Improving Local Search with Open Geographic Data

Chuankai An, Dan Rockmore
Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA
{chuankai, rockmore}@cs.dartmouth.edu

ABSTRACT

Local search helps users find certain types of business units (restaurant, gas stations, hospitals, etc.) in the surrounding area. However, some merchants might not have much online content (e.g. customer reviews, business descriptions, opening hours, telephone numbers, etc.). This can pose a problem for traditional local search algorithms such as vector space based approaches. With this difficulty in mind, in this paper we present an approach to local search that incorporates geographic open data. Using the publicly available *Yelp* dataset we are able to uncover patterns that link geographic features and user preferences. From this, we propose a model to infer user preferences that integrates geographic parameters. Through this model and estimation of user preference, we develop a new framework for \local" (in the sense of geography) search that offsets the absence of contexts regarding physical business units. Our initial analysis points to the meaningful integration of open geographic data in local search and points out several directions for further research.

Keywords

Local Search, Geographic Open Data, Preference Estimation

1. INTRODUCTION

Local search supports the query for a certain type of \target" in the vicinity of a user's geographic location. Related online information providers include *Yellow Pages*, *Yelp* and *FourSquare*. In traditional web search, since some local business units do not contain a long text as a part of their online presence, information retrieval models based on word-document relationship may not work well. While current search engines generally are able to return satisfying results, new business units that lack a significant online description/presence are still challenged by this \partial availability problem". Presuming that many of these business units would present excellent options for potential users, this also presents a problem/challenge for enhancing user experience.

With this in mind we aim to find external freely available resources (\open data") to augment this kind of potential scarcity of information and build an enhanced model for local search. For instance, the keywords in an advertisement of a local shop can be used as a proxy for the basic description of the shop. Basic geographic open data is also very useful. The locations of both user and business units enable the computation of distance between them. However, after we can query local locations from an open database to get basic attributes such as the name of address, how can we use that extra information? In this paper, we discuss the specific question of how local search can be improved with geographic open data.

Current works tend to analyze a user's search log to improve local search. Teevan [9] conducts a survey about mobile local search and describes the user's desired target in terms of distance and time. Lv [7] considers several user-related signals in ranking for mobile local search. Dragut [4] merges similar search results in a local area with a consideration of user's ratings. Bernerich [2] exploits direction requests, browsing logs and mobile search logs to refine search ranking. Meanwhile, Zhou [11] designs new tree structures for location-based web search. Ahlers [1] introduces the \entity retrieval system" for Yellow Pages. Several papers note the geographic factors in search problems: Gan [5] investigates the properties of geo-queries and develops a new taxonomy for such queries. Lymberopoulos [8] predicts click behaviors with high-level location features, such as states and zip codes.

The methods in the above papers have some limitations. Query log-based methods cannot perform well when a new user executes a query for a new local store. For the local search problem (in this paper local search does not refer to the same-named optimization strategy in artificial intelligence), when the history records are not complete, improvements are derived from incorporating open data into the search model. Moreover, the user-oriented analysis should also take advantage of more detailed geographic and practical features beyond simple distance. The geographic data that we request from open databases are details about local business units, such as the name of a store, the street address and the locations (accurate longitude and latitude). The more features we get, the more may be able to improve local search. Other useful sources of information can also be included, including competitors, size of target stores and

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890482 .

business categories, since these can a ect a user's decision when choosing among several shops.

In this paper, we start by investigating the relationship between geographic features from open datasets and user's expressed interests from applications (e.g. $Yelp$) concerning local business units (Section 2). We analyze a $Yelp$ dataset [10] by searching locations from the open database to show relationships between geographic features and user preferences (such as check-in and review, here check-in refers to the claim of visiting a nearby business unit with short comments or rating on a mobile application) among all business units. Using conservative choices for the geographic parameters, in Section 3 we build a model to infer a user's preference for business units to serve as the basis for creating a ranking list. In Section 4, we describe a further improved framework of local search that incorporates knowledge of user preference. Finally we conclude with a summary of the results and indications for further research.

