Grayscale Image-based Top-k Spatial Dataset Search Processing

Mingyue Zhang[1], Hua Dai[1,2], Pengyue Li[1], Sheng Wang[3], Bohan Li[4], Hao Zhou[1], and Geng Yang[1]

[1] Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China
1023041008, daihua, 1222045825, 2020040120, yangg@njupt.edu.cn
[2] The State Key Laboratory of Tibetan Intelligence, Nanjing, Jiangsu, China
[3] Wuhan University, Wuhan, Hubei, China
swangcs@whu.edu.cn
[4] Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China
bhli@nuaa.edu.cn

Abstract. In the data-driven era, dataset search has become a critical task in data science and engineering. Traditional spatial dataset search methods primarily rely on keyword or range queries, which are insufficient for capturing user intent expressed through exemplar datasets. To address this gap, this paper investigates the problem of top-k spatial dataset search using exemplar datasets as input. A novel grayscale image-based similarity model is first proposed, which maps the spatial distribution of datasets into grayscale images to capture detailed distribution features. Based on this model, a baseline search scheme (GIDS) is proposed. To further improve the search efficiency, an optimized search scheme (GIDS +) is introduced, which incorporates two key optimization strategies: a Morton code-based strategy to accelerate similarity calculations and a ω-MSDtree-based strategy to enable efficient pruning during candidate filtering. Experiments conducted in two real-world spatial data repositories demonstrate that the proposed methods outperform existing approaches in search efficiency, providing a new solution for spatial dataset search.

Keywords: Dataset Discovery · Top-k Dataset Search · Grayscale Image · Search Index.

1 Introduction

With the rapid growth of structured and semi-structured data in repositories such as data lakes and web tables, data discovery has become a crucial task for data scientists and engineers. Dataset search, as a key enabling technology, empowers users to locate relevant datasets within vast data repositories efficiently. Numerous search engines and platforms have been developed to facilitate data discovery [3, 6, 8], including prominent examples such as Google Dataset Search [4] and Zenodo [10], which significantly enhance data accessibility through user-friendly search functionalities and extensive dataset coverage.

Searches based on spatial datasets, a critical component of real-world data, have attracted growing global attention and found applications across diverse industries, including smart city development, environmental monitoring, and disaster management [2]. As such, advancing the study of spatial dataset search is vital for unlocking and utilizing the inherent value of spatial datasets.

This paper focuses on the top-k spatial dataset search, which allows users to describe their search intent using an exemplar dataset [9]. The user uploads a search request using an exemplar dataset, and the dataset search system returns the k datasets most similar to the users by calculating the similarity between each dataset and the exemplar dataset [12]. The spatial dataset similarity model, which measures the similarity between datasets, is crucial in search processing. An approach is to use the size of the overlap area between the minimum bounding rectangles (MBRs) that enclose all points in each dataset [7]. Another method is the Hausdorff distance, which calculates the maximum distance between nearest-neighbor points in the datasets [1]. The Earth Mover's Distance (EMD) transforms datasets into distributions to measure similarity [11]. However, MBR does not account for the point distribution within the rectangle, the Hausdorff distance is sensitive to outliers [5], and EMD is computationally expensive for large datasets. In summary, balancing the effectiveness and efficiency of similarity models remains a challenge.

In this paper, two efficient spatial dataset search schemes are proposed to address the aforementioned issues. First, a novel grayscale image-based spatial dataset similarity calculation model is introduced, which converts the distribution information of spatial datasets into grayscale images and uses these images for similarity calculation. Then, a baseline grayscale image-based spatial dataset search scheme (GIDS) is introduced. To further improve search efficiency, the Morton code-based optimization strategy and the ω MSDtree-based optimization strategy are used in the optimized grayscale image-based spatial dataset search scheme (GIDS+). Furthermore, experimental results in two real-world data repositories demonstrate the search efficiency and space cost of the proposed schemes.

