

ETGP: Top-K Geography-Text P/S Approach without Threshold

Zongmin CUI, Hongbo LI, Hong ZHU

Abstract: Social media are more and more popular. Subsequently, geography-text data has caused wide attention. Different from the traditional publish/subscribe (P/S), geography-text data is published and subscribed in the form of dynamic data flow in the mobile network. The difference raises higher demands for facility. However, previous top-k geography-text P/S approaches want to set a set of thresholds. A user should take time to set a threshold for each subscription, which is not facile enough. The threshold yields many weaknesses to users. Therefore, we herein propose an efficient top-k geography-text P/S approach that excludes the threshold, called ETGP. Our approach does not need users to set any threshold. Subsequently, the ETGP returns the highest score results to the subscriber without setting a threshold. Therefore, our approach can lessen redundant computations, promote the query integrity rate, and make P/S system easier for the user to use. Comprehensive experiments prove the efficiency of the proposed approach with high facility.

Keywords: geography-text data query; none threshold; P/S; publish/subscribe system; top-k highest score

1 INTRODUCTION

Nowadays, with the popularity of intelligence devices equipped with GPS (Global Position System) receivers [1], a large amount of geography-text data is published and subscribed in the form of dynamic data flow in the mobile network [2]. Subsequently, geography-text P/S systems have attracted widespread attention [3]. Hence, the geography-text P/S has been used in lots of mobile network scenarios, such as social networks and geography-based product [4].

In a geography-text P/S system, subscribers (i.e. users, data visitors, buyers, etc.) register their interests as geography-text subscriptions (for example, the "food promotion" near the subscriber's location "ShanghaiNorth Street") [5]. Then the system will send information quickly and accurately (such as "pizza 20% discount in Shanghai North Street No. 18 pizza store") to the relevant subscribers [6]. The information is published by publishers (i.e. data owners, data providers, sellers, etc.) [6]. The mobile computing resource of the user is valuable. In general, we only return the top-k results to the user. This returning way has become the typical approach of data query.

When handling top-k issues, previous geography-text P/S approaches [7, 8] need users to set a threshold τ . The threshold is the bridge between the geography and the text. For the deficiency, here are some examples to illustrate this more explicitly.

(1) Too large threshold. For example, a subscriber only wants to return a publication with the highest score of geography-text matching (that is, return to the Top-1 publication). However, the system needs the subscriber to set a threshold between 0 and 1. It is very difficult to set a suitable threshold. If the subscriber has set a two large threshold, the system computes 10,000 results based on the threshold. Then the system returns the highest score publication to the subscriber. Obviously, in this process, the extra 9,999 intermediate results waste the valuable computing resources of the system.

(2) Too small threshold. For example, a subscriber wants to get Top-1000 publications. However, the subscriber sets a too small threshold. Then the system computes only 5 results based on the threshold, which misses the other 995 results. Obviously, in this process, the

system misses many results. The missing reduces the system's accuracy, which cannot meet subscriber's actual requirements.

In short, threshold is both inexplicable and unintuitive. This makes it very difficult for an ordinary subscriber to set an appropriate threshold. Thus, threshold reduces the facility of the P/S system.

Existing approaches [7, 8] need to set the threshold k (i.e., the number of returned highest score results) and preference parameter δ (the decision of whether the geography or text is more important). To remove the above problem, we propose a top-k geography-text P/S approach that excludes the threshold (called ETGP). Our approach permits users to input only k and δ . Subsequently, P/S system returns the top-k results based on only k and δ . Therefore, the ETGP can lessen the redundant computing, promote the query integrity rate, and make P/S system easier for the user to use.

The remainder of this paper is organized as follows. Section 2 provides the basic concepts. Section 3 demonstrates our idea. Section 4 shows the experiment results. Section 5 discusses the related works. Finally, section 6 concludes the paper.

2 BASIC CONCEPTS

To accurately and formally define our problem, we provide the following definitions of ETGP.

Definition 1. (Subscription). A subscription is denoted as $s = (s_T, s_G, \delta, k)$. (1) In above equation, s_T denotes s 's text description which is composed of a set of keywords $\{t_1, t_2, \dots, t_{|s_T|}\}$. Each keyword t_i is associated with a weight $w(t_i)$. $w(t_i)$ is TF-IDF [9] weight of keyword t_i in s . (2) s_G denotes s 's geography description which is composed of latitude and longitude. (3) δ is a preference parameter. If $\delta > 0.5$, text description is more important than geography description. (4) k is the number of returned results of subscription s .

