Location-Aware Query Recommendation for Search Engines at Scale

Zhipeng Huang[†] and Nikos Mamoulis[‡]

[†]The University of Hong Kong
zphuang@cs.hku.hk
[‡]University of Ioannina
nikos@cs.uoi.gr

Abstract. Query recommendation is a popular add-on feature of search engines, which provides related and helpful reformulations of a keyword query. Due to the dropping prices of smartphones and the increasing coverage and bandwidth of mobile networks, a large percentage of search engine queries are issued from mobile devices. This makes it possible to provide better query recommendations by considering the physical locations of the query issuers. However, limited research has been done on location-aware query recommendation for search engines. In this paper, we propose an effective spatial proximity measure between a query issuer and a query with a location distribution obtained from its clicked URLs in the query history. Based on this, we extend two popular query recommendation approaches to our location-aware setting, which provides recommendations that are semantically relevant to the original query and their results are spatially close to the query issuer. In addition, we extend the bookmark coloring algorithm for graph proximity search to support our proposed approaches online, with a spatial partitioning based approximation that accelerates the computation of our proposed spatial proximity. We conduct experiments using a real query log, which show that our query recommendation approaches significantly outperform previous work in terms of quality, and they can be efficiently applied online.

1 Introduction

Keyword search, which allows a user to express her query with a few keywords, has become a fundamental tool in Web search engines. Recently, lots of interest from both research and industry has been drawn to the topic of *query recommendation*, which is closely related to keyword search. Besides ranking the Web pages according to their relevance to the query provided by the user, a search engine may provide several alternative formulations of the query, which can be more focused and interesting to the user. Query recommendation, as an add-on function to a search engine, has provided significant benefit to search engine users [11] in terms of providing better user experience.

Most of the existing work on query recommendation focuses on the analysis of *query logs*, which contain large amounts of historical information of search

engine users, including what query was issued by whom at what time and which URLs were subsequently clicked [5,6,8,14]. The query logs are often represented as *graphs* of queries and other related components, allowing graph analysis techniques to perform relevance search on query logs. For example, [5] built a graph of *queries* based on query logs. The weight of a graph edge that connects two queries is proportional to the times that the two queries are issued by a same user within a short period (i.e., the queries are in the same *search session*). Query recommendation is performed by applying Personalized PageRank proximity search on this graph starting from the original query.

Nowadays, as mobile devices are ubiquitous, many keyword search queries are expressed by mobile users and have *spatial intent*, i.e., the users require results that are physically close to their locations. However, very few studies have considered location information when performing query recommendation. The most recent work [17] proposes a solution based on a bipartite graph that connects queries to their clicked documents (or URLs). The edges of the graph are adjusted based on the location of the query issuer and then Personalized PageRank proximity search is applied to obtain the recommendations. The work of [17] only considers the locations of documents to derive the proximity between queries. In this work, we propose a more sophisticated approach that generates a spatial distribution for each query (based on its clicked URLs) and uses it to directly measure the proximity between each query and the query issuer. We extend two popular query recommendation approaches, i.e., query-flow graph [5] and term-query-flow graph [6], to our location-aware setting. The basic idea is to give higher preference to queries that have larger spatial proximity to the user during the random walk-based recommendation process, which finally leads to recommendations that are more spatially relevant compared to those suggested by the method of [17], as shown in our experimental results.

The main technical challenge is efficiency; search engines should provide instant responses to users; hence, query recommendation should also be conducted in sub-seconds. However, differently from the traditional query recommendation setting, which ignores user locations, we need to consider the spatial proximity between queries and the user, which can only be obtained online after the user issues her query. We first adopt Bookmark Coloring Algorithm (BCA) [4], a classic method for online Personalized PageRank based proximity search, to support our recommendation method. Then, we design an approximate version of the algorithm, which is based on a spatial partitioning and greatly reduces the cost without sacrificing the quality of the results.

We perform extensive experiments on a real query log from to verify the performance of our methods. We first conduct a user study, which shows that the users prefer recommendations generated by our proposed spatial proximity measure compared to four alternatives. Then, we compare our proposed location-aware query recommendation approaches with previous work, and show that our approaches achieve significantly better recommendations in terms of both semantic relevance and spatial proximity.

The contributions of our paper are summarized as follows:

– We consider the spatial proximity between a query and a search engine user in location-aware query recommendation. We propose an effective spatial proximity measure, shown to outperform alternatives in our user study.
– We extend two popular query recommendation approaches to support location-aware recommendation.
– We evaluate our proposed location-aware query recommendation methods, demonstrating that our methods outperform previous work significantly and that they can be applied online (within 300ms).

The rest of the paper is organized as follows. Section 2 introduces some preliminaries and definitions. Section 3 presents our location-aware query recommendation model. Section 4 includes our algorithm for efficient query recommendation. Section 5 presents our experimental results. Section 6 reviews related work and Section 7 concludes the paper.

2 Preliminaries and Definitions

In this section, we introduce some necessary preliminaries and definitions, including *query logs* (Section 2.1), how we model and obtain the relevance of queries to locations (Section 2.2), and two popular location-agnostic query recommendation methods (Sections 2.3 and 2.4).

2.1 Query Log

The query log of a keyword search engine is typically modeled as a set of records (q_i, u_i, t_i, C_i), where q_i is a query submitted by user u_i at time t_i, and C_i is the set of clicked URLs by u_i after q_i and before the user issues another query.

