QRDI-index-based Spatial Dataset Range Search

Jing Luo
1024041121
Nanjing University of Posts and Telecommunications
Nanjing, China

***Abstract*—As the number of open spatial datasets continues to grow, there is a corresponding increase in demand for the ability to efficiently identify spatial datasets that align with users' specific requirements. It has become a significant issue, resulting in the necessity for a variety of spatial dataset search requirements, including the need for spatial dataset range search. In this paper, we propose an efficient spatial dataset range search processing based on the quadtree-based region-dataset inverted index (QRDI-index). A relevance measurement between a spatial dataset and a search range is first presented, which is inspired by a common text similarity model. To provide efficient searches, the QRDI-index is designed, which combines the inverted index, quadtree and spatial datasets. By using the index, we propose an efficient search processing algorithm that can filter the minimum tree nodes in QRDI-index, and the search space is narrowed into these nodes. Experimental results on three real-world spatial data repositories validate the accuracy and efficiency of the proposed search scheme.**

***Index Terms*—Dataset search, spatial dataset, range search, search index**

I. Introduction

The increasing availability of open datasets from governments, businesses, and non-profit organizations provides individuals with valuable data for analysis through online platforms. For example, many governments have established open government data systems, such as Data.gov Home [1], National Earth System Science Data Center [2], etc, which serve as central hubs for retrieving and accessing data produced by their agencies [3]. These open datasets serve multiple purposes, including supporting decision-making and application development, as well as facilitating the training, validation, and improvement of machine learning [4]. Therefore, the increasing demand for dataset search has led to the development of various search engines [5]–[7] in recent years. Moreover, statistics indicate that at least 60% of open datasets contain spatial information [8], which makes spatial dataset search one of the most important issues within the domain of dataset discovery. As one of the fundamental spatial dataset search methods, it is essential to design accurate and efficient spatial dataset range search processing.

A spatial dataset range search is to obtain the spatial datasets that have the top-k highest relevance to a given search range. Fig. 1 shows an example of the spatial dataset range search, where a search range r is submitted, and k spatial datasets from the spatial data repository $\mathcal{S} = \{S_1, S_2, \cdots, S_m\}$ with the highest relevance scores to r are returned as the search result. In recent years, a few researches have been conducted on the subject of range search for spatial datasets. Ghosh et al. [5] used R-Grove [9] as the index of a system for spatial datasets. Auctus [7] is capable of supporting a multitude of search operations, including those based on keywords, searches defined by specific spatial and temporal parameters, and data integration searches. Nevertheless, as far as we know, the above demo literature only discusses spatial dataset range search but lacks systematic introductions about indexes and search algorithms. In addition, the literature [10]–[13] propose the spatial data item search schemes that can obtain a set of spatial items (location points) located in a given search range, and other literature [14]–[17] present the exemplar spatial dataset search schemes that can obtain k spatial datasets with the highest similarity to a given exemplar dataset. However, these search schemes cannot solve the spatial dataset range search problem. As a result, a comprehensive investigation of efficient spatial dataset range search processing is required.

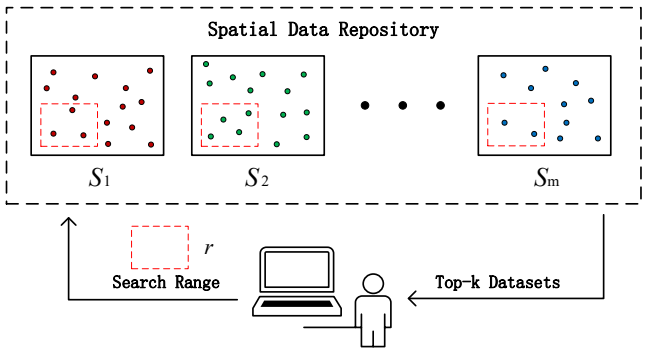


Fig. 1: An example of spatial dataset range search.

