ISSN No: - 2456 - 2165

Spatial Inverted Index for Spatial Queries

Dr. M Nagartna
Assistant Professor
Computer Science and Engineering
JNTUH College of Engineering
Hyderabad, India

E.Saraswathi
Computer Science and Engineering
JNTUH College of Engineering
Hyderabad, India

***Abstract*—Today's applications request to find spatial objects closer to user's location or within some range which satisfies user-supplied set of keywords. The result is a combination of a location and textual information based on the IR2 tree. But IR2-tree is not capable of providing real-time effective answers. In this paper an index method is proposed called spatial inverted index, to overcome the drawbacks of IR2-tree. Spatial inverted index based on conventional inverted index. It is compressed version of an I-index which consumes much less space and used to minimize the size of an inverted index.**

For example, in case of emergency instead of considering all hospitals, the nearest neighbor query would return hospital that is the closest among those whose equipment list contain "X-ray viewer, ICU, ECG Monitor" all at the same time.

Keywords— Spatial Inverted Index; Spatial Query; keyword search

I. INTRODUCTION

With the increasing popularity of GPS (geo-positioning system) and GIS (Geographic Information System), demand on exploiting various location-based information increases rapidly on the Web. Data objects associated with location information and textual descriptions are extensively available on the Web. People can locate these spatial objects such as restaurants, hotels, stores, hospitals through a Web query interface. The spatial keyword querying has received much attention from both the database and web research community. The role of spatial databases is continuously increasing in many modern applications. Spatial database systems (SDBS) are designed to handle spatial databases. Spatial data mining is the branch of data mining that deals with spatial data. It differs from regular data mining in parallel with the differences between spatial data and non-spatial data. It has wide applications in Geographic Information Systems (GIS), image databases exploration, medical imaging, and other areas where spatial data are used.

A GIS (geographic information system) is a used to capture, store, manipulate, analyze, manage, and present spatial or geographic data. Vector and raster data are the two primary data types used in GIS. There has been lot of commercial interest in Location based information for restaurants, hospitals, theaters etc. As a result, Google Earth and Yahoo Maps, as well as other geographic applications, queries in spatial databases have become increasingly popular in recent years. Vector and Raster data are the two primary data types used in GIS. Raster data consists of bit maps or pixel maps, in two or more dimensions. 2D-Raster image of a satellite image is an example for raster data. It may also include the location of the image, specified by the latitude and longitude of its corners, and the resolution, specified by the total number of pixels. Raster data is often represented as tiles, each covering a fixed sized area. A larger area can be displayed by displaying all the tiles that overlap with the area. Raster data can also be 3D, third dimension could be temperature or time. Temperature is measured with the help of a satellite. In Vector model, space is not quantized into discrete grid cells like Raster model. Vector data models use XY coordinates to represent the vertices. Lines and polygons are other types of vector data.

There has been lot of interest in Location based information for restaurants, hospitals, theaters etc. As a result, Google Earth and Yahoo Maps, as well as other geographic applications, have become increasingly important in recent years. Consider the following example Query:

Find hospitals with emergency medical facilities such as ICU and X-ray viewer near to user specified location. The information that above queries seek is significant during emergency operations. Such information is contained in GIS databases explicitly created for such purposes. Location-based information present in GIS databases fundamentally comprise of two components:

1) spatial or location information
2) Textual information.

Spatial information refers to the geographic location, size and shape of an entity. The shape could be a point, or extended in space such as a line or a polygon. The result to query consists of a set of data objects that satisfy certain location constraints and keyword matching condition. Basically, there are more than one candidates and the query processing algorithms will return top-k items according to a certain ranking scheme. In general, one can define different constraints and conditions on the query location and query keywords to represent various query semantics. Thus, spatial keyword query has a high flexibility and is attracting more and more study from both the database community and the Web community.

