International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.4,No.7, July 2015 DOI:10.15693/ijaist/2015.v4i7.118-121

Spatial Indices: Methods for Spatial Information Retrieval and Its Analytical Performance Evaluation

Eldhose Paul[1]
[1](Computer Science and Engineering, ASIET Kalady , India,mr.eldhose@gmail.com)

Ierin Babu[2]
[2](Computer Science and Engineering ,ASIET, Kalady,India, ierinbabu@gmail.com)

***Abstract* - Many applications require finding objects closest to a specified location that contains a set of keywords. For example, online yellow pages allow users to specify an address and a set of keywords. In return, the user obtains a list of businesses whose description contains these keywords, ordered by their distance from the specified address. The problems of nearest neighbor search on spatial data and keyword search on text data have been extensively studied separately. However, to the best of our knowledge there is no efficient method to answer spatial keyword queries, that is, queries that specify both a location and a set of keywords. In this work, we present an efficient method to answer top-k spatial keyword queries. To do so, we introduce an indexing structure called IR^2-Tree (Information Retrieval R-Tree) which combines an R-Tree with superimposed text signatures. We present algorithms that construct and maintain an IR^2-Tree, and use it to answer top-k spatial keyword queries. Our algorithms are experimentally compared to current methods and are shown to have superior performance and excellent scalability.**

Index terms - Spatial inverted list , spatial queries, IR ^{2}tree , multidimensional data.

I. INTRODUCTION

Information retrieval is the activity of discovering information from a pool of information resources. Searches can be based on metadata or on full-text indexing. An information retrieval starts when a user enters a query into the system. User queries are matched against the database information. User queries is a broad term that actually denote to non-spatial queries[1][2] and spatial query .The former refer to the are names, phone numbers, email addresses of people whereas the latter refers to spatial elements such as points and regions. A spatial database handle multidimensional objects and provides quick access to those objects based on different selection criteria.An increasing number of applications require the efficient execution of nearest neighbor (NN) queries constrained by the properties of the spatial objects. Due to the popularity of keyword search, particularly on the Internet, many of these applications allow the user to provide a list of keywords that the spatial objects (henceforth referred to simply as objects) should contain, in their description or other attribute. For example, online yellow pages allow users to specify an address and a set of keywords, and return businesses whose description contains these keywords, ordered by their distance to the specified address location. As another example, real estate web sites allow users to search for properties with specific keywords in their description and rank them according to their distance from a specified location. We call such queries spatial keyword queries. A spatial keyword query consists of a query area and a set of keywords. The answer is a list of objects ranked according to a combination of their distance to the query area and the relevance of their text description to the query keywords. A simple yet popular variant, which is used in our running example, is the distance-first spatial keyword query, where objects are ranked by distance and keywords are applied as a conjunctive filter to eliminate objects that do not contain them.

II. RELATED WORKS

A spatial index, in contrast to a **B+ tree[1] ,** utilizes some kind of spatial relationship to organize data entries, with each key value seen as a point (or region, for region data) in a k-dimensional space, where k is the number of fields in the search key for the index. In a B+ tree index [1], the two-dimensional space of (age, salary) values is linearized. In contrast, a spatial index stores data entries based on their proximity in the underlying two-dimensional space. A **space-filling curve [1]** executes a linear ordering on the domain. The curve used represents the Z-ordering curve for domains with two-bit representations of attribute values. Consider the point with X = 01 and Y = 11. The point has Z-value 0111, obtained by interleaving the bits of the X and Y values; we take the first X bit (0), then the first Y bit (1), then the second X bit (1), and finally the second Y bit (1). In decimal representation, the Z-value 0111 is equal to 7, and the point X = 01 and Y = 11 has the Z-value 7.**Grid files [1][4]** rely upon a grid directory to identify the data page containing a desired point. When searching for a point, first find the corresponding entry in the grid directory. The grid directory entry identifies the page on which the desired point is stored, if the point is in the database.

The key idea of the **R-Tree [5] [7]** is to collect near objects and denote them with their minimum bounding rectangle in the next higher level of the tree; the "R" in R-tree stands for rectangle. Since all objects lie within this bounding rectangle, a query that does not cross the bounding rectangle also cannot cross any of the contained objects. At the leaf level, each rectangle defines a single object; at higher levels the combination of an increasing number of objects. This can also be seen as an increasingly coarse approximation of the data set. **An R+ tree [6]** is a method for looking up data using a location, often (x,y) coordinates, and often for locations on the surface of the earth. An R+ tree is a tree data structure, a modified form of the R tree, used for indexing spatial information. R+ trees are a concession between R-trees and kd-trees: they escape overlapping of internal nodes by introducing an object into multiple leaves if necessary. Coverage is the total area to cover all connected rectangles.

