

Privacy Preserving Publish/Subscribe for Geo-Textual Data Streams

Ya Gao, Qiuyuan Ai, Jie Song[†]

College of Engineering, Peking University No.5 Yiheyuan Road, Beijing 100871, China

Keywords: Publish-Subscribe Systems; Spatio-Textual; Spatial Keyword Queries

Citation: Gao Y., Ai Q.Y., Song J.: Privacy Preserving Publish/Subscribe for Geo-Textual Data Streams. *Data Intelligence* 6(4), 994-1013 (2024). doi: <https://doi.org/10.3724/2096-7004.di.2024.0014>

Submitted: March 11, 2024; Revised: April 11, 2024; Accepted: May 29, 2024

ABSTRACT

In the era of abundant location-based data, an increasing number of mobile users seek access to timely local information tailored to their individual interests. Consequently, the development of an efficient publish/subscribe system becomes pivotal, allowing a vast user base to seamlessly receive geo-textual objects based on their specific preferences from data streams. However, a notable challenge arises as mobile users often hesitate to disclose their personal interests, requirements, and locations to service providers in location-based publish/subscribe systems, giving rise to substantial data privacy concerns. In this light, we propose a privacy-preserving publish/subscribe framework, which not only facilitates real-time delivery of geo-textual objects to a large-scale audience of location-based subscribers, but also ensures the utmost privacy of subscribers' locations and query keywords. Through experiments conducted on two real-life datasets, our proposed privacy-preserving publish/subscribe system demonstrates its capability to produce real-time matching results. The system can simultaneously handle millions of privacy-enhanced subscription queries over a stream of geo-textual objects.

1. INTRODUCTION

The ongoing advancement of location-based social media platforms enables a vast number of online users to access a diverse range of geo-textual objects. These objects are typically represented as triples comprising a geographical location, a textual description, and a timestamp. Geographical locations can be specified either as geographical coordinates or as geographic terms. A textual description often consists of short messages (e.g., tweets, reviews of points-of-interest (POIs)) or articles (e.g., local news). A

[†] Corresponding author: Jie Song (E-mail: jie.song@pku.edu.cn; ORCID: 0000-0002-4916-8272).

timestamp indicates the publication time of the geo-textual object accordingly. The ubiquity of geo-textual data facilitates the discovery of local event news and trending public opinions from location-based social media. Geo-textual data serves as a rapid source of firsthand information for breaking events [1] and acts as an early indicator of local emerging opinions and evolving trends. Consequently, ongoing research efforts focus on the development of location-aware publish/subscribe frameworks and continuous spatial keyword query processing methods. These approaches aim to provide subscribers with timely and relevant information based on their preferences across a stream of geo-textual objects.

However, existing proposals regarding continuous spatial keyword query processing and location-based subscription query processing algorithms often fail to deliver satisfactory results to users in a safe and timely manner due to the following two reasons. Firstly, the influx of geo-textual objects is naturally overwhelming. For instance, X (Twitter) has 528.3 million monthly active users in 2024 [2], who collectively publish at least 500 million tweets on average per day. Consequently, subscribers may find themselves inundated with a deluge of relevant messages within a short time frame, particularly during bursty trends. Secondly, subscribers often express reluctance to publicly disclose their locations and preferences to others. However, existing location-based publish/subscribe systems and continuous spatial keyword query processing mechanisms overlook the privacy concerns of subscribers during subscription registration. Consequently, service providers may gain access to subscribers' precise locations and preferences without considering their privacy. Therefore, whether the privacy of subscribers can be ensured or not relies entirely on the trustworthiness of the service provider.

In this light, we develop a privacy-preserving location-aware publish/subscribe system that is capable of processing a large number of Spatial Keyword Subscription (SKS) queries while addressing the privacy concerns of subscribers. In developing our publish/subscribe system, we must prioritize two critical factors: efficiency and privacy. Both elements are fundamental to the system's effectiveness and user trust, demanding careful consideration and strategic integration into our framework. Existing location-aware query processing methods employ encrypted spatial indexing structures (e.g., encrypted tree indices [3, 4] and R-tree with Power Diagrams [5]), spatial transformation techniques [6], blockchains [7, 8], or differential privacy technologies [9] to enhance user privacy. While the discussed privacy-preserving spatial keyword search frameworks effectively address privacy concerns related to user location and keywords, their performance in terms of efficiency and scalability falls short in the publish/subscribe context. This environment demands the real-time processing of massive-scale subscription queries, a task these frameworks struggle with due to the complex nature of their privacy-aware data structures and search algorithms. Therefore, there is a pressing need to devise a lightweight approach that bolsters user privacy without relying on existing privacy-aware data structures and the extensive array of current privacy-preserving search methodologies.

For this purpose, we model each location-aware subscriber as a Privacy-enhanced SKS (P-SKS) query, which is defined by a spatial region and a collection of subscription keywords connected by semantics "AND" or "OR". Figure 1 presents examples of SKS queries and their P-SKS queries. Specifically, the left SKS query, posted by a web user, consists of keywords "fever covid". Its query region is denoted by