Following common practice [5, 6, 14], we can partition a query log into task-oriented *sessions*, where each session is a contiguous sequence of query records from the same user. Two contiguous queries are put in the same session if their time difference is at most t_θ (typically, t_θ is 30 minutes). Within the same session, we can assume that the user's search intent remains unchanged.

2.2 Obtaining Locations from a Query Log

Location Distribution of URLs. A webpage, corresponding to a URL, may contain information about one or more spatial locations. The *location distribution* of a URL d is a probability distribution p_d over a set of locations $L = \{(lat, lon) \mid lat, lon \in R\}$, where $\sum_{l \in L} p_d(l) = 1$. For the purposes of this paper, for each URL in the query log, we fetch the document and parse the content using GeoDict[1], a simple library/tool for pulling location information from unstructured text. This provides us with the location distribution for each URL. Alternatively, other methods for extracting locations from text can be applied [9].

[1] https://github.com/petewarden/geodict

Location Distribution of Queries. We also model the location distribution of a query q_i as a probability distribution p_{q_i}, and we can obtain it from a linear combination of the distributions of the clicked URLs for q_i. Formally,

$$p_{q_i}(l) = \frac{\sum_{d_j \in C_{q_i}} p_{d_j}(l)}{\sum_{l' \in L} \sum_{d_j \in C_{q_i}} p_{d_j}(l')},$$

where C_{q_i} is the set of clicked URLs for query q_i. In this paper, we use the location distributions of the queries to facilitate the problem of recommending queries to a search engine user u, that are not only semantically relevant to the query issued by u, but also are spatially close to the physical location of u.

2.3 Query-Flow Graph

One of the most promising directions for performing query recommendation relies on the extraction of behavioral patterns in query reformulation from query logs. The query-flow graph (QFG in short) [5] is a graph representation of query logs, capturing the "flow" between query units. Intuitively, a QFG is a directed graph of queries, in which an edge (q_i, q_j) with weight w indicates that query q_j follows query q_i in the same session of the query log with probability w.

More formally, QFG is defined as a directed graph $G_{qf} = (Q, E, W)$, where Q is the set of nodes, with each node representing a unique query in the log, $E \subseteq Q \times Q$ is the set of edges, and W is a weighting function assigning a weight $w(q_i, q_j)$ to each edge $(q_i, q_j) \in E$. In G_{qf}, two queries q_i and q_j are connected if and only if there exists a session in the query log where q_j follows q_i. Figure 1 illustrates a QFG with three query nodes.

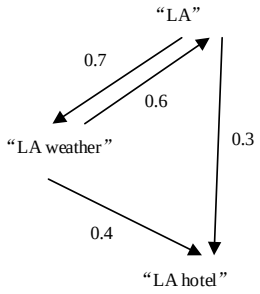


Fig. 1: QFG

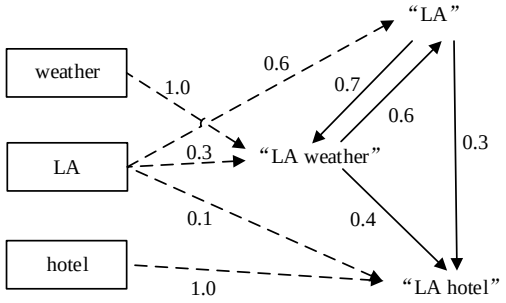


Fig. 2: TQGraph

Recommendation via QFG. Given a query $q \in Q$, the top-k recommendations for q can be obtained by a random walk with restart (RWR) [13] process starting from q, as suggested in [5]. At each step of the RWR, the random walker moves to an adjacent node with probability $1 - \alpha$ via the transition matrix W, or

teleports to the original node q with probability α. In this way, a RWR process defines a Personalized PageRank score $PPR(q, q', W)$ for each node q' as the probability that the RWR starting from q reaches node q'. In this way, the top-k recommendations can be the set of k nodes q' in Q, which have the maximum PPR scores w.r.t. q in QFG. In other words, the recommendation score $rec_q(q')$ for each $q' \in Q$ is defined as:

$$rec_q(q') = PPR(q, q', W), \tag{1}$$

where W is the transition matrix for the PPR, and queries having the top-k scores are recommended.

However, QFG has an obvious disadvantage for query recommendation; that is, it cannot make any recommendation to an input query q, if $q \notin Q$. In other words, if a query has not appeared in the query log before, QFG fails to generate any recommendation.

2.4 Term-Query-Flow Graph

Another popular method for query recommendation is the term-query-flow graph (TQGraph) [6], which basically extends the QFG method by considering a center-piece subgraph induced by terms contained into queries.

Formally, a TQGraph is a directed graph $G_{tqg} = G_{qf} \cup G_{tq}$, where G_{qf} is the QFG as described in Section 2.3, and G_{tq} is a bipartite graph of *term* nodes and *query* nodes. Specifically, the set of nodes in the TQGraph is $V_{tq} = Q \cup T$, where Q is the set of queries and T is the set of terms. Edge (t, q) exists in E_{tq}, if the term t is contained in query q. Figure 2 illustrates a TQGraph with three query nodes and three term nodes.

Recommendation via TQGraph. Given a query $q \in Q$, the top-k TQGraph recommendations for q are obtained by ranking all $q' \in Q$ based on their aggregate PPR scores w.r.t. each term $t \in q$. In other words, the recommendation score $rec_q(q')$ for each $q' \in Q$ is defined as follows:

$$rec_q(q') = \prod_{t \in q} PPR(t, q', W \cup E_{tq}) \tag{2}$$

We can see that TQGraph can generate recommendations for query q, as long as all the terms within q appear in the query log. Empirically, TQGraph has a much better *query coverage* compared to QFG, because it can also be used for queries that are asked for the first time.