Overlap is the total area which is contained in two or more nodes . **R* trees [7]** is another modified form of R-trees used for indexing spatial information. R*-trees have a little higher implementation cost than standard R-trees, as the data may need to be reinserted; but the resulting tree will usually have an improved query performance.

NEAREST NEIGHBOR SEARCH TECHNIQUES

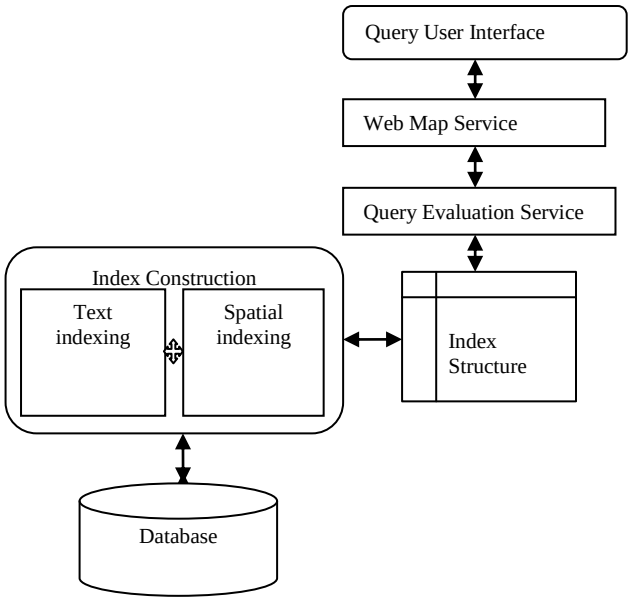

Fig1: Spatial Index Architecture

IR-Tree

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 [10]

IR^2-Tree

IR^2-Tree [10] is combination of two concepts: R-tree, a standard spatial index [8], and signature file [9], a better method for keyword-based document retrieval. By doing so they develop a structure called the IR^2-tree [10], which has the powers of both R-trees and signature files. Like R-trees, the IR^2- tree preserves objects spatial proximity, which is the key to solving spatial queries efficiently. As with many new solutions, the IR^2-tree also has a few disadvantages that affect its efficiency. The most important one of all is that the number of false hits can be really large when the object of the final result is far away from the query point, or the result is simply empty. 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 net 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.

Spatial Inverted Index and Minimum Bounding Method

So, new access technique spatial inverted access technique [11] 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 and R tree. 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.

TABLE 1
COMPARISON OF TECHNIQUES

Sl.no	Techniques	Advantages	Disadvantages
1	B+ Tree[1]	• It can be used for searching a point.	• Group objects only along one dimension,so it is not preserve spatial proximity
2	Space Filling Curves[1]	• It can be used for searching a point and region. • Group objects only along any dimension.	• Group objects only along two dimensions. • Order of the Z-curve affects performance.

3	Grid files [1][4]	• It is effective than space filling curves. • No special computations are required only the right records are retrieved	• Only for searching a point. • Imposes space overhead
4	R Tree[5][7]	• If the data storage created in the form of tree then space required is less also time needed for searching the keyword is less	• Since the larger the overlapping, the larger is the expected number of paths followed for a query.
5	R+ Tree[6]	• Avoids multiple paths during searching. • MBRs of nodes at same tree level do not overlap	• R+ tree can be larger than an R tree built on same data set. • Construction and maintenance of R+ trees is more complex
6	R* Tree[7]	• Area covered by a rectangle should be minimized.	• R*-trees have slightly higher construction cost.
7	IR^2-Tree[8][9][10]	• IR^2-tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined. It provide keyword search	• The IR^2-tree also has a few drawbacks that affect its efficiency, the most serious one of all is that the number of false hits can be really large
8	Inverted Index with R Tree [10] [11]	• Listing the words per article in the index. • The inverted index data structure is developed which lists the documents per word It is free from false hits	• The drawback is when keyword size is has only a single word, the performance of I-index is very bad. I-index consumes more space, that more time .
9	Spatial Inverted List[11]	• Not only that the SI-index is fairly space economical, but also it has the ability to perform keyword-augmented nearest neighbor search in time • Less space complexity than inverted index .	• It is focus on dimensionality two. • It uses R-tree index structure, it has large overlapping.