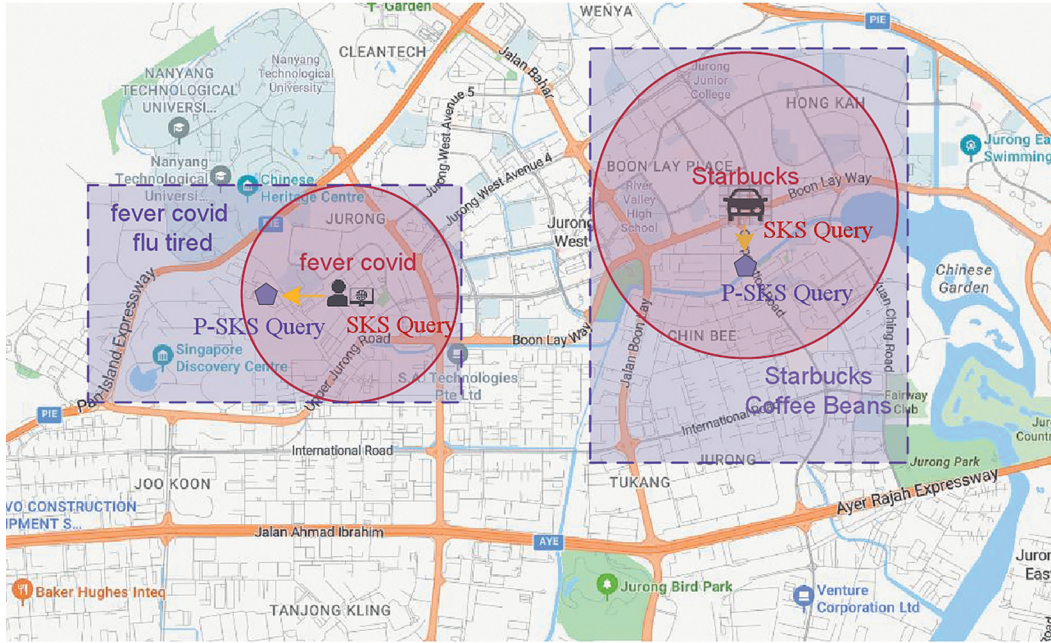


Figure 1. Example of SKS and P-SKS Queries.^①

the red circle. In order to protect the privacy of the web user, we generate its P-SKS query by generating a rectangular region centered at the purple pentagon and adding keywords that have similar or related meanings to the original keyword set. The right SKS query, issued by a driver, consists of keyword “Starbucks”. Likewise, its P-SKS query is illustrated by the right purple rectangle with keywords “Starbucks Coffee Beans”. Given a stream of geo-textual objects, the original SKS query continuously receives matching geo-textual objects such that: (1) they are located inside the subscription region, and (2) their textual descriptions satisfy the subscription keyword predicate. When a subscriber has a particular subscription requirement, the subscriber may register an SKS query to a location-aware publish/subscribe system. However, subscribers must trust the publisher/subscribe service provider to keep their personal locations and input subscription keywords confidential. To offer subscribers peace of mind, it is crucial to implement measures that prevent the service provider from accessing subscribers’ precise locations and preferences. For this, we develop a *subscription privacy enhancing scheme*. This scheme aims to enhance privacy by converting an SKS query into a P-SKS query before it is transmitted. In particular, the subscription privacy enhancing scheme consists of two privacy enhancing algorithms, namely location privacy enhancing algorithm and keyword privacy enhancing algorithm.

The location privacy enhancing algorithm operates as follows: Given an SKS query q , it takes the subscription region of q as input. The algorithm then outputs an altered region by shifting its center by

^① is a snapshot taken from Google Maps, depicting a region within Singapore.

a distance of δ_i and enlarging the region size to ensure that the altered region fully encompasses the original region. Such alteration guarantees that the SKS query q will not miss any matching geo-textual objects. Similarly, the text privacy enhancing algorithm operates as follows: Given an SKS query q , it takes its subscription keyword predicate as input. The algorithm then outputs an altered subscription keyword predicate by adding m keywords connected by “OR” semantic. Such alteration also guarantees that the SKS query q will not miss any matching geo-textual objects. After performing the aforementioned two privacy enhancing algorithms, we have converted SKS queries to P-SKS queries. We aim to convert massive-scale SKS queries into P-SKS queries effectively and process P-SKS queries over a stream of geo-textual objects. This task faces the following three challenges.

- *Conversion efficacy.* Each SKS query is expected to be converted into P-SKS query on the basis of subscription privacy enhancing scheme. We need to effectively balance the trade-off between subscriber privacy and subscription accuracy while performing location privacy enhancing algorithm and text privacy enhancing algorithm.
- *Matching efficiency.* After conversion, each P-SKS query is expected to be matched with geo-textual objects that meet the subscription requirements in a real-time manner. The proposed matching algorithm is supposed to be capable of delivering matching geo-textual objects immediately to corresponding subscriptions.
- *Subscription scalability.* In real-life applications, it is very common that the publish/subscribe system need to handle a huge volume of P-SKS queries. The proposed publish/subscribe framework should be able to process at least millions of P-SKS queries simultaneously over a stream of geo-textual objects.

To address the challenges, we develop a Privacy-preserving Location-aware Publish/Subscribe (P²-LPS) framework to process a large number of SKS queries efficiently while ensuring the location and keyword privacy of subscribers. The P²-LPS framework consists of a two-stage subscription processing mechanism. The first stage is subscription privacy enhancement. During this stage, each incoming SKS query serves as input. We execute spatial privacy enhancement algorithms and text privacy enhancement algorithms to modify both the subscription region and keyword predicate of the query, respectively. Here, for each SKS query we aim to minimize the proportion of “unqualified matching” given distance shifting parameter δ_i and additional keyword parameter m . For this purpose, we devise a cost model to calculate the expected numbers of unqualified matching and qualified matching based on historical statistical data regarding geo-textual data streams. The second stage is subscription matching. During this stage, we take the underlying data stream and a collection of P-SKS queries, which are generated via subscription privacy enhancement of SKS queries, as input. For each incoming geo-textual object, we regard the object as “query” and find a subset of P-SKS queries that match the new object. To facilitate efficient matching between new object and P-SKS queries, we develop a P-SKS query indexing structure (P-index). The P-index is originally a Quad-tree where each tree node is associated to a rectangular spatial region, and is augmented by inverted file that indexes the keyword information of subscriptions whose regions overlap with the region associated to the node.

To sum up, the main contributions of this paper are summarized below.

- We define a new type of subscription query, P-SKS query that continuously receive geo-textual objects based on subscription location and keyword requirements over geo-textual data streams while addressing the subscribers' location and preference privacy concern.
- We propose a subscription privacy enhancing scheme consisting of location privacy enhancing algorithm, keyword privacy enhancing algorithm, and a purposeful subscription cost model to ensure subscribers' privacy by introducing spatial and keyword modifications of original STS queries while maintaining a low ratio of unqualified matching.
- We develop a Privacy-preserving Location-aware Publish/Subscribe (P²-LPS) framework that is capable of processing a large number of P-SKS queries efficiently in a real-time manner.
- We conduct comprehensive experiments with real-life datasets. Our experiment results verify the superiority of our proposal in terms of efficiency, scalability, and matching accuracy.

The remainders are organized as follows. Section 2 defines geo-textual object, SKS query, P-SKS query, and presents our problem formulation. Section 3 presents our P²-LPS framework, including the subscription privacy enhancement scheme, the P-index, and the matching algorithm. Section 4 presents our empirical study. Section 5 investigates related work. Section 6 presents our conclusion.

2. PROBLEM STATEMENT

This Section presents the definitions of geo-textual object, SKS query, P-SKS query, and the publish/subscribe problem we target.

Definition 1: Geo-textual object. A geo-textual object is defined by $o = \langle p, \psi, t \rangle$, where $o.p$ denotes geographical location, which is generally represented as a spatial coordinate with latitude and longitude, $o.\psi$ denotes a textual description, and $o.t$ denotes a timestamp indicating the publish time of o .

In real-life applications, geo-textual objects can be modeled as a data stream, such as tweets, local posts, reviews of commercial sites, and geo-tagged photos with descriptions.