3 Location-aware Query Recommendation

In this section, we introduce our location-aware query recommendation models. Our goal is to provide recommendations that are spatially close to the query issuer. For this, we should first define spatial proximity between a user located at l_u and a query with location distribution p_q. Some of the alternatives are:

– (i) Expected Distance (ED). Formally, $ED(p_q, l_u) = \sum_{l \in L(q)} p_q(l) \times dist(l, l_u)$.
– (ii) Min Distance (MinD). Formally, $MinD(p_q, l_u) = min_{l \in L(q)} dist(l, l_u)$.
– (iii) Max Distance (MaxD). Formally, $MinD(p_q, l_u) = max_{l \in L(q)} dist(l, l_u)$.
– (iv) Mean Distance (MeanD). Formally, $MeanD(p_q, l_u) = dist(mean(p_q), l_u)$.

However, these four distance-based measures are not fully consistent with our goal of finding spatially close queries to the user. For example, suppose we have two queries q_1 and q_2 with the following location distribution:

q_1	Hong Kong : 0.6, New York : 0.3, Los Angeles : 0.1
q_2	Beijing: 0.8, Los Angeles : 0.2

One would argue that q_1 is more spatially relevant to a user u_1 located in Hong Kong, compared to q_2, as a great portion of q_1's distribution is very close to u_1, which means that the majority of URLs related to query q_1 contain information that is spatially relevant to u_1. However, using ED, MaxD or MeanD might not select q_1, because after considering all locations of q_1, its overall distance to Hong Kong is quite far. At the same time, MinD does not distinguish q_1 and q_2 for a user u_2 located in Los Angeles, because MinD neglects the support probability of each location. Hence, we should define a more appropriate spatial proximity measure that better captures the distance between a query and a search engine user. In this direction, we propose the following measure:

Definition 1. Spatial Proximity sim_s**.** *Given the location of the user $l_u \in L$, a range threshold r, and a location distribution p_q of a query q, the spatial proximity between l_u and q is the portion of p_q within distance r from l_u, i.e.,*

$$sim_s(q, l_u) = \sum_{dist(l_u, l') < r} p_q(l')$$

The range threshold r models the distance that the user is willing to travel in order to use a service offered by the query results. In the example above, assuming that we select r as a within-city travel distance, we will have: $sim_s(q_1, l_{u_1}) = 0.6$, $sim_s(q_2, l_{u_1}) = 0$, $sim_s(q_1, l_{u_2}) = 0.1$, $sim_s(q_2, l_{u_2}) = 0.2$. This is consistent with the intuition that u_1 is more related to query q_1, and u_2 is more related to query q_2. In our experiments, we use $r = 100km$ by default and compare the performances of our methods for different values of r.

After obtaining the spatial proximity between the user and the queries, we can adjust the weights on the edges of QFG to give higher preference to queries that have larger sim_s to the user.

Definition 2. Spatially Adjusted Weights. *Given a query $q \in Q$ issued at location l_u, the spatially adjusted weight for an edge of a QFG $(q_i, q_j) \in E$ is defined as:*

$$\tilde{w}(q_i, q_j) = \beta \times w(q_i, q_j) + (1 - \beta) \times sim_s(q_j, l_u)),$$

where β is a parameter that controls the relative importance of spatial proximity and $w(q_i, q_j)$ is the original weight of the edge (q_i, q_j) in the QFG.

With the linear function in Definition 2, we obtain a location-aware transition matrix $\tilde{W}$ from the original matrix W. Then, we can perform location-aware query recommendation based on $\tilde{W}$. In other words, the recommendation processes for spatial QFG (SQFG) and spatial TQGraph (STQGraph) are the same as their location-agnostic counterparts, except that we use $\tilde{W}$ instead of W.

Recommendation via SQFG. Given a query $q \in Q$ issued at location l_u, the top-k SQFG recommendations for q can be obtained by a location-aware Personalized PageRank w.r.t. node q, i.e.,

$$rec_q(q') = PPR(q, q', \tilde{W}), \tag{3}$$

where $rec_q(q')$ is the recommendation score for query q' and $\tilde{W}$ is the location-aware transition matrix for the PPR.

Recommendation via STQGraph. Given a query $q \in Q$ issued at location l_u, the top-k STQGraph recommendations for q can be obtained by ranking all $q' \in Q$ based on their aggregate PPR scores w.r.t. each term $t \in q$. In other words, the recommendation score $rec_q(q')$ for each $q' \in Q$ is defined as follows:

$$rec_q(q') = \prod_{t \in q} PPR(t, q', \tilde{W} \cup E_{tq}) \tag{4}$$

4 Location-aware PPR

Both our SQFG and STQGraph models require computation of location-aware PPR over the spatially adjusted transition matrix $\tilde{W}$. However, this can be expensive, as the transition matrix $\tilde{W}$ depends on the location l_u of the search engine user, which can only be known at query time. Thus, traditional indexing techniques for efficiently computing PPR cannot be used in our setting. In this section, we first introduce a basic solution, by extending the Bookmark Coloring Algorithm (BCA) [4]. Then, we propose a more efficient, approximate version of the algorithm, which is based on a spatial partitioning approach.

4.1 BCA with Online Transition Matrix $\tilde{W}$

We extend the famous Bookmark Coloring Algorithm (BCA) [4] to compute the top-k PPR results based on the location-aware transition matrix $\tilde{W}$ on query time as our basic method. The basic idea of BCA is to model the RWR process as a bookmark coloring process, in which some portion of the ink in a processed node is sent to its neighbors, while the remaining ink is retained at the node.

