.05	-0.06	-0.04
Computers&Electronics	0.01	0.06	0.00
Government&Community	-0.02	-0.02	-0.02
Food&Dining	-0.01	-0.05	-0.01
Home&Garden	-0.04	-0.02	-0.05
Legal&Financial	0.00	-0.04	0.02
Retail shopping	0.01	-0.02	0.02
Professional services	-0.01	-0.07	0.01
Business to business	-0.01	-0.05	0.01
Travel&Lodging	-0.02	-0.01	-0.02
Recreation&Sports	0.00	0.03	-0.01
Real estate	-0.08	-0.17	-0.05
Health&Beauty	-0.03	-0.04	-0.02
Other	-0.05	-0.06	-0.05

Table 4: *Pearson correlation between LCTR and log(distance) by category and rating.*

4. LCTR PREDICTION

The learning task considered in this work can be formalized in two different ways, either as a regression on the true value of LCTR or as binary classification using presence or absence of clicks as labels, as both types have been employed in previous work [1][2][18].

Given a set of training examples $\mathcal{T}$ we aim at learning a function $\vartheta^*(q, q_{loc}, b) \rightarrow [0, 1]$ that minimizes the quadratic loss between the observed LCTR and the predicted one:

$$\vartheta^* = \arg\min_{\vartheta} \sum_{e=\langle q, q_{loc}, b\rangle \in \mathcal{T}} \left(\vartheta(e) - LCTR(e)\right)^2 . \quad (1)$$

Alternatively, we can model the problem as a binary prediction problem and aim to learn a function that minimizes the binomial log-likelihood:

$$\vartheta^* = \arg\min_{\vartheta} 2 \sum_{e=\langle q, q_{loc}, b\rangle \in \mathcal{T}} \log\left(1 + exp\left(-2\vartheta(e)\mathcal{I}\{LCTR(e) > 0\}\right)\right) , \quad (2)$$

where $\mathcal{I}\{\cdot\}$ is the indicator function. We make use of both formulations and in Section 5 we report results for both prediction and classification.

To address these learning problems, we resort to a feature-based approach and design a collection of features that are expected to capture correlations between different aspects of the matching query-business and LCTR. More specifically, we represent each training example by a feature vector $\phi(q, q_{loc}, b) \in \mathbb{R}^F$, where F denotes the number of features, and then we use this vector as input for the prediction function ϑ. This approach allows us to develop a large number of features and to estimate relationships between the values of such features and LCRT by employing techniques suited for learning at large scale, such as generalized linear models or tree-based algorithms.

We propose three classes of features. The first one is computed by analyzing the text similarity between the query and the information available on the business. The second group focuses on attributes of the business, such as the category and rating. Finally, the last one consists of features that are computed by using the location information. We intentionally omit some other features that are often considered for the task for predicting CTR. For example, we do not consider query/business-dependent estimates that can be computed on historical data (such as average CTR, the frequency in the log, etc.) or treat the id of the business explicitly as a feature. In fact, we assume a cold-start setting for pairs query-business and in this context historical estimates would not be available. Similarly, we do not consider the position of the business into the list provided to the user for two reasons. First, while in click modeling the position at which a document is displayed introduces a bias on the probability that the user will click on it [13], in local search all the results are displayed on a map and we can not assume that all results will be viewed by ranking order. Secondly, by ignoring rank-dependent features we could reuse this model to produce a query dependent ranking.

4.1 Text-based features

The main goal of this group of features is to estimate the degree of alignment between the intent of the query and the product/service provided by the business. Among different choices for modeling textual similarity, we ignore the ordering of words and opt for a bag-of-words representation.

Okapi BM25. This is one of the most widely used scoring methods in information retrieval. In our setting, each document represents the concatenation of all the textual information available for the business (title, description and url), and IDF dictionary is computed over the overall collection of such documents. The textual similarity between the query and different parts of the business textual information might have different degrees of importance. To account for

this effect, we also use *BM25*F, which is a weighted aggregation of *BM25* scores each one computed between the query and a different part of the business textual information [22].