In this paper, we propose an efficient spatial dataset range search processing based on QRDI-index. First, we present a measurement method for the relevance between a spatial dataset and a search range, which is inspired by the BM25 [18], [19] model used for measuring text similarity. To support efficient search processing, we designed a quadtree-based region-dataset inverted index (QRDI-index), which combines the inverted index [20], quadtree [21] and spatial datasets. By using the QRDI-index, we propose an efficient search processing algorithm. It first filters the minimum tree nodes in QRDI-index having the candidate result, which significantly narrows the search space, and then it obtains the search result

This work was supported by the National Natural Science Foundation of China under grants Nos. 62372244, 62202338, and 62272238; and the Jiangsu Province Postgraduate Scientific Research Innovation Program under grand No. KYCX24_1221. The corresponding author is Hua Dai.

by processing these tree nodes. Comprehensive experiments are conducted on three real-world spatial data repositories, and the experimental results show that the proposed scheme performs well in terms of search accuracy and efficiency.

Overall, the contributions of this paper are summarized as follows:

- We propose a relevance measurement between a spatial dataset and a search range, which is inspired by the text similarity model BM25.
- We design a novel quadtree-based region-dataset inverted index (QRDI-index) by combining the inverted index, quadtree and spatial datasets. Using the index, we propose an efficient spatial dataset range search algorithm.
- We conduct comprehensive experiments on real-world spatial data repositories to validate the accuracy and efficiency of the proposed scheme.

II. Problem Formulation

Definition 1. Spatial Dataset. A spatial dataset is a set of location points, denoted as $S_i = \{l_{i,1}, l_{i,2}, \cdots, l_{i,n_i}\}$, where the location $l_{i,j}$ consists of longitude and latitude.

Definition 2. Spatial Data Repository. A spatial data repository is a set of spatial datasets, denoted as $\mathcal{S} = \{S_1, S_2, \cdots, S_m\}$.

Definition 3. Spatial Dataset Range search. A spatial dataset range search $Q = (r, k, \mathcal{S})$ is to obtain the top-k spatial datasets in $\mathcal{S}$ which have the highest relevance score between the spatial dataset and search range r. Assuming that $\mathcal{R}$ is the search result of Q, it should satisfy the following conditions:

- $|\mathcal{R}| = k$
- $\forall S_i \in \mathcal{R}, S_j \in \mathcal{S} - \mathcal{R} \rightarrow Score(S_i, r) > Score(S_j, r)$

where $Score(S_i, r)$ represents the relevance between the spatial dataset S_i and search range r. The detailed calculation of the relevance score will be explained in Section III.A.

The goal of this paper is to design an accurate and efficient spatial dataset range search processing. The evaluation metrics of the proposed scheme are described as follows:

- **Efficiency.** The search time cost of a spatial dataset range search is the search efficiency metric.
- **Accuracy.** The search accuracy of a spatial dataset range search is the ratio of the search result to the correct result.

III. Spatial Dataset Range Search Processing

In this section, we describe in detail the spatial dataset range search scheme. First, we design a spatial dataset, search range relevance measurement, and a new hybrid index structure. Finally, We present a search method based on this index.

A. Spatial Dataset and Search Range Relevance Measurement

For spatial dataset range search, we can't simply take the number of locations of each dataset in the search range as the search condition. We should consider the relevance between the search range and the dataset as our search criteria. For spatial datasets, their relevance to the search range needs to take into account the ratio of the number of location points in the search range to the total number of location points in the spatial dataset, as well as the scale characteristics of the spatial dataset. Therefore, we adopt a method similar to the BM25 text similarity model for text length information and combine it with the distributional characteristics of spatial datasets to calculate the relevance between spatial datasets and search ranges. We present the definition of relevance as shown in Definition 4.

