Geo-Spatial Index Structure for Myanmar Keyword Query

Myat Thiri Khine and Myint Myint Sein
University of Computer Studies, Yangon, Myanmar
myatthirikhine@ucsy.edu.mm, myintucsy@gmail.com

Abstract

Geographic databases are used for both storage of Spatial and Non-Spatial attributes. It can be used for indexing data structures to provide fast response to spatial queries. Spatial query can select geographical features based on location or spatial relationship, and a Nearest Neighbor search can be used to find the nearest object of a query object. The search process in the spatial database takes much time as the database size increases. There has been developed many index structure in recent years to quickly retrieve the geo-information. In this paper, a new index structure is proposed to retrieve the desired nearest information with Myanmar language on the mobile devices. There has been proposed the index structure that combines K-d tree and inverted file which is considered on the Myanmar keywords queries. In K-d tree, the data points are scattered all over the tree. In this proposed paper, a new index structure is constructed by using Hilbert space filling curve and B-tree and also combines the inverted file to reduce the researching time. Myanmar 3 Unicode is used for keyword search.

Keywords: *Location-Based Service, Spatial Query, Hilbert Curve, B-tree, Index Structure, Inverted File*

1. Introduction

Although the effective Location-based services have been developed in the developed countries, it still lacks to develop the effective and efficient one in the developing countries as Myanmar. Location-based services application for Myanmar has been developed but most of the applications mainly depend on the web services. So, it still needs to develop it on the mobile devices to quickly search the desired location anywhere and anytime.

Geospatial database uses Geographic Information System (GIS) to locate and access data quickly and efficiently [19] [20]. Geospatial data is defined as geographically referenced data that describes both the locations and characteristics of spatial features such as roads, land parcels, and vegetation stands on the Earth's surface, it is also called geospatial data and it is stored in spatial databases.

Spatial data are data that have a location (spatial) and mainly required for Geographic Information Systems (GIS) whose information is related to geographic locations. Geographical information system stores spatial data and retrieves the geo-information from existing spatial data.

Many index structures have been proposed in recent years to quickly retrieve the geo-

information. R-tree is mainly used and that combine with inverted file, namely the families of IR-tree [12, 11, 6, 7, 9, 8, 10]. R-tree is used for spatial (latitude/longitude) index and inverted file is used for textual index. As the data objects in the R-tree can be overlapping and covering each other, the search process in the R-tree might suffer from unnecessary node visits and higher IO cost [13]. Hybrid index structure that combines the K-d tree and the inverted file for spatial keyword search with the minimum IO costs and CPU costs has been proposed [1][2][3]. Moreover, an index structure that combines K-d tree and inverted file to process spatial keyword queries with Myanmar language is also developed [17]. In K-d tree, the data points are scattered all over the tree.

This paper presents a new index structure that is constructed by using Hilbert space filling curve and B-tree and also combines the inverted file to effectively and efficiently retrieve the desired nearest location with Myanmar language. It can also reduce the searching time.

The remainder of the paper is organized as follows. Section 2 reviews the related works. Hilbert Curve is explained in section 3. Section 4 describes B-tree. In section 5, the proposed index structure is described. The experimental results is showed in section 6. Section 7 concludes the paper with directions for future work.

2. Related Works

There are many index structure has been developed. Mostly, R-tree [10, 11, 6, 9, 8] is used and its variants as spatial index and inverted file for text index. They all combine both indices depending on the combination schemes [5]. T.Wang, G. Li, J. Feng [14] proposed a new index structure, spatial keyword R-tree, called SKR-Tree which extended from the R-tree with an R-tree node storing both spatial and keyword information. Hariharan et al. R. G¨obel, A. Henrich, R. Niemann, and D. Blank [9] presented the KR*-tree. This paper proposed a framework for GIR systems and focus on indexing strategies. I. D. Felipe, V. Hristidis, and N. Rishe [8] uses R*-tree for spatial index and inverted file for text index. The IR tree [6] creates each nodes of the R-tree with a summary of the text content of the objects in the corresponding subtree. X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu [4] proposed S2I index structure based on R-tree and inverted file.