Embeddings. Among several choices for defining the embedding space for computing the similarity between queries and documents, we opt for *LSI* [8, 9] as it has two main advantages. First, the truncated SVD decomposition of the term-business matrix $\mathbf{B}$ (where $\mathbf{B}_{w,b}$ represents the frequency of word w into the textual information available for the business b)

$$\mathbf{B} \approx U_k \Sigma_k V_k^T,$$

can be computed efficiently even on large scale corpus; secondly, the embedding of a new query/document into the latent dimensional space can be determined in a *exact way* by applying a fast *fold-in* procedure:

$$\tilde{q} = \Sigma_k^{-1} U_k^T \vec{q},$$

where $\vec{q}$ is the vector-representation of the input query in the dictionary space and the projection matrix $\left(\Sigma_k^{-1} U_k^T\right)$ can be pre-computed once.
Similarity features between a query and a business are computed as the *cosine* between the respective projections into the embedding space. As above, we compute different *LSI* scores for all the textual information available for the business and each textual information part (*LSI-F*).

4.2 Business-dependent features

This group of features is used to profile the characteristics of a given venue. We consider mainly two signals: the category of the business and an indicator of the quality of the service. Different services may provide this information at different levels of granularity, i.e., the complete list of categories/sub-categories and reviews that apply to each business. We consider the most general setting by assuming that only the main category (*Category*) and the average rating (*Rating*) are available for each business. To take into account category specific rating patterns (e.g. users' bias in assigning ratings on different categories), we also compute the difference between the average rating of the business and the one corresponding to the category (*RatingDev*).

4.3 Geographical features

This collection of features is designed to capture relationships between LCRT and: (i) distance between the user and the business; (ii) properties of the neighborhood where the query was issued and (iii) where the business is placed.

Explicit local query indicator. The categorization of the query into explicit vs. implicit has a direct implication when computing the distance between the geographical area of interest of the user and the location of the business. In addition, the explicit area of interest could be much broader (e.g. "nyc") than the implicit geolocation extracted from the query (e.g. some zip code in New York City). To address these situations we introduce two features:

- *EQ*: a binary indicator that says if the query contains a geolocation entity or not;
- *GeoLocID*: an identifier of the geolocation entity, if present and -1 otherwise.

To determine if the query contains a geolocation we string match it against a gazetteer that contains a list of locations, acronyms and common abbreviations. If the query contains multiple locations, then we consider the most specific (e.g. "pizza manhattan nyc" $\rightarrow$ "manhattan"). We found this simple technique to work well in the context of New York City, but in general the usage of a geoparser should be preferred to disambiguate ambiguous references like "Springfield".

Zip codes. The availability and granularity of the location information depends on the user's privacy setting; in this work we assume it is available at a coarse-grained level, i.e. the zip code. We introduce two features:

- *QZip*: an identifier of the zip code from which the query was issued. This feature will account for biases related to the neighborhood of the user (e.g. if combined with distance it can model the tendency of users from this neighborhood to travel more/less distance, etc.).
- *BZip*: an identifier of the zip code corresponding to the location of the business. This is expected to model direct interactions with LCTR (e.g. average LCTR for the considered neighborhood), as well as more complex interactions when combined with other factors (e.g. likelihood of users to cover a given distance when traveling to this neighborhood).

Distance. According to the intuition that, other factors being equal, users tend to prefer business that are located near to them, distance signals are expected to play an important role in discriminating which business entry will be clicked for each local search query. We consider the following features:

- *Distance*: this is computed as Haversine distance between the location of the user/business;
- *LogDist*: a logarithmic transformation of the raw distance values;
- *LkDist*: this is the likelihood of observing the distance value d according to a log-normal model $\mathcal{N}(\log(d); \mu; \sigma)$ where the parameters of such model are estimated on distance values recorded for clicked entries.
- *LkDistCat*: represents the likelihood of observing distance d according to a log-normal model that depends on the current category c of the business $\mathcal{N}(\log(d); \mu_c; \sigma_c)$. In this case the parameters of the model are estimated by considering the distance for clicked events on business of each category.

Zip profile. This last group of features is used to profile the neighborhood where the business is located, by aggregating business information at the level of zip-code. This information is likely to model the bias of different aspects of the neighborhood (like density of businesses) on LCTR, as well as competition between the considered venue and other businesses in the same area.