Definition 4. Relevance between Spatial Dataset and Search Range. Given a spatial dataset $S_i \in \mathcal{S}$ and a search range r, the relevance score between S_i and r is calculated by Eq.(1),

$$Score(S_i, r) = \frac{N_{S_i,r}}{N_{\mathcal{S},r}} \cdot \frac{\frac{N_{S_i,r}}{|S_i|} \cdot (h+1)}{\frac{N_{S_i,r}}{|S_i|} + h \cdot (1 - b + b \cdot \frac{|S_i|}{avg\_|S|})} \quad (1)$$

where

- $N_{S_i,r}$ is the number of location points in S_i within the search range r.
- $N_{\mathcal{S},r}$ is the total number of location points in $\mathcal{S}$'s spatial datasets within the search range r.
- $|S_i|$ is the number of location points in S_i.
- $avg\_|S|$ is the average number of location points of the spatial datasets in $\mathcal{S}$.
- h is the saturation impact parameter of spatial datasets, where $h \geq 0$.
- b is a moderator that adjusts the influence of the number of location points in the spatial dataset on the relevance score, where $0 \leq b \leq 1$.

Moreover, h and b can be adjusted to make the relevance score calculation fit their actual needs. The larger the value, the higher the relevance of the dataset with more points in the search range. Therefore, h regulates the effect of the number of location points of the dataset in the search range on the relevance. The larger the value of b, the lower the relevance of datasets with a higher number of location points. Therefore, the value of b determines the effect of dataset size on relevance score.

B. Search Index Design

Quadtree is one of the important methods in spatial data searching, which can represent the spatial range more accurately and improve the efficiency of searching location points. Based on this, QRDI-index is designed. It can represent the distribution of location points in spatial datasets and greatly improve search efficiency.

Definition 5. Quadtree-based Region-dataset Inverted Index (QRDI-index). A QRDI-index is a quadtree recording location point numbers of spatial datasets in partitioned regions, denoted as $\mathcal{T}$. Each node in $\mathcal{T}$ is a triple,

$$< rect, plist, ptr[1 \cdots 4] >,$$

where

- $rect = (p_{\mathrm{L}}, p_{\mathrm{R}})$ is a rectangle region, where p_{L} and p_{R} are lower left and upper right coordinates.

- $plist = \{(S_i, num_i) | S_i \in \mathcal{S} \wedge num_i > 0\}$ is a posting list, where num_i is the number of location points of S_i located in the region $rect$;
- $ptr[1 \cdots 4]$ are four pointers of a quadtree node pointing to four child nodes which correspond to four equal regions generated via horizontal and vertical lines through the midpoint of the region $rect$, respectively.

According to Definition 5, each node in the QRDI-index corresponds to a rectangle region which consists of four equal rectangle regions corresponding to its four child nodes.

To construct the QRDI-index of the spatial data repository $\mathcal{S}$ with the layer threshold ω, each spatial dataset in $\mathcal{S}$ is processed, where each location point in the dataset is recursively inserted into a target node, and the corresponding posting list is updated. The details of index construction are presented in Algorithms 1 and 2.

Algorithm 1: $BuildIndex(\mathcal{S}, \omega)$

Input: the data repository $\mathcal{S}$, and the QRDI-index layer threshold ω.
Output: the QRDI-index $\mathcal{T}$.

```
Initial the QRDI-index T and set T.root = ∅;
for each S_i ∈ S do
    for each l_{i,j} ∈ S_i do
        InsertLoc(l_{i,j}, S_i, T.root, ω, 0);
Return T;
```

Algorithm 2: $InsertLoc(l_{i,j}, S_i, u, \omega, d)$

Input: the inserted location point $l_{i,j}$, the spatial dataset S_i where $l_{i,j}$ belongs to, the node u in QRDI-index, and the QRDI-index layer threshold ω.
Output: the QRDI-index after inserting $l_{i,j}$.

```
if l_{i,j} ∈ u.rect then
    if S_i is in u.plist then
        Get the pair (S_i, num_i) from u.plist, and set num_i = num_i + 1;
    else
        Set num_i = 1, and add the pair (S_i, num_i) into u.plist;
    if depth < ω then
        if u.ptr = ∅ then
            u.ptr ← Divide the coverage area of node u into four equal-sized rectangle regions and generate the corresponding child nodes;
        for each u.ptr[i], i ∈ {1,2,3,4} do
            if l_{i,j} is in u.ptr[i].rect then
                InsertLoc(l_{i,j}, S_i, u.ptr[i], ω, d + 1);
```

In Algorithm 1, we insert the location points of each spatial dataset recursively into each child node starting from the root node of the QRDI-index. In Algorithm 2, we insert $l_{i,j}$ of S_i into the node u. After inserting the point $l_{i,j}$ into the node u, we update the $u.plist$. If the depth of u is less than ω and $u.ptr$ is empty, the node u is divided into four child nodes with equal regions, and inserts $l_{i,j}$ into one of the four nodes.