Definition 2: Spatial Keyword Subscription (SKS) query. An SKS query is defined by $q = \langle r, w \rangle$, where $q.r$ denotes a subscription region and $q.w$ denotes a set of keywords connected by semantic "AND" or "OR". Given a geo-textual object o , we regard o as a matching of q (a.k.a., q matches o) iff: (1) $o.p \in q.r$, and (2) $o.\psi$ satisfies $q.w$.

According to existing studies, the subscription region is generally a circle or rectangle. However, our proposal is capable of handling regions of arbitrary shape. We proceed to present the definition of our P-SKS query.

Definition 3: Privacy-enhancing Spatial Keyword Subscription (P-SKS) query. A P-SKS query associated to SKS query q is defined by $p(q) = \langle r, \delta_r, w, m \rangle$, where r denotes the subscription region, δ_r denotes the

drifting distance of the center of subscription region, w denotes the keyword predicate, and m denotes the number of additional keywords in comparison to $q.w$. Given SKS query q , its P-SKS query $p(q)$ must satisfy the following conditions:

- (1) The distance between the center of $q.r$ and the center of $p(q).r$ is δ_r ,
- (2) $p(q).r$ fully covers $q.r$, and
- (3) $p(q).w$ contains $p(q).m$ more terms connected by “OR” semantic.

Theorem 1: Given a geo-textual object o , SKS query q and its P-SKS query $p(q)$, if q matches o then $p(q)$ matches o as well.

Proof Sketch: Regarding the spatial aspect, if q matches o , $q.r$ must cover $o.p$. Because $p(q).r$ covers $q.r$, we may deduce that $p(q).r$ must cover $o.p$. Regarding the textual aspect, if q matches o , $o.w$ must satisfy $q.w$. Because $p(q).w$ contains additional term(s) connected by “OR” semantic, we may deduce that $o.w$ satisfies $p(q).w$. Thus, we complete the proof.

Theorem 1 shows that $p(q)$ may never miss a matching result of q . In this paper, we investigate the problem of processing a large number of P-SKS queries over a stream of geo-textual objects.

3. THE P²-LPS FRAMEWORK WITH PRIVACY ENHANCEMENT

In this Section, we first present our subscription privacy enhancing scheme (cf. Section 3.1). Next, we present our subscription matching algorithm based on the P-SKS query indexing structure (cf. Section 3.2).

3.1 Subscription Privacy Enhancing Scheme

Our subscription privacy enhancing scheme comprises two components: location privacy enhancement and keyword privacy enhancement. The two components introduce minor modifications to the subscription region and subscription keywords of an SKS query, respectively, to safeguard the privacy of subscribers’ locations and preferences. Before presenting the details of our privacy enhancement scheme, we first introduce our relevant data structures to facilitate the privacy enhancement.

3.1.1 Location privacy enhancement

Given SKS query q , let us consider an example illustrated by Figure 2. At the beginning, we partition the underlying space into an $n \times n$ grid ($n = 8$ in Figure 2). Next, we record the number of arrived geo-textual objects located in each grid cell within a specified timespan, which is called *sliding window*. Consider the yellow and gray cells in Figure 2 as an example. by the figures to the right. Assume that the timespan of sliding window is set to ΔT , then we see that the yellow cell receives 1 object within the sliding window while the gray cell receives 4 objects. Likewise, we record the number of arrived objects for each grid cell. When a new object o arrives, based on the center drifting parameter δ_r we aim to generate a modified subscription region r' based on $q.r$. The modified region must satisfy the following three conditions: (1) r'

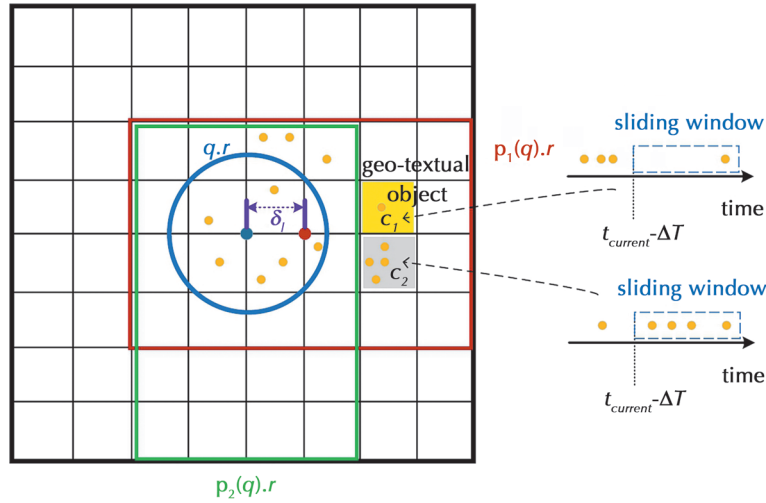


Figure 2. Grid Index for Location Privacy Enhancement.

is rectangular and consists of a collection of grid cells; (2) The distance between the center of $q.r$ and the center of r' is no less than δ_i ; (3) r' fully covers $q.r$.

Here, the red rectangle and the green rectangle, which are denoted by $p_1(q).r$ and $p_2(q).r$ respectively in Figure 2, are two valid examples of r' . Specifically, there are 14 geo-textual objects in total that fall inside the area of $p_1(q).r$ within the time sliding window. In contrast, only 9 objects fall inside the area of $p_2(q).r$ within the same timeframe. Thus, region $p_2(q).r$ is considered to be a more appropriate modified subscription region in comparison to $p_1(q).r$. The rationale is that $p_2(q).r$ is less likely to receive additional geo-textual objects that fall outside $q.r$ in the future compared with $p_1(q).r$. The location privacy enhancement aims to find an optimal valid modified region with the least chance of receiving unqualified objects.

3.1.2 Keyword privacy enhancement

Given SKS query q and parameter m indicating the number of additional keywords, the idea of keyword privacy enhancement is to generate a modified subscription predicate by adding m additional keywords to $q.w$ and connect them by “OR” semantic. Similarly, to reduce the occurrence of receiving irrelevant geo-textual objects following keyword predicate modification, we propose the following data structures:

Geo-textual Object Inverted File: We maintain an inverted file to index the text information of all geo-textual objects within the sliding window. This enables efficient retrieval of relevant objects based on their textual content, ensuring that only qualified objects are considered.

Semantic-Based Keyword Graph: Additionally, we employ a semantic-based keyword graph to facilitate the identification of keywords with similar semantic meanings. This graph is used for querying related keywords, thereby enhancing the relevance of retrieved geo-textual objects while minimizing the likelihood of including irrelevant ones.

By leveraging inverted file and keyword graph, we aim to optimize the keyword privacy enhancement process by reducing the occurrence of irrelevant geo-textual objects, thereby enhancing the overall effectiveness of the keyword predicate modification.