Let P be a set of multidimensional points. The points in P have integer coordinates, such that each coordinate ranges in [0, t], where t is a large integer. A place $p \in P$ has two attributes: $<\lambda, \psi_q>$, where λ is geo-location and ψ_q is a text value (set of keywords)

It returns a point in Pq that is the nearest to query point q, where Pq is defined as

$$Pq = \{p \epsilon P \mid \Psi q \subseteq \Psi p\}$$

In other words, Pq is the set of objects in P whose documents contain all the keywords in ψ_q.

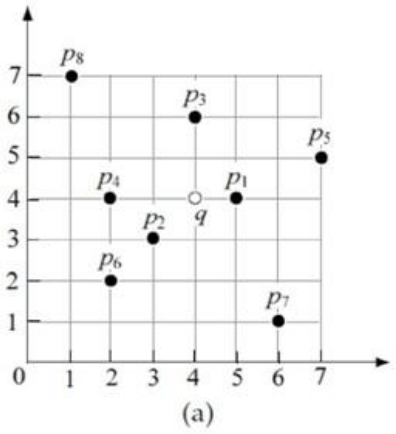

Fig. 1. Locations of points

p	W_p
p_1	$\{a, b\}$
p_2	$\{b, d\}$
p_3	$\{d\}$
p_4	$\{a, e\}$
p_5	$\{c, e\}$
p_6	$\{c, d, e\}$
p_7	$\{b, e\}$
p_8	$\{c, d\}$

Fig. 2. Points and their associated keywords

Let q is the query point from which all nearest neighbors are to be found which consists of "a, e" as keywords, i.e. Ψq= {a, e} nearest neighbor will find P_4 as answer. If the keywords are c, d then it returns P_6. If the value of k is 2 then it returns both P_6 and P_8.

II. RELATED WORK

A. Spatial Indexing Scheme

- R Tree

This approach first finds Top-1 nearest neighbor from current location within all available locations. R-Tree [5][10] approach contains only pointers of location not detailed information like keyword data. After getting Top-1 nearest neighbor its corresponding text description is loaded to check with query keywords if this fails then that object get discarded. Then Top-2 nearest neighbor is found this process is continued till locations are found which satisfies query keywords. The final result will be Top-k spatial keyword objects. Splitting algorithms such as the linear split, the quadratic split are proposed. Guttman suggested using the quadratic algorithm as a good compromise to achieve reasonable retrieval performance. In earlier work [3] R-tree based indices loosely combine the R-tree and inverted files to organize the spatial and text data separately. The R*-tree [10]of Beckmann , has better performance than Guttman R-tree "quadratic split". The main idea in the R*-tree is the concept of forced re-insert, resulting in a R-tree with better structure. R+ Tree is another variant R tree. The difference between R tree and R+ tree is that MBR's does not intersect with one another which saves the disk accesses when compared to R-Tree. The drawback of this approach is it leads to retrieval of objects which does not satisfy query keywords. In worst case complete tree needs to be traversed and each object needs to be inspected.

- Kd-tree:

Each level of a k-d tree [13] partitions the space into two. The partitioning is done along one dimension at the node at the top level of the tree, along another dimension in nodes at the next level, and so on, cycling through the dimensions. By partitioning at each node, approximately one-half of the points stored in the sub tree fall on one side and one-half fall on the other side. Partitioning stops when a node has less than a given maximum number of points.

- Space Filling curve(Z-curve):

Space-filling curves map points of a multidimensional space to one-dimensional values [12]. A point p, is a triplet (id_p, x_p,y_p), consisting of identifier of location and x and y coordinates of p. Gap-keeping method can be applied on only one attribute of the triplet requires a sorted order, therefore it requires sorted order of attributes. SFC converts a multidimensional point to a 1D value such that if two points are close in the original space, their 1D values also tend to be similar. For example, consider a set of integers {2, 3, 6, 8}, the gap-keeping method stores values such as {2, 1, 3, 2} illustrate in [1]. To calculate gap between ids and coordinates of points, first it takes binary form of id and coordinates of points. Then merge binary numbers to get converted values of points based on Z-curve. Let (x, y) be the point, then first apply f(x) and f(y) on x and y values, where f is a function which converts decimal to binary. for example , p6(2,2) f(2)=010, f(2)=010.Then let consider output of f(x) or f(y) as set of binary bits{x_1 x_2……x_n} where x_i is bit having value output.