- *CntBusiness(Bzip)*: it represents the number of businesses located in this zip code, which is assumed as a proxy for the density of businesses in the area;
- *CntBusinessByCat(Bzip)*: it is a set of features, each one of them represents the number of businesses belonging to the particular category that are located in the considered zip code.
- *ProbCat(Bzip)*: is a set of features, representing the probability of each category in the considered zip-code. It is obtained by normalizing *CntBusinessByCat* and it

	Training	Test
# distinct queries	49K	21K
# distinct business	20K	13K
# of entries	142K	62K
avg # queries by business (clicked)	0.36	0.38
avg # business query (clicked)	0.41	0.41
# zip (queries)	7.1K	5K
# zip (business)	1.2K	924
avg # of queries by zip	19.84	12.49
avg # of business by zip	16.05	14.41
avg distance (clicked entries)	392.73	393.81
avg rating per business	3.55	3.57
avg rating per zip	3.62	3.66

Table 5: *Main properties of train and test data.*

is used as a proxy for profiling the overall distribution of business categories in the neighborhood;

- *CntBusinessCurrCat(Bzip)*: represents the number of businesses located in the zip code for the same category of the considered venue; it is used as a proxy for the density of potential competitors in the area;
- *ProbCurrCat(Bzip)*: it is the probability of observing businesses belonging to the same category of the considered venue in this neighborhood;
- *EntropyCat(Bzip)*: it represents the entropy of categories in the neighborhood, i.e.

$$-\sum_{cat} P(cat|BZip)\log P(cat|BZip).$$

It is a proxy for the degree of "specialization" of the considered area;

- *AvgRating(Bzip)*: it is the average of ratings computed on businesses located in the considered area;
- *AvgRatingByCat(Bzip)*: it is a set of features representing the average rating of businesses located in this area for each category;
- *AvgRatingCurrCat(Bzip)*: represents the average rating of businesses located in this area for the same category of the considered venue.

5. EXPERIMENTAL EVALUATION

In this section we report our main findings for the regression and classification tasks outlined in Section 4, along with error analysis and discussion on the effect of features and model performance across query types.

Data. To assess the predictive accuracy in the task of predicting CTR on local search result, we further preprocess the data discussed in Section 3 to avoid poor LCTR estimates due to a small number of impressions. More specifically, we discarded triples (b, q, q_{loc}) that occur less than 5 times in the considered query log, leaving roughly 204K entries. The following evaluation is based on query-based train-test random split (70-30). The main characteristics of the dataset are given in Table 5.

Learners. We note that in order to learn a model (Section 4) one can employ a readily available machine learning toolkit. For the problem at hand we consider two options: Gradient Boosting Decision Trees (GBDT) [10] and Logistic Regression. GBDT can solve both the regression (Eq. 1) and classification (Eq. 2) losses, whereas logistic regression is often used in classification tasks. The main results of this section are performed using GBDT and we add a comparison between both methods with a simple explanation of their different behavior at the end of this section.

Regarding baselines, we use: (i) models that focus on strong textual relevance features derived from BM25 (description only) and BM25F using the different fields (as proposed in [21]), (ii) an adaptation (BL) of the state of the art methods proposed by Kang et al.[14] and Berberich et al.[3] which incorporates text relevance, distance and ratings. In previous works, a discretization of the distance into three classes is used to perform label aggregation and to derive a final score for a query and business pair. In our setting, best performances are achieved by incorporating directly the raw features into the learner. Furthermore, considering all textual fields available, rathen than just description, improved the baseline performance by over 10%.

Learners meta-parameters (number of trees, number of nodes, BM25's b and k_1) for all models have been selected independently in an external development set, whereas for LSI we set $k = 100$ for all the experiments.

Performance measures. We use two different families of metrics: RMSE for evaluating the *true* regression performance of the different classifiers and standard micro and macro precision, recall and F measures (macro is averaged per unique query). The two sets of metrics are evaluated on different label sets, in particular the classification based metrics are reported for a classifier trained with labels converted into 0 and 1 (if CTR is greater than zero or not). We report on statistical significance using a standard two-sided t-test with significance level $\alpha = 0.01$.

Features. Throughout all the experiments, features described in Section 4 are grouped in the following way:

- *BUSINESS_INFO*: *Category*,*Rating*, *RatingDev*;
- *DISTANCE*:*EQ*,*GeoLocID*,*QZip*, *Bzip*,*Distance*,*LogDist*, *LkDist*,*LkDistCat*;
- *ZIP_PROFILE_CAT*: *CntBusinessByCat(Bzip)*, *ProbCat(Bzip)*, *CntBusinessCurrCat(Bzip)*, *ProbCurrCat(Bzip)*, *EntropyCat(Bzip)*, *AvgRatingByCat(Bzip)*, *AvgRatingCurrCat(Bzip)*;
- *ZIP_OVERALL*: *CntBusiness(Bzip)*, *AvgRating(Bzip)*;
- *ALL*: includes all previous features.