Specifically, starting from the query node q with 1.0 units of ink, BCA keeps α portion in q and distributes the remaining $1 - \alpha$ portion to q's neighbors in the graph using the weights of the outgoing edges to determine the percentage of ink sent to each neighbor. The process is repeated for each node that receives ink, until the residue ink to be redistributed becomes a very small percentage of the original 1.0 units. Different from traditional PPR computation using BCA,

Algorithm 1: BCA

Input: Transition matrix W, starting node q, user location l_u
Output: Apprximated PPR vector rec_q

PriorityQueue $que \leftarrow \emptyset$
Add q to que with $q.ink \leftarrow 1.0$
$R \leftarrow \emptyset$;
$Cache \leftarrow \emptyset$;
while $que \neq \emptyset$ *and* $que.top.ink \geq \epsilon$ **do**
    Deheap the first entry top from que;
    $R[top] \leftarrow R[top] + top.ink \times \alpha$;
    for $q' \in top.neighbors$ **do**
        if $q' \in Cache$ **then**
            $sim_s(q', l_u) \leftarrow Cache[q']$;
        else
            Compute $sim_s(q', l_u)$ using Definition 1;
            $Cache[q'] \leftarrow sim_s(q', l_u)$;
        Compute $\tilde{w}(top, q')$ using Definition 2;
        $q'.buf \leftarrow top.ink \times (1 - \alpha) \times \tilde{w}(top, q')$;
        if $q'.buf \geq \epsilon$ **then**
            Add q' to que with ink $q'.buf$;
            $q'.buf \leftarrow 0$;
return R

our transition matrix $\tilde{W}$ can only be obtained online by Definition 2, after we know the location of the query issuer l_u.

In our implementation, the spatially adjusted weights of each edge (q_i, q_j) are also computed online based on l_u, at the time when the query node q_i is distributing ink. This means that the computation of $\tilde{W}$ is done during BCA simulation. A node distributes ink only if the quantity of the ink exceeds a threshold ϵ (typically, $\epsilon = 10^{-5}$). BCA terminates when there are no more nodes to distribute ink.

We adopt the following two optimizations in our BCA implementation.

- **Lazy Updating Mechanism.** In the original BCA, a node distributes its ink aggressively, i.e, each neighbor q' of node top will be pushed into the priority queue que after receiving some portion of $top.ink$. On the other hand, we only care about the nodes with ink greater than ϵ. Based on these two observations, a lazy updating mechanism can reduce the number of non-necessary pushing without changing the final results; the pushing a node q' into the priority queue que is delayed until the ink it receives is greater than ϵ. If the amount of received ink is less than ϵ, q' only accumulates it in in a buffer; as soon as the buffer's ink exceeds ϵ, q' is pushed into que.
- **Spatial Proximity Caching**. Every time when we need to distribute ink to a query node q', we need to compute the spatial proximity between q' and

the location of the user l_u. However, the same query node may be processed multiple times in a single BCA call. In view of this, we cache the spatial proximities between the location of the user l_u and the query nodes that have been computed so far. By doing this, we only need to compute the spatial proximity for a query q' once during a BCA call.

Algorithm 1 details our implementation of BCA, including the two optimizations mentioned above. Priority query *que* maintains the nodes to be processed in descending order of their ink (Line 1). *que* initially contains only one node with ink amount 1.0 (Line 2), i.e., the starting node of the PPR. The nodes that have some retained ink are kept in a dictionary R, which is initially empty (Line 3). Termination conditions are checked at each iteration (Line 5). Within each iteration, we first dequeue from *que* the node with the most ink to distribute (Line 6). Then we leave α portion to its result (Line 7), and distribute the rest to its neighbors with weights $\tilde{w}$ (Lines 8-18). We first check the spatial cache whether q' has been computed before (Line 9). If so, the spatial proximity between q' and l_u can be directly got from the cache (Line 10). Otherwise we need to compute $sim_s(q', l_u)$ (Line 12) and save to the cache (Line 13). Finally, for each of the neighbors, if the received ink is greater than a threshold ϵ (Line 16), the corresponding query node will be pushed into *que* (Line 17) and the corresponding buffer is cleared (Line 18). Finally, the dictionary R is returned. In SQFG, where a single RWR search is applied, the k query nodes in R with the most retained ink are recommended. In STQGraph, the Rs of the RWR searches from all terms are summed up at each query node before computing and returning the top-k query nodes.

Partitioning Based Approximation. The previous two optimizations guarantee the same results as the original BCA and improve its performance. However, the cost of computing the spatial proximity $sim_s(q', l_u)$ between a query q' and the location of the user l_u at each iteration is still the bottleneck of the algorithm. Recall from Definition 1 that we need to enumerate all locations of the query q' in order to accumulate the distribution. To reduce this high cost, we propose to compute a partitioning based approximation of $sim_s(q, l)$ as follows:

$$\hat{sim}_s(q, l) = \sum_{cir(l_u) \text{ intersects } c} p_q(c), \tag{5}$$

where c is a spatial partition of locations, $cir(l_u)$ is the circle with l_u as center and r as radius, and $p_q(c) = \sum_{l' \in c} p_q(l')$ is the location distribution of q that falls into partition c.

