ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Survey On: Nearest Neighbour Search With Keywords In Spatial Databases

SayaliBorse[1], Prof. P. M. Chawan[2], Prof. VishwanathChikaraddi[3], Prof. Manish Jansari[4]

P.G. Student, Dept. of Computer Engineering& IT, V.J.T.I., Mumbai, Maharashtra, India[1]

Associate Professor, Dept. of Computer Engineering& IT, V.J.T.I., Mumbai, Maharashtra, India[2]

Assistant Professor, Dept. of Computer Engineering & IT, V.J.T.I., Mumbai, Maharashtra, India [3]

Assistant Professor, Dept. of Computer Engineering & IT, V.J.T.I., Mumbai, Maharashtra, India [4]

ABSTRACT:Users may search for different type of things from anywhere. But Search results depend on the user entered query which has to satisfy their searched properties that is stored in the spatial database. Due to rapid growth of users it become essential to optimize search results based on nearest neighbour property in spatial databases. Conventional spatial queries, such as range search and nearest neighbour retrieval, involve only geometric properties of objects which satisfies condition on geometric objects. Now a days many modern applications aim to find objects satisfying both a spatial condition and a condition on their associated texts which is known as Spatial keyword search.For example, the search engine can be used to find those nearest restaurant which offers Chinese Dishes like "dumplings, pastry, stew and porridge" or hospitals or chemists all at the same time. There are some straight forward approach which first deals with spatial condition and then on the non-spatial condition as a process of reduction. But these approaches are not good with complex queries. Solution to such queries is based on the IR2-tree, but IR2-tree also have some drawbacks which reduces the efficiency of IR2-tree badly.

KEYWORDS:Spatial database, Data mining, Spatial queries, Spatial keyword search, Spatial inverted index

I. INTRODUCTION

Spatial databases are the databases used to hold geographical information. They contain data which are multidimensional in nature. A spatial database manages the multidimensional objects such as points and rectangles, etc. Such databases can be used for obtaining information based on queries and provides fast access to which objects based on the different the selection criteria. Now a days, the general use of the search engines had made it possible to write the spatial queries in the new way. Conventionally, queries much focus on the object's geometric properties only, for example, whether the point is in the rectangle, or how close two points are from the each other in the rectangle likewise query can be made on hotels, locations, schools, hospitals, lakes, parks and so on with various combinations of criteria. In the world there is need oflocation based services to search nearest neighbour. The NN search needs to process huge amount of geography information. As the number of geographic properties are more and multi-dimensional in nature, developing such systems is considered difficult task. Search speed is an important factor in such systems. With respect to a web application for making searches on geographical data, speed is an essential element. The application should produce desired results in short span of time. This is a challenging problem needs to be addressed.

Nearest neighbour search is to find the nearest object contains the set of keywords. For example if the users are interested to find the nearest restaurants that offers Chinese Dishes like "dumplings, pastry, stew and porridge" all the same time or chemists offering particular brand medicines . There are two approaches to implement such queries that combine spatial and text features. In the first approach, we could first fetch all the restaurants whose menu contains the set of keywords {Chinese Food, dumplings, pastry, stew and porridge} and then from retrieved restaurants, find the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

nearest one. Similarly in the second approach same query result can also find out reversely by targeting first spatial condition and then textual condition. IR2-tree is used in the existing system for providing best solution for finding nearest neighbour but this method has few deficiencies.

II. RELATED WORK

Nearest neighbour search (NNS), also known as closest point search, similarity search. It is an optimization problem for finding closest (or most similar) points. We can search closest point by giving keywords as input; it can be spatial or textual.

Cao et al. [1] presented paper on the new problem for retrieving a spatial object's groups that are nearest to the query point location and each associated with a set of keywords. They collective known as spatial keyword query.

G. Cong, C.S. Jensen, and D. Wu [2] proposed an approach that computes the relevancy of a query result by means of language models and a probabilistic ranking function. This relevance is then incorporated with the Euclidean distance between object and query to calculate an overall similarity of object to query.

Zhang and Chee [3] introduced hybrid indexing structure bR*-tree, that combines the R*-tree and bitmap indexing to process the m-closest keyword query that returns the spatially closest objects matching m keywords.

In [4] Ian De Felipe et al. focused on finding objects closest to a specified location contains set of keywords. A method to answer top k spatial queries is effectively presented the method has tight of data structure and algorithms used in spatial database search and Information Retrieval.

In [5] Chen li and BijiHore focused on location based information retrieval. A framework is proposed for Geographical Information Retrieval (GIR) system and focus on indexing strategies can process Spatial Keywords (SK) queries effectively. Two type of indexing mechanism is used. First method is separate index for spatial and text attribute. Second method is hybrid indices techniques combine the spatial and inverted file indices.

III. SPATIAL DATA STRUCTURES

1. Quad Tree:

The quad tree is used for indexing 2-Dimensional space. Each internal node of the quad tree splits the 2D space into four disjoint subspaces according to the axes [6]. Further each subspaces is split until there is one object or no object inside each of subspaces recursively. The quad tree is not balanced tree. Its balance depends on the data distribution among subspaces and the order of inserting the data points.