Category information is obtained directly from the Yahoo local search engine, while as Gazetteer we use the *NYS GIS Data Set*.[3]

Evaluation on CTR prediction The first columns of Table 6 show the main results on evaluating CTR regression prediction and the rest report the results on click classification. Results are consistent independently if they are averaged by query or the pair (query, business) for both tasks and all metrics, except for a minor disagreement between the order of BM25 and BM25+LSI. All the differences between our methods and every baseline are statistically significant (p values around zero), as well as the differences between the feature combination themselves (with respect to all metrics).

With respect to the baselines, the best results are obtained by combining textual information (BM25F), which incorporates all text fields (title, description and url) into the model

[3]http://gis.ny.gov/

	μ-RMSE	RMSE	μ-P	μ-R	μ-F	AUC	P	R	F
BM25	0.2206	0.2129	0.5975	0.4950	0.5414	0.5997	0.2705	0.3229	0.2849
BM25F	0.1847	0.1941	0.7256	0.5593	0.6317	0.6859	0.3290	0.3710	0.3419
BM25+LSI	0.2165	0.2150	0.6512	0.5072	0.5702	0.6331	0.2745	0.3216	0.2878
BM25F+LSI	0.1720	0.1978	0.7618	0.5799	0.6585	0.7095	0.3465	0.3857	0.3580
BL	0.1666	0.1911	0.7637	0.6021	0.6733	0.7185	0.3660	0.3985	0.3738
BM25F+BUSINESS_INFO	0.1612	0.1841	0.7833	0.5966	0.6773	0.7251	0.3629	0.3957	0.3727
BM25F+DISTANCE	0.1453	0.1827	0.7820	0.6382	0.7028	0.7402	0.3916	0.4220	0.3989
BM25F+ZIP_PROFILE_CAT	0.1431	0.1750	0.8033	0.6400	0.7124	0.7505	0.3991	0.4284	0.4081
BM25F+ZIP_OVERALL	0.1609	0.1810	0.7797	0.6131	0.6865	0.7298	0.3706	0.4040	0.3807
BM25F+BUSINESS_INFO+DIST	0.1351	0.1755	0.8005	0.6485	0.7165	0.7526	0.4025	0.4297	0.4091
ALL	**0.1294**	**0.1738**	**0.8077**	**0.6594**	**0.7261**	**0.7601**	**0.4145**	**0.4407**	**0.4211**

Table 6: *Evaluation results for regression and classification on different feature sets for GBDT. μ-metric denotes the metric averaged over the total number of examples, whereas the metric is averaged over the different unique queries otherwise. We report on Precision, Recall, F, Area Under the roc Curve and Root Mean Squared Error.*

		μ-P	μ-R	μ-F	AUC	P	R	F
LR	BM25F	0.5739	**0.6070**	0.5900	0.6037	**0.3374**	**0.4016**	**0.3554**
	BL	0.5904	0.4942	0.5381	0.5951	0.2663	0.3148	0.2797
	BM25F+ BUSINESS_INFO + DIST	0.6428	0.5750	0.6070	0.6458	0.3206	0.3689	0.3329
	ALL	**0.6472**	0.5793	**0.6115**	**0.6497**	0.3231	0.3705	0.3356

Table 7: *Logistic regression performance for classification on different feature sets. μ-metric denotes the metric averaged over the total number of examples, whereas the metric is averaged over the different unique queries otherwise.*

alongside with business rating and distance. The first thing that stands-out is that textual information alone is able to produce relatively good results (especially when combining field and embeddings information) but they underperform when business and distance information are incorporating into the model, confirming the results from [14] and [3]. Similarly, adding business category information is able to boost the performance by around 10-15% depending on the metric of interest (RMSE, AUC or F). The model that uses all the features outperforms BL over 22% in μ-RMSE and over 8% in terms of μ-F. With respect to feature group combination, the best model is the one that incorporates business information and distance signals, followed somehow closely by BM25F+ZIP_PROFILE_CAT.

Remarkably, the different feature groups proposed are able to impact the performance significantly, outperforming all the baselines even if those feature groups are different in nature (business information, distance or zip code for profile categories). In general, distance and zip code profile category features, considered individually, seem to produce higher impact on the performance.

Error analysis. We now explore the errors produced by the best performing method (BM25F+ALL). Figure 5 shows a breakdown of the performance taking into account different characteristics of the examples classified, namely business category, query frequency(low, mid and high frequency queries) and business rating.