Next we merge f(x) and f(y) using M (f(x), f(y)), Where, M is the merger function will do bit by bit merging of f(x) and f(y) output. If f(x)={ x_1,x_2……x_n }, f(y)={ y_1,y_2 …… y_n }

M (f(x), f(y)) = {$x_1y_1x_2y_2$ …… x_ny_n }

Let d(M) be the function which converts bits to decimal. For example,

f (2)=010

f (2)=010

M (f(2),f(2))=001100 Then d(M)=12.

P_6	P_2	P_8	P_4	P_7	P_1	P_3	P_5
12	15	23	24	41	50	52	59

B. Text Indexing Schemes

- Inverted Index

This is an effective method for the search based on Keyword matching. In this approach, there is inverted list maintained for each word that is entered. This basically consists of Ids of all the location that consists of the particular keyword. Inverted list maintains all Point Ids corresponding to the keyword which allows an efficient merging step. In this method,

1. For each keyword, spatial objects containing the keyword are identified.
2. Intersect them
3. For each object compute distance to query point
4. Sort and return to user

For Example, finding points that have keywords b and d both at the same time then it is possible to take intersection of two inverted list i.e. P_2.

TABLEI. Inverted List

Word	Inverted List
A	P_1 P_4
B	P_1 P_2 P_7
C	P_5 P_6 P_8
D	P_2 P_3 P_6 P_8
E	P4 P5 P6 P7

- Signature file:

Signature files were introduced by Faloutsos and Christodoulakis [11] as a technique to efficiently search a group of text documents. Signature file is hashing based framework which is also known as Superimposed Coding (SC). It basically performs the membership test. It checks whether a given word w belongs to set of words ψ. Superimposed Coding [11] is the same concept as Bloom Filter. It checks the positions of 1's of the query word to the ψ keywords against the database and check all 1's are in the same position for both query and existing keywords if yes then it directly returns "yes" otherwise "no". "w" to a string of l bits, and so taking the disjunction of all bit strings. For instance, allow us to denote by h(w) the bit string of a word w.

1) L bits of h(w) are initialized to zero

2)Sc repeats m times : i.e. randomly choose a bit and set it to 1. Randomization must use w because the same w should always end up with an identical h(w).

For example consider l = 5 and m = 2.

TABLEII. Signature file

WORD	HASHED BIT STRING
a	00101
b	01001
c	00011
d	00110

The bit signature of a set ψ of words simply ORs the bit strings of all the members of ψ. For example, the signature of a set of keywords consisting of { a, c, d} equals 00111, and that of {b, c, d} equals 01111. For a keyword w, SC performs the membership test in the set of keywords ψ, by checking whether signature of ψ and all the 1's of h(w)are at the same positions or not. If the signature does not matches then it is guaranteed that w cannot belong to ψ. Otherwise, scan of keywords should be performed, whether w is actually present or not. A *false hit* occurs if the scan reveals that the keywords does not contain w.

III. PROPOSED WORK

The best method to date for spatial queries with keywords is IR^2 –Tree, based on R-tree. A signature [11]is added to each node of the IR2 - Tree .Like R-trees, the IR2- tree preserves objects spatial proximity. The IR2-tree also has the drawback of signature file known as false hits (false drop or false alarm). That is, the search on objects is done, even though all the keywords specified by user are not present. Therefore verification of object is needed and requires loading of its full-text description, which is expensive.

In this paper, new access technique named as *spatial inverted index* (SI-index) [1] access technique is used to get rid of the drawbacks of previous strategies such as false hits. We are able to combine various lists much like merging natural inverted lists with the help of ids. The compressed spatial inverted index has a triplet to represent both ID and x, y coordinates of each point. In order to apply gap-keeping effectively, converting 2D coordinates into only one attribute. A space filling curve (Z-curve) is constructed for efficient processing, this is called 2D Z-curve generation. For better sorting procedure, Pseudo ids are used instead of real id and form 3D Z-curve. P_2, P_3, P_4, P_5 Whose Z-values are15, 52, 24, 59 respectively, with pseudo-ids being 1, 6, 3, 7respectively. With gap-keeping, the Z-values and pseudo-ids are recorded as 15, 9, 28, 7 and 1, 2, 3, 1 respectively. So we can precisely capture the four points with four pairs: {(1, 15), (2, 9), (3, 28), (1, 7)}.