Definition 2. (Publication). A publication is denoted as $p = (p_T, p_G)$. (1) In above equation, p_T denotes p 's text description which is composed of a set of keywords $\{t_1, t_2, \dots, t_{|p_T|}\}$. Each keyword t_i is associated with a weight $w(t_i)$. $w(t_i)$ is TF-IDF [9] weight of keyword t_i in p . (2) p_G denotes

p 's geography description which consists of longitude and latitude.

Example 1. Tab. 1 provides an example with a subscription set and a publication set.

Table 1 The example of subscription set and publication set

Subscription		Publication	
Name	Content	Name	Content
s_0	$(s_{T_0}, s_{G_0}, 0.8, 2)$	p_0	(p_{T_0}, p_{G_0})
s_1	$(s_{T_1}, s_{G_1}, 0.3, 1)$	p_1	(p_{T_1}, p_{G_1})
s_2	$(s_{T_2}, s_{G_2}, 0.5, 2)$	p_2	(p_{T_2}, p_{G_2})
s_3	$(s_{T_3}, s_{G_3}, 0.6, 3)$	p_3	(p_{T_3}, p_{G_3})
s_4	$(s_{T_4}, s_{G_4}, 0.7, 5)$	p_4	(p_{T_4}, p_{G_4})
s_5	$(s_{T_5}, s_{G_5}, 0.4, 4)$	p_5	(p_{T_5}, p_{G_5})
s_6	$(s_{T_6}, s_{G_6}, 0.2, 1)$	p_6	(p_{T_6}, p_{G_6})
s_7	$(s_{T_7}, s_{G_7}, 0.1, 3)$	p_7	(p_{T_7}, p_{G_7})

A geography-text P/S system needs to return each publication to related subscribers in a timely manner. Thus, we quantitatively define the similarity between subscriptions and publications as follows.

Definition3. (Geography Similarity). The geography similarity, $GSIM(s, p)$ between a subscription s , and a publication p , is defined as Eq. (1).

$$GSIM(s, p) = MAX\left(0, 1 - \frac{DIST(s_G, p_G)}{MAXDIST}\right) \quad (1)$$

In Eq. (1), $DIST(s_G, p_G)$ is the Euclidian distance between s_G and p_G . $MAXDIST$ denotes the max Euclidean distance that the subscriber can tolerate (The publication that exceeds the distance will not be related to the subscription) [2].

Definition 4. (Text Similarity). The text similarity, $TSIM(s, p)$ between a subscription s , and a publication p , is defined as Eq. (2).

$$TSIM(s, p) = \frac{\sum_{t \in s_T \cap p_T} w(t)}{\sum_{t \in s_T} w(t)} \quad (2)$$

In Eq. (2), $w(t)$ is the weight of keyword t .

Definition 5. (Geography-Text Similarity). The geography-text similarity $SIM(s, p)$ between a subscription s , and a publication p , is defined as Eq. (3).

$$SIM(s, p) = \delta \times TSIM(s, p) + (1 - \delta) \times GSIM(s, p) \quad (3)$$

In Eq. (3), $\delta \in [0, 1]$ is a preference parameter. If δ is bigger than 0.5, the weight of text similarity is heavier than geography similarity.

In the next sections, $SIM(s, P)$ represents the geography-text similarity set between a subscription s , and a publications set P .

Example 2. Tab. 2 shows the geography-text similarity set $SIM(s_2, P)$ between a subscription s_2 and all publications.

Table 2 Examples of geography-text similarities based on Tab. 1

Geography-Text Similarity	Value	Geography-Text Similarity	Value
$SIM(s_2, p_0)$	0.2	$SIM(s_2, p_1)$	0.4
$SIM(s_2, p_2)$	0.9	$SIM(s_2, p_3)$	0.5
$SIM(s_2, p_4)$	0.3	$SIM(s_2, p_5)$	0.6
$SIM(s_2, p_6)$	0.1	$SIM(s_2, p_7)$	0.8

3 P/S SYSTEM

The System structure of ETGP is shown in Fig. 1. Publishers (i.e. data owners, data providers, sellers, etc.) upload their publications to the system (i.e. smart phone, i-pad, internet, cloud, etc.). Subscribers (i.e. users, data visitors, buyers, etc.) upload their subscriptions to the system. Then the system computes the set of geography-text similarities. Finally, the system returns the top-k results (that have the highest k geography-text similarity scores) to the subscriber. The subscriber can get the top-k results wherever he/she is. For example, he/she can sit in a plane, train, car, ship, etc.