TABLE 2
USABILITY RESULT

Sl.no	Techniques	Space complexity
1	Space Filling Curves[1]	O(d*l log m) bits
2	Grid files [1][4]	O(m) bits
3	R Tree[5][7]	O(d*l log m) bits

4	R+ Tree[6]	O(d*l log m) bits
5	R* Tree[7]	O(d*l log n) bits
6	IR^2-tree[8][9][10]	O(d*l log n) bits
7	Inverted Index with R Tree [10] [11]	O(l(log p + d log m)) bits
8	Spatial Inverted List[11]	O(l(log(p/l)+log(m^d/l))) bits

Theoretical analysis: Theoretical analysis of various spatial indices, its advantages and disadvantages As the handling of each spatial indices is shown in Table 2, it suffices to focus on only one of them, denoted as L. Let us assume that the entire data set has p>=1 points and l of them appear in L. To make

analysis general, take the dimensionality d into account. Also, recall that each coordinate ranges from 0 to m, where m is a large integer. Naively, each pseudo-id can be represented with log p bits, and each coordinate with log m bits .

III. CONCLUSION

This paper presents the survey of varied techniques for nearest neighbor seek for spatial information. As in the previous ways there have been several drawbacks. The present solutions incur too expensive area consumption or they are unable to present real time answer. Therefore to beat the drawbacks of previous ways, new technique relies on variant of inverted index and R-tree and formula of minimum bounding technique is employed to scale back the search space. This technique can increase the potency of nearest neighbor search too.

ACKNOWLEDGMENT

The authors would like to thank HOD, Prof. R Rajaram, Department of Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady for his moral and technical support**.**

REFERENCES

[1] Raghu Ramakrishnan, Johannes Gehrke, "Database Management Systems", Chapter 26 ,pp. 777-795, McGraw Hill, Third Edition, 2004.

[2] X. Cao, L. Chen, G. Cong, C.S. Jensen, Q. Qu, A. Skovsgaard, D.Wu, and M.L. Yiu, "Spatial Keyword Querying," Proc. 31st Int'l Conf. Conceptual Modeling (ER), pp. 16-29, 2012.

[3] M Ester, HP Kriegel, J Sander "Spatial Data Mining," A Database Approach, Springer 1997.

[4] J. Nievergelt, H. Hinterberger, K. C. Sevcik: "The grid file: An adaptable, symmetric multikey file structure," ACM Trans. on Database Sys. 9, 1, 38-71 (1984).

[5] A Guttman "R-trees a dynamic mdex structure for spatial searching," Proc ACM SIGMOD Int Conf on Management of Data, 47-57, 1984

[6] Timos K. Sellis, Christos Faloutsos " The R+-Tree- A Dynamic Index for Multi-Dimensional Objects ," In ACM Trans,1987

[7] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, "The R*-tree: An Efficient and Robust Access Method for Points and Rectangles," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 322-331, 1990.

[8] C. Faloutsos and S. Christodoulakis, "Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation," ACM Trans. Information Systems, vol. 2, no. 4, pp. 267-288, 1984.

[9] I.D. Felipe, V. Hristidis, and N. Rishe, "Keyword Search on Spatial Databases," Proc. Int'l Conf. Data Eng. (ICDE), pp. 656-665, 2008.

[10] J. Zobel, A. Moffat, K. Ramamohanarao "Inverted Files Versus Signature Files for Text Indexing," In ACM Trans. Database Syst. 23(4): 453-490 (1998)

[11] Yufei Tao and Cheng Sheng ," Fast Nearest Neighbor Search with Keywords", IEEE, April 2014.

Authors Profile

[1]**Mr.Eldhose Paul**, PG Scholar in Adi Shankara Institute of Engineering and Technology, Kalady, did his B.Tech in IT from Viswajyothi College of Engineering and Technology, Muvattupuzha.

[2]**Mrs.Ierin Babu,**Assistant Professor of IT in Adi Shankara Institute of Engineering and Technology, Kalady, Completed M.Tech (CSE) from Anna University and B.Tech (CSE) degree from MG University . She has 6 years of Teaching experience.