The objects in [4] are stored differently according to the document frequency and infrequency of the term. Cary et al, [11] proposed SKI that combines and R-tree with an inverted index by the inclusion of spatial references in posting lists. The posting list of term contains all its term bitmaps rather than documents.

X. Chen, C. Zhang, B. Ge, W. Xiao porpose two index structures, TUR-tree and TUA-tree for accelerating query process. Query processing algorithms are designed for the three queries in social network, aiming to explore temporal dimension in users, relationships and social activities [18].

X.Cao, G.Cong, Christian S. Jensen, Jun.J. Ng, BengC.Ooi, N.T. Phan, D. Wu [5] proposes a Web Object Retrieval System (SWORS) that is capable of efficiently retrieving spatial web objects that satisfy spatial keyword queries. This

system use IR tree and inverted file for index. It supports two types of queries that are location aware top-k text retrieval (Lkt) query and spatial keyword group (SKG) query.

3. Hilbert Curve

The Hilbert Curve is space-filling curve which visits every point within a two dimensional space. The basic curve has the shape of an upside down "U". A square is initially divided into 4 quadrants and a first-order curve is drawn through their centre points. The quadrants are ordered such that any two which are adjacent in the ordering share a common edge. The top vertices are replaced by the previous order, and the bottom vertices suffer a rotation. The bottom left vertex is rotated 90 degrees clockwise, and the bottom right rotates 90 degrees counterclockwise. Figure 1 shows the Hilbert curve orders one, two and three. In this figure, the curve starts on the lower left corner and ends on the lower right corner, but this can be changed as long as the curve keeps the ”U” shape.[15]

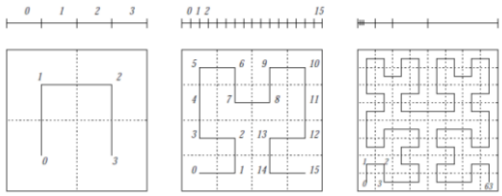

Figure 1. Hilbert Curve Orders 1, 2, and 3 respectively.

4. B-Tree

B-tree is balanced search tree and is similar to red-black trees but they are better at minimizing disk I/O operations. Many database system use B-trees, or variants of B-trees, to store information. B-trees keep values in every node in the tree, and may use the same structure for all nodes. Leaf nodes never have children. Unlike a binary-tree, each node of a b-tree may have a variable number of keys and children. Each key has an associated child that is the root of a sub tree containing all nodes with keys less than or equal to the key but greater than the preceding key. A node also has an additional rightmost child that is the root for a subtree containing all keys greater than any keys in the node.

A B-tree of order m (the maximum number of children for each node) is a tree which satisfies the following properties:

1. Every node has at most m children.
2. Every node (except root and leaves) has at least m⁄2 children.
3. The root has at least two children if it is not a leaf node.
4. All leaves appear in the same level, and carry information.
5. A non-leaf node with k children contains k–1 keys. [16]

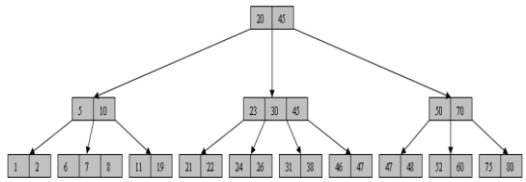

Figure 2. B-Tree of Order 2

5. Proposed Index Structure

The proposed index structure combines B-tree with inverted files. Before creating the B-

tree, two dimensional coordinate points are converted to the single value by using the Hilbert curve. Then, B-tree that combines the inverted file is constructed according to the value from the Hilbert curve (h-values) and services.

The algorithm to compute the h-values of the two-dimensional Hilbert curve on a 2^n*2^n grid is:

Step 1: Read in the (n-bit) binary representation of the x and y coordinates.

Step 2: Interleave bits of the two binary numbers into one string, i.e., the same way as for the Peano curve.

Step 3: Divide the string from left to right into 2-bit strings, si for i=1,...,N.

Step 4: Give a decimal value, di, for each two bit string according to the following chart

a. '00' equals 0
b. '01' equals 1
c. '10' equals 3
d. '11' equals 2

and put into an array in the same order as the strings occurred. (This gives the h-values of the basic Hilbert curve.)

