Semantic Location in Email Query Suggestion

John Foley[*]
University of
Massachusetts Amherst
jfoley@cs.umass.edu

Mingyang Zhang
Google, Inc.
mingyang@google.com

Michael Bendersky
Google, Inc.
bemike@google.com

Marc Najork
Google, Inc.
najork@google.com

ABSTRACT

Mobile devices are pervasive, which means that users have access to web content and their personal documents at all locations, not just their home or office. Existing work has studied how locations can influence information needs, focusing on web queries. We explore whether or not location information can be helpful to users who are searching their own personal documents.

We wish to study whether a users' location can predict their queries over their own personal data, so we focus on the task of query suggestion. While we find that using location directly can be helpful, it does not generalize well to novel locations. To improve this situation, we explore using semantic location: that is, rather than memorizing location-query associations, we generalize our location information to names of the closest point of interest. By using short, semantic descriptions of locations, we find that we can more robustly improve query completion and observe that users are already using locations to extend their own queries in this domain.

We present a simple but effective model that can use location to predict queries for a user even before they type anything into a search box, and which learns effectively even when not all queries have location information.

ACM Reference Format:
John Foley, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2018. Semantic Location in Email Query Suggestion. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210116

1 INTRODUCTION

Users at particular locations typically have information needs based on their immediate geographic context. For example, a user at a restaurant engaging with a search system is likely to be searching for that restaurant's menu. Recent works have studied this kind of contextual information, even going so far as to consider zero-query ranking [4, 15]. These works focus on the web, where query log mining can provide an understanding of trends and global behavior. In the personal search domain, the challenge becomes more difficult: one cannot simply learn location-wise trending behavior due to the privacy constraints of personal search.

Using a set of anonymized email search logs with location information provided by Google, we explore whether location information can be leveraged for query auto-completion. Since we are unable to submit new queries, we explore a simulated task on this raw log data: user-independent query suggestion. We ask whether we can predict the queries a user is going to issue based upon 1. any characters they have already entered (possibly none), and 2. the location information.

The contributions of this paper are as follows:

- We validate that location information is valuable for personal search by demonstrating the ability to predict queries using location information.
- We validate that semantic location information is valuable, using a non-parametric Click-Context model that allows us to learn location information from queries and documents with and without location associations.
- We observe that users often manually expand their personal search queries with their location context, indicating that it is a strong signal for relevance.

We demonstrate our first two contributions by focusing on a query prediction or suggestion task: using minimal or no query information, we try to use the location information in our log to predict the queries. In doing so, we explore a handful of models and look at their ability to generalize and perform on this task.

We find that hashes of GPS location provide evidence that location is helpful, but the coverage of this technique is not ideal: the majority of unique locations in our test set remain unseen even though our training set is larger. With much more data we would expect this problem to dissipate, but we look to a better opportunity: semantic location information. We annotate our query logs with geographic entity look-up: that is, for every latitude/longitude point, we perform a search of the nearest point of interest item using the Google Places Web API, and include the title of this point in our extended logs. These titles provide the basis for our generalization.

Finally, we analyze our performance on query completion and find some surprising behavior in this task. Our core observation is that users manually expand their queries with location, and hypothesize that will be difficult to beat this "human-expansion" baseline if we were to look at improving search satisfaction directly (until users realize they no longer need to manually include location names).

2 RELATED WORK

At the core of our work is the hypothesis that location will be helpful for personal search tasks, much like it has been in other domains and other tasks, e.g., for web search [5]. We will discuss the history of semantic location, general uses of locations in queries, query-completion, and personal search methods.

[*] Work done while at Google.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210116

2.1 Semantic Locations

Liu et al. were the first to attempt to automatically assign semantic locations (e.g., home, work, grocery store, etc.) to physical locations as output by GPS units [18]. Since then, it has become common for cell phones and other smart devices to have GPS chips or to use Wifi to infer the location of a user [12]. We leverage the Google Places API[1] to assign names to our raw location coordinates, and use the names and category markers from their API as our semantic location.

2.2 Queries and Locations

Exploring location and geographical relevance began as a need to understand which queries and documents were relevant globally or locally to users [11, 14]. Sanderson and Kohler studied a sample of 2,500 queries from the 2001 Excite query log and found that 18.6% of their queries had a geographic term [21]. Location has also been studied more recently, focusing on mobile search. Benetka et al. study query and information needs before, during, and after activities as a way to motivate location-aware search systems [4]. Qi et al. infer locations on top of the AOL query log by using a geocoding service along with retrieved URLs and they generate a more recent dataset by using location-stamped tweets from the NYC area [20]. They show that location is an important dimension to consider for query suggestion in general, though they study web-search.

2.3 Query Completion & Query Suggestion

Although our focus in this work is on the use of location for a query suggestion, we do not delve into particularly complex query suggestion or autocompletion models, although they exist in related work. We focus on this task as an example of how location may be leveraged in a probabilistic suggestion model. For a deeper study of query auto-completion, we direct the interested reader to a recent survey by Cai & de Rijke [6]. A similar approach to our own models appears in methods for using user history [2], but we focus on shorter prefixes.

Traditionally, query suggestion or query completion is done by mining a query log for suggestions [6, 7, 19, 22, 23]. A lot of work in query completion looks at spelling correction and user reformulation over time to learn to complete queries [13, 17]. Because we do not have session data or have typing data, existing approaches are less relevant to our approach.

2.4 Email & Personal Search Methods

Recent work in the email and personal search domain addresses learning from attributes rather than direct data in order to better generalize [3], leveraging user demographics [8]. A closely-related work generates suggestions using query logs from similar users and settles on a combination of many approaches [16], and recently location has been successfully incorporated into email ranking systems [24].

3 DATASET, MEASURES & NOTATION

All of our findings, models, and experiments are built upon our analysis of email query logs, so we will discuss our log in advance. Due to the privacy constraints of personal search – unlike many other query logs – we have no information that can discriminate either sessions or users.

3.1 Training Splits & Parameter Tuning

Overall, we use 14 million queries for training. Most of these queries were issued through a web client, and therefore do not have any exact location indicators. However, we also include a sample of 300 thousand queries that were issued through a mobile app that has access to user location, and therefore these queries have location information associated with them. We use another sample of 150 thousand queries with location data for testing & evaluation purposes. The train/test query splits described are based on time, and all queries used had clicks associated with them, but only strictly-anonymized queries are available.

3.2 Relative Evaluation Measures

Although our results reflect only our experiments with location and do not reflect any production systems, we present relative evaluation measures, in order to prevent inaccurate speculation about production system behavior. For instance, when we talk about mean-reciprocal rank or (MRR), we will present results between a treatment t and a baseline b as $\Delta\text{MRR}(t, b) = \frac{\text{MRR}(t) - \text{MRR}(b)}{\text{MRR}(b)}$. The baseline will always be identifiable as the method achieving "1.00x" performance, and treatments will vary accordingly to their relative performance.

3.3 Notation and Contents

We notate our query log as a list of tuples: $\mathbb{L}$. Each entry in this list is a 4-tuple: (Q, h, L, S) where Q is a set of query n-grams, h is a (possibly null) GPS location hash, L is a possibly-empty set of location n-grams, and S is a possibly empty set of subject n-grams from the email the user clicked.

We will use the function $\mathbb{1}(\phi)$ as shorthand for the "truth" function. Since each of (Q, L, S) is a set of n-grams, we need notation to express indexing into these sets to express some computations. We draw the reader's attention to the fact that our log truly contains *sets*, and not bags: if a user submitted a query "hello hello hello", we would have a set of the unique n-grams, e.g., $Q = \{$[hello], [hello hello], [hello hello hello]$\}$ from the original query.

Since we have sets and not bags, we can use our truth function to skip a summation when we represent counts by using containment. For example, counting the occurrences of an n-gram q in a particular set Q can be done as a summation over the elements of Q or directly: $\sum_{q_i \in Q}[\mathbb{1}(q = q_i)] \equiv \mathbb{1}(q \in Q)$ and this works because we know the count of $q \in Q$ will be exactly 0 or exactly 1. We will use an explicit sum when many query n-grams may match a condition, e.g., exactly 3 n-grams match our prefix: "he" in our "hello" example.

4 QUERY SUGGESTION MODELS

In this section, we have one research question: can location predict the personal search queries a user will issue?

[1] https://developers.google.com/places/

We take a hash function applied to the raw location information at a reasonable granularity, and consider this as the h in our log. The question, more formally, is whether or not a hash h can be used to predict a query Q. We note that there are a number of schemes for converting GPS location information into tokens or hashes, but we only wish to answer our basic research question, so we used a string-based hash function.

Although we see some improvement based upon simple hashes of locations (and may see better results with a hash designed for GPS locations), we introduce a model later based on semantic location that clearly improves on any technique that only uses GPS coordinates instead of semantics for particular locations.

4.1 Query Prefix Model (QPM)

Given a query term q and a prefix p, we can calculate a baseline memorization probability $P(q|p)$ for every q in our query log. The efficacy of this baseline will naturally depend upon p. For instance, a user searching for the word q ="coupon" will get better results after typing more characters (p ="cou") than with only one (p ="c") or none: $p = \epsilon$. In fact, when no prefix is available, this popularity baseline becomes exactly that: merely a ranked list of all query n-grams issued, by popularity.

We estimate the probability of a query term completion as the number of times a query term q occurs divided by possible completions: any occurrence of any query term that starts with p.

$$P(q|p) = \frac{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(q \in Q)}{\sum_{(Q,h,L,S)\in\mathbb{L}} \sum_{q_i \in Q} \mathbb{1}(\text{StartsWith}(q_i, p))}$$

This is our baseline probabilistic model; it does not use location.

4.2 Direct Location Model (DLM)

To incorporate location directly, we use our hashed latitude/longitude representation h. We compute the probability of all candidate completions: if we sample randomly from the queries issued at location h', what is the probability it is our candidate completion q?

$$P(q|h') = \frac{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(q \in Q \wedge h' = h)}{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(h' = h)}$$

Note that since these are probabilities, we can make an independence assumption, and combine them when we have both a useful prefix $p \neq \epsilon$ and location information.

$$P(q|h,p) = P(q|p) \cdot P(q|h)$$

We explored calculating the joint probability of observing both a query starting with a prefix and a hash. Although this technique offers more exact calculation, it is more sparse, and therefore more difficult to estimate accurately. We present only our the results from our independent models for space reasons.

4.3 Semantic Location Representation

In order to improve our ability to generalize and to analyze this data, we used a reverse-geocoding API to assign location names to each location point in our logs. This provides us with a new textual field, $L = \{l_0, l_1 \ldots l_{|L|}\}$ with which we can model probabilities. These location names were tokenized in the same manner as queries and subjects, into n-grams. The n-grams in this field pose a noise and detail-challenge that the GPS locations we used previously lacked (e.g., some are extremely specific, "University of Oz, 999 Yellow Brick Rd" and others are broad, e.g., "Emerald City Airport", or only the name of a store chain), but they allow us, generally, to describe semantic similarity and partial matches.

4.4 Textual Location Model (TLM)

Our modeling of location n-grams is probabilistic: of all the query n-grams that co-occur with this location term, what is the chance of a particular one?

$$P(q|l,p) = \frac{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(q \in Q \wedge l \in L)}{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(l \in L)} \cdot P(q|p)$$

At this point, we borrow the term independence assumption from probabilistic retrieval models [10] in order to calculate an overall probability for query completions given an observed location string.

4.5 Click-Context Location Model (CCLM)

Under the hypothesis that locations are relevant to a query, we developed a new model: what if we considered more information from our logs as possible evidence for locations, i.e., a query containing the term "restaurant" (like "restaurant menu") is probably more informative of what users are likely intend to search in an restaurant, like "menu".

Recall that our query log has four separate feature spaces: queries, GPS locations, semantic location, and subject n-grams (Q, h, L, S). Because any or all of Q, L, and S may be empty sets for any given log entry, it becomes difficult to learn relationships and meaning from any of the spaces independently.

Therefore, we propose that we consider this "click-context" information jointly, rather than independently. We wish to better understand a relationship between query attributes, so putting both document attributes S and query attributes Q, L into the same space serves our objectives.

$$P_{CC}(q|l,p) = \frac{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(q \in Q \wedge l \in (L \cup Q \cup S))}{\sum_{(Q,h,L,S)\in\mathbb{L}} \mathbb{1}(l \in (L \cup Q \cup S))} \cdot P(q|p)$$

5 RESULTS

Full results, with both precision (MRR, P@1) and recall-oriented (mAP) measures are presented in Table 1.

While our results demonstrate that location is effective, GPS location (DLM) features have an obvious weakness: they lack semantic similarity. This is obviously implied by a basic example: we expect that users in airports are likely to issue similar queries. Suggesting popular airlines is likely to be a very strong baseline for any queries submitted in *any airport*, even though by their nature there will be great dissimilarities in their latitude/longitude coordinates, and they will definitely be assigned a different location hash. This over-specificity is solved in our TLM and CCLM models.

The gains from using location are impressive, especially in the zero-query scenario ($p = \epsilon$). In addition, our CCLM models P_{CC} provide a nice boost over directly memorizing location terms (TLM) and what query terms they predict. These improvements become less sizable, but remain significant until $|p| = 3$, when the user has mostly disambiguated what they are typing, and even then, using location is significantly better than not having it available,

Table 1: Query Prediction Results: This table presents the effectiveness of models relative to the popularity baseline. Approaches using location, and especially semantic location show strong gains in both recall-oriented measures like mean Average Precision (mAP), and precision oriented measures (MRR, P@1). At $\lvert p \rvert = 4$ (not included) only improvements in mAP remain significant at weaker levels.

			$p = \epsilon; \lvert p \rvert = 0$			$\lvert p \rvert = 1$			$\lvert p \rvert = 2$			$\lvert p \rvert = 3$		
			MRR	mAP	P@1	MRR	mAP	P@1	MRR	mAP	P@1	MRR	mAP	P@1
QPM	§4.1	$P(q \mid p)$	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
DLM	§4.2	$P(q \mid h, p)$	1.88x†	3.47x†	2.64x†	1.09x†	1.09x†	1.15x†	1.03x†	1.03x†	1.04x†	1.01x†	1.01x†	1.01x†
TLM	§4.4	$P(q \mid L, p)$	4.08x†	7.71x†	5.21x†	1.19x†	1.20x†	1.31x†	1.07x†	1.07x†	1.10x†	1.02x†	1.02x†	1.03x†
CCLM	§4.5	$P_{CC}(q \mid L, p)$	4.51x†	8.78x†	5.91x†	1.22x†	1.22x†	1.35x†	1.08x†	1.08x†	1.11x†	1.02x	1.02x	1.03x

† Represents statistical significance with $p < 0.0001$ with a pairwise randomization test over the entry in the previous row.

it is merely that the baseline has risen sufficiently that the more sophisticated uses of location provide fewer advantages.

5.1 Query-Log Analysis

We find that a large fraction of queries include some term that is also part of the name or title of their location (L in our query log). These types of queries occur independent of the frequency of the particular query or the popularity of the location involved.

What are users doing? They are including the name of their current location. In typical email search, like real-time search, one of the key features used to present results or to rank is recency [1, 9]. If we consider the example of a user submitting the generic query "coupon", a typical system would probably retrieve poor results, given the frequency of promotional email – the most recent "coupon" you've received is not necessarily correlated with your immediate desire for a coupon. It appears that users are aware of this phenomenon and are compensating by including the brand name or product name of their request to aid in disambiguation.

At a hypothetical supermarket, "Food & Stuff" a user is likely to submit queries relevant to that location, i.e., "food and stuff coupon". We observed that the ideal queries which would showcase location as a useful disambiguator, i.e., "coupon", "rewards", "flight" were almost non-existent in our actual log, presumably because users know these queries are unlikely to be successful in existing systems.

6 CONCLUSION

In this work, we find that location can be helpful for query completion, but it is more helpful when treated semantically, and merged with other textual features as in our Click-Context models. We present observations of learned user behavior that show most users have learned to manually expand their queries with location keywords. Our strong results with no characters available suggest future directions in personal search: such as incorporating these features directly in ranking or even in zero-query scenarios, when we can pre-emptively present relevant documents to a user.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Qingyao Ai, Susan T. Dumais, Nick Craswell, and Dan Liebling. 2017. Characterizing Email Search Using Large-scale Behavioral Logs and Surveys *(WWW)*. 1511–1520.

[2] Ziv Bar-Yossef and Naama Kraus. 2011. Context-sensitive Query Auto-completion. In *WWW*. 107–116.

[3] Michael Bendersky, Xuanhui Wang, Donald Metzler, and Marc Najork. 2017. Learning from User Interactions in Personal Search via Attribute Parameterization *(WSDM)*. 791–799.

[4] Krisztian Benetka, Jan R .and Balog and Kjetil Nørvåg. 2017. Anticipating Information Needs Based on Check-in Activity *(WSDM)*. 41–50.

[5] Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz. 2011. Inferring and Using Location Metadata to Personalize Web Search *(SIGIR)*. 135–144.

[6] Fei Cai and Maarten de Rijke. 2016. A survey of Query Auto Completion in Information Retrieval. *Foundations and Trends in Information Retrieval* 10, 4 (2016), 273–363.

[7] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data *(KDD)*. 875–883.

[8] David Carmel, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, and Ariel Raviv. 2017. The demographics of mail search and their application to query suggestion *(WWW)*. 1541–1549.

[9] David Carmel, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, and Ariel Raviv. 2017. Promoting Relevant Results in Time-Ranked Mail Search *(WWW)*. 1551–1559.

[10] W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice.* Pearson Education.

[11] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. 2000. Computing Geographical Scopes of Web Resources *(VLDB)*. 545–556.

[12] Olivier Dousse, Julien Eberle, and Matthias Mertens. 2012. Place Learning via Direct WiFi Fingerprint Clustering. In *Proc. of the 13th International Conference on Mobile Data Management (MDM)*. 282–287.

[13] Huizhong Duan and Bo-June (Paul) Hsu. 2011. Online Spelling Correction for Query Completion *(WWW)*. 117–126.

[14] Luis Gravano, Vasileios Hatzivassiloglou, and Richard Lichtenstein. 2003. Categorizing web queries according to geographical locality *(CIKM)*. 325–333.

[15] Liangjie Hong, Yue Shi, and Suju Rajan. 2016. Learning Optimal Card Ranking from Query Reformulation. arXiv:1606.06816

[16] Michal Horovitz, Liane Lewin-Eytan, Alex Libov, Yoelle Maarek, and Ariel Raviv. 2017. Mailbox-Based vs. Log-Based Query Completion for Mail Search *(SIGIR)*. 937–940.

[17] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning User Reformulation Behavior for Query Auto-completion *(SIGIR)*. 445–454.

[18] Juhong Liu, Ouri Wolfson, and Huabei Yin. 2006. Extracting semantic location from outdoor positioning systems. In *Proc. of the 7th International Conference on Mobile Data Management (MDM)*. 73–73.

[19] Edgar Meij, Peter Mika, and Hugo Zaragoza. 2009. An Evaluation of Entity and Frequency Based Query Completion Methods *(SIGIR)*. 678–679.

[20] Shuyao Qi, Dingming Wu, and Nikos Mamoulis. 2016. Location aware keyword query suggestion based on document proximity. *IEEE Transactions on Knowledge and Data Engineering* 28, 1 (2016), 82–97.

[21] Mark Sanderson and Janet Kohler. 2004. Analyzing geographic queries. In *SIGIR Workshop on Geographic Information Retrieval.* 8–10.

[22] Yang Song and Li-wei He. 2010. Optimal Rare Query Suggestion with Implicit User Feedback *(WWW)*. 901–910.

[23] Yang Song, Dengyong Zhou, and Li-wei He. 2012. Query suggestion by constructing term-transition graphs *(WSDM)*. 353–362.

[24] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational Context for Ranking in Personal Search *(WWW)*. 1531–1540.