Example.1 We illustrate the QRDI-index in Fig. 2. Given a spatial data repository $\mathcal{S} = \{S_1, S_2\}$ and assuming the layer threshold of QRDI-index, $\omega = 3$, we can obtain the QRDI-index $\mathcal{T}$. The location point distribution of spatial datasets in $\mathcal{S}$ is shown in Fig. 2a, and the QRDI-index is shown in Fig. 2b.

Based on the QRDI-index, we can obtain the nodes located in the search range, which contain all the location points in the search range r, and by using the posting list of these nodes, the relevance score of the spatial dataset to the search range can be quickly calculated.

C. Search Processing Algorithm

In this section, we propose a search algorithm based on QRDI-index. In this search algorithm, when the user gives the search $Q = (r, k, \mathcal{S})$, we can generate the minimum covering node set MCS based on the search range r and QRDI-index, and then count the number of location points for each dataset within r by MCS and return the top-k datasets.

Definition 6. Minimum Covering Node Set (MCS). For a given search $Q = (r, k, \mathcal{S})$, the minimum covering node set of Q is a node set that contains the minimum number of nodes of the QRDI-index covering the search range r.

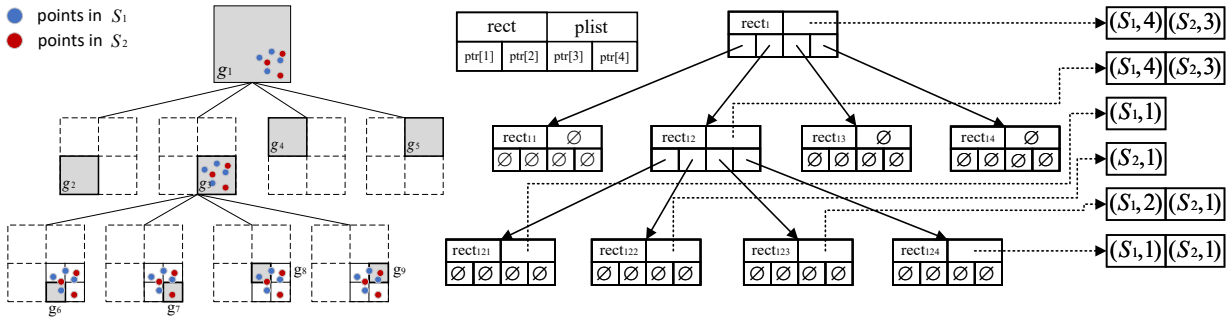
In order to improve the search efficiency, we need to use the minimum number of nodes to represent the search range r. So we design an algorithm for generating MCS by QRDI-index and r. The details of generating MCS are presented in Algorithm 3, where $u.rect \subseteq r$ means that the rectangle area $u.rect$ is in the search range r, and $u.rect \cap r$ is to get the intersected area between $u.rect$ and r.

Algorithm 3: $GenMCS(r, u)$

Input: The search range r, the node of QRDI-index u.
Output: The minimum covering node set MCS.

```
if u.rect ⊆ r then
    Add the u into MCS;
else if u.rect ⊈ r ∧ u.rect ∩ r ≠ ∅ then
    if u.plist ≠ ∅ then
        if u.ptr[i] ≠ ∅, i ∈ {1,2,3,4} then
            for each u.ptr[i], i ∈ {1,2,3,4} do
                MCS = GenMCS(r, u.ptr[i]);
        else
            Add the u into MCS;
Return MCS;
```

Example 2. Fig. 3 shows the example of the minimum covering node set. Given a QRDI-index $\mathcal{T}$ with $\omega = 5$ and a



(a) (b)

Fig. 2: An Example of QRDI-index.

search range represented by the dashed rectangle, then all the nodes in MCS are marked by color. Based on the *plist* in these nodes, we can obtain the number of location points for each dataset in the search range. This is the main content of the search algorithm.

