ISSN: 2321-7782 (Online)
Volume 3, Issue 11, November 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Survey on Fast nearest Neighbor Search Using K-D Tree and Inverted Files

Swapnali M. Mahadik[1]
Computer Department
Tssm's Bhivarabai Sawant College of Engineering And Research, Pune, India

Prof. Sucheta M. Kokate[2]
Computer Department
Tssm's Bhivarabai Sawant College of Engineering And Research, Pune, India

Abstract: ***Spatial queries, such as range search and nearest neighbor retrieval, involve only conditions on objects geometric properties. A spatial database manages multidimensional objects such as points, rectangles, etc. This type of database provides fast access to those objects based on different selection criteria. Now-a-days many applications call a new form of queries to find the objects that satisfying both a spatial predicate, and a predicate on their associated texts. For example, instead of searching all the restaurants, a nearest neighbor query would ask for the restaurant that is the closest among those whose menus contain the specified keywords all at the same time. Existing system develop a new access method called the spatial inverted index that extends the conventional inverted index. In our proposed work we present the k-d tree for location search and inverted files for the fastest keyword search. In this approach, we propose the inverted files in combination with k-d tree to accomplish the current user needs. It avoids searching in overlapping area.***

I. INTRODUCTION

The Nearest neighbor search also known as closest point search or similarity search .Nearest neighbor search returns the nearest neighbor of a query point in a set of points, it is an important and widely studied problem in many fields, and it has wide range of applications. We can search closest point by giving set of keywords as input; it can be spatial or textual. A spatial database use to manage multidimensional objects i.e. points, rectangles, lines etc. Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. The search engines available today provide keyword search on top of sets of documents. When a set of query keywords is provided by the user, the search engine returns all documents that are associated with these query keywords. The best solution to queries is based on the IR2-tree. But Efficiency of IR2-tree is badly impacted because of some drawbacks in it. The solution for overcoming this problem should be searched. Existing system develop a new access method called the spatial inverted index that extends the conventional inverted index.

In our proposed work we present the k-d tree for location search and inverted files for the fastest keyword search. In our proposed approach, we propose the inverted files in combination with k-d tree to accomplish the current user needs. It presents hybrid index structure for range keyword query searching with minimum IO cost and CPU cost. It avoids searching in overlapping area. So it can reduce searching time in overlap area.

II. LITERATURE SURVEY

The Literature survey is the most important step in software development process. Before developing any tool it is required to determine the time factor, economy and company strength. Once these things are satisfied, then the next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the too then he/she needs lot of external support. This support can be obtained from senior programmers, book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

Cao et al. [1] projected collective spatial keyword queries, they conferred the new down side of retrieving a group of spatial objects, and every related to a collection of keywords. They develop approximation algorithms with provable approximations bounds and precise algorithms to solve the two issues.

Lu et al. [2]combined the notion of keyword searches with reverse nearest neighbor queries. They propose a hybrid index tree known as IUR-tree (Intersection-Union R-Tree) to answer the Reverse spatial textual k Nearest Neighbor (RSTkNN) question that effectively combines location proximity with matter similarity. They style a branch-and-bound search rule that relies on the IUR-tree. To more increases the question process, they proposed to associate degree improved variants of the IUR-tree known as cluster IUR-tree and two corresponding improvement rules.

Zhang and Chee [3] introduced hybrid categorization structure BR*-tree, that mixes the R*-tree and bitmap indexing to method the m-closest keyword question that returns the spatially nearest objects matching m keywords. They utilized a priority based mostly to search strategy that successfully cut backs the search area and additionally planned two monotone constraints, distance mutex and keyword mutex to help effective pruning.

Ian DE Flipe [4] given in economical technique to answer top-K spatial keyword question. They planned in index structure IR2-tree that mixes signature files and R-tree to allow keyword search on spatial knowledge objects that every have restricted range of keywords. Exploitation the IR2-tree inefficient progressive rule is given to answer the spatial keyword queries

G. Cong, C.S. Jensen, and D. Wu dialect [5] planned to associate approaches that compute the relevancy between the documents of associate objects and a question. This relevancy is then incorporated with the geometrician distance between an object and question to calculate associate overall similarity of object to query.

III. Nearest Neighbor Search Technique

a) IR-Tree, Approximation Algorithmic Rule and Precise Algorithm.