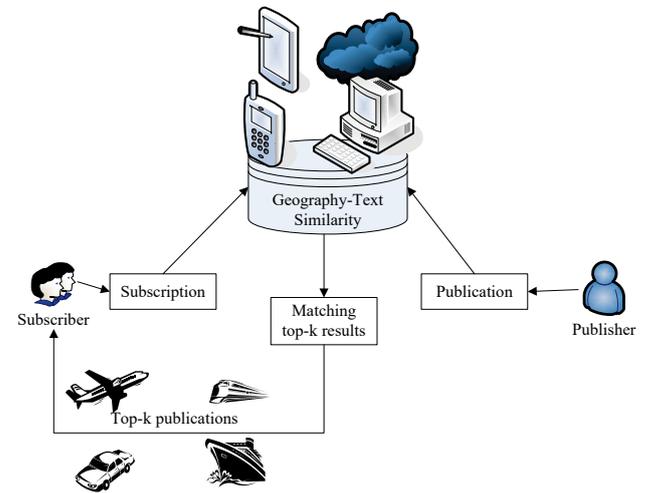


Figure 1 System structure

The "algorithm 1 Query" regulates the P/S query process of the ETGP. The algorithm takes a subscription s and the geography-text similarity set $SIM(s, P)$ as the input. Meanwhile, the algorithm takes the top-k query results N_k as the output.

If N has the maximum geography-text similarity with s , we insert N into N_k (Steps 3-8). Subsequently, algorithm 1 removes N from $|SIM(s, P)|$ (Steps 9-11).

Similarly, the algorithm next determines the maximum value among the remaining $|SIM(s, P)|$ (Steps 5-6). At last, we insert the k highest score publications into N_k . The k highest score publications are the top-k results. These results have the highest score of geography-text similarities.

Algorithm 1: Query	
Input: $s, SIM(s, P)$	
Output: N_k	
1:	$N_k := \phi$
2:	For all $i \in [1, k]$ do
3:	$N := SIM(s, p_i)$
4:	For all $j \in [2, SIM(s, P)]$ do
5:	If $SIM(s, p_j) > SIM(s, p_i)$ then
6:	$N := SIM(s, p_j)$

7:	End if
8:	End for
9:	Let the publication of N is p_m
10:	$N_k := N_k \cup p_m$
11:	$SIM(s, P) := SIM(s, P)/N$
12:	End for
13:	Return (N_k)

Example 3. Based on Tab. 1 and Tab. 2, we perform $Query(s_2, SIM(s_2, P))$. The top-k result set is $N_2 = \{p_2, p_7\}$.

First, from Tab. 1, we obtain s_2 's $k = 2$. Thus, the number of top-k results is 2. The maximum geography-text similarity in Tab. 2 is $SIM(s_2, p_2) = 0.9$. Thus, $N_2 = \{p_2\}$. Subsequently, we remove $SIM(s_2, p_2) = 0.9$ from Tab. 2. Next, the maximum geography-text similarity among the remaining $SIM(s_2, P)$ is $SIM(s_2, p_7) = 0.8$. Thus, $N_2 = \{p_2, p_7\}$. That is, the final top-2 results of s_2 are $N_2 = \{p_2, p_7\}$.

Based on the illustrations above, our approach ETGP shows three advantages as follows.

- (1) The ETGP does not require a user to set the threshold, which improves the practicability of the system.
- (2) The ETGP does not require the ranking of results. Only the highest top-k scores of the matching results are to be determined, which improves the efficiency.
- (3) When the computed number of results is less than k , the computed results are sent to the user without any other operation in the ETGP. Thus, the ETGP reduces the redundant computations.

By analysing algorithm 1, the computing cost of the ETGP is $O(k \cdot |SIM(s, P)|)$. IGPT [7, 8] requires computing more than k results to be trimmed and sorted based on the threshold. Therefore, the computing cost of the IGPT is $O(k \cdot |SIM(s, P)| \cdot \log_2 |SIM(s, P)|)$. Obviously, our approach is more efficient than the IGPT.