2 Problem Formulation

Definition 1 (Spatial Dataset). *A spatial dataset D_i contains a set of location points, $D_i = \{(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \cdots, (x_{i,n_i}, y_{i,n_i})\}$, where $(x_{i,j}, y_{i,j})$ is latitude and longitude of D_i in spatial space.*

Definition 2 (Spatial Data Repository). *A spatial data repository $\mathcal{D}$ contains a set of spatial datasets, $\mathcal{D} = \{D_1, D_2, \cdots, D_m\}$.*

Definition 3 (Top-k Spatial Dataset Search). *Given an exemplar dataset D_q, the top-k spatial dataset search $Q = (k, D_q, \mathcal{D})$ is to obtain the k most similar spatial datasets in $\mathcal{D}$ to D_q, and the search result R should satisfy*

$$|R| = k \wedge \forall D_i \in R, D_j \notin R \rightarrow Sim(D_q, D_i) > Sim(D_q, D_j),$$

where $Sim(D_q, D_i)$ is the similarity of D_q to D_j, described in Section 3.

The goal of this paper is to design a top-k spatial dataset search processing that can effectively and efficiently perform top-k dataset searches in the spatial data repository.

3 Grayscale Image-based Spatial Dataset Similarity Model

Definition 4 (Grid-based Spatial Representation). *Given a spatial space $\mathcal{S}$ and a space partition threshold σ, $\mathcal{S}$ is equally divided into $2^\sigma \times 2^\sigma$ grids. The grid in x-th row and y-th column is denoted as $g_{x,y}$, and thus $\mathcal{S}$ can be represented by a set of grids, i.e., $\mathcal{S} = \{g_{x,y} | x, y \in \{0, 1, \cdots, 2^\sigma - 1\}\}$.*

Definition 5 (Grid-based Minimum Bounding Rectangle, GMBR). *Given a spatial dataset $D_i \in \mathcal{D}$, the GMBR of D_i is a minimum set of grids, denoted as G_i, which can form a rectangle just covering all location points of D_i. Assumging that the bottom left grid of G_i is $g_{x,y}$ and G_i covers $u \times v$ grids, thus we have*

$$G_i = \{g_{x,y} \,|\, g_{x,y} \in \mathcal{S} \wedge x \in \{x, \cdots, x+u-1\} \wedge y \in \{y, \cdots, y+v-1\}\}.$$

After performing a grid partition, we discovered an interesting fact: If we treat each grid cell as a pixel in an image and set the grayscale value of the pixel based on the number of points within the grid, we can use a grayscale image to represent the distribution characteristics of the spatial dataset. Based on this idea, we will define a grayscale image representation of the spatial datasets.

Definition 6 (Grayscale Image Representation of Spatial Dataset). *Given a spatial dataset D_i and its GMBR G_i, the grayscale image representation of D_i is denoted as I_i,*

$$I_i = \{(g_{x,y}, p^i_{x,y}) | g_{x,y} \in G_i\}. \tag{1}$$

and $p^i_{x,y} \in [0, 255]$ is the grayscale value of grid $g_{x,y}$ in G_i,

$$p^i_{x,y} = \left\lceil \frac{255 \cdot N(g_{x,y})}{\max\{N(g_{x,y}) \mid g_{x,y} \in G_i\}} \right\rceil, \tag{2}$$

where $N(g_{x,y})$ is the number of points of D_i that fall within the grid $g_{x,y}$, and $\max\{*\}$ is to get the maximum from a set.

For the data repository $\mathcal{D}$, the GMBR set and the grayscale image set corresponding to $\mathcal{D}$ are denoted as $\mathcal{G} = \{G_i | D_i \in \mathcal{D}\}$ and $\mathcal{I} = \{I_i | D_i \in \mathcal{D}\}$, respectively. We give an example of generating grayscale images in Fig. 1. The spatial dataset D_i and its GMBR G_i are shown in Fig. 1(a). After calculating the grayscale values of the grids in G_i according to Eq. (2), we have the grayscale image I_i as shown in Fig. 1(b).

According to Definition 5 and 6, a spatial dataset can be represented by a grayscale image. Thus, we adopt a proven grayscale image similarity calculation method, which measures the similarity between spatial datasets by measuring the relative difference between two grayscale images as follows.