This technique is employed to retrieve a bunch of spatial internet objects specified the query's keywords measure cowl by group's keywords and objects area unit around the question location and have very cheap bury object distances. This technique addresses two internal representation of the cluster keyword question. First is searching out the cluster of objects that cowl the keywords such that the add of their distances to the question is minimized. Second is searching out a bunch of objects that cowl the keywords specified add of the most distance among associate object in cluster of objects and question and maximum distance among two objects in cluster of objects is reduced. Each of those sub issues area unit NP-complete. Greedy rule is employed to produce associate approximation solutions to the matter that utilizes the spatial keyword index IR-tree to scale back the search house. However in some application question doesn't contain an oversized range of keywords, for this actual rule is employed that uses the dynamic programming [1]

b) IUR-Tree (Intersection union R-Tree):

Geographic objects related to descriptive texts area units becoming common. This provides importance to special keyword queries that take each the situation and text description of content. This system is employed to research the problem of reverse spatial and matter k- nearest neighbor search that is finding objects that takes the question object in concert of their spatial matter similar objects. For this kind of search hybrid index structure are employed that with success merge the situation proximity with matter similarity.

For searching, the branch and sure algorithmic program is employed. Additionally to increase the speed of question process a variant of IUR- tree and two improvement algorithmic programs is employed. To reinforce the IUR-tree text cluster is employed, during this object of all the information base is cluster into clusters in keeping with their text similarity. Every node of the tree is extended by the cluster information to make a hybrid tree that is termed as cluster IUR-tree. To boost the search

performance of this tree two improvements way is employed, an initial relies on outlier detection and extraction and second technique is predicated on text entropy. [2]

c) BR*-Tree:

This hybrid index structure is employed to look m-closest keywords. This method finds the nearest tuples that match the keywords provided by the user. This structure combines the R*-tree and classification assortment to method the m-closest keyword question that returns the spatially nearest objects matching m keywords to scale back the search house apriority based mostly search strategy is employed. Two monotone constraints are employed as priority properties to facilitate efficient pruning that is named as distance mutex and keyword mutex. However this approach isn't appropriate for handling ranking queries and during this range of false hits is large.[3]

d) IR2-Tree:

The growing range of applications needs the efficient execution of nearest neighbor queries that is constrained by the properties of spatial objects. Keyword search is extremely common on the internet therefore these applications allow users to present list of keywords that spatial objects should contain. Such queries known as a spatial keyword query. This can be consisted of question space and set of keywords. The IR2-tree is developed by the mix of R-tree and signature files, wherever every node of tree has spatial and keyword data. This technique with its efficiency provides solution to top-k spatial keyword queries. Also it facilitates the signature is added to each nodes of the trees. Associate in a position rule is employed to answer the queries exploitation the tree. Progressive nearest algorithm is employed for the tree traversal and if root node signature doesn't match the question signature then it prunes the whole sub trees. However the IR2-tree has some drawbacks such as false hits ratio, wherever the thing of ultimate result's isolated from the question or this can be not appropriate for handling ranking queries.[4]

e) Spatial Inverted Index and Minimum Bounding Method:

New access technique spatial inverted access technique is used to get rid of the drawbacks of previous strategies such as false hits. This technique is that the variant of inverted index using for two-dimensional points. This index stores the spatial region of information points and on each inverted list R-tree is built. Minimum bounding technique is employed for traversing the tree to prune the search area.[6]

IV. IMPLEMENTATION

a) Contribution and Objectives:

1. To create index structure which combine K-d tree and inverted file to increase efficiency of spatial keyword queries with minimum time.
2. To develop a range keyword search algorithm using proposed index structure to explore useful and important information for user queries.

b) Proposed Work:

In spatial databases the most of the queries are range queries and nearest neighbor queries. In text retrieval, queries may be based on the Boolean or ranking based returning the top k result matching to the query.

But now a days the user interest has been increased in spatial databases and expect the text based search results along with the closer geographical location. In order to accomplish the user requirements we propose a new framework which combines the text based search with location based search using proposed index structure.

The *Boolean range* queries are receiving more attention *query* $q = (p, K_w)$, where p is a spatial region and $K_w$is a set of keywords, returns all places that arelocated in region pand that contain all the keywords in K_w.

 ISSN: 2321-7782 (Online) ISSN: 2347-1778 (Print)

The *Boolean kNN query q=(Pl,k_w,k)* takes three arguments, where Plis a point location, $K_w$is as above, and *k* is the number of places to return.