3.1.3 Privacy Enhancing Algorithm

We now present the details of our Privacy Enhancing (PE) algorithm. Algorithm 1 presents the pseudocode of our privacy enhancing algorithm. The algorithm takes original SKS query q , center drifting distance δ_r , number of additional keywords m , and grid index G as input. To begin with, we set a to be the center of original subscription region $q.r$ and set c to be the grid cell that covers a (Lines 1-2). Then we initialize I by the four grid intersection points associated to c (Line 3). Next, we create priority queue PQ to store grid intersection points sorted ascendingly by the distance between a and points, and set CP to store center point candidates (Lines 4-5). We add all points in I to PQ (Lines 6-7). After initialization, we iteratively select the top element p in PQ and calculate the distance between p and a . If the distance is smaller than δ_r , we expand our visit and retrieve adjacent grid intersection points of p and add them to PQ (Lines 12-15); Otherwise, we regard p as a center point candidate and add p to CP (Lines 16-17); After generating CP , for each center point candidate c_a in CP we evaluate its corresponding minimum rectangle-shaped grid cell set R centered at c_a . Specifically, we calculate the total number of geo-textual objects falling within the spatial region indicated by R over the time sliding window by summing up the object count associated to each grid cell of R (Lines 23-25). We find the grid cell set R_{min} that has the minimum sum of object count and regard R_{min} as the modified subscription region of q , which is denoted by $p(q).r$ (Lines 29-30). Next, we perform keyword privacy enhancement. We initialize set AW to be the m most similar words to keywords in $q.w$ (Line 31). For each word w_i in AW we connect the original subscription keyword predicate with w_i by “OR” semantic (Lines 32-33). Finally, we return $p(q)$, which is the privacy enhancing query of q , as the result (Line 34).

Algorithm 1. Privacy Enhancing (PE) algorithm.

Data: SKS query q , drifting distance δ_r , number of additional keywords m , Grid index G

Result: P-SKS query $p(q)$

- 1 $a \leftarrow$ the center of $q.r$;
 - 2 $c \leftarrow$ the cell in G that covers a ;
 - 3 $I \leftarrow$ grid intersection points associated to c ;
 - 4 Initialize PQ ;
-

Algorithm 1. Continued.

```

5   $CP \leftarrow \emptyset$ ;
6  for each  $p \in I$  do
7       $PQ.add(p)$ ;
8  while  $PQ$  is not empty do
9       $p \leftarrow PQ.top()$ ;
10      $PQ.pop()$ ;
11      $d \leftarrow Dist(p, a)$ ;
12     if  $d < \delta_i$  then
13          $N \leftarrow$  unvisited adjacent grid intersection points of  $p$ ;
14         for each  $p_i \in I$  do
15              $PQ.add(p_i)$ ;
16     else
17          $CP.add(p)$ ;
18      $N_{min} \leftarrow +\infty$ ;
19     Initialize  $R_{max}$ ;
20     for each  $c_a$  in  $CP$  do
21          $R \leftarrow$  minimum rectangle-shaped grid cell set centered at  $c_a$ ;
22          $N_R \leftarrow 0$ ;
23         for each grid cell  $c_i$  in  $R$  do
24              $n_i \leftarrow$  number of received objects within sliding window;
25              $N_R \leftarrow N_R + n_i$ ;
26         if  $N_R < N_{min}$  then
27              $N_{min} \leftarrow N_R$ ;
28              $R_{min} \leftarrow R$ ;
29      $p(q).r \leftarrow$  region of  $R_{min}$ ;
30      $p(q).w \leftarrow q.w$ ;
31      $AW \leftarrow m$  most similar words to keywords in  $q.w$ ;

```

Algorithm 1. *Continued.*

```

32  for each word  $w_i$  in  $AW$  do
33       $p(q).w.concat("OR"+w_i);$ 
34  return  $p(q);$ 

```

3.2 Matching between Queries and Geo-textual Objects

We now present our matching process between P-SKS queries and incoming geo-textual objects. The matching algorithm is designed to efficiently handle incoming geo-textual objects from the data stream and identify a subset of P-SKS queries whose subscription criteria are met by the incoming object. Before presenting the matching algorithm, we first introduce our P-index, which lays foundation for our matching algorithm.

3.2.1 P-index

The P-index serves as an indexing structure tailored for efficiently managing a large volume of P-SKS queries. It operates within both spatial and textual domains:

Spatial domain. The spatial domain of the P-index is represented by a Quad-tree, where each node corresponds to a rectangular spatial region. Registration Process: Upon registration of a new P-SKS query q , the Quad-tree is traversed in a top-down manner to locate the node at the lowest level whose spatial region entirely encompasses the spatial region of q . This ensures accurate spatial indexing of queries.

Textual domain. The textual domain of the P-index employs an inverted file, where each inverted list is associated with a specific keyword. It is worth mentioning that we apply different ways to handle semantic connections. To be specific, if the subscription keyword predicate of a query is connected by the “OR” semantic, the query ID is stored in the inverted lists corresponding to all keywords within the predicate. Conversely, if the subscription keyword predicate is connected by the “AND” semantic, the query ID is stored in only one inverted list associated with any of the keywords. By leveraging this combined spatial and textual indexing approach, the P-index efficiently organizes and manages P-SKS queries, facilitating rapid query processing and retrieval based on both spatial and textual criteria.

3.2.2 Matching algorithm

Algorithm 2 presents the pseudocode of our matching algorithm. The inputs are the incoming geo-textual object o from the data stream, the root node n_{root} of the P-tree, and the P-SKS query set Q . The output is the set of P-SKS queries Q_r whose spatial and keyword requirements match o . At the beginning, we set n to be the root node of the P-index, which is a Quad-tree structure in spatial domain (Line 1). Then we evaluate the queries indexed under n by visiting the inverted lists associated to the words in $o.\psi$ (Line 4). For each visited query q in inverted lists, we check whether $q.r$ covers $o.p$. If $q.r$ covers $o.p$, we

Algorithm 2. Matching Algorithm.

Data: Geo-textual object o , Root node $n_{root'}$, Indexed P-SKS query set Q
Result: Matched query set Q_r

```

1   $n \leftarrow n_{root'}$ ;
2  do
3      for each keyword  $w$  in  $o.\psi$  do
4           $CQ \leftarrow \text{InvertedList}(n, w)$ ;
5          for each query  $q$  in  $CQ$  do
6              if  $o.p$  is covered by  $q.r$  then
7                   $Q_r.\text{add}(q)$ ;
8              if  $n$  is not a leaf node then
9                  for each child node  $n_c$  of  $n$  do
10                     if  $n_c$  covers  $o.p$  then
11                          $n \leftarrow n_c$ ;
12 while  $n$  is not a leaf node;
13 return  $Q_r$ ;

```

regard q as a matching of o and add q into the matching result set Q_r (Lines 5-7). Next, we evaluate the child node of n whose region covers $o.p$ and perform the same evaluation until we reach the leaf node. Finally, we return Q_r as the matching result query set (Line 13).