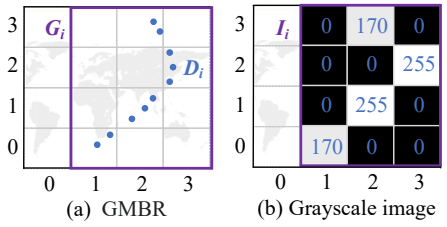


Fig. 1: An example of grayscale image

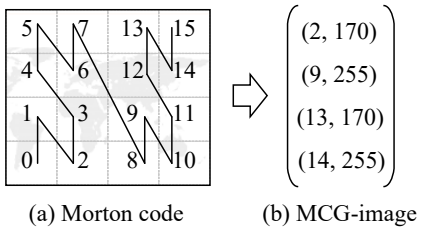


Fig. 2: An example of MCG-imag

Definition 7 (Grayscale Image-based Spatial Dataset Similarity). *Given two spatial datasets D_i and D_j, the corresponding grayscale images are I_i and I_j, the similarity between D_i and D_j is denoted as $Sim(I_i, I_j)$, which is calculated by Eq. (3),*

$$Sim(I_i, I_j) = 1 - \frac{1}{2|M|} \sum_{g_{x,y} \in M} Rel_{I_i,I_j}(g_{x,y}), \tag{3}$$

where

$$M = \left\{ g_{x,y} \mid g_{x,y} \in G_i \cup G_j \wedge (p^i_{x,y} \neq 0 \vee p^j_{x,y} \neq 0) \right\}, \tag{4}$$

$$Rel_{I_i,I_j}(g_{x,y}) = \begin{cases} \frac{2|p^i_{x,y} - p^j_{x,y}|}{p^i_{x,y} + p^j_{x,y}} & g_{x,y} \in G_i \cap G_j \\ 2 & otherwise \end{cases}, \tag{5}$$

G_i and G_j are the GMBRs of D_i and D_j. In Eq. (3), if $g_{x,y} \notin G_i$, then $p^i_{x,y} = 0$.

Lemma 1. *Assuming that D_i and D_j are two spatial datasets with no overlapping area in their GMBRs, $Sim(D_i, D_j)$ will be 0.*

Proof. According to Definition 7, for each grid $g_{x,y}$ in M, $Rel_{I_i,I_j}(g_{x,y})$ equals 2, which implies that $Sim(I_i, I_j)$ equals 0 and Lemma 1 is proved. ■

In conclusion, the proposed grayscale image-based similarity model can describe the distribution characteristics of spatial datasets by using the grayscale image representation and applied to spatial dataset search.

4 Top-k Spatial Dataset Search Processing

4.1 The Baseline Search Processing

The baseline grayscale image-based spatial dataset search processing (GIDS) is presented by adopting the proposed grayscale image-based similarity model to measure the similarity between spatial datasets. In search processing, the candidate spatial datasets are first identified using the GMBR overlapping detection, and then similarity calculations are performed between the candidate datasets and the exemplar dataset to obtain the search result. The details of the baseline search processing are shown in Algorithm 1.

Algorithm 1: *GIDS*(k, D_q, $\mathcal{D}$, $\mathcal{G}$, $\mathcal{I}$)

Input: The number of requested spatial datasets k; the exemplar spatial dataset D_q; the spaital data repository $\mathcal{D}$; the GMBR set $\mathcal{G}$ and grayscale image set $\mathcal{I}$ corresponding to $\mathcal{D}$.
Output: The search result R.

```
Initialize the search result R = ∅ and a priority queue Q = ∅ for storing similarity and spatial dataset pairs ordered by the similarity value;
Generate the GMBR G_q and the grayscale image I_q for D_q.
foreach D_i ∈ D do
    Get the GMBR G_i and the grayscale image I_i from G and I, respectively;
    if G_i ∩ G_q ≠ ∅ then
        Calculate the similarity sim_i = Sim(I_i, I_q);
        if |Q| < k then
            Add the pair (sim_i, D_i) into Q;
        else
            Get the pair (sim_j, D_j) from Q where sim_j is the minimum;
            if sim_i > sim_j then
                Remove (sim_j, D_j) from Q and add (sim_i, D_i) into Q;
    Get all the spatial datasets stored in Q, and add them into R;
return R;
```

The time complexity of the GIDS algorithm is $O\left(m + \eta(\alpha + \beta + \log k)\right)$, where m is the number of spatial datasets in the repository, η is the average number of candidate datasets having overlaps in GMBR, and α and β are the average numbers of grids in grayscale images of exemplar datasets and candidate datasets, respectively.