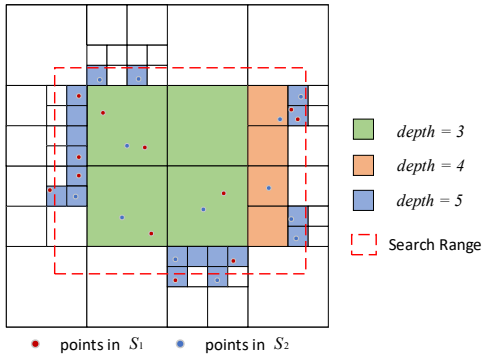


Fig. 3: An example of Minimum Covering Node Set.

To obtain the datasets with the highest relevance score to the search range, when the user gives a search $Q = (r, k, S)$, the minimum covering node set corresponding to the search range r is obtained by Algorithm 3, and then the total number of location points of each dataset within the search range $N_{S_j,r}$ is obtained by traverse MCS. Finally, the relevance score of each dataset with the search range is calculated. The specific process is shown in Algorithm 4.

Complexity Analysis. According to the execution process of Algorithm 3 and Algorithm 4, we analyze the time complexity of the search algorithm. The search process is divided into two steps, the first step is to get MCS, and the second step is to calculate the relevance by traversing the MCS. In the first step, We obtain MCS by recursively traversing the QRDI-index, assuming that the total number of nodes in the QRDI-index is N, and the time complexity of this process is $O(logN)$. In the second step, assuming that the average number of the nodes traversed is M, the time complexity is $O(M)$. So the time complexity of search processing is $O(M + logN)$.

Algorithm 4: $Search(r, k, \mathcal{S}, \mathcal{T})$

Input: The search range r, the number of requested spatial datasets k, the spatial data repository $\mathcal{S}$, and the QRDI-index $\mathcal{T}$.

Output: The search result $\mathcal{R}$.

Initial two pair sets $PS = \varnothing$ and $RS = \varnothing$, and initial the minimum covering node set $MCS = \varnothing$;
$MCS = GenMCS(r, \mathcal{T}.root)$;
for *each* $u_i \in MCS$ **do**
  for *each* $(S_j, num_j) \in u_i.plist$ **do**
    if S_j *is not in* PS **then**
      Add the pair (S_j, num_j) into PS;
    else
      Get the pair $(S_j, N_{S_j,r})$ of PS, and set $N_{S_j,r} = N_{S_j,r} + num_j$;
for *each* $(S_i, N_{S_i,r}) \in PS$ **do**
  Set $N_{\mathcal{S},r} = N_{\mathcal{S},r} + N_{S_i,r}$;
for *each* $(S_i, N_{S_i,r}) \in PS$ **do**
  Calculate $a = Score(S_i, r)$ according to Eq.(1);
  Add the pair (a, S_i) into RS;
$\mathcal{R} \leftarrow$ Get k spatial datasets with the highest relevance scores from RS;
Return $\mathcal{R}$;

IV. Performance Evaluation

In this section, we evaluate the performance of the proposed scheme using real-world spatial data repositories. The search efficiency and search accuracy of the proposed scheme under different parameter settings are evaluated. We compare the proposed scheme with the naive search scheme that determines the search result by traversing each spatial dataset and calculating the relevance score between the dataset and the search range. We evaluated the proposed scheme on three real spatial data repositories, $Identifiable$, $Public$, and $Trackable$, which are collected from OpenStreetMap [16] .

The experimental environment consists of an Intel computer.

V. Conclusion

The spatial dataset range search has become one of the important dataset search issues, as the number of open datasets continues to grow and most open datasets contain spatial information. In this paper, we propose a QRDI-index-based spatial dataset range search processing. A relevance measurement between a spatial dataset and a search range is first presented. A quadtree-based region-dataset inverted index is designed for accelerating search processing. By using the index, an efficient search processing algorithm is proposed. Experimental results on three real-world spatial data repositories show that the proposed search scheme performs well in search accuracy and efficiency.

