ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 7, July 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

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

Swapnali M. Mahadi[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 conditions on only the objects geometric properties. A spatial database manages multidimensional objects. Then spatial database provides fast access to the objects based on different selection criteria. There are many applications used a new form of queries to find the objects that satisfying both a spatial predicate and a predicate on their associated texts. Consider the example of restaurant, 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(SI) index that extends the conventional inverted index. Existing system uses IR Tree index method. IR tree uses binary tree data structure which requires rebalancing. It increases the I/O cost. In proposed work the k-d tree is used for location search which does not require rebalancing and uses inverted files for the fastest keyword search. The proposed approach reduces the overhead caused by the existing approach. It also avoids searching in overlapping area. Smart phones and other trendy mobile wearable devices are rapidly becoming the dominant sensing, computing and communication devices in human being's daily lives. Hence the proposed system of data mining is used with the mobile computing.

Keywords: Nearest neighbour search, keyword search, K-d tree and inverted files.

I. INTRODUCTION

The Nearest neighbors search also known as the closest point search or also known as similarity search. Nearest neighbor search returns the nearest neighbor of a query point in a set of points, it has wide range of applications. We can search closest point by giving set of keywords as input to the query; keywords can be spatial or textual. A spatial database use to manage multidimensional objects like points, rectangles etc. Keyword search is the most popular information discovery method because the user does not need to know either a query language or structure of the data. 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 such type of query is based on the IR2-tree and signature files. But IR2 has many drawbacks and it is not sufficient to give real time answers.The solution to this problem should be searched. Existing system develop a new access method called the spatial inverted (SI) index which is the combination of IR tree and inverted files.

In this proposed the k-d tree is used for location search and inverted files for the fastest keyword search. In this work the inverted files used in combination with k-d tree to accomplish the current user needs. It presents hybrid index structure for range keyword query searching with minimum CPU cost. It avoids searching in overlapping area. Which is the drawback of existing ststem.

II. LITERATURE SURVEY

Cao et al. [1] proposed the 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 developed an 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 query that effectively combines location proximity with matter similarity. They used a branch-and-bound search rule that relies on the IUR-tree. To more increases the query 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 used to help effective pruning.

Ian DE Flipe [4] proposed in economical technique to answer top-K spatial keyword question. They proposed an 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. But IR2 tree is inefficient to give real time answer to the spatial keyword queries.

G. Cong, C.S. Jensen, and D. Wu dialect [5] proposed an approach 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 and provide the answer to question.

III. NEAREST NEIGHBOR SEARCH TECHNIQUE

This technique is proposed to retrieve a bunch of spatial internet objects specified the query's keywords The 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 the cluster of objects that cowl the keywords such that the address of their distances to the question is minimized. Second is searching a bunch of objects that cowl the keywords specified address 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 used to produce associate approximation solutions to the matter that utilizes the spatial keyword. The 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].

A. 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 proposed 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 used that merge the situation proximity with matter similarity.

For searching, the branch and sure algorithmic is used. 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]

B. *BR*-Tree*

This hybrid index structure is proposed to look m-closest keywords. This method finds the nearest tuples that match the keywords provided by the user with query. 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 used. Two 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]

C. *IR2-Tree*

The growing range of applications needs the efficient execution of nearest neighbor queries. The Keyword search is extremely common on the internet therefore these applications allow users to present list of keywords that spatial objects should satisfies. Such type of queries known as a spatial keyword query.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 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. The 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]

D. *Spatial Inverted Index and Minimum Bounding Method*

New access technique spatial inverted access technique is used to overcome 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. MOBILE COMPUTING FOR DATA MINING

Mobile computing for desired data mining is the discipline for creating an information management platform, which is free from spatial and temporal constraints. The freedom from these constraints allows its users to access and process desired information from anywhere in the space. The state of the user, static or mobile, does not affect the information management capability of the mobile platform. A user can continue to access and manipulate desired data while traveling on plane, in car, on ship, etc. Thus, the discipline creates an illusion that the desired data and sufficient processing power are available on the spot, where as in reality they may be located far away. Otherwise Mobile computing is a generic term used to refer to a variety of devices that allow people to access data and information from where ever they are.

V. PROPOSED APPROACH

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.

A. MODULES

1. Keyword search.
2. User Location Search

3. Google Map.

1. **Keyword Search**:

   Search restaurants by keywords of the restaurants. Keyword such as spice, drinks, bars. These keywords embedded with restaurants, used as the key for the restaurant.

2. **User Location Search:**

   Filter the results of the keyword, user can choose the kilometer range. That is user expect the results around the radius of the kilometer.

3. **Google map**:

   Google map is implemented to show the user location and the results of the restaurants.

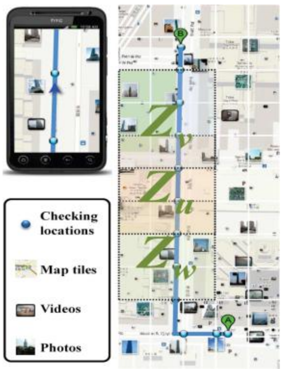


Fig-1 System architecture

B. ALGORITHM OF PROPOSED SYSTEM

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.