4. EXPERIMENT

This Section presents our experimental results and analyses the efficiency, efficacy, and scalability of baseline and our proposed method.

4.1 Experimental Settings

4.1.1 Datasets

We use the following two real-world datasets in our experiments.

FQ dataset. The FQ dataset is collected from Foursquare, which is comprised of 1 million world-wide check-ins. Each check-in is formulated by a spatial point and a short textual description.

XT dataset. The XT dataset is collected from X (formerly known as Twitter). The XT contains 5 million tweets with location information. Specifically, each tweet is represented by a spatial coordinate and a short text.

We consider all check-ins in FQ and tweets in XT to be geo-textual objects. We generate original SKS queries as follow. We first randomly select geo-textual object o from FQ or XT. Next, we regard $o.p$ as the center of the subscription region. The radius of region is determined based on our specific setting, which can be either a random value or a pre-defined value. Finally, we select a pre-defined number of words from $o.\psi$ and regard them as subscription keywords.

Our parameter settings are presented in Table 1.

Table 1. Parameter Settings.

Parameter	Value	Default
Number of Subscription Keywords	1–5	random 1–5
Radius of Subscription Region	1km–100km	random 1km–100km
Center Drifting Parameter δ_i	1km–9km	5km
Number of Additional Keywords m	1–4	1
Number of SKS Queries	depending on scenarios	FQ 500K XT 2M

4.1.2 Baseline

We use Direct Subscription Matching (DSM) as our baseline. The DSM algorithm does not have indexing structure. When a new geo-textual object from the data stream arrives, we traverse the P-SKS query collection and check if the new object satisfies the spatial and keyword requirements of each P-SKS query. If the new object satisfies the corresponding requirements indicated by a P-SKS query, we add the query into the matching list associated to the new object.

4.2 Experimental Result

4.2.1 Varying the number of subscription keywords

First, we evaluate the effect of the number of subscription keywords in subscription keyword predicates. From Figure 3, we see that when we increase the number of keywords from 1 to 5, the runtime of both DSM and PPLPS increases. The reason is that more subscription keywords incurs more checking time in textual domain. We also find that the runtime increasing trend of PPLPS is less significant in comparison against DSM. This can be explained by the fact that due to the indexing scheme applied by the P-index in textual domain, keyword predicates connected by “AND” semantic may not induce additional checking

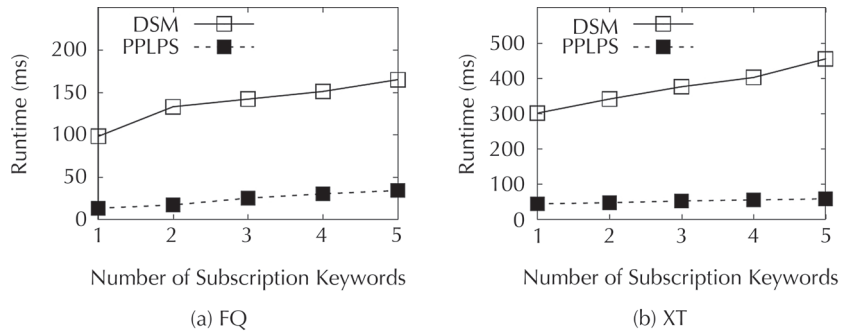


Figure 3. Effect of the number of subscription keywords.

cost. In addition, our inverted file is capable of filtering out P-SKS queries that do not have textual overlap with each incoming geo-textual object at an early stage.

4.2.2 Varying the radius of subscription region

This set of experiments evaluate the runtime as we vary the radius of subscription region. From Figure 4, we see that when we increase the radius of subscription region from 1km to 100km, the runtime costs both algorithms are relatively stable. In particular, PPLPS performs consistently better than DSM by a factor of 4-8.

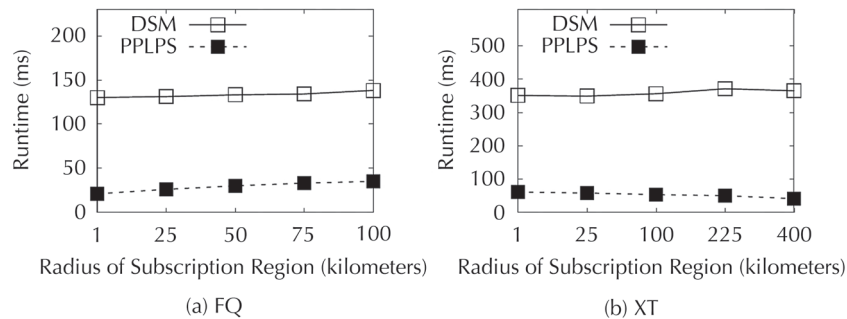


Figure 4. Effect of subscription region radius.

4.2.3 Varying the center drifting distance of subscription region

Next, we evaluate the ratio of qualified results as we vary the center drifting distance, parameter δ_i . A higher value of δ_i indicates a higher deviation of the subscription region of P-SKS query in comparison to the region of its original SKS query. Given an SKS query q and its P-SKS query $p(q)$, denoted by $R(p(q))$ the ratio of qualified results is computed by Equation 1.

$$R(p(q)) = \frac{|M(q)|}{|M(p(q))|}, \quad (1)$$

where $|M(q)|$ denotes the number of matching geo-textual objects received by q and $|M(p(q))|$ denotes the number of matching geo-textual objects received by $p(q)$. From Figure 5, we observe that when we change δ_i from 1km to 100km, the ratio of qualified results of PPLPS exhibits a slight decreasing trend, underlying the effectiveness of our privacy enhancing scheme.

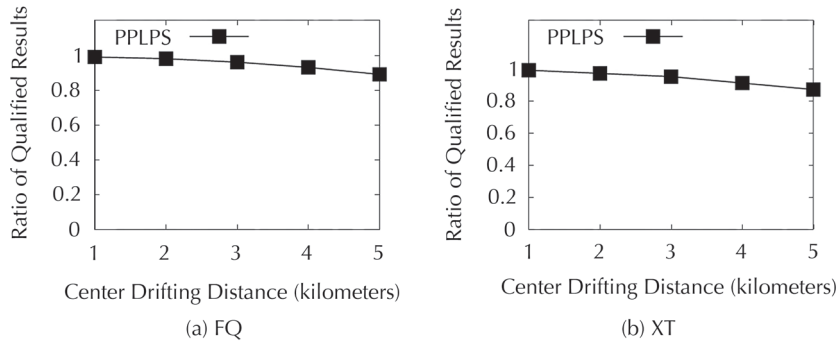


Figure 5. Effect of center drifting distance of subscription region.