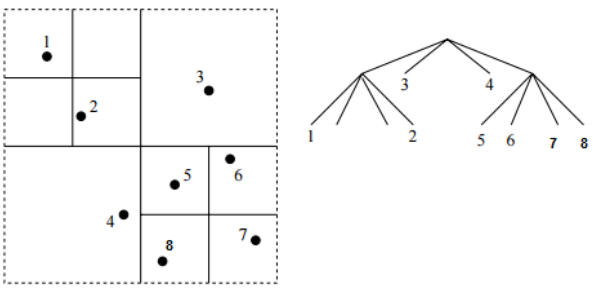

Fig.1 Quad Tree

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

2. k-d Tree:

In k-d tree method a binary tree is used to split k–dimensional space [7]. According to one of the coordinates of the splitting point k-d tree splits the space into two subspaces as shown in below figure. Let level(n) be the length of the path from the root to the node n and suppose the axes are numbered from 0 to k − 1. At the level level(n) in every node the space is split according to the coordinate number (level(n) mod k). Insertion and searching of the data are similar to the binary trees. According to the coordinate number (level(n) mod k) we only have to compare nodes.
Disadvantage: It is very sensitive to the order of inserting the objects.

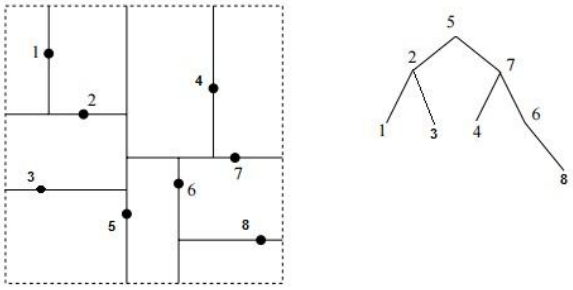

Fig.2 k-d Tree

3. R-Tree:

The R–tree [7] is the modified B-tree for spatial data. R-tree is balanced tree. In R-tree spaces are splits into the rectangles which can overlap unlike quad tree. Except root node each node contains ‘n’ to ‘N’ children, where 2 <= n <= N/2. The root node contains 2 or more children unless it is a leaf. Below figure shows an example of order 3 R-tree. Each node is represented by the minimum bounding rectangle contains all the objects of its subtree. Each node is split recursively. In the leaf node pointers are stored which points to the data objects.

Because of the overlapping of bounding rectangles it could be necessary to search more than one branch of the tree. Therefore, it is important to separate the rectangles as much as possible. This problem must be solved by the operation INSERT which uses some kind of heuristic. It finds such a leaf that inserting a new object into it will cause as small changes in the tree as possible. The splitting operation is also important. We want to decrease the probability that we will have to search both new nodes. Testing all the possibilities has exponential complexity, so the algorithms for approximate solution are used. The R-tree is one of the most cited spatial data structures and it is very often used for comparison with new structures.

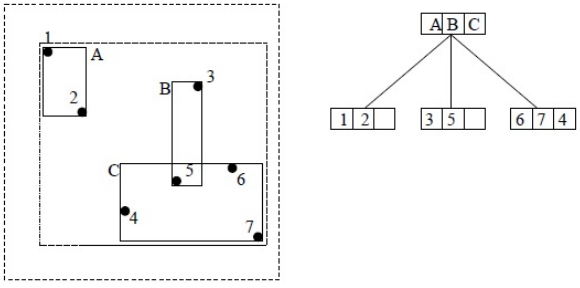

Fig.3 R-Tree

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

4. R*-Tree:

R*-tree [8] is a modified R–tree. It uses different kinds of heuristic for inserting data into the tree like R-tree. R*–tree combines the criteria of minimizing the area of all nodes of the tree like in the R-tree with the area under the bounding rectangle, the margin of a rectangle and the overlap between rectangles. The objective of decreasing the area under the bounding rectangle is to decrease the dead space, which means the space covered by the bounding rectangle but not by the enclosed rectangles. This decreases the number of branches that are searched unnecessarily. Minimization of the margin (the sum of the lengths of the edges) of a bounding rectangle prefers the squares. Minimization of the overlap between rectangles decreases the number of paths that must be searched. The implementation of this method is harder, but R*–trees are more effective than R-trees.

5. R+ -Tree:

R+–tree is an extension of the R–tree. In the R+–tree bounding rectangles of the nodes at one level can overlap. This decreases the number of searching branches of the tree and reduces the time for searching. R+–tree allows splitting of data objects due to which different parts of one object can be stored in more than one node on the same level of the tree. If in case a rectangles overlaps we decompose them into a group of non-overlapping rectangles which cover the same data objects. This increases a space consumption. But due to zero overlap of the nodes it reduces the time consumption.