4.2 The Optimized Search Processing

We first present two optimization strategies to speed up the baseline dataset search and then propose the optimized spatial dataset search scheme (GIDS+) based on these two optimization strategies.

• **Morton Code-based Optimization Strategy.** The Morton code is used for encoding spatial data, which supports encoding grids with unique numerical codes. We adopt the Morton code to optimize the representation of grayscale images, which enables efficient similarity calculations and are defined below.

Definition 8 (Morton Code-based Grayscale Image, MCG-image). *Given a spatial dataset $D_i \in \mathcal{D}$ and its grayscale image I_i, the MCG-image of D_i, denoted as $\hat{I}_i$, is a sequence of Morton code and grayscale value pairs, ordered in ascending Morton code,*

$$\hat{I}_i = \left\{(c_{x,y}, p^i_{x,y}) \mid (g_{x,y}, p^i_{x,y}) \in I_i \wedge p^i_{x,y} \neq 0\right\} \tag{6}$$

where $c_{x,y}$ is the Morton code of grid $g_{x,y}$. The MCG-image set corresponding to $\mathcal{D}$ is denoted as $\hat{\mathcal{I}} = \{\hat{I}_i | D_i \in \mathcal{D}\}$.

Definition 8 indicates that an MCG-image stores only the grids with non-zero grayscale values. An example of the MCG-image is shown in Fig. 2, using the dataset from Fig. 1. Fig. 2(a) shows the Morton code for each grid in a 4×4 grid space. Fig. 2(b) illustrates the generated MCG-image, which records the Morton codes and their corresponding non-zero grayscale values, sorted in ascending order of the Morton codes.

Definition 9 (MCG-image-based Spatial Dataset Similarity). *Given two spatial datasets D_i and D_j, the corresponding MCG-images are $\hat{I}_i$ and $\hat{I}_j$, the similarity between D_i and D_j is denoted as $Sim(\hat{I}_i, \hat{I}_j)$, which is calculated by Eq. (7).*

$$Sim(\hat{I}_i, \hat{I}_j) = 1 - \frac{1}{2|M_c|} \sum_{g_{x,y} \in M_c} Rel_{\hat{I}_i, \hat{I}_j}(g_{x,y}) \tag{7}$$

where

$$M_c = \left\{ g_{x,y} \mid g_{x,y} \in \mathcal{S} \wedge \left((c_{x,y}, p^i_{x,y}) \in \hat{I}_i \vee (c_{x,y}, p^j_{x,y}) \in \hat{I}_j \right) \right\}, \tag{8}$$

$$Rel_{\hat{I}_i, \hat{I}_j}(g_{x,y}) = \begin{cases} \frac{2|p^i_{x,y} - p^j_{x,y}|}{p^i_{x,y} + p^j_{x,y}} & g_{x,y} \in G_i \cap G_j \\ 2 & otherwise \end{cases}, \tag{9}$$

and M_c and $Rel_{\hat{I}_i, \hat{I}_j}(g_{x,y})$ have similar connotations to M and $Rel_{I_i, I_j}(g_{x,y})$ in Definition 7, respectively.