4 EXPERIMENTS

The IGPT [7, 8] is the most classic and closely related approach to our approach. Therefore, we compared our approach, ETGP, to IGPT to verify the efficiency with high facility.

4.1 Experimental Setup

Each procedure is programmed by Visual C++ 6.0. Two computers are used in our experiments. One is for system server. Another one is for subscribers and publishers. They both have 3.4GHz dual-core CPU and 32GB memory. Following the traditional settings (for example, [6, 7]) of the P/S system, we assume that the memory stores all indexes for realtime responses.

P/S datasets are randomly generated by the system. All geography-text similarities are randomly generated in the interval [0, 1]. During each comparison, IGPT and ETGP always have the same dataset.

In our experiments, the keyword number increases from 10 to 50. Each keyword follows TF-IDFweight [9]. k value (ie the number of results returned to the subscriber) increases from 2 to 10. Subscription number $|S|$ increases from 10M (million) to 50M. Publication number $|P|$

increases from 2M to 10M. The details are shown in Tab. 3.

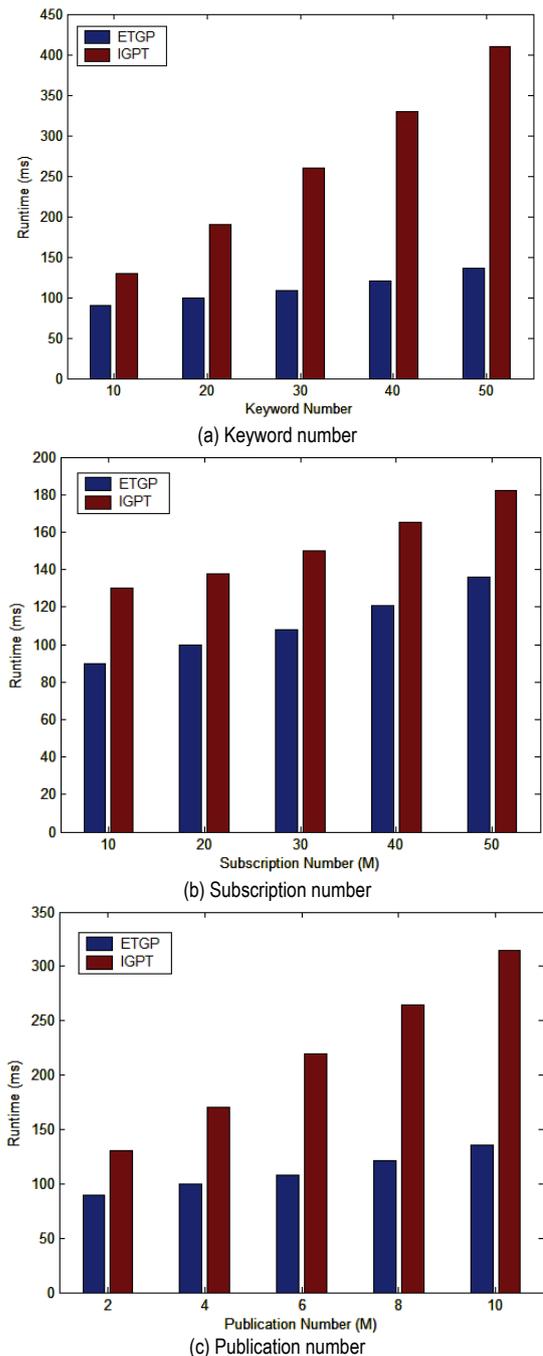


Figure 2 The comparison of similarity

Table 3 Experimental parameters

Keyword number	10	20	30	40	50
k	2	4	6	8	10
$ S $	10M	20M	30M	40M	50M
$ P $	2M	4M	6M	8M	10M

4.2 Similarity

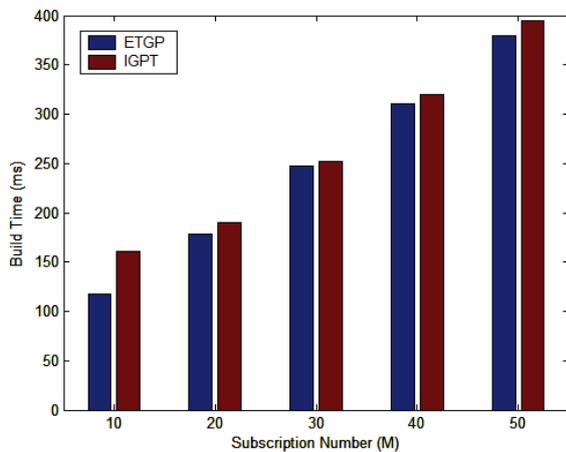
The geography-text similarity is related to subscription number $|S|$, publication number $|P|$ and keyword number. Thus, the experiment results are shown in Fig. 2, where Y -axis denotes runtime whose unit is millisecond (ms). Meanwhile, X -axes of Fig. 2(a), 2(b) and 2(c) denote

keyword number, subscription number $|S|$ and publication number $|P|$ respectively.