References

[1] "Data.gov Home." https://data.gov/.

[2] "National Earth System Science Data Center." http://www.geodata.cn/.

[3] R. J. Miller, "Open data integration," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 2130–2139, 2018.

[4] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, and P. Groth, "Dataset search: A survey," *The VLDB Journal*, vol. 29, no. 1, pp. 251–272, 2020.

[5] S. Ghosh, T. Vu, M. A. Eskandari, and A. Eldawy, "UCR-STAR: The UCR spatio-temporal active repository," *SIGSPATIAL Special*, vol. 11, no. 2, pp. 34–40, 2019.

[6] Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada, "Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach," in *2021 IEEE 37th International Conference on Data Engineering (ICDE 2021)*, pp. 456–467, 2021.

[7] S. Castelo, R. Rampin, A. Santos, A. Bessa, F. Chirigati, and J. Freire, "Auctus: A dataset search engine for data discovery and augmentation," *Proceedings of the VLDB Endowment*, vol. 14, no. 12, pp. 2791–2794, 2021.

[8] A. Degbelo and B. B. Teka, "Spatial search strategies for open government data: A systematic comparison," in *Proceedings of the 13th Workshop on Geographic Information Retrieval (GIR '19)*, (New York, NY, USA), pp. 1–10, Association for Computing Machinery, 2019.

[9] T. Vu and A. Eldawy, "R-grove: growing a family of r-trees in the big-data forest," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2018)*, pp. 532–535, 2018.

[10] R. H. Güting, "An introduction to spatial database systems," *The VLDB Journal*, vol. 3, no. 4, pp. 357–399, 1994.

[11] G. Xu, H. Li, Y. Dai, K. Yang, and X. Lin, "Enabling efficient and geometric range query with access control over encrypted spatial data," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 870–885, 2018.

[12] R. Guo, B. Qin, Y. Wu, R. Liu, H. Chen, and C. Li, "MixGeo: Efficient secure range queries on encrypted dense spatial data in the cloud," in *Proceedings of the International Symposium on Quality of Service (IWQoS 2019)*, (New York, NY, USA), pp. 1–10, Association for Computing Machinery, 2019.

[13] Y. Miao, Y. Yang, X. Li, Z. Liu, H. Li, K.-K. R. Choo, and R. H. Deng, "Efficient privacy-preserving spatial range query over outsourced encrypted data," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 3921–3933, 2023.

[14] P. Li, H. Dai, J. Sun, S. Wang, W. Yang, and G. Yang, "Edss: An exemplar dataset search service over encrypted spatial datasets," in *2024 IEEE International Conference on Web Services (ICWS 2024)*, pp. 1344–1346, IEEE, 2024.

[15] P. Li, H. Dai, S. Wang, W. Yang, and G. Yang, "Privacy-preserving spatial dataset search in cloud," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*, pp. 1245–1254, 2024.

[16] W. Yang, S. Wang, Y. Sun, and Z. Peng, "Fast dataset search with earth mover's distance," *Proceedings of the VLDB Endowment*, vol. 15, no. 11, pp. 2517–2529, 2022.

[17] W. Yang, S. Wang, Y. Sun, Z. Chen, and Z. Peng, "Efficient spatial dataset search over multiple data sources," *arXiv preprint arXiv:2311.13383*, 2023.

[18] B. He and I. Ounis, "Term frequency normalisation tuning for bm25 and dfr models," in *European conference on information retrieval (ECIR 2005)*, pp. 200–214, Springer, 2005.

[19] A. Trotman, A. Puurula, and B. Burgess, "Improvements to bm25 and language models examined," in *Proceedings of the 19th Australasian Document Computing Symposium*, pp. 58–65, 2014.

[20] G. E. Pibiri and R. Venturini, "Techniques for inverted index compression," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–36, 2020.

[21] H. Samet, "The quadtree and related hierarchical data structures," *ACM Computing Surveys (CSUR)*, vol. 16, no. 2, pp. 187–260, 1984.