Lemma 2. *Assuming that I_i and I_j are two grayscale images and $\hat{I}_i$ and $\hat{I}_j$ are the corresponding MCG-images, we have*

$$Sim(I_i, I_j) = Sim(\hat{I}_i, \hat{I}_j). \tag{10}$$

Proof. According to Definition 7, 8 and 9, we can deduce that $M = M_c$ and $Rel_{I_i, I_j}(g_{x,y}) = Rel_{\hat{I}_i, \hat{I}_j}(g_{x,y})$, thus $Sim(I_i, I_j) = Sim(\hat{I}_i, \hat{I}_j)$ holds and Lemma 2 is proved. ■

Lemma 2 indicates that the MCG-image representation is the same as the original grayscale image representation in the spatial dataset similarity measurements. Because the MCG-image-based similarity calculation has no need to consider grids with a grayscale value of 0, the MCG-image-based similarity calculation is faster than the grayscale image-based similarity calculation, which benefits the improvement in search efficiency.

• **ω-MSDtree-based Optimization Strategy.** We present a multi-way spatial dataset tree with degree ω in this subsection, which is abbreviated as ω-MSDtree. The ω-MSDtree is designed by combining the multi-way search tree with GMBRs and MCG-images of spatial datasets, which can be used to filter candidate datasets efficiently in search processing.

Definition 10 (ω-MSDtree). *An ω-MSDtree is a multi-way spatial dataset tree, denoted as $\mathcal{L}_\omega$, where each node except the root contains at least $\lceil \omega/2 \rceil$ and at most ω entries. Each node u in the ω-MSDtree is a pair,*

$$u = (EL, pptr),$$

where $u.EL = \{e_1, e_2, \ldots, e_s\}$ is an entry list ($\lceil \frac{\omega}{2} \rceil \leq s \leq \omega$), and $u.pptr$ is a pointer pointing to its parent node. Specifically, the number of root's child nodes should be in $[0, \omega]$. The structure of u depends on its position in $\mathcal{L}_\omega$:

- ***(Leaf Node).*** *If u is a leaf node, each entry $e_i \in u.EL$ is named as the leaf entry corresponding to a spatial dataset, and it is a triple, $e_i = (id, gmbr, mcg)$, where $e_i.id$, $e_i.gmbr$ and $e_i.mcg$ are the unique identifier, GMBR and MCG-image of the corresponding spatial dataset, respectively.*
- ***(Internal Node).*** *If u is an internal node, each entry $e_i \in u.EL$ is name as the internal entry, and it is a pair, $e_i = (gmbr, cptr)$, where $e_i.cptr$ is a pointer pointing to u's i-th child node, and $e_i.gmbr$ is the GMBR covering all entries' GMBRs in its child nodes, i.e., $e_i.gmbr = \biguplus_{e_j \in e_i.cptr.EL} e_j.gmbr$.*

Here, $G_i \uplus G_j$ is a minimum set of grids that can form a rectangle that only covers G_i and G_j. The details of ω-MSDtree construction are presented as follows. For each spatial dataset in $\mathcal{D}$, a leaf entry is created and inserted into a target leaf node which has the minimum increment in GMBR when adding the leaf entry to the leaf node. If the number of entries in the leaf node is not less than ω, then the leaf node is split into two new leaf nodes, and an internal entry is generated to insert into the leaf node's parent node. The internal entry insertion in the parent node is similar to that in the leaf node. The entry insertion is performed from a bottom leaf node to an upper internal node (including the root) until the number of entries in a node is less than ω.

The ω-MSDtree will be used as the search index for accelerating the dataset filtering with logarithmic complexity, which can improve the search efficiency.

• **The Optimized Search Algorithm** During the above optimization strategies for the baseline search scheme, the Morton code-based optimization strategy is used to accelerate similarity calculation, and the ω-MSDtree is used to speed up the acquisition of candidate datasets.

Details of the optimized dataset search processing are presented as follows.

1) The GMBR and MCG-image of the exemplar dataset are first generated.

2) The candidate datasets with GMBR overlaps with the exemplar dataset are obtained from ω-MSDtree by using the Algorithm 2.

3) The similarity between each candidate dataset and the exemplar dataset is calculated according to Definition 9.

4) The priority queue storing the k highest similarity datasets is maintained, and the search results are returned to the user.

The time complexity of the search algorithm is $O(\log m + \eta(\alpha' + \beta' + \log k))$, where m, η and k are the same as that in the baseline search scheme, and α' and β' are the average number of grids in the MCG-images of the exemplar datasets and candidate datasets, respectively.