Next, the *top-k range query q = (p, K_w, k)* where p,k_w, and *k* are as above, returns up to *k* places that are located in the query region p, now ranked according to their text relevance to K_w. Finally, the *top-k kNN query* takes the same arguments as the Boolean *k*NN query. It retrieves *k* objects ranked according to a score that takes into consideration spatial proximity and text relevance.

Let D is a spatial database which contains D= {O_1,O_2,O_3…..,O_n } such that every object o in D has many attributes; <O_{id}, O_l, O_d> where O_{id} is an identifier of an object, O_l is a spatial location that contain latitude and longitude and O_d is an text document of each object for keyword querying.

Let q=<q_k, q_r> be a Boolean keyword range query where q_k is user required keywords w_1…..w_m and $q_r$is the user desired range A query q return allobjects in D that contain all keywords q_k={w_1, w_2,…., w_m} and belong to the range q_r .

$$Ans(q) = \begin{cases} o\epsilon\ qr\ ; O\ is\ contained\ in\ qr \\ o\epsilon qk; \quad All\ w\ belongs\ to\ qk \end{cases}$$

c) Proposed Algorithm:

We propose an algorithm to find the nearest restaurant location with the specified menu list. We propose an algorithm which uses the above proposed index structure. Our approach uses the k-d tree to search the nearest location and inverted files for the fast and nearest keyword search. The combination of these two methods improves the output results in terms of the time and speed.

Input: Location, Range, Keywords.

Output: Object list nearest to query object.

1. Begin
2. Get the user query with required details.
3. Find results in the user specified range.
4. Filter results found in the above steps using the specified keyword set.
    a. Perform sorting according to the similarity scores.
    b. Output the top results with user requirements.
5. Stop.

The proposed system integrates the text and location index to process spatial keywords queries. K-d tree is loosely combined with the inverted file for text information retrieval. For each node of K-d tree, an inverted file is created for indexing the text components of objects contained in the node.

V. CONCLUSION

As K-d trees represent a disjoint partition, the proposed system can't cause more IO costs and also K-d trees don't need to rebalance the textual information so the proposed can reduce update cost (CPU costs). Most geo-textual indices use the inverted file for text indexing. Inverted file can be used to check the query keywords contain or not. K-d tree structure is known as point indexing structures as it is designed to index data objects which are points in a multi-dimensional space. It can be used efficiently for range queries and nearest neighbor queries.

 ISSN: 2321-7782 (Online) ISSN: 2347-1778 (Print)

ACKNOWLEDGMENT

I have taken efforts in this survey on fast nearest neighbor search using K-d tree and inverted files. However, it would not have been possible without the kind support and help of many individuals. I am highly indebted to Prof. S.M.Kokate for her guidance and constant supervision as well as for providing necessary information regarding this approach.

References

1. X. Cao, G. Cong, C.S. Jensen, and B.C. Ooi, "Collective Spatial Keyword Querying," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 373-384, 2011.
2. J. Lu, Y. Lu, and G. Cong, "Reverse Spatial and Textual k nearest Neighbor Search," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 349-360, 2011.
3. D. Zhang, Y.M. Chee, A. Mondal, A.K.H. Tung, and M. Kitsuregawa, "Keyword Search in Spatial Databases: Towards Searching by Document," Proc. Int'l Conf. Data Eng. (ICDE), pp. 688-699, 2009.
4. G. Cong, C.S. Jensen, and D. Wu, "Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects," PVLDB, vol. 2, no. 1, pp. 337- 348, 2009.
5. I.D. Felipe, V. Hristidis, and N. Rishe, "Keyword Search on Spatial Databases," Proc. Int'l Conf. Data Eng. (ICDE), pp. 656-665, 2008.
6. Yufei Tao and Cheng Sheng "Fast nearest Neighbor Search with Keywords" IEEETransactionsOn Knowledge And Data Engineering, Vol.26, No.4, April2014.

AUTHOR(S) PROFILE

Swapnali Mahadik, is currently pursuing M.E (Computer) from Department of Computer Engineering, Bhivarabai Sawant College of Engineering and research, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. She received her B.E (Computer Science and Engineering) Degree from KIT's college of Engineering , Kolhapur, India. Shivaji University, Kolhapur, Maharashtra, India -416113. Her area of interest is Data Mining.

Sucheta Kokate, received the M.Tech (CST) degree from the Department of Computer Engineering, Shivaji University, Kolhapur, Maharashtra, India in 2014. He is currently working as Asst. Professor with Department of Computer Engineering, Bhivarabai Sawant College of Engineering and Research, Pune, MAH, India.