We use a grid to partition the space. Hence, locations that fall into the same cell belong to the same partition. If the length of each grid cell is a, to compute $\hat{sim}_s$, we only need to accumulate $p_q(c)$ for at most $\lceil \frac{2r}{a} + 1 \rceil^2$ partitions. In our experiments, we use $a = r$, so the computational cost is much lower than computing the exact sim_s, which requires enumeration of all locations.

Figure 3 illustrates an example of our partitioning based approximation. The dots with number next to them represent the location distribution of a query q.

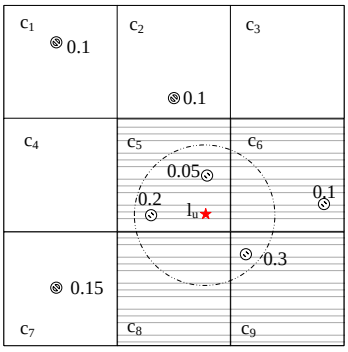


Fig. 3: Example of partitioning

Table 1: Location distribution approximation

cell	$p_q(c)$	cell	$p_q(c)$	cell	$p_q(c)$
c_1	0.1	c_2	0.1	c_3	0
c_4	0	c_5	0.25	c_6	0.1
c_7	0.15	c_8	0	c_9	0.3

Suppose a user is located at the starred location and the circle is defined by that location and the range threshold r. The shaded cells are those which intersect the circle and according to Definition 1, $sim_s(q, l_u) = 0.2 + 0.05 + 0.3 = 0.55$. After using our spatial partition, we can obtain an approximation of the location distribution as in Table 1. Then, an approximation of the spatial proximity is computed as $\hat{sim}_s(q, l_u) = 0.25 + 0.1 + 0 + 0.3 = 0.65$.[2]

5 Experimental Evaluation

5.1 Dataset

We use AOL in all our experiments. AOL is a well-known public query log from a major commercial search engine, which consists of Web search queries collected from 657k users over a two months period in year 2006. This dataset is sorted by anonymous user ID and sequentially arranged, containing 20M query instances corresponding to around 9M distinct queries. After we sessionize the query log with θ_t = 30min, we obtain a total of 12M sessions.

5.2 Methodology

We adopt the automatic evaluation process described in [14], to assess the performance of the tested methods. In a nutshell, we use part of the query log as training data to generate recommendations for a kept-apart query log fragment (the test data). In the test query log, we denote by $q_{i,j}$ the j_{th} query in session s_i. We assume that all $\{q_{i,j} \mid j > 1\}$ are good recommendations for query $q_{i,1}$ which, in accordance to previous work [14].

[2] We do not further refine to get an exact result by looking into the locations within the cells, because we believe that those locations near the range r from the user are still spatially relevant (see the location in cell c_6 of Figure 3).

Specifically, we use 90% of the query log for training, which contains $11M$ sessions and $8.4M$ distinct queries. We use the remaining 10% of the query log to generate testing queries. We first extract sessions with at least two queries, and randomly sample 10K queries as our testbed. We take the first query of each session as input and the queries that follow as the ground truth recommendations. Formally, the ground truth for input query $q_{i,1}$ is $\{q_{i,j} \mid j > 1\}$, where $q_{i,j}$ is the j_{th} query appearing in the i_{th} session. While the objective of this evaluation approach may not necessarily be aligned with what a good recommendation could be on particular instances, by being entirely unsupervised and applied on a large number of sessions, it is a strong indicator of the techniques' performance. Note that we randomly assign the location of the query issuer l_u, as the dataset does not contain the location information of about the users.

We use the following three metrics to evaluate the performance of each method:

- $coverage$. This is the percentage of input queries that can be served with at least one recommendation.
- $precision@k$. This is the percentage of recommended queries in the top-k lists that are in the ground truth as described previously. Formally, $precision@k = \frac{\#HIT}{k \cdot \#query}$, where $\#HIT$ is the total number of recommended queries that are part of the ground truth, and $\#query$ is the number of input queries.
- $sim_s@k$. This is the average spatial proximity (see Definition 1) between the recommended queries in the top-k lists to the location of the query issuer l_u.

Competitors

- **LKS** [17]. LKS is the most recently proposed location-aware keyword suggestion approach. It first builds a bipartite graph of queries and URLs using the query log, and then performs location-aware random walk over the graph during online recommendation. We use the default settings of LKS, i.e., the restart probability α_{LKS} and the edge weight adjustment parameter β_{LKS} are both set to 0.5.
- **SQFG**. SQFG is our spatial QFG method as described in Section 3. By default, we set the spatial radius threshold r = 100km, the restart probability $\alpha = 0.5$, and the spatial adjustment factor $\beta = 0.5$.
- **STQGraph**. STQGraph is our proposed spatial TQGraph approach as described in Section 3. We use the same default settings as in SQFG.
- **STQGraph***. STQGraph* is our spatial partitioning based approximation approach. By default, we use 100km as the length of each grid cell, and all the other parameters are same as STQGraph.