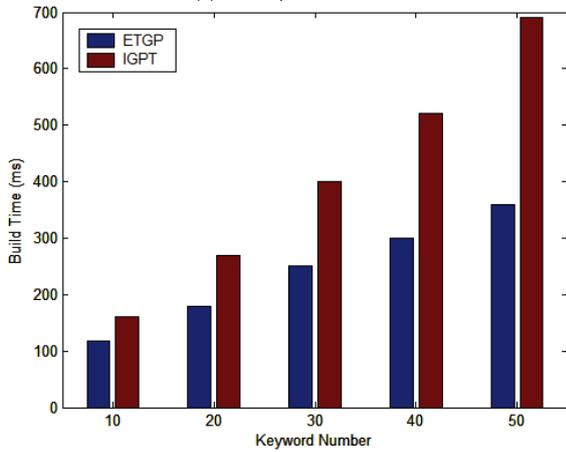
(1) Fig. 2(a). In this case, IGPT pays the higher cost for the similarity. The runtime of ETGP is fewer than IGPT. ETGP removes threshold. Thus the runtime increase of ETGP is much slower than IGPT. IGPT uses a Quadtree for the entire geography-text partition. Thus, IGPT consumes more runtime.

(2) Fig. 2(b). When subscription number increases from 10M to 50M, the runtime increases of the two approaches are all not fast. This means that the two approaches are all suitable for the scenarios with a large number of subscriptions. However, we remove the threshold. Thus ETGP has the better performance.

(3) Fig. 2(c). IGPT does not remove the threshold. Obviously, there are many redundancies computing in IGPT. Therefore, when publication number increases from 2M to 10M, ETGP consumes less runtime.



(a) Subscription number



(b) Keyword number

Figure 3 Comparison of index construction

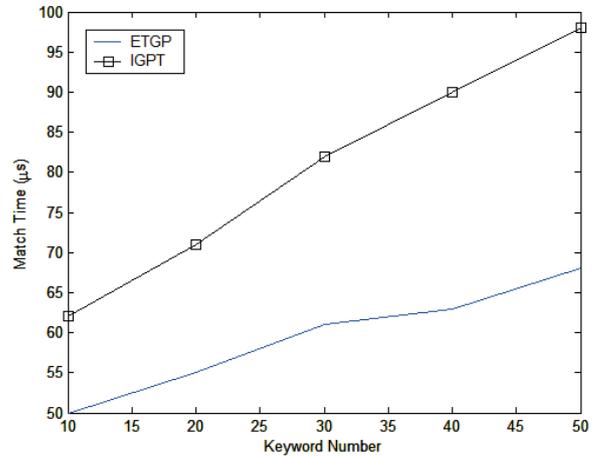
4.3 Index Construction

Index construction is primarily affected by the subscription number and keyword number. The experiment results are shown in Fig. 3, where the Y-axis denotes the build time of index construction, whose units are milliseconds (ms). Meanwhile, the X-axes denote the subscription number and keyword number, respectively.

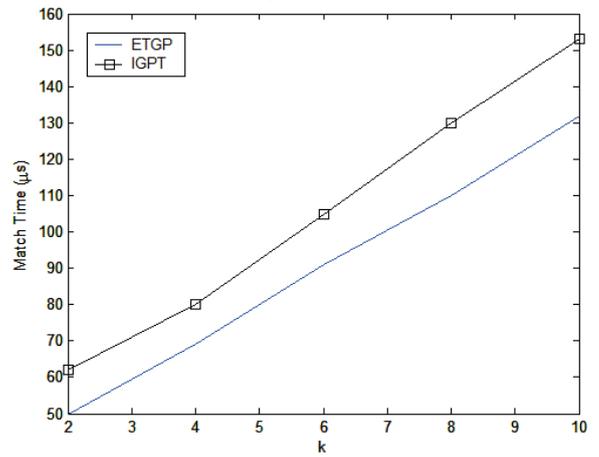
(1) Fig. 3(a). The IGPT uses pruning technology with threshold. Therefore, the build time of the IGPT increases

more quickly than the ETGP. That is, the ETGP is slightly more efficient than the IGPT.