Regarding business category, although all different categories obtain similar error values, those categories with less businesses (e.g. Landmarks or Legal services) tend to obtain better results. This observation holds as well with respect to business ratings: business with ratings between 3 and 4.5 account for the largest error, and those classes comprise the most frequent slice of the dataset. With respect to the query frequency, the queries in the tail of the distribution produce slightly worse results than queries more frequent, whereas torso-head queries perform somehow similarly.

Linear vs non-linear classifiers. Logistic regression (LR) for click prediction is a popular learner for feature-based models [21, 19]. Note that LR is a generalized linear model and is not able to model feature interactions; on the contrary GBDT can learn relatively complex decision rules among groups of features. The results on Table 7 show that LR is a good option when using only textual features whereas when incorporating features that require non-linear combinations (business category and distance) it fails to capture the richer feature dependencies. Overall, GBDT was found to be the most reliable learner for the problem at hand, improving, except for BM25F, all the measures evaluated (e.g. μ-F1 up to 10%).

Feature importance. Aside of the insights already reported for different feature classes and their combinations, we provide a finer grained feature importance analysis. Figure 6 provides a graphical representation of the features that contribute the most to the regressed GDBT model and their relative importance,[4] computed with Friedman's method [10]. This takes into account individual feature correlations.

This analysis suggests that BM25F plays the main role in estimating *LCTR*, followed by business category. Surprisingly, Bzip and even rating seem to be more important than the actual distance between the venue and the user, which suggests that the learner favors area and business modeling over distance. This may indicate that once we select a business category, knowing where it is located is more important than how far away it is.

[4]For features that are defined as a group, we report the importance of the top one.

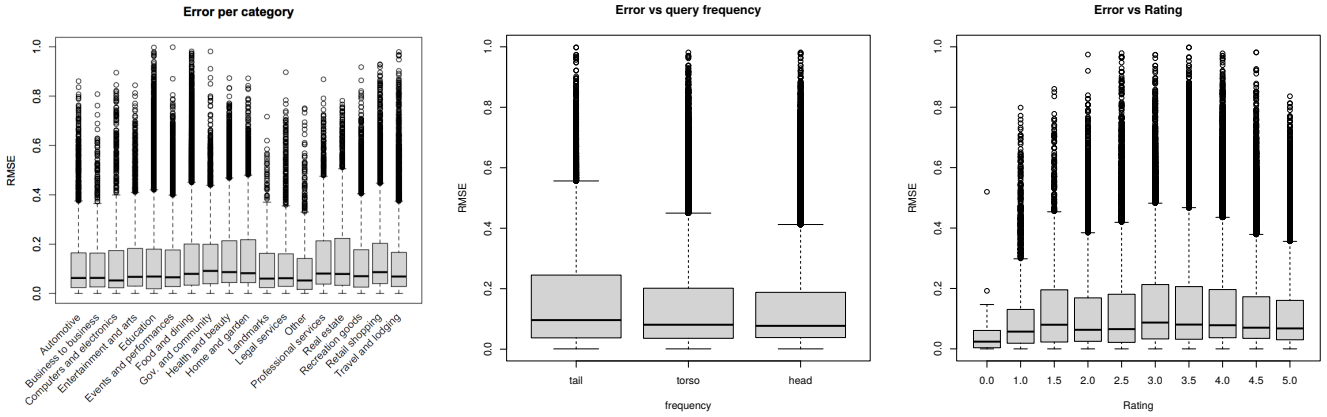

Figure 5: *RMSE breakdown per business category, rating and query frequency (low, mid and high frequency).*

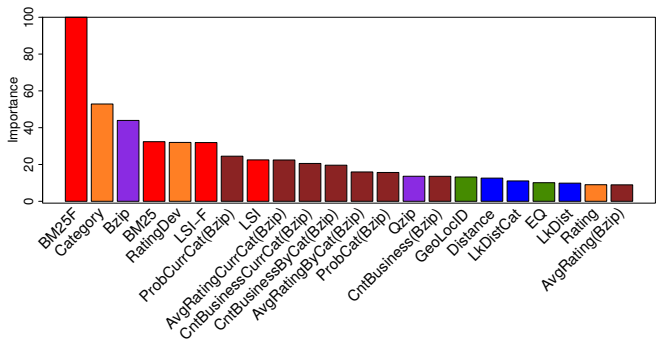

Figure 6: *Feature importance. Colors represent feature-groups: textual-relevance (red), business-dependent (orange), explicit query indicator (green), zip-codes (violet), distance (blue) and zip profile (brown).*