Algorithm:

Output: Nearest Neighbor to user specified location.

1) Read three values as (id_p, x_p, y_p).

Where, id_p - id of place or word.

X_p - Position of the x-axis of place.

Y_p - Position of the y-axis of place

//2D gap-keeping

2) Apply gap-keeping on x and y first as following,

a) Read x

b) Let 2D = {b1b2b3.....bn}

c) Convert x to binary values as, b1= binary(x);

Repeat step a and b for y and create b2.

d) Merge b1and b2 bit by bit and store in b3.

e) Convert b3 to decimal.

f) Generate Z-curve using gap-keeping.

3) Repeat step 1 and 2 for all places.

4) Generate sorted 2D Z-curve using gap-keeping.

//3D gap keeping

5) Repeat 1 and 2 for id as x and value from set 2D as y

will store merge results

Let, 3D = {c1, c2.....cn}

6) Generate 3D Z-curve using gap-keeping on set 3D.

7) Apply 2-level gap-keeping on set 3D.

IV. DATASET INFORMATION

A geographic coordinate system (GCS) uses a three-dimensional spherical surface to define locations on the earth. The location of spatial objects is based on its longitude and latitude values. Longitude and latitude are angles measured from the earth's center to a point on the earth's surface. The angles often are measured in degrees (or in grads). To find the nearest hospital system needs to track the current user location. It can be tracked by using GPS location tracker

Sample dataset for finding nearest hospital, with emergency facilities near to user specified location.

V. CONCLUSION

In this paper we mainly focus on spatial data mining technique. Spatial inverted Index structure is used to deal with the problem of IR2-tree. Compression of SI- index has done using Gap-keeping method. This method can't be applied on triplet. So firstly consider 2D Z-curve values and then 3D Z-curve values. And to calculate these Z-curve values there are two steps that is binary representation and merging.

References

[1] Yufei Tao and Cheng Sheng, "Fast Nearest Neighbor Search with Keywords," IEEE Transactions on Knowledge and Data Engineering, 2013, p1-13.

[2] 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

[3] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, "Hybrid index structures for location-based web search," In Proc. of Conference on Information and Knowledge Management (CIKM), pages 155–162, 2005

[4] 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.

[5] I. D. Felipe, V. Hristidis, and N. Rishe, "Keyword search on spatial databases," In Proc. of International Conference on Data Engineering (ICDE), pages 656– 665, 2008.

[6] X. Cao, G. Cong, and C. S. Jensen, "Retrieving top-k prestige-based relevant spatial web objects," VLDB, 3(1):373–384, 2010.

[7] Yen.-Yu Chen, Torsten Suel, and Alexander Markowetz, "Efficient query processing in geographic web search engines," In Proc. of ACM Management of Data (SIGMOD) , pages 277–288, 2006.

[8] 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.

[9] A. Guttman. R-trees: A dynamic index structure for spatial searching. In *SIGMOD*, pages 47–57, 1984.

[10] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The r*-tree: an efficient and robust access method for points and rectangles. ACM SIGMOD, pages 322- 331, May 1990.

[11] Christos Faloutsos, Stavros Christodoulakis: Design of a Signature File Method that Accounts for Non-Uniform Occurrence and Query Frequencies. In VLDB 1985: 165-170

[12] C. Ohm, G. Klump, and H. Kriegel. Xz-ordering: A Space- Filling Curve for Objects with Spatial Extension. In *Proc. Ofthe 6th Intl. Symp. on Advances in Spatial Databases*, pages75-90, July 1999.

[13] Bentley, J.L.Multidimensional binary search trees used for associative searching. Comm. ACM 18,9 (sept. 1975), 509-517.