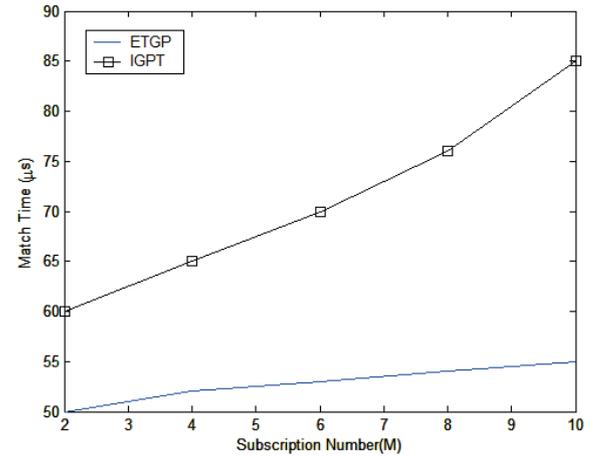
(2) Fig. 3(b). The IGPT adopts a Quadtree to set the text partition. However, the ETGP has a faster build time for losing a parameter than the IGPT. Thus, the ETGP is more efficient than the IGPT.



(a) Keyword number



(b) k



(c) Subscription number

Figure 4 Comparison of matching efficiency

4.4 Matching Efficiency

Faster matching yields a quicker returned query to users. Thus, matching efficiency is the key performance indicator of a P/S system. In our experiments, the match time is the time consumption of a subscriber querying a set of publications. The match time is primarily affected by the

keyword number, k value, and the subscription number. The experiment results are shown in Fig. 3, where the Y -axis denotes the match time, whose units are microseconds (μs). Meanwhile, the X -axes denote the keyword number, k value, and subscription number, respectively.

(1) Fig. 4(a). The final matching number of the ETGP is only k . However, the final matching number of the IGPT may be more than k . Therefore, when the keyword number increases from 10 to 50, the ETGP is more efficient than the IGPT.

(2) Fig. 4(b). When the k value increases from 2 to 10, the IGPT has more complex cutting operations than the ETGP. The ETGP has fewer matching operations than the IGPT. Therefore, the ETGP exhibits better performance than the IGPT.

(3) Fig. 4(c). In this case, the ETGP has fewer parameters than the IGPT. The IGPT has a more complex index structure than the ETGP based on the TAAT paradigm. Therefore, the time consumption rate of the ETGP is slower than that of the IGPT.

4.5 Update

The publication update is the dynamic performance of the system. Thus, its performance must be tested.

In our experiments, the update time is the time consumption of a publication update. The update time is primarily affected by the keyword number, k value, and subscription number. The experiment results are shown in Fig. 5, where the Y -axis denotes the update time, whose units are microseconds (μs). Meanwhile, the X -axes denote the keyword number, k value, and subscription number, respectively.

(1) Fig. 5(a). When the computed number of results is less than k , the computed results are sent to the user without any other operation in the ETGP. Thus, the ETGP is more efficient than the IGPT when the keyword number increases from 10 to 50.

(2) Fig. 5(b). The IGPT is almost not influenced by k . Therefore, when the k value increases from 2 to 10, the IGPT's growth speed of update time is lower than that of the ETGP.

(3) Fig. 5(c). When a publication is updated, the ETGP does not require the ranking of results. Only the related publication is to be obtained. The IGPT requires obtaining the best publication based on the k -skyband cost model. Thus, the ETGP exhibits better performance than the IGPT.

Based on the experiments above, the following conclusions are drawn:

(1) The comprehensive performance of the ETGP is obviously better than that of the IGPT.

(2) Owing to its good performance based on various parameters, the ETGP can be widely used for various application scenarios, such as large data, distributed computing, e-commerce, mobile platform, and cloud computing.

(3) The ETGP does not require a user to set the threshold; thus, the ETGP yields a more convenient user operation, which improves the availability of the geography-text P/S system.