Step 5: For each number i in the array, if i=0 then switch every following occurrence of 1 in the array to 3 and every following occurrence of 3 in the array to 1; i=3 then switch every following occurrence of 0 in the array to 2 and every following occurrence of 2 in the array to 0;

(This makes up for the rotation and reflection of the curves of order higher than 1.)

Step 6: Convert each number in the array to its binary representation (two-bit strings), concatenate all the strings in order from left to right, and calculate the decimal value [15].

The proposed index structure is shown in Figure 3. The Example Dataset is shown in Table 1.

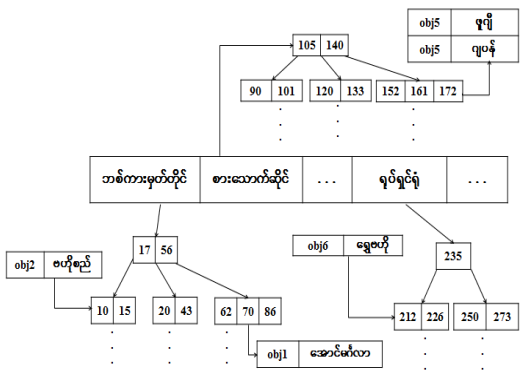

Figure 3. Proposed Index Structure

Table 1. Example Dataset

Id	Latitude	Longitude	Keywords	Services
Obj1	16.796433	96.176803	အောင်မင်္ဂလာ	ဘစ်ကား မှတ်တိုင်
Obj2	16.779908	96.140056	ဗဟိုစည်	ဘစ်ကား မှတ်တိုင်
Obj3	16.800442	96.162225	ဗိုလ်ချုပ်	ပန်းခြံ
Obj4	16.829281	96.155644	ဆီဒိုးနား	ဟိုတယ်
Obj5	16.816497	96.127464	ဖူဂျီ, ဂျပန်	စားသောက် ဆိုင်
Obj6	16.810881	96.176419	ရွှေဗဟို	ရုပ်ရှင်ရုံ

6. Experimental Results

In this paper, the desired nearest location is searched based on the current location and service. Current location is acquired by GPS and the desired service is chosen by the user. It takes these two inputs and search in the proposed index structure. This system is considered on the mobile devices and is tested on the Yangon Region which has 46 townships. In this system, it is mainly focused on the 20 townships. It provides the user with 72 services and there are 3000 data in the database. Figure 4 shows Input required query for searching. In figure 4, user needs to choose the desired service.

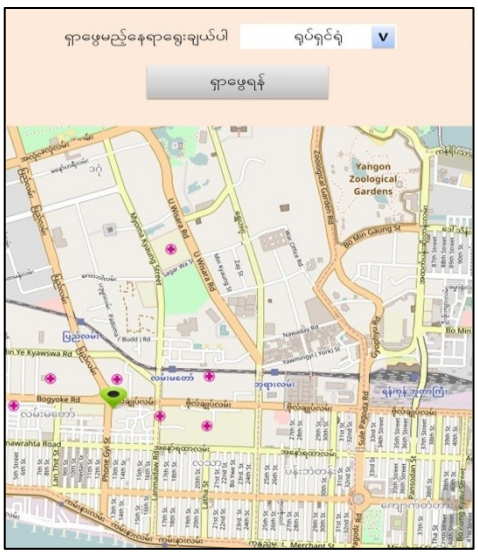

Figure 4. Input Required Query for Searching

In Figure 5, it shows the nearest place which is searched by user based on the current location. Figure 6 compares the searching time (second) between using proposed index structure and R-tree and combined K-d tree with inverted file. Searching time using proposed index structure is faster than R-tree and K-d tree with inverted file.