6. CONCLUSIONS AND FUTURE WORK

In this paper, we present an extensive study to characterize and predict online users' behavior on local search. Such analysis is based on a large sample (>600k click events) of user interactions with the Yahoo search engine when looking for business located in NYC, which provides interesting insights on the interplay between clicking behaviour and distance-to-business, business reputation and category. Then, we focus on the problem of estimating the click-through rate on the different local search results according to two complementary perspectives (regression and binary classification). To address these problems we design a collection of features that are expected to model different aspects of the matching query-business. Our evaluation shows that the overall approach proposed in this paper outperforms a geolocation-aware baseline that combines textual information, rating and distance by 22% on μ-RMSE and 8% on μ-F1.

This work can be extended in several directions. First, it would be interesting to analyze user's behaviour on a larger geographical scale. Secondly, we plan to study the integration of new features, such as contextual factors (e.g. the time at which the query is issued), or alternative definitions of distance that can account for the way people deal with space (such as the *Space Syntax* framework [12]). Alternative machine learning methods that model complex iteractions among features, such as Factorization machines [20], can be also considered for the task of predicting click-through-rate. Finally we plan to evaluate the effectiveness of different groups of features on a ranking-based setting.

Acknowledgments

Fidel Cacheda acknowledges the support from the Spanish Government, Ministerio de Educación, Cultura y Deporte (PR2015-00231) and Ministerio de Economía y Competividad (TIN2015-70648-P).

7. REFERENCES

[1] D. Agarwal, B.-C. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 21–30, 2009.

[2] S. Balakrishnan, S. Chopra, and I. D. Melamed. The business next door: Click-through rate modeling for local search. *Machine Learning in Online Advertising*, page 14, 2010.

[3] K. Berberich, A. C. König, D. Lymberopoulos, and P. Zhao. Improving local search ranking through external logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 785–794, 2011.

[4] J. Bian and Y. Chang. A taxonomy of local search: Semi-supervised query classification driven by information needs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2425–2428, 2011.

[5] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1–10, New York, NY, USA, 2009. ACM.

[6] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 247–256, 2009.

[7] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

[9] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human factors in computing systems*, pages 281–285, 1988.

[10] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002.

[11] A. Henrich and V. Luedecke. Characteristics of geographic information needs. In *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval*, GIR '07, pages 1–6, 2007.

[12] B. Hillier and J. Hanson. *The social logic of space*. Cambridge university press, 1984.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, 2005.

[14] C. Kang, X. Wang, Y. Chang, and B. Tseng. Learning to rank with multi-aspect relevance for vertical search. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 453–462, 2012.

[15] R. Kumar, M. Mahdian, B. Pang, A. Tomkins, and S. Vassilvitskii. Driven by food: Modeling geographic choice. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 213–222, 2015.

[16] N. D. Lane, D. Lymberopoulos, F. Zhao, and A. T. Campbell. Hapori: Context-based local search for mobile phones using community behavioral modeling and similarity. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 109–118, 2010.

[17] H. Lu and J. Caverlee. Exploiting geo-spatial preference for personalized expert recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 67–74, 2015.

[18] D. Lymberopoulos, P. Zhao, C. Konig, K. Berberich, and J. Liu. Location-aware click prediction in mobile local search. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 413–422, 2011.

[19] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: A view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1222–1230, 2013.

[20] S. Rendle. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57, 2012.

[21] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International World Wide Web Conference(WWW-2007)*, 2007.

[22] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, Apr. 2009.

[23] D. Saez-Trumper, D. Quercia, and J. Crowcroft. Ads and the city: Considering geographic distance goes a long way. In *Proceedings of the 6th ACM Conference on Recommender Systems*, RecSys '12, pages 187–194, 2012.

[24] T. Sohn, K. A. Li, W. G. Griswold, and J. D. Hollan. A diary study of mobile information needs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 433–442, 2008.

[25] J. Teevan, A. Karlson, S. Amini, A. J. B. Brush, and J. Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 77–80, 2011.

[26] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

[27] R. West, R. W. White, and E. Horvitz. Here and there: Goals, activities, and predictions about location from geotagged queries. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 817–820, 2013.

[28] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th International ACM SIGIR*, SIGIR '11, pages 325–334, 2011.