4.2.4 Varying the number of additional subscription keywords

We proceed to evaluate the efficacy performance when we vary the number of additional subscription keywords, which is denoted by parameter m . From Figure 6 we see that when we increase m from 1 to 5, the performance of PPLPS exhibits a moderate decreasing trend. The reason is that a higher value of m indicates a higher deviation of the subscription keyword predicate of P-SKS query in comparison to the keyword predicate of its original SKS query.

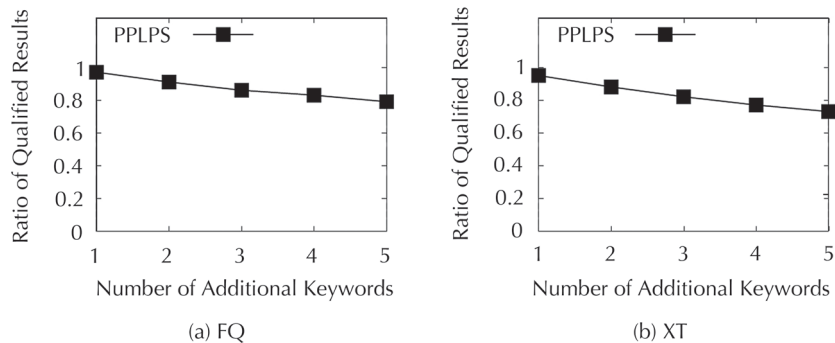


Figure 6. Effect of the number of additional keywords.

4.2.5 Varying the number of P-SKS queries

Finally, we evaluate the scalability of DSM and PPLPS. Figure 7 demonstrates that when we increase the number of P-SKS queries from 500K to 2M and from 2M to 8M respectively on FQ and XE, the runtime of DSM exhibits a linear increasing trend. In contrast, the performance of PPLPS remains consistent. This is resulted from the pruning power of the P-index, which has the capability of filtering out most unqualified query candidates.

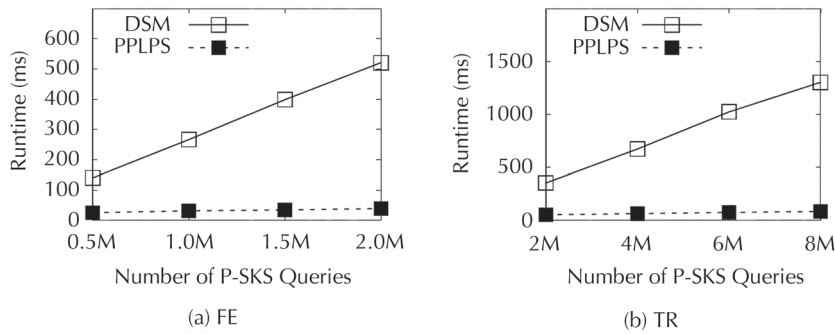


Figure 7. Effect of the number of P-SKS queries.

5. RELATED WORK

The problem of privacy-preserving location-aware publish/subscribe can be regarded as an optimization of traditional location-based publish/subscribe, location-based search, and spatial keyword query processing.

5.1 Location-based Publish/Subscribe

Although location-based publish/subscribe can be defined in different forms, it contains two major components: “publisher” and “subscriber”. A publisher can be modeled as a data stream. In the context of location-based publish/subscribe, each object in the data stream consists of spatial, textual, and temporal information. It is often called geo-textual object or spatio-temporal object. A subscriber can be modeled as a continuous query. In the context of location-based publish/subscribe, each continuous query consists of both location and keyword predicates. Specifically, a location predicate can be a spatial region [10, 11, 12, 13, 14], a similarity constraint [15], or a “top- k ” similarity ranking function [16]. Note that some publish/subscribe frameworks apply a similarity measurement that combines both spatial proximity and text relevancy [17], others regard one domain as a hard constraint and the other domain as an optimization criteria [18, 19, 20]. In addition, the problems of continuous search over moving objects [21, 22], continuous trajectory search [23, 24], and trajectory similarity computation [25, 26, 27] can be served as key components in location-based publish/subscribe systems. However, the aforementioned existing studies do not consider the privacy concerns of location-based subscribers.

5.2 Privacy Preserving Spatial Keyword Query Processing

As far as we are aware, there is no current research that specifically focuses on the problem of location-aware publish/subscribe systems while also taking into account the privacy of user-provided location and keyword information. Existing methods for processing privacy-preserving spatial keyword queries can be categorized based on their approaches to addressing privacy concerns as follows. The first category of methods adopts encrypted spatial indexing structures to ensure data privacy. To be specific, Su et al. [3, 4] propose encrypted tree indices to facilitate privacy-preserving top- k spatial keyword queries. Chen et al. [5] combine the R-tree and Power Diagram indexes to handle location-based top- k queries. Yiu et al. [6] propose a method that involves a spatial transformation that re-distributes locations within the underlying space. Additionally, they develop a cryptographic-based transformation for enhanced security. Furthermore, blockchains [7, 8] and differential privacy technologies [9] can be used to further enhance the user privacy. Although the mentioned frameworks for privacy-preserving spatial keyword search successfully mitigate privacy issues pertaining to user location and keywords, they lack efficiency and scalability, particularly within the context of publish/subscribe systems.

5.3 General Spatial Keyword Search

The problem of spatial keyword query processing has been extensively investigated over the past couple of decades. Spatial keyword search can be regarded as a fundamental operation for downstream tasks, including location-based recommendations [28, 29, 30, 31, 32], trajectory data analytics [27, 33, 34, 35, 36], etc. A traditional spatial keyword query comprises a spatial predicate and a keyword predicate. To be specific, the spatial predicate can be either a spatial range constraint or a spatial proximity constraint. Keyword predicate is generally formulated as a Boolean predicate or simply a set of words. Existing studies regarding spatial keyword search generally consist of two phases: object indexing phase, and query processing phase. In the object indexing phase, a collection of geo-textual objects are indexed by a hybrid indexing structure, which effectively organize geo-textual objects based on their spatial and textual domains [37, 38, 39, 40, 41, 42, 43]. Some comprehensive surveys investigate the comparison and contrast of different types of hybrid indices [44, 45]. However, the aforementioned studies regard the input data as a static collection of geo-textual objects. They do not have the capability of processing a stream of geo-textual objects.