2. GEO FEATURES VS. PREFERENCES

In Section 2, we introduce an open dataset, available on $Yelp$ [10], and several geographic features that we wish to relate to user's choice. Our data analysis reveals several patterns linked to user preference.

***Yelp* dataset.** The $Yelp$ dataset contains $1.6M$ reviews by $366,000$ users for $61,000$ business units. After applying a lter for the set of cities, which is that a city must have at least 10 business units of any kind listed, we are left with 96 cities in North America and Europe and a total of $60,503$ business units. For each business unit, the database provides the name, address, and its accurate location (latitude and longitude). Included are also reviews and and linked ratings by customers. Though review content is also available, we do not dig into the natural language processing in this paper (it presents a further consideration and potential opportunity). We can use the business unit location as the input for a secondary query to get more information from a Geocoder [6] database, then generate geographic features of business units, such as neighboring business units density and others in the following paragraph.

Features of interest. We explore the interactions of ve features with the $Yelp$ user ratings and number of reviews (#reviews) per business unit. They are all available from an open geographic database [6]. And user's query and other actions do not change the value of those ve features, because they are geographic features rather than personal attributes. Several papers [3, 7, 9] demonstrate the importance of incorporating a user's current location into mobile local search. While we account for that as well, we additionally consider the following information in our preference estimation model.

1. **Significance of ratings and #reviews in an area.** For all business units in a city, we compute the average of all ratings and #reviews. The result will show whether local area matters in terms of user's opinion. More statistical methods and criteria should be applied here in future research, such as weighted average and analysis of distribution about ratings and #reviews. Besides, the size of city has a wide range so perhaps the big cities should be divided into smaller districts.

2. **Average distance between a given business unit and the other business units in all types within the same city.** This is e ectively a measure of business unit neighbor centrality. When a store is far from others, it looks like an outlier away from the central business area of a city. More advanced metrics and methods in the detection of outlier can be applied.

3. **Density of neighboring business units.** For a store, we count the number of neighbors of all types within a certain radius. We do not lter with the same type of business when counting the number neighbors because di erent types of business might attract customers for each other. In general, high density may be linked to the existence of shopping centers or prosperous business areas.

4. **Number of roads within a certain distance to a business unit.** This re ects the availability of local transportation. We query the address [6] of several points nearby the business unit. The points are located on the circumference (without the limit of address query quota, we can set discrete values of the radius so as to get the address names of more points within a certain distance from the center point) of a circle at equal angle intervals, whose center is the business unit. Then we analyze the returned addresses to see the diversity (number of di erent roads by comparing road names) of roads nearby the target. Considering the example situation in Figure 1, There we see stores A and B along the road $R1$. We draw two circles of a manually given radius around the two stores. For the points with equal angle interval on the two circles, the database query with their longitudes and latitudes based on the longitude and latitude of the centers. The results of the query would include the full names of roads (e.g. $R2$ and $R3$) nearby.

5. **Location of the business unit in a street or road.** \Location" is an attribute such as \middle" or \end". The attribute here is a relative concept. For the pur-

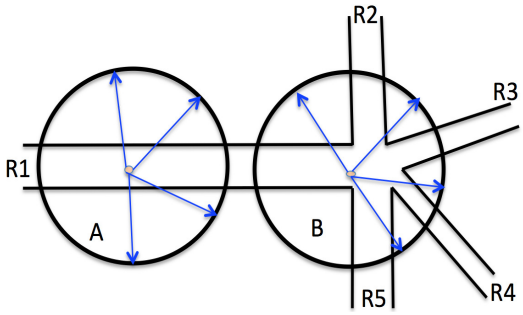

Figure 1: "Surrounding roads information". We look to incorporate the road address around certain center points of circles area in a road. We use the longitude and latitude of center point of a circle to get the output of address names of points on the circumference. In Figure 1, given a radius of interest, we can determine circles of interest, A and B, with different centers along the road $R1$. We can compute the location of points on the circumference and query the corresponding address names from geographic database. If we set the radius with different values, we can get the road names of more points near the center business unit.