5.3 User Study

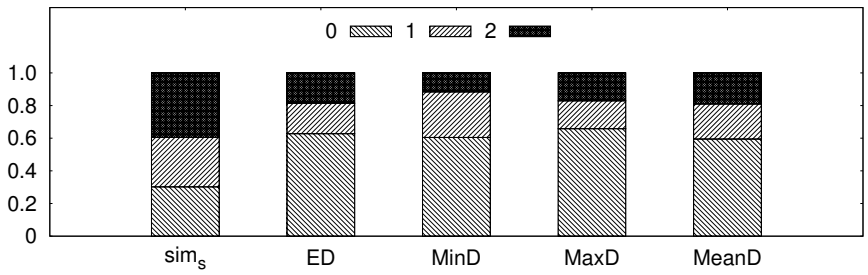
We first conducted a user study to compare our proposed spatial proximity sim_s in Definition 1 with the four alternatives mentioned in Section 3. We first used our STQGraph to generate top-1 recommendations for 100 random queries with

different spatial proximity measures. Then, for each of the recommended queries, we showed its location distribution as well as the location of the query issuer l_u to the participants. They were asked to rate the spatial relevance between the recommended query and the query issuer, using one of the following rating levels: 0 for not related at all, 1 for somehow related and 2 for very related. The recommended queries were shuffled before given to the participants, so that they could not know which spatial proximity measure was used to generate the recommended query. We asked 15 participants (HKU students) to rate the recommended queries, and each of the queries were given at least 3 ratings.

The results are shown in Figure 4. We can see that our sim_s has the largest percentage of 2s, which means that sim_s is acknowledged to be the best measure of spatial proximity. Out of the four alternatives, ED and MeanD received relatively better user feedback. This is because a smaller ED or MeanD implies smaller overall distance to the query issuer. MinD got the worst user feedback because MinD only considers the location l which is the closest to l_u, but not the probability $p_q(l)$, which could be too small.

From this user study, we can conclude that users prefer sim_s over the other four proximity measures. In the rest of our experiments, we use sim_s to evaluate the spatial quality of recommended queries.

Fig. 4: User Study on Different Spatial Proximity Measures



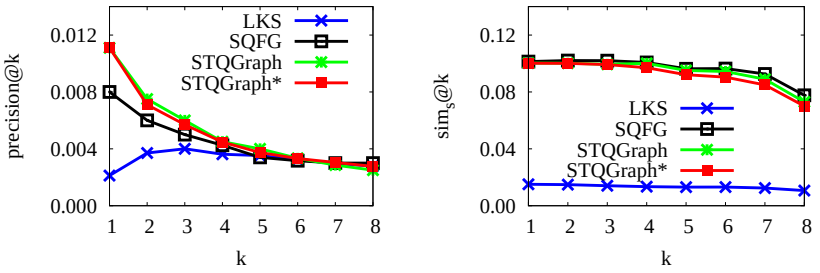
5.4 Effectiveness

We first compare the four tested methods. Then we test the parameter sensitivity of our STQGraph method to different parameter values.

Comparison Results. Table 2 compares all methods in terms of their *coverage*. We can see that SQFG has relatively low *coverage* compared the other three approaches. This is expected, because SQFG has the same disadvantage as QFG, i.e., it cannot provide recommendations to any previously unseen queries. STQGraph and LKS have similar *coverage*, much higher than that of SQFG. Note

Table 2: *coverage* results

method	LKS	SQFG	STQGraph	STQGraph_P
coverage	36.8%	27.9%	37.1%	37.1%



(a) *precision*@*k* (b) sim_s@*k*

Fig. 5: Comparison.

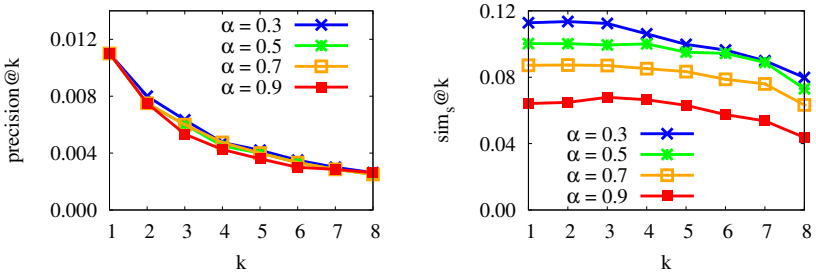
that STQGraph and STQGraph* have the same *coverage*, since our spatial partition based approximation only affects the ranking of the recommended queries.

Figure 5(a) shows the *precision*@*k* of all methods. Since typical search engines (e.g. Google and Yahoo!) show eight recommendations, we test values of k from 1 to 8. Observe that our STQGraph and STQGraph* methods have significantly larger *precision*@*k* compared to LKS. As k becomes larger, *precision*@*k* becomes smaller. This means our recommendation methods rank the recommended queries reasonably, as those with smaller ranks are more precise. STQGraph* has almost the same *precision*@*k* as STQGraph, which means that our spatial partition based approximation does not harm the recommendation quality in terms of semantic quality. The precision of SQFG is lower than that of STQGraph for small values of k.

Figure 5(b) shows the results of sim_s@*k*. Similar to the case of *precision*@*k*, sim_s@*k* drops as k increases. LKS has very poor sim_s@*k* result compared with our approaches. This is because LKS tends to recommend queries that share the same clicked URLs with the input query, without directly considering the location distribution of the recommended queries. SQFG, STQGraph* and STQGraph achieve almost the same sim_s@*k*.

Parameter Sensitivity. We now test the sensitivity of our STQGraph approach to the values of its parameters. As *coverage* result is only related to the connectivity of STQGraph and is not sentitive to the parameters, we only show the *precision*@*k* and sim_s@*k* results.