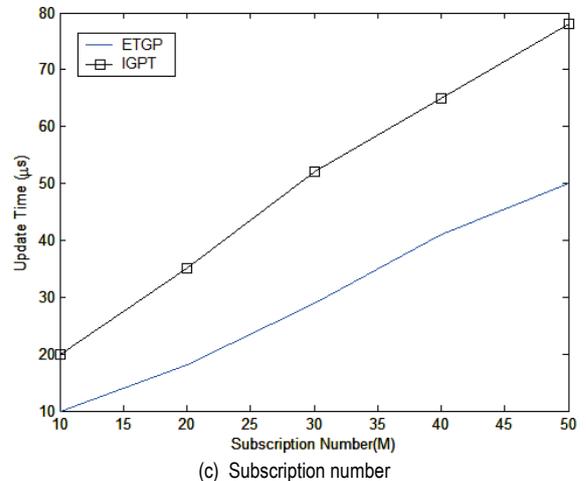
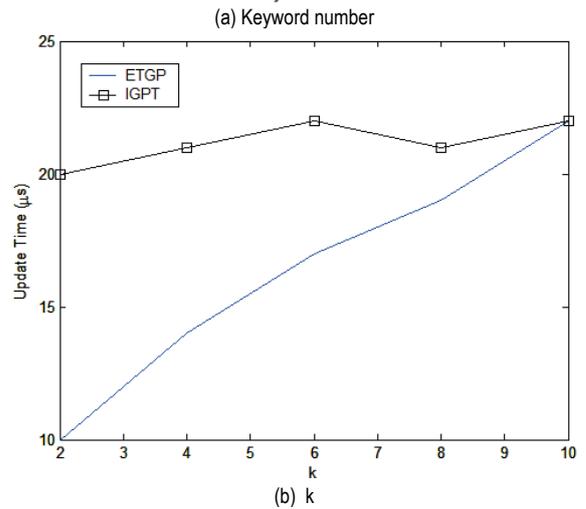
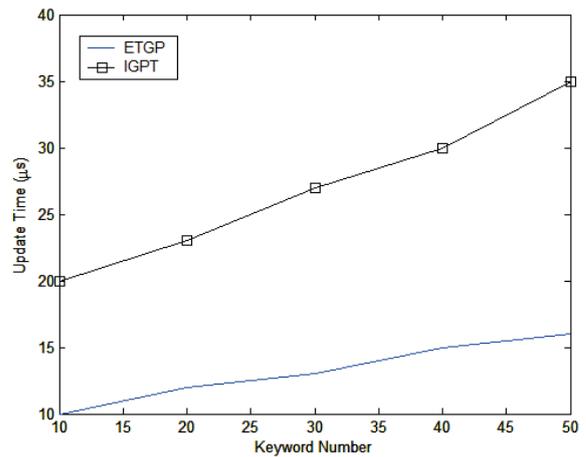


Figure 5 Comparison of publication update

5 RELATED WORKS

5.1 P/S System

The user registers his/her interest as a long-running query in a P/S system. When the information meets the user's interest, information will be sent to the user as flow text. Nevertheless, many existing P/S approaches do not consider the geography information [1, 10]. However, we focus on the use and extension of the geography information.

5.2 Geography Keyword Query

Geography keyword query has been widely studied in literature [4]. These studies are based on Boolean matching. It retrieves a set of geography text objects based on the joint spatial index (e.g., R-Tree, Quadtree) and reverses the text index [5]. Some works [2] summarized the query process of the geography keyword. [6] extended the geography keyword. We noticed that a geography keyword query is in fact a snapshot query. However, our questions focus on the text-similarity matching query.

5.3 Geography-text Query

Recently, the geography-text P/S system has been studied [3]. Nevertheless, these works do not unite geography similarity and text similarity. They cannot quantify the geography-text similarity. Therefore, these works are essentially different from our work.

The existing approach, IGPT [7, 8] is close to our work. The approach also supports the top-k geography-text P/S system.

The IGPT uses the Quadtree structure. If each subscription has the same cell coverage, a text boundary is to be precomputed for each subscription. A text boundary is a cell. Based on the subscription ID, a sorting file is created to manage the subscriptions assigned to the cells. For a new message, the IGPT index structure is designed for the TAAT paradigm. The IGPT integrates the advanced pruning technology based on the threshold. Therefore, the core parameter of the IGPT is the threshold. We discovered that including the threshold results in three problems. Our approach removes the three problems based on excluding the threshold.