Algorithm 2: $Filtering(N_r, G_q)$

Input: The node of ω-MSDtree N_r, the GMBR G_q of D_q.
Output: The pair set CR for storing the candidate spatial dataset identifier and corresponding MCG-image pairs.

```
Initialize CR = ∅;
foreach e_i ∈ N_r.EL do
    if G_q intersects with e_i.gmbr then
        if N_r is not a leaf node then
            Filtering(e_i.cptr, G_q);
        else
            Add (e_i.id, e_i.mcg) into CR;
return CR;
```

5 Performance Evaluation

In this section, we perform a comprehensive evaluation of the proposed GIDS and GIDS+ and compare them with the EMD-based search scheme [11]. The evaluation metrics include search time cost and space cost.

In the experiment, we use two real-world spatial data repositories, *Identifiable* and *Public*, which are collected from OpenStreetMap [11]. The default parameters of our evaluations are the number of clusters $c = 20$, the number of search results $k = 10$, the space partition threshold $\sigma = 14$, the number of datasets $n = 100,000$, and the order parameter of ω-MSDtree $\omega = 16$.

5.1 Search Time Cost Evaluation

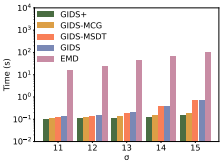
In Figs. 3-6, GIDS-MCG and GIDS-MSDT are the optimized search schemes based on Morton code only and ω-MSDtree only.

Figs. 3 and 4 both indicate that the search time cost of five approaches all grow as σ and n increase. The reason is that an increase in σ adds more grids to the similarity calculation, while an increase in n involves more datasets, both of which raise the search time cost.

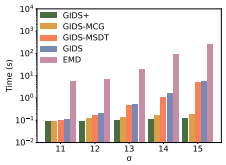
Fig. 5 indicates that the search time costs of GIDS, GIDS-MCG, GIDS-MSDT and GIDS+ are almost constant, while for EMD it grows. The reason is that EMD depends on k for filtering during search, unlike the other methods.

Fig. 6 shows the search time cost and the filtering time cost versus ω. It indicates that ω has a lower effect on the search time. Additionally, from the filtering time cost represented by the line graph, it can be seen that when $\omega = 16$ in both spatial data repositories, the filtering time is minimal.

In all experiences. the search time cost of GIDS and GIDS+ are lower than those of EMD. Furthermore, GIDS and GIDS+ are approximately 128 and 435 times faster than EMD, respectively. These experimental results indicate that GIDS and GIDS+ are efficient in spatial dataset searches.

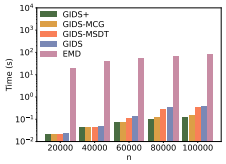


(a) Public

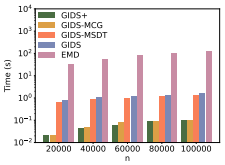


(b) Identifiable

Fig. 3: Search time versus σ

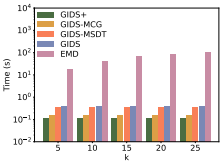


(a) Public

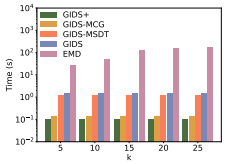


(b) Identifiable

Fig. 4: Search time versus n

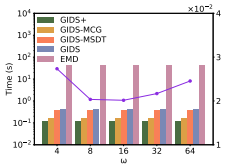


(a) Public

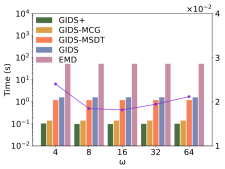


(b) Identifiable

Fig. 5: Search time versus k



(a) Public



(b) Identifiable

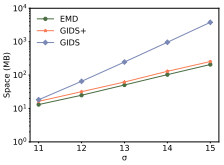
Fig. 6: Seach time versus ω

5.2 Space Cost Evaluation

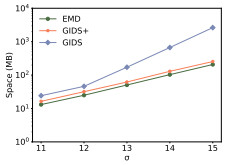
We evaluate the index space costs of GIDS, GIDS+ and EMD impacted by σ and n, and the experiment results are shown in Figs. 7 and 8.