VI. Stepwise Screenshots of Proposed Application

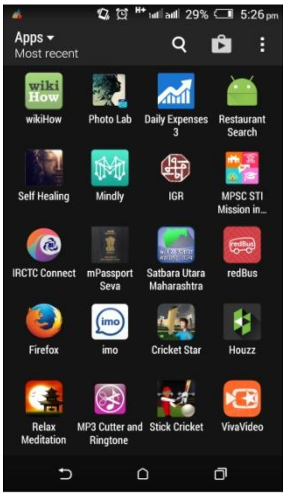


Fig- Home screen of proposed application

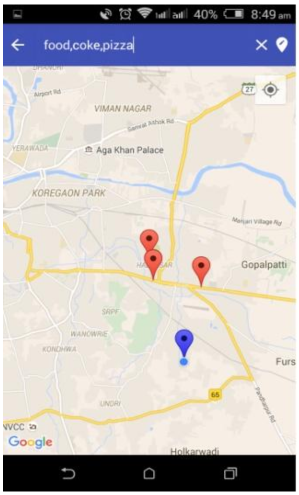
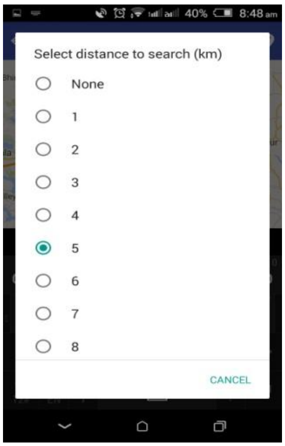

food,coke,pizza

Fig- Search keyword


Select distance to search (km)

○ None
○ 1
○ 2
○ 3
○ 4
● 5
○ 6
○ 7
○ 8

Fig- Selection of location

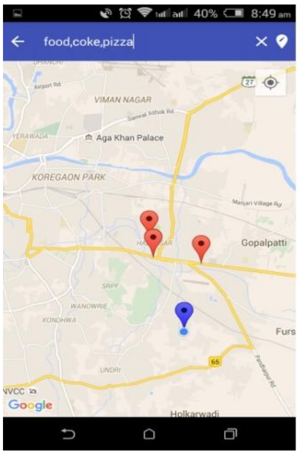

food,coke,pizza

Fig- Search Result

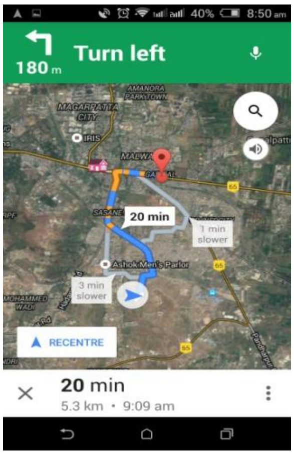
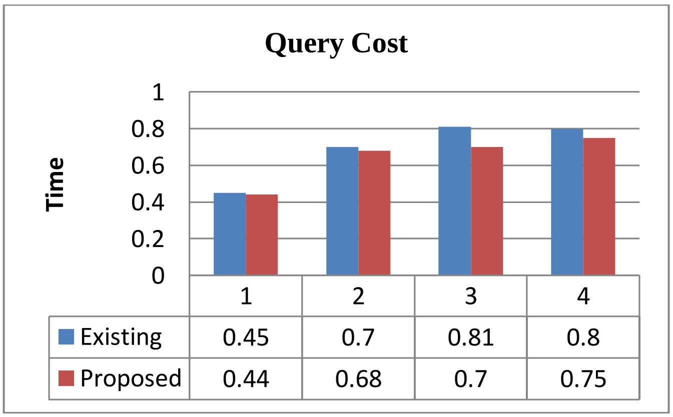

Turn left
180 m
20 min
5.3 km • 9:09 am

Fig- Navigation for resulted location



	1	2	3	4
🟦 Existing	0.45	0.7	0.81	0.8
🟥 Proposed	0.44	0.68	0.7	0.75

Fig- Query time versus the number of keywords

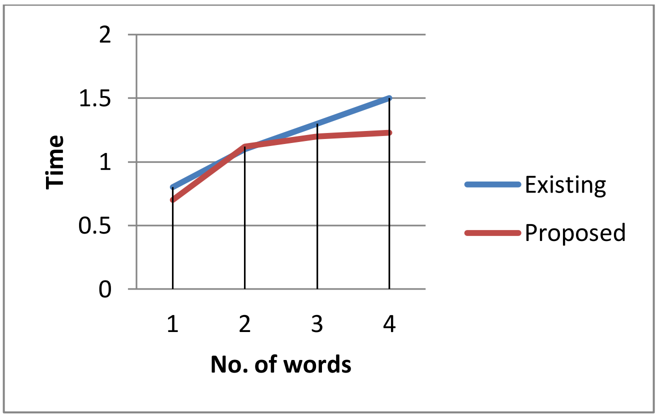


Fig- Comparison of system using Query time & No. of keywords.

VII. 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. K-d tree uses disjoint partition hence it requires less time compared to existing system. It can be used efficiently for range queries and nearest neighbor queries.

ACKNOWLEDGEMENT

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. 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

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.