6. CONCLUSION AND FUTURE DIRECTIONS

To the end, we define and study the problem of privacy-preserving location-based publish/subscribe. Based on a widely applied location-based subscription query type, SKS query, we develop a subscription privacy enhancing scheme consisting of location privacy enhancement and keyword privacy enhancement. Based on the privacy enhancing scheme, we convert each SKS query into a privacy-enhancing SKS query, the P-SKS query, which avoids the disclosure of subscriber's exact location and keyword information while keeping subscriber's subscription requirements to a great

extent. To process massive-scale P-SKS queries efficiently over a stream of geo-textual objects, we develop the P-index and corresponding matching algorithm to facilitate efficient matching between new object and P-SKS queries. Extensive experiment results show that our proposal is capable of achieving high efficiency, scalability, and accuracy of publishing. The proposed privacy-preserving publish/subscribe framework could fundamentally enhance a broad spectrum of location-based applications. This includes the provision of high-quality, personalized user data for the pre-training of Large Language Models (LLMs), facilitating real-time, localized searches for personalized trending topics, and offering real-time spatio-temporal data cleaning for location-based services while maintaining confidentiality. These capabilities under-score the multifaceted utility of our research in improving the efficacy and privacy of location-based technologies.

In the development of location-based publish/subscribe systems and real-time location-aware query processing mechanisms, balancing the trade-off between privacy protection and efficiency of query processing is of paramount importance. In future work, we plan to design a more precise model to find the optimal balance between these two aspects. Developing continuous query processing algorithms that can both protect user privacy and maintain high efficiency and accuracy is key to building user trust and ensuring the system's sustainable operation.

AUTHOR CONTRIBUTIONS

Ya Gao: index design, algorithm design, solution implementation, and paper writing

Qiuyuan Ai: baseline implementation and paper writing

Jie Song: problem definition and project coordination

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China under Grant No.72131001.

REFERENCES

- [1] Shraer, A., Gurevich, M., Fontoura, M., Josifovski, V.: Top-k publish-subscribe for social annotation of news. *Proceedings of the VLDB Endowment* 6(6), 385–396 (2013)
- [2] Shubham Singh: Twitter (X) User Statistics 2024. Available at: <https://www.demandsage.com/twitter-statistics> (2024). Accessed 3 March 2024
- [3] Su, S., Teng, Y., Cheng, X., Wang, Y., Li, G.: Privacy-preserving top-k spatial keyword queries over outsourced database. In: *Database Systems for Advanced Applications*, pp. 589–608 (2015)
- [4] Su, S., Teng, Y., Cheng, X., Xiao, K., Li, G., Chen, J.: Privacy-preserving top-k spatial keyword queries in untrusted cloud environments. *IEEE Transactions on Services Computing* 11(5), 796–809 (2018)
- [5] Chen, Q., Hu, H., Xu, J.: Authenticating top-k queries in location-based services with confidentiality. *Proceedings of the VLDB Endowment* 7(1), 49–60 (2013)

- [6] Yiu, M. L., Ghinita, G., Jensen, C. S., Kalnis, P.: Enabling search services on outsourced private spatial data. *The International Journal on Very Large Data Bases* 19(3), 363–384 (2010)
- [7] Xu, C., Zhang, C., Xu, J.: vchain: Enabling verifiable boolean range queries over blockchain databases. In: *ACM SIGMOD International Conference on Management of Data*, pp. 141–158 (2019)
- [8] Wang, H., Xu, C., Zhang, C., Xu, J., Peng, Z., Pei, J.: vchain+: Optimizing verifiable blockchain Boolean range queries. In: *IEEE International Conference on Data Engineering*, pp. 1927–1940 (2022)
- [9] Liu, Z., Liu, Q., Wei, J., Miao, D., Wang, J.: Location privacy-preserving query scheme based on the moore curve and multi-user cache. *Information* 13(9), 417 (2022)
- [10] Li, G., Wang, Y., Wang, T., Feng, J.: Location-aware publish/subscribe. In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 802–810 (2013)
- [11] Chen, L., Cong, G., Cao, X.: An efficient query indexing mechanism for filtering geo-textual data. In: *ACM SIGMOD International Conference on Management of Data*, pp. 749–760 (2013)
- [12] Chen, L., Cui, Y., Cong, G., Cao, X.: SOPS: A system for efficient processing of spatial-keyword publish/subscribe. *Proceedings of the VLDB Endowment* 7(13), 1601–1604 (2014)
- [13] Wang, X., Zhang, Y., Zhang, W., Lin, X., Wang, W.: Ap-tree: efficiently support location-aware publish/subscribe. *The International Journal on Very Large Data Bases* 24(6), 823–848 (2015)
- [14] Chen, Z., Cong, G., Zhang, Z., Fu, T. Z. J., Chen, L.: Distributed publish/subscribe query processing on the spatio-textual data stream. In: *IEEE International Conference on Data Engineering*, pp. 1095–1106 (2017)
- [15] Hu, H., Liu, Y., Li, G., Feng, J., Tan, K.: A location-aware publish/subscribe framework for parameterized spatio-textual subscriptions. In: *IEEE International Conference on Data Engineering*, pp. 711–722 (2015)
- [16] Chen, L., Cong, G., Cao, X., Tan, K.: Temporal spatial-keyword top-k publish/subscribe. In: *IEEE International Conference on Data Engineering*, pp. 255–266 (2015)
- [17] Wang, X., Zhang, W., Zhang, Y., Lin, X., Huang, Z., 2017. Top-k spatial-keyword publish/subscribe over sliding window. *The International Journal on Very Large Data Bases* 26(3), 301–326 (2017)
- [18] Chen, L., Shang, S., Jensen, C. S., Xu, J., Kalnis, P., Yao, B., Shao, L.: Top-k term publish/subscribe for geo-textual data streams. *The International Journal on Very Large Data Bases* 29(5), 1101–1128 (2020)
- [19] Chen, L., Shang, S.: Approximate spatio-temporal top-k publish/subscribe. *World Wide Web* 22(5), 2153–2175 (2019)
- [20] Chen, L., Shang, S., Zhang, Z., Cao, X., Jensen, C. S., Kalnis, P.: Location-aware top-k term publish/subscribe. In: *IEEE International Conference on Data Engineering*, pp. 749–760 (2018)
- [21] Zhang, J., Yu, J., Shang, S., Chen, L., Feng, S.: Continuous frequent contact detection over moving objects. *GeoInformatica* 28(2), 271–290 (2024)
- [22] Li, K., Chen, L., Shang, S., Wang, H., Liu, Y., Kalnis, P., Yao, B.: Towards controlling the transmission of diseases: Continuous exposure discovery over massive-scale moving objects. In: *International Joint Conferences on Artificial Intelligence*, pp. 3891–3897 (2022)
- [23] Chen, L., Shang, S., Feng, S., Kalnis, P.: Parallel subtrajectory alignment over massive-scale trajectory data. In: *International Joint Conferences on Artificial Intelligence*, pp. 3613–3619 (2021)
- [24] Chen, L., Shang, S., Jensen, C. S., Yao, B., Kalnis, P.: Parallel semantic trajectory similarity join. In: *IEEE International Conference on Data Engineering*, pp. 997–1008 (2020)
- [25] Chen, Z., Zhang, D., Feng, S., Chen, K., Chen, L., Han, P., Shang, S.: KGTS: contrastive trajectory similarity learning over prompt knowledge graph embedding. In: *AAAI Conference on Artificial Intelligence*, pp. 8311–8319 (2024)
- [26] Chen, Z., Li, K., Zhou, S., Chen, L., Shang, S.: Towards robust trajectory similarity computation: Representation-based spatio-temporal similarity quantification. *World Wide Web (WWW)* 26(4), 1271–1294 (2023)