6 CONCLUSIONS

Most of recent Top-k geography-text P/S approaches need the user to set thresholds. Based on thresholds, the number of system's scored publications is difficult to be equal to the number of user's required publications. Actually, many publications should not be scored. On the one hand, the user wastes time to set the parameter. On the other hand, the system wastes precious resources to compute redundant scores.

Therefore, we propose an efficient approach that has a parameter (i.e., threshold) less compared to the previous approaches. By removing threshold, we reduce the redundancy computing, promote the query integrity rate and enhance the facility. Extensive experiments prove that the proposed approach has good efficiency with high facility.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Nos. 61762055 and 61772215); the Jiangxi Provincial Natural Science Foundation of China [Nos. 20161BAB202036 and 20181BAB202014]; and the Jiangxi Provincial Social Science "13th Five-Year" (2016) Planning Project of China (No. 16JY19) and the Doctoral Fund of Daqing Normal University (No. 15ZR05).

7 REFERENCES

- [1] Ghafouri, M., Wang, X., Yuan, L., Zhang, Y. & Lin, X. (2018). Maintaining Boolean Top-K Spatial Temporal Results in Publish-Subscribe Systems. *Australasian Database Conference*, 147-160. https://doi.org/10.1007/978-3-319-92013-9_12
- [2] Wang, B., Zhu, R., Yang, X., & Wang, G. (2018). Top-K representative documents query over geo-textual data stream. *World Wide Web*, 21(2), 537-555. <https://doi.org/10.1007/s11280-017-0470-0>
- [3] Jia, M., Zhao, Y., Zheng, B., Liu, G., & Zheng, K. (2018). A Time-Aware Path-Based Publish/Subscribe Framework. *International Conference on Database Systems for Advanced Applications*, 511-528. https://doi.org/10.1007/978-3-319-91452-7_33
- [4] Chen, L., & Shang, S. (2018). Approximate spatio-temporal top-k publish/subscribe. *World Wide Web*, 1-23. <https://doi.org/10.1007/s11280-018-0564-3>
- [5] Zhang, D., Li, Y., Cao, X., Shao, J., & Shen, H. T. (2018). Augmented keyword search on spatial entity databases. *The International Journal on Very Large Data Bases*, 27(2), 225-244. <https://doi.org/10.1007/s00778-018-0497-6>
- [6] Nakamura, S., Ogiela, L., Enokido, T., & Takizawa, M. (2018). An information flow control model in a topic-based publish/subscribe system. *Journal of High Speed Networks*, 24(3), 243-257. <https://doi.org/10.3233/JHS-180593>
- [7] Wang, X., Zhang, Y., Zhang, W., Lin, X., & Huang, Z. (2016). Skype: top-k spatial-keyword publish/subscribe over sliding window. *VLDB Endowment*, 9(7), 588-599. <https://doi.org/10.14778/2904483.2904490>
- [8] Wang, X., Zhang, Y., Zhang, W., 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. <https://doi.org/10.1007/s00778-016-0453-2>
- [9] Rayner, M., Tsourakis, N., & Gerlach, J. (2017). Lightweight spoken utterance classification with CFG, TF-IDF and dynamic programming. *International Conference on Statistical Language and Speech Processing*, 143-154. https://doi.org/10.1007/978-3-319-68456-7_12
- [10] Li, S., Sun, Y., & Soergel, D. (2018). Automatic Decision Support for Clinical Diagnostic Literature Using Link Analysis in a Weighted Keyword Network. *Journal of medical systems*, 42(2), 27. <https://doi.org/10.1007/s10916-017-0876-3>

Contact information:

Zongmin CUI, Assoc. Prof., PhD
School of Information Science and Technology, Jiujiang University,
No. 551, Qianjin East Road, Jiujiang, Jiangxi 332005, China
E-mail: cuizm01@gmail.com

Hongbo LI, PhD candidate
(Corresponding author)
School of Computer Science and Information Technology,
Daqing Normal University,
Xibin Xi Road, Ranghulu District, Daqing, Heilongjiang Province, China
E-mail: lihongbodaqing@163.com

Hong ZHU, Prof. PhD
School of Computer Science and Technology,
Huazhong University of Science and Technology,
1037 Luoyu Road, Hongshan District, Wuhan, Hubei 430074, China
E-mail: whzhuhong@gmail.com