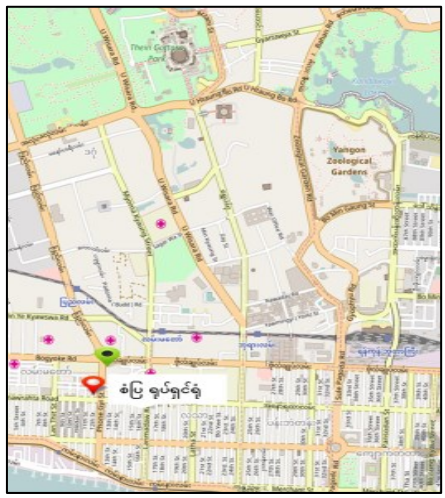

Figure 5. Result After Searching

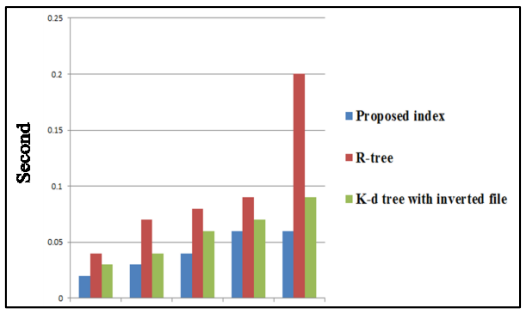

Figure 6. Searching Time Compare with Proposed Index, R-Tree and K-d tree with Inverted File

7. Conclusions

This paper presents a new index structure that is constructed by using Hilbert space filling curve and B-tree and also combines the inverted file to retrieve the desired nearest location with Myanmar language. It is tested on Yangon region.

This application is considered on the mobile devices. As a further extension, we will consider the system that will search the desired location with both English and Myanmar Language on the mobile devices and will work in an offline.

References

[1] S. N. Aung, M. M. Sein, "Hybrid Geo-Textual Index Structure for Spatial Range Keyword Search", *Computer Science & Engineering: An International Journal (CSEIJ)*, Vol. 4, No.5/6, December 2014.

[2] S. N. Aung and M. M. Sein, "K-Nearest Neighbours Approximate Keyword Search for Spatial Database", in *Proceedings of 9thInternational Conference on Technological Advances in Electrical, Electronics and Computer Engineering (ICTAEECE)*, Bangkok, Thailand, 7th February 2015, pp. 65-68.

[3] S. N. Aung and M. M. Sein, "Index Structure for Nearest Neighbors Search with Required Keywords on Spatial Database", the *9th International Conference on Genetic and Evolutionary Computing (ICGEC 2015)*, Yangon, Myanmar, 26-28 August 2015.

[4] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu, "Spatial keyword querying", in*ER*, 2012, pages 16–29.

[5] X. Cao, G. Cong, Christian S. Jensen, Jun.J. Ng, BengC.Ooi, N.T. Phan, D. Wu, "SWROS: A System for the Efficient Retrieval of Relevant Spatial Web Objects".
Available:
http://www.ntu.edu.sg/home/gaocong/papers/*vldb12swors.pdf*

[6] A. Cary, O. Wolfson, and N. Rishe, "Efficient and scalable method for processing top-k spatial Boolean queries", in*SSDBM*, 2010, pages 87–95.

[7] G. Cong, C. S. Jensen, and D. Wu, "Efficient retrieval of the top-k most relevant spatial web objects", *PVLDB*, 2009, 2(1):337–348.

[8] I. D. Felipe, V. Hristidis, and N. Rishe, "Keyword search on spatial databases", in *ICD*E, 2008, pages 656–665.

[9] R. G¨obel, A. Henrich, R. Niemann, and D. Blank, (2009), "A hybrid index structure for geo-textual searches", in *CIKM*, 2009, pages 1625–1628.

[10] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, (2007), "Processing spatial-keyword (sk) queries in geographic information retrieval (gir) systems", in *SSDBM*, 2007, page 16.

[11] Z. Li, K. C. K. Lee, B. Zheng, W.-C. Lee, D. L. Lee, and X. Wang, "Ir-tree: An efficient index for geographic document search", *IEEE TKDE*, 2011, 23(4):585–599.

[12] J. B. Rocha-Junior, O. Gkorgkas, S. Jonassen, and K. Nørv°ag, "Efficient processing of top-k spatial keyword queries", in *SSTD*, 2011, pages 205–222.

[13] Y. Theodoridis, T. Sellis, "Optimization Issues in R-tree Construction", Technical Report KDBSLAB-TR-93-08.

[14] T. Wang, G. Li, J. Feng, "Efficient Algorithms for Top-k Keyword Queries on