Fig. 7 indicates that the index space costs of GIDS, GIDS+, and EMD grow as σ increases. The reason is that the larger σ is, the more grids are needed to store in indexes for a spatial dataset, and the higher the index space costs are.

Fig. 8 indicates that the index space costs of GIDS, GIDS+ and EMD grow as n increase. The reason is that larger n means more spatial datasets in the indexes, leading to higher costs. Additionally, EMD and GIDS+ have lower index space costs than GIDS, with only a 6% difference between GIDS+ and EMD. This indicates that GIDS+ offers more efficient spatial dataset searches, with index space costs similar to EMD.

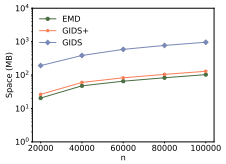


(a) Public

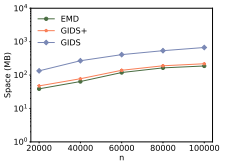


(b) Identifiable

Fig. 7: Index space cost versus σ



(a) Public



(b) Identifiable

Fig. 8: Index space cost versus n

6 Conclusion

In this paper, we proposed two top-k spatial dataset search schemes, GIDS and GIDS+. The key novelty of our schemes lies in representing spatial datasets as grayscale images and leveraging an image-based similarity calculation model. Then, we introduced a Morton code-based optimization strategy to accelerate similarity computations and a ω-MSDtree-based pruning strategy to improve candidate filtering efficiency. The experimental results demonstrated that our schemes can effectively and efficiently perform top-k spatial dataset search, outperforming existing approaches in search effectiveness and efficiency.

References

1. Adelfio, M.D., Nutanong, S., Samet, H.: Similarity search on a large collection of point sets. In: Proceedings of the 19th ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL). pp. 132–141 (2011)
2. Atitallah, S.B., Driss, M., Boulila, W., Ghézala, H.B.: Leveraging deep learning and iot big data analytics to support the smart cities development: Review and future directions. Computer Science Review **38**, 100303 (2020)
3. Bogatu, A., Fernandes, A.A., Paton, N.W., Konstantinou, N.: Dataset discovery in data lakes. In: Proceedings of 36th IEEE International Conference on Data Engineering (ICDE). pp. 709–720 (2020)
4. Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: Proceedings of the 2019 World Wide Web Conference (WWW). pp. 1365–1375 (2019)
5. Cao, C., Li, M.: Generating mobility trajectories with retained data utility. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (SIGKDD). pp. 2610–2620 (2021)
6. Castelo, S., Rampin, R., Santos, A., Bessa, A., Chirigati, F., Freire, J.: Auctus: A dataset search engine for data discovery and augmentation. Proceedings of the VLDB Endowment **14**, 2791–2794 (2021)
7. Degbelo, A., Teka, B.B.: Spatial search strategies for open government data: a systematic comparison. In: Proceedings of the 13th Workshop on Geographic Information Retrieval (GIR). pp. 1–10 (2019)
8. Dong, Y., Takeoka, K., Xiao, C., Oyamada, M.: Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach. In: Proceedings of the 37th IEEE International Conference on Data Engineering (ICDE). pp. 456–467 (2021)
9. Li, P., Dai, H., Wang, S., Yang, W., Yang, G.: Privacy-preserving spatial dataset search in cloud. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM). pp. 1245–1254 (2024)
10. Sicilia, M.A., García-Barriocanal, E., Sánchez-Alonso, S.: Community curation in open dataset repositories: Insights from zenodo. Procedia Computer Science **106**, 54–60 (2017)
11. Yang, W., Wang, S., Sun, Y., Peng, Z.: Fast dataset search with earth mover's distance. Proceedings of the VLDB Endowment **15**, 2517–2529 (2022)
12. Yew, J.X., Liao, N., Mo, D., Luo, S.: Example searcher: A spatial query system via example. In: Proceedings of the 39th IEEE International Conference on Data Engineering (ICDE). pp. 3635–3638 (2023)