- [27] Zhou, S., Li, J., Wang, H., Shang, S., Han, P.: GRLSTM: trajectory similarity computation with graph-based residual LSTM. In: AAAI Conference on Artificial Intelligence, pp. 4972–4980 (2023)
- [28] Liu, Y., Pham, T. N., Cong, G., Yuan, Q.: An experimental evaluation of point-of-interest recommendation in location-based social networks. *Proceedings of the VLDB Endowment* 10(10), 1010–1021 (2017)
- [29] Yin, F., Liu, Y., Shen, Z., Chen, L., Shang, S., Han, P.: Next POI recommendation with dynamic graph and explicit dependency. In: AAAI Conference on Artificial Intelligence, pp. 4827–4834 (2023)
- [30] Han, P., Shang, S., Sun, A., Zhao, P., Zheng, K., Zhang, X.: Point-of-interest recommendation with global and local context. *IEEE Transactions on Knowledge and Data Engineering* 34(11), 5484–5495 (2022)
- [31] Rao, X., Chen, L., Liu, Y., Shang, S., Yao, B., Han, P.: Graph-flashback network for next location recommendation. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1463–1471 (2022)
- [32] Han, P., Li, Z., Liu, Y., Zhao, P., Li, J., Wang, H., Shang, S.: Contextualized point-of-interest recommendation. In: International Joint Conferences on Artificial Intelligence, pp. 2484–2490 (2020)
- [33] Yang, C., Chen, L., Wang, H., Shang, S.: Towards efficient selection of activity trajectories based on diversity and coverage. In: AAAI Conference on Artificial Intelligence, pp. 689–696 (2021)
- [34] Shang, S., Chen, L., Zheng, K., Jensen, C. S., Wei, Z., Kalnis, P.: Parallel trajectory-to-location join. *IEEE Transactions on Knowledge and Data Engineering* 31(6), 1194–1207 (2019)
- [35] Shang, S., Chen, L., Wei, Z., Jensen, C. S., Zheng, K., Kalnis, P.: Parallel trajectory similarity joins in spatial networks. *The International Journal on Very Large Data Bases* 27(3), 395–420 (2018)
- [36] Shang, S., Chen, L., Jensen, C. S., Wen, J., Kalnis, P.: Searching trajectories by regions of interest. *IEEE Transactions on Knowledge and Data Engineering* 29(7), 1549–1562 (2017)
- [37] Felipe, I. D., Hristidis, V., Risse, N.: Keyword search on spatial databases. In: ICDE, pp. 656–665 (2008)
- [38] Cong, G., Jensen, C. S., Wu, D.: Efficient retrieval of the top-k most relevant spatial web objects. In: *Proceedings of the VLDB Endowment*, pp. 337–348 (2009)
- [39] Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., Nørnvåg, K.: Efficient processing of top-k spatial keyword queries. In: International Symposium on Spatial and Temporal Databases, pp. 205–222 (2011)
- [40] Zhang, D., Tan, K.-L., Tung, A. K. H.: Scalable top-k spatial keyword search. In: International Conference on Extending Database Technology, pp. 359–370 (2013)
- [41] Zhang, C., Zhang, Y., Zhang, W., Lin, X.: Inverted linear quadtree: Efficient top k spatial keyword search. In: IEEE International Conference on Data Engineering, pp. 901–912 (2013)
- [42] Yang, C., Chen, L., Shang, S., Zhu, F., Liu, L., Shao, L.: Toward efficient navigation of massive-scale geo-textual streams. In: International Joint Conferences on Artificial Intelligence, pp. 4838–4845 (2019)
- [43] Li, M., Chen, L., Cong, G., Gu, Y., Yu, G.: Efficient processing of location-aware group preference queries. In: ACM International Conference on Information and Knowledge Management, pp. 559–568 (2016)
- [44] Chen, L., Cong, G., Jensen, C. S., Wu, D.: Spatial keyword query processing: an experimental evaluation. In: *Proceedings of the VLDB Endowment*, pp. 217–228 (2013)
- [45] Chen, L., Shang, S., Yang, C., Li, J.: Spatial keyword search: a survey. *Geoinformatica* 24(1), 85–106 (2020)

AUTHOR BIOGRAPHY

Ya Gao is currently pursuing a Doctor of Engineering degree in Peking University.
E-mail: gaoya@stu.pku.edu.cn; ORCID: 0000-0002-9785-922X

Qiuyaun Ai is currently pursuing a Master's Degree in Peking University.
E-mail: 2101213262@stu.pku.edu.cn; ORCID: 0000-0002-1192-7675

Jie Song is currently a Ph.D. advisor in the Department of Industrial Engineering and Management of the College of Engineering at Peking University. She is also a Distinguished Professor under the Ministry of Education's "Changjiang Scholars Program." Prof. Song graduated with a bachelor's degree in Applied Mathematics at Peking University in 2004. She then obtained her Ph.D. in Management Science and Engineering from the Department of Industrial Engineering at Tsinghua University in 2010. Prof. Song has conducted postdoctoral and visiting scholar research at several prestigious institutions, including Georgia Tech, the University of Wisconsin-Madison, and Columbia University. Her research focuses on stochastic optimization modeling and algorithm design, and data privacy. Her research has been supported by grants from the National Natural Science Foundation of China, the Ministry of Science and Technology, and other agencies. She has published over 60 papers in internationally renowned SCI journals and major international conferences. In addition to her research, Prof. serves as an editor for three international SCI journals, including IEEE Automation Science and Engineering. She is also the chair of the IEEE RAS International Medical Automation Management Committee.
E-mail: jie.song@pku.edu.cn; ORCID: 0000-0002-4916-8272