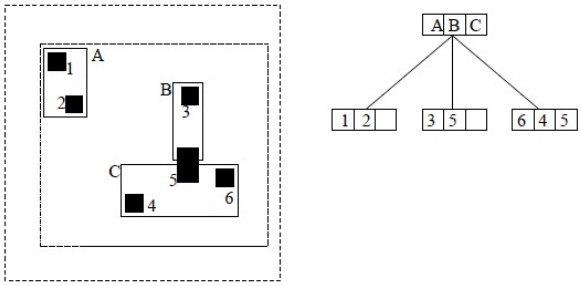

Fig.4 R+ - Tree

IV. NEAREST NEIGHBOUR SEARCH TECHNIQUE

1. IR-Tree:

IR-tree is used to retrieve a group of spatial web objects. Those group of objects contains query's keywords and objects are near to the query location and have the lowest inter object distances. This method addresses the two illustration of the group keyword query. In the first illustration we have to find the group of objects that cover the query keywords such that the sum of their distances to the query point is minimum. In the second illustration we find a group of objects that cover the query keywords such that sum of the maximum distance among an object in group of objects and query and maximum distance among two objects in group of objects is minimum. Both of these sub problems are NP-complete. An approximate solution to the above problem is provided by Greedy algorithm that utilizes IR-tree (spatial keyword index) to reduce the searching space. But in some application query contain a small number of keywords, for this exact algorithm is used that uses the dynamic programming.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

2. IUR-Tree (Intersection union R-tree):

IR-tree is used to retrieve a group of spatial web objects. Those group of objects contains query's keywords and objects are near to the query location and have the lowest inter object distances. This method addresses the two illustration of the group keyword query. In the first illustration we have to find the group of objects that cover the query keywords such that the sum of their distances to the query point is minimum. In the second illustration we find a group of objects that cover the query keywords such that sum of the maximum distance among an object in group of objects and query and maximum distance among two objects in group of objects is minimum. Both of these sub problems are NP-complete. An approximate solution to the above problem is provided by Greedy algorithm that utilizes IR-tree (spatial keyword index) to reduce the searching space. But in some application query contain a small number of keywords, for this exact algorithm is used that uses the dynamic programming.

3. BR*-Tree:

This hybrid index structure is used to search m-closest keywords [9]. This technique finds the closest tuples that matches the keywords provided by the user. For processing the m-closest keyword query BR*-tree structure combines the features of R*-tree and bitmap indexing and returns the spatially closest objects matching m keywords as a results. A priori based search strategy is used to reduce the search space. To facilitate efficient pruning, two monotone constraints are used as a priori properties. But this method has some disadvantages while handling ranking queries and in this number of false hits is large.

4. IR2-Tree:

IR2-tree is a combination of R-tree and signature files [10]. R-trees and the best-first algorithm for NN search are well-known techniques in spatial databases. In general, Signature file refers to a hashing-based framework which is known as superimposed coding (SC), which is more effective than other instantiations. It is designed to perform membership tests: determine whether a query word't' exists in a set 'T' of words. SC is conservative, if it says "no", then w is definitely not in W. On the other hand, if SC returns "yes", the true answer can be either way, in which case the whole W must be scanned to avoid a false hit.

V. CONCLUSION

This paper presents the survey of various spatial data structures and techniques for nearest neighbor search for spatial database. There were many drawbacks in the previous methods. The current existing solutions for finding nearest neighbor keyword are unable to give real time answer and have drawback of more space consumption.

REFERENCES

[1] X. Cao, G. Cong, C.S. Jensen, and B.C. Ooi, "Collective Spatial Keyword Querying", in Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 373- 384, 2011.
[2] 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.
[3] D. Zhang, Y.M. Chee, A. Mondal, A.K.H. Tung, and M. Kitsuregawa, "Keyword Search in Spatial Databases: Towards Searching by Document",in Proc. Int'l Conf. Data Eng. (ICDE), pp. 688-699, 2009.
[4] D. Felipe, V. Hristidis, and N. Rishe.,"Keyword search on spatial databases", in Proc. of International Conference on Data Engineering (ICDE), pp.656–665, 2008.
[5] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems", in Proc. of Scientific and Statistical Database Management (SSDBM), 2007.
[6] Ester M., Kriegel H. P., Sander J.,"Spatial Data Mining: A Database Approach",in Proc. of the Fifth Int'l Symposium on Large Spatial Databases (SSD '97), Berlin, Germany, Springer, 1997.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

[7] Gaede V., G¨ unther O.,"Multidimensional Access Methods",ACM Computing Surveys, Vol. 30, No. 2, pp. 170 – 231, June 1998.
[8] Beckmann N., Kriegel H. P., Schneider R. and Seeger B.,"The R*-tree: an efficient and robust access method for points and rectangles",in Proc.of the ACM SIGMOD International conference on Managementof data, Atlantic City, NJ USA, pp. 322-331, 1990.
[9] Yufei Tao and Cheng Sheng, "Fast Nearest Neighbor Search with Keywords", IEEE transactions on knowledge and data engineering, VOL. 26, NO. 4, APRIL 2014.
[10] Cao,Cong, G., and Jensen, C. S.,"Retrieving top-k prestige based relevant spatial web objects", PVLDB, 3(1), pp. 373–384, 2010.