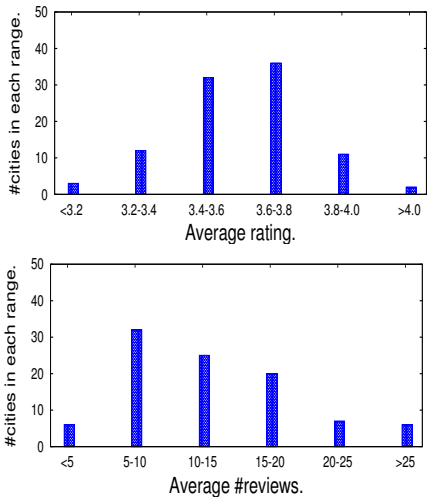

Figure 2: Histograms of ratings data per city. Top: average rating. Bottom: average #reviews.

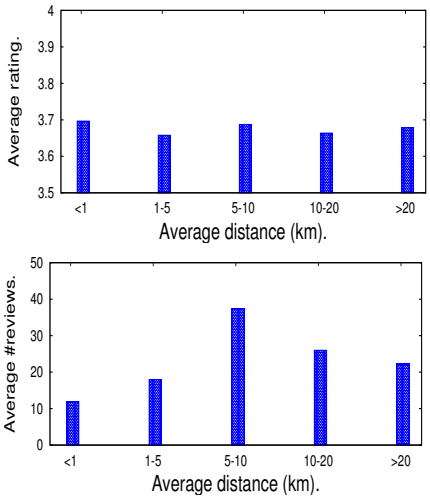

Figure 3: Effect of average distances. Top: average rating. Bottom: average #reviews.

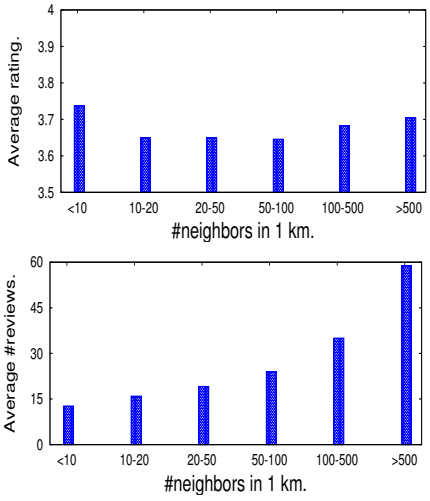

Figure 4: Effect of neighbors density. Top: Average rating. Bottom: average #reviews.

poses of modeling we assume the business unit is located along a straight road, rather than some types of roads (e.g. highway and roundabout) without too many business units. Suppose a person is walking through blocks to find a store as the target of shopping, the in-street location might affect the possibility of seeing the store. We still query several points around the store, and count how many points on the circumference are on the same street with the center business unit. Considering again Figure 1, we introduce two circles. For the locations around a center which is in the middle of road $R1$, almost all points on the circumference are along the road $R1$, where the center of circle A is located in. As a comparison, the center of circle B is close to the right end of road $R1$, so the points on the circumference of circle B are located in different roads, thus less points are in the road $R1$. By this difference, we can approximately judge the location of a business unit in a road.

Patterns. The *Yelp* dataset offers the name, location, reviews and other information about a business unit. With the accuracy location (longitude and latitude) of a point as the input, open geographic database such as Geocoder [6] will return the full address of the point. Combined with the two data resources, we investigate the previous five features and get the following histograms.