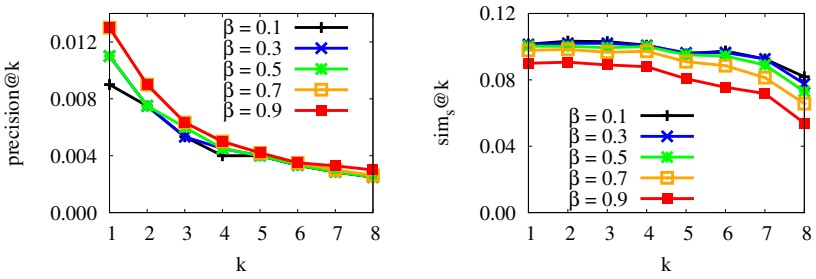
• **Varying** α. Figure 6 shows the quality of STQGraph for various values of α. Observe that α does not influence *precision*@*k* very much. In addition, larger α leads to smaller sim_s@*k*. This is because a larger α gives higher weight to



(a) *precision*@k (b) sim_s@k

Fig. 6: Varying α

the adjacent queries to the input query in the graph, whereas potentially better queries exist at a larger distance.

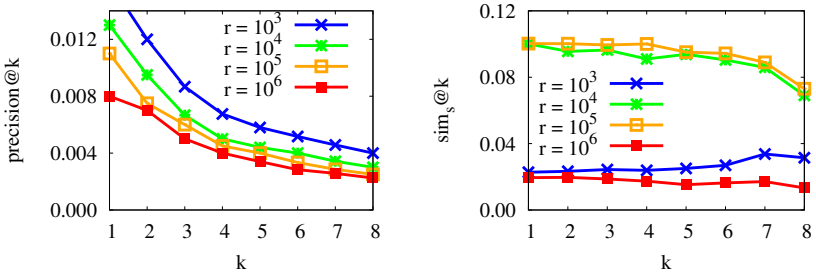


(a) *precision*@k (b) sim_s@k

Fig. 7: Varying β

• **Varying** β. Figure 7 shows the quality of STQGraph for various values of β. We can see that larger β values lead to slightly higher *precision*@k. However, larger β values lead to smaller sim_s@k. This is because a larger β gives higher weight to the semantic relevance between queries and lower weight to the spatial proximity. Overall, a value of β close to 0.5 strikes a balance between the two factors giving good *precision*@k and sim_s@k at the same time.

• **Varying** r. Figure 8 plots the quality of STQGraph for various values of r. A larger r leads to a smaller *precision*@k. This is because a larger r will result in larger spatial proximity in general, which eventually puts less emphasis to the original weights on the edges of QFG. From Figure 8(b), we observe that too large and too small r values lead to worse spatial proximity results. When we use a very small r, we get very small spatial proximity in general, leading to a worse



(a) *precision*@*k* (b) sim_s@*k*

Fig. 8: Varying r

sim_s@k result. When we use a very large r, we cannot distinguish queries that are actually close to the user, also leading to a worse sim_s@k. This result shows that we should choose an appropriate r for our method. Empirically, $r = 10^6$ (100 km) gives good results.

5.5 Efficiency

Now we test the efficiency of our optimizations and the approximation technique. We compare the overall running time of our STQGraph recommendation method, implemented with four versions of BCA for this purpose:

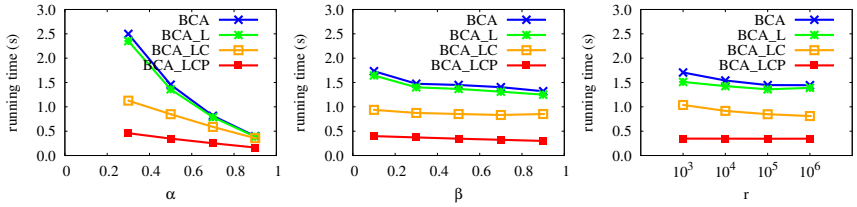
- **BCA**. The basic BCA algorithm without any optimizations.
- **BCA_L**. The BCA algorithm with the lazy update mechanism.
- **BCA_LC**. The BCA algorithm with the lazy update mechanism and spatial caching.
- **BCA_LCP**. The BCA algorithm with the lazy update mechanism, spatial caching and spatial partitioning based approximation. Note that this method corresponds to our STQGraph* method, which returns slightly different recommendations to STQGraph.

• **Varying** α. Figure 9(a) shows the average running time of STQGraph using the different versions of BCA for different values of α. We can see that all four versions terminate faster for larger values of α, which is consistent with our intuition. BCA_LCP is significantly faster than all other versions. When $\alpha = 0.5$, it takes only 0.3s for a query, which indicates that our STQGraph* can provide instant query recommendations.

• **Varying** β. Figure 9(b) shows the running times for different values of β. A first observation is that the cost of the different versions of BCA is not much sensitive to β, as β only determines how much importance we put to spatial proximity. For the default values of α and r, BCA_LC takes around 1.0s for each query, while BCA_LCP needs only 0.3s. Considering that STQGraph* achieves

similar effectiveness to STQGraph, as shown in our previous experiments, STQGraph* (which uses BCA_LCP) is more suitable for real-time applications.

• **Varying** r. Figure 9(c) shows the running times for different values of r. Observe that the runtimes for all methods are not sensitive to the change of r. This is because r only influences the values of spatial proximity sim_s.



(a) varying α (b) varying β (c) varying r

Fig. 9: Running time.

6 Related Work

Query autocompletion and query recommendation both aim at providing accurate query reformulation suggestions on-the-fly. In query autocompletion, only the most relevant *expansions* of the input query are shown, typically while the user is typing the query. Most existing works apply prefix-based recommendations and use trie-like index structures [3, 7, 18, 19]. In this paper, we focus on query recommendation, where the suggestions are not necessarily expansions of the input query. There have been many works on query recommendation, and most of them rely on analyzing query logs to extract useful patterns that model user behavior. All these works boil down to modeling the similarity between queries, often using random walk based proximity measures on graphs that may include users, terms, queries and URLs.