Figure 2 shows the histograms of the average ratings and the average number of reviews for all business units, per city. The number of business units in a city varies from 11 to 13600. Though the majority of the average per city ratings are in range [3.4, 3.8], the ratings have a obvious difference among cities, since the range of rating is an integer from 0 to 5 and users rarely give a rating lower than 3.2 (from the most left bin in the top histogram). We also find the uneven distribution of #reviews in the bottom histogram. Except Figure 2, other value of bin size might be acceptable. Here we pick up the bin size as with the knowledge of the mode and range. To sum up, Figure 2 tends to support the assumption that business unit location matters in terms of user's ratings and reviews on business units, so we should consider location in city scale (and perhaps with the other smaller scale geographic features) for user preference modeling.

Figure 3 shows the distributions of average rating and #reviews, when the average distances from one business unit to the others changes. Average distance is a form of geographic centrality. The average rating does not have a clear trend with the uniform distribution, but the uneven distribution of #reviews seems to match a normal distribution or else. We find that if a business unit has an average distance of 5-10 km to others, it tends to receives the most reviews.

By Figure 4, we explore the relationship between #neighbors (number of neighbors) and reviews in terms of rating and #reviews. When a business unit has the least or the most neighboring business units within 1 km radius as the top subfigure shows, the ratings seem to be better than others. Though we are not sure why the low or high neighbor density might relate to a higher rating, the possible relationship suggests us consider the number of neighboring business units when predicting the user's reviews of a business unit. In ad-

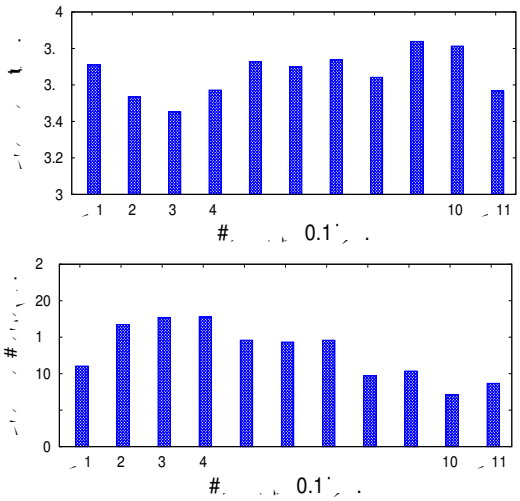

Figure 5: Effect of #roads nearby. Top: average rating. Bottom: average #reviews.

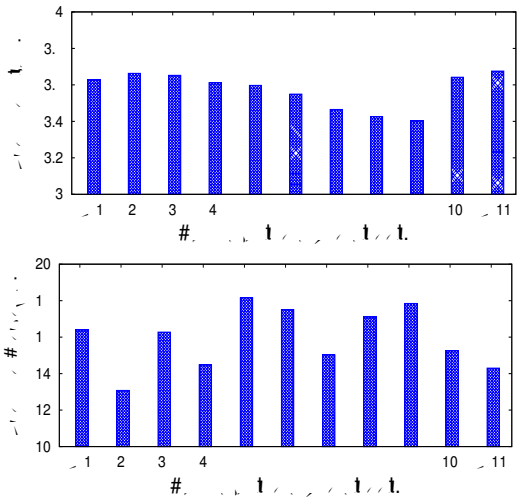

Figure 6: Effect of #roads in the same street. Top: average rating. Bottom: average #reviews.

dition, the bottom sub gure illustrates that more neighbors (high neighbor density) will bring more reviews. This corresponds with intuition: more people are attracted by more business units, and post more reviews there.

Figure 5 shows the relationship between #roads around a business unit and our statistics. We detect 12 points around a business unit with the same radius and equal interval (30 degree) between angles. The radius is 0.1 km and angles are (0, 30, 60 ... 330) degrees in anticlockwise direction starting from the x axis in a 2-D virtual plane. A comparison of the left half (0-5) and the right half (6-11) in the top sub gure nds a a general trend of more roads locate around a business unit correspond with higher average rating. In addition, when #roads becomes larger, #reviews decreases a little for some reason. One assumption is that a person might choose another way to go and miss some business units at road intersections, so a business unit with more surrounding roads might receive less reviews. To sum up, the #roads nearby seems to vary with the review statistics.

Figure 6 illustrates the relationship between location of a business unit in its street and our statistics. Limited by the query request quota of Geocoder [6], we randomly sample about three thousand business units from 27 cities and display the distributions in Figure 5 and Figure 6. Each business unit is treated as the center of a circle as Figure 1 shows, and we need the query of 12 points around the center to analyze the road information nearby, so it becomes a bottleneck given the limited times of address query. We nd that business unit location in the middle or at the ends of a street corresponds weakly with a higher average rating. There is not much of a recognizable trend in #reviews.

The above gures suggest that geographic features do have some relationships with user ratings and #reviews. This in turn suggests that open geographic data can make an important contribution to local search.

3. USER PREFERENCE MODEL

Since geographic factors a ect rating values and the number of check-in/review (check-in refers to special comments with short words about a nearby business unit on some mobile applications) in Section 2, we propose in Section 3 a model to quantify user preference among surrounding (relative to the user's location at time of query) business units. A better estimate of user preference, will produce a ranking list closer to a user's needs and thus enhance user experience. We do not treat those geographic features in Section 2 as the entire feature list about user preference, since more environmental or personal factors might also change user's mind. As a result, we analyze the possible factors about user's preference and describe it with the following Equation 1.

Let f_i denote the preference for business unit i and assume it has the form in Equation 1. It requires two kinds of known locations: business unit and user location. It is a simple case since it does not include other neighboring business units as competitors. A larger value of f_i means the user is more likely to choose the business unit i in mind.

$$f_i = \frac{l^{\alpha_{l;i}} * t^{\alpha_{t;i}} * s^{\alpha_{s;i}} * g^{\alpha_{g;i}}}{c^{\beta_{c;i}}} \quad (1)$$

The values α_l, α_t, α_s, α_g and β_c are positive parameters re ecting user sensitivity, which is similar to weights in a linear function. Each item in the numerator should have its own exponent (α). The exponents denote the weights of several parameters. Users might be able to input their initial values, and the search algorithm can adapt the parameters with response of query results. We also include several variables/functions in the numerator that might have positive correlation with user's preference.

- l captures the city's environmental bias factors. Different city/towns have their own standard of rating and review style as Figure 2 and the previous work [8] show.

- t represents the text matching result. If the semantics of query inputs matches the type of business unit, t will be a larger value. For example, if a user would like to have a meal and input \Where to eat", then restaurants will have a higher value of all business units. Some open semantic data with latent vectors (such as Word2Vec[1]) might improve the matching performance.

[1]https://code.google.com/archive/p/word2vec/

- s means the score of inner attributes for a business unit, including but not limited to the size of store, the cleanliness, the opening hour, the quality of service. Other kinds of open data, such as customer's review and introduction on Yellow Pages, can also be added to determine the value of s.

- g represents the score based on geographic factors. The density of neighbors, the feasibility of transportation (number of nearby roads) and the location of store in a road (middle or end) are possible attributes. The Figures in Section 2 illustrate a possible relationship between geographic factors and user feedback, so we add this item into the model.

- c represents the cost of traveling from the current location to the business unit, which has (generally) a negative correlation (when cost for a business unit increases, the user will grade the business unit with a lower preference score) with user's preference so it is in the denominator. Limits on nancial budget and time can a ect user's choice. So it contains at least two parts, the time cost and money cost, depending on the way of tra c from user's current location to the target business unit. The user's current location is a key factor in computation of c, since it determines the distance to the business unit. Several papers [2, 7, 9] point out the importance of distance in providing relevant recommendations.

Equation 1 is not the only possible form of a relationship between external factors and user preference. It might also take the form of a weighted sum, but this fractional form better re ects which factors have positive or negative correlation with the preference. In addition, there are several functions behind t, s, g and c. Each function deals with factors of an aspect (e.g., geographic factors) and set corresponding values to describe user's preference.

In any real case, if the user has a general goal (e.g., a shopping mall) rather than a clear query with a name (e.g., Walmart), the user might wander around a local business area. Equation 2 considers this case and encodes the e ect of neighbors. The neighbors of a store might compete with the store, or they might sell complementary goods.

$$F_i = p * f_i + \sum_{k \in N_i} (1-p) * u(t_k, t_i) * \frac{f_k}{|N_k|} \qquad (2)$$

Here are the de nitions of variables:

- F_i is total preference value with neighbor's contribution, which might work in the ranking part of a search engine.

- f_i and f_k result from Equation 1. Store k is a neighbor of store i.

- N_i represents the set of store i's neighbors. $|N_k|$ means the size of the set. In traveling, a user might be attracted by other stores nearby, so the interest of a particular store can be a ected either positively or negatively by a neighbor.

- p is the probability of staying focused on the original target. It is a personal attribute about purchase behavior.

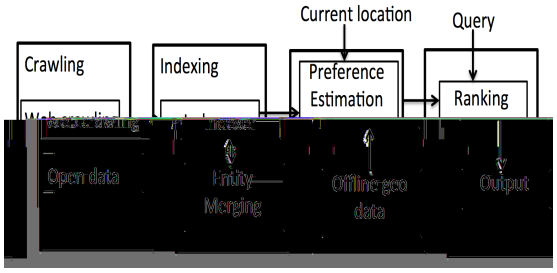

Figure 7: Revised structure of local search. It shows the data flow of the revised local search. After the crawling and indexing, one additional layer of preference estimation is added. It requires user's current location and query from open geographic data. The additional layer can give a estimation of user's preference. Finally, an input query will trigger the module of ranking combined with the estimation of preference.

- $u(t_k, t_i)$ describes the relationship between two business units. They may be cooperators or competitors.

Here the preference value F_i depends on two parts, the simple point-to-point interests and the neighbors' e ects. To de ne the set of neighbors N_i, geographic open data must o er the locations of surrounding business units and the road information. To get $u(t_k, t_i)$, a comparison of keywords is necessary. For the personal parameters (α, β and p), the model should learn them with user's choices about query results. Over time, the search algorithm might provide a customization according to user history. After the collection of query logs with user's location track, we can evaluate the model. The model puts more weights on features from open data, so even when the documents (set of words used in traditional information retrieval model) of business units are not complete, the revised model with geographic features might generate a list of preference values for a better ranking result.

4. IMPROVED LOCAL SEARCH

Section 2 shows the potential utility of including geographic factors in local search, especially when the document data of business units are not complete. In Section 3 we give a high level description of a user preference model incorporating geographic features. Here we describe the possible change in the structure of local search to comply with the model.

Difference from traditional search method. The obvious di erence is the additional geo-analysis component with external o ine geographic open data. Suppose the task is to estimate user preference with user current location. The possible features include the city's attributes, the distances between a user and all business units, feasibility of transportation to all business units, competition between a business unit and its neighbors. Of course, the basic schema we propose is not limited by the above features discussed in Section 3. Other business and travel related geographic features can be added, too.

The use of open data. Though we mainly focus on the new incorporation of geographic open data, other types of open data can also contribute to a better (in terms of user

experience and performance) ranking result in the structure. The classical structure of searching includes three submodules, which are *crawler, indexer,* and *query*. *Crawler* downloads webpages and *indexer* build indices for those words in webpages, then *query* responds to user input by returning the most related pages.

With this basic model in hand, here is a possible usage scenario of open data in searching. In the first step of crawling, the crawling from both online and offline open data (such as geographic databases, Yellow Pages brochure, social network reviews, etc.) should be performed. Since the local area often has a limited range of business unit candidates within a certain radius, it is possible to collect information from multiple aspects and resources, when the single resource cannot generate a large enough document set about business units. The second step is indexing. This necessitates execution of the challenging task of merging multiple descriptions of the same entity, acquired from diverse information resources. The third step is the modified design that incorporates local search. Since GPS on mobile devices enables a real-time location record, a user's current location can trigger the preference estimation model given in Section 3. The model will use the information of surrounding business units acquired from geographic open data. Semantic open data can also work in the matching of query and business category. The model will then produce a list of nearby business units with their preference values for a user. The last step is the response to user's query. Traditional ranking results relate the semantic similarity between the input string and the candidate document with the index. Here we have another preference list based on the additional estimation model in Section 3. A suitable mix of the two methods should improve the searching performance.

Advantages. The use of open data in preference estimation could solve the problem of insufficient web-available information about local business units. Besides, the added step of geographic analysis can also serve for a local recommendation system before the user's query. Meanwhile, the structure leaves room for incorporating other types of open data.

5. CONCLUSION

In this paper, we analyze the patterns of relationships between geographic features derived from open geographic data and user preference, and describe a preference model that incorporates several detailed geographic features. We discuss the potential improvement about the structure of local search for better preference estimation. The initial analysis tend to support that open data, and especially geographic open data, can be a powerful factor to estimate user preference, and local search incorporating a parser of geographic features might overcome a lack of descriptive words about business units.

Future possible directions of work include: (1) Collecting real query logs that track movement and evaluate the preference model and the revised local search. (2) Finding and determining more helpful geographic features. (3) Mining the patterns encoding the relationship geographic features in Section 2 and the preferences. At the same time, working on different scales of local data in terms of the size of a city and the radius of address query around a business unit. (4) Finalizing several sub-parts in local search algorithm, including the merge of information about the same entity, the cooperation of preference estimation model and traditional ranking method.

6. REFERENCES

[1] Ahlers, D. Business entity retrieval and data provision for yellow pages by local search. In *IRPS Workshop* (2013).

[2] Berberich, K., König, A. C., Lymberopoulos, D., and Zhao, P. Improving local search ranking through external logs. In *Proc. of the SIGIR* (2011), ACM, pp. 785–794.

[3] Church, K., and Smyth, B. Who, what, where & when: a new approach to mobile search. In *Proceedings of the 13th international conference on Intelligent user interfaces* (2008), ACM, pp. 309–312.

[4] Dragut, E. C., Dasgupta, B., Beirne, B. P., Neyestani, A., Atassi, B., Yu, C., and Meng, W. Merging query results from local search engines for georeferenced objects. *ACM Transactions on the Web (TWEB) 8*, 4 (2014), 20.

[5] Gan, Q., Attenberg, J., Markowetz, A., and Suel, T. Analysis of geographic queries in a search engine log. In *LocWeb Workshop* (2008), ACM, pp. 49–56.

[6] Geopy-1.11.0. Python geocoding toolbox. `https://pypi.python.org/pypi/geopy/`. Accessed Dec.15, 2015.

[7] Lv, Y., Lymberopoulos, D., and Wu, Q. An exploration of ranking heuristics in mobile local search. In *Proc. of the SIGIR* (2012), ACM, pp. 295–304.

[8] Lymberopoulos, D., Zhao, P., Konig, C., Berberich, K., and Liu, J. Location-aware click prediction in mobile local search. In *Proc. of the CIKM* (2011), ACM, pp. 413–422.

[9] Teevan, J., Karlson, A., Amini, S., Brush, A., and Krumm, J. Understanding the importance of location, time, and people in mobile local search behavior. In *Proc. of the MobileHCI* (2011), ACM, pp. 77–80.

[10] Yelp. Yelp data challenge dataset. `https://www.yelp.com/dataset_challenge/dataset/`. Accessed Dec.2, 2015.

[11] Zhou, Y., Xie, X., Wang, C., Gong, Y., and Ma, W.-Y. Hybrid index structures for location-based web search. In *Proc. of the CIKM* (2005), ACM, pp. 155–162.