Early approaches rely on clustering similar queries [1, 20], where the similarity is defined using the query-URL graph or the term-vector representations of queries obtained from the clicked URLs. Later, [22] proposed the extraction and analysis of search sessions from the query log that capture the causalities between queries, and combined this with content-based similarity. In [2], the authors introduced the concept of *cover-graph*, a bipartite graph between queries and Web pages, where links indicate the corresponding clicks. [12] proposes recommending queries in a structured way for better satisfying exploratory interests of users, and [21] proposes a context-aware query recommendation model considering the relationship between queries and their clicks.

[5] and [6] are two of the most influential works in query recommendation. Both of them exploit flow patterns in query logs, and use graph-based methods to perform query recommendation. [5] builds a graph of queries, termed the query-flow graph (QFG), in which the links model the transition probabilities

between queries. [6] further extends the QFG to a term-query-flow graph (TQ-Graph), which also include nodes representing terms within queries. In this way, TQGraph can provide recommendations even for queries that never appeared in the query log. In both works, the top-k recommendations are obtained by performing random walk with restart (RWR) in the graphs.

Although many keyword search queries are sent from mobile users who have spatial search intent, there is limited research on location-aware query recommendation. In [16], the similarity between two queries is considered high if there are groups of similar users issuing these queries from nearby places at similar times. Google [15] keeps track of the locations where past queries are issued and determines the similarity between queries by also considering the proximity between the locations of the corresponding query issuers. [23] apply a learning model on the tensor representation of the user-location-query relations to predict the user's search intent. The most recent related work [17] proposes a location-aware keyword suggestion (LKS) method, which extends the idea of [10]. However, LKS only considers the location information for the documents (URLs), without considering that of queries. As we argue in this paper, it is more important to consider the spatial proximity between the user and the queries than the documents, because it is the queries we recommend to the user in the task of query recommendation. We experimentally compare our proposed methods with LKS and show that our methods provide better recommendations.

7 Conclusion

We study the problem of location-aware query recommendation for search engines. We first propose a spatial proximity measure between a keyword search query and a search engine user. Then, based on this proximity measure, we extend two popular query recommendation approaches (i.e., QFG and TQGraph) to apply for the location-aware setting. In this way, we can generate recommendations that are not only semantically relevant, but also spatially close to the query issued by a user at a specific location. In addition, we extend the Bookmark Coloring Algorithm to support efficient online query recommendation. We also propose an approximate version of the algorithm that uses spatial partitioning to accelerate the computation of our proposed spatial proximity measure. Experiments on a real query log show that our proposed methods significantly outperform previous work in terms of both semantic relevance and spatial proximity, and that our method can be applied to providing recommendations within only a few hundreds of milliseconds.

Acknowledgements

We thank the reviewers for their valuable comments. This work is partially supported by GRF Grant 17205015 from Hong Kong Research Grant Council. It has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 657347.

References

1. R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology - EDBT Workshops*, 2004.
2. R. A. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *KDD*, 2007.
3. Z. Bar-Yossef and N. Kraus. Context-sensitive query auto-completion. In *WWW*, 2011.
4. P. Berkhin. Bookmark-coloring algorithm for personalized pagerank computing. *Internet Math*, 3:41–62, 2006.
5. P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *CIKM*, pages 609–618. ACM, 2008.
6. F. Bonchi, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini. Efficient query recommendations in the long tail via center-piece subgraphs. In *SIGIR*, pages 345–354. ACM, 2012.
7. F. Cai, S. Liang, and M. de Rijke. Time-sensitive personalized query auto-completion. In *CIKM*, 2014.
8. H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In *KDD*, pages 875–883, 2008.
9. Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In *SIGMOD*, pages 277–288, 2006.
10. N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR*, pages 239–246. ACM, 2007.
11. D. Downey, S. T. Dumais, and E. Horvitz. Heads and tails: studies of web search with common and rare queries. In *SIGIR*, 2007.
12. J. Guo, X. Cheng, G. Xu, and H. Shen. A structured approach to query recommendation with social annotation data. In *CIKM*, pages 619–628. ACM, 2010.
13. T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, pages 517–526. ACM, 2002.
14. Z. Huang, B. Cautis, R. Cheng, and Y. Zheng. Kb-enabled query recommendation for long-tail queries. In *CIKM*, pages 2107–2112, 2016.
15. J. Myllymaki, D. Singleton, A. Cutter, M. Lewis, and S. Eblen. Location based query suggestion, Oct. 30 2012. US Patent 8,301,639.
16. X. Ni, J. Sun, and Z. Chen. Mobile query suggestions with time-location awareness, May 31 2012. US Patent App. 12/955,758.
17. S. Qi, D. Wu, and N. Mamoulis. Location aware keyword query suggestion based on document proximity. *TKDE*, 28(1):82–97, 2016.
18. M. Shokouhi. Learning to personalize query auto-completion. In *SIGIR*, 2013.
19. M. Shokouhi and K. Radinsky. Time-sensitive query auto-completion. In *SIGIR*, 2012.
20. J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. In *WWW*, 2001.
21. X. Yan, J. Guo, and X. Cheng. Context-aware query recommendation by learning high-order relation in query logs. In *CIKM*, pages 2073–2076. ACM, 2011.
22. Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *WWW*, pages 1039–1040, 2006.
23. Z. Zhao, R. Song, X. Xie, X. He, and Y. Zhuang. Mobile query recommendation via tensor function learning. In *IJCAI*, pages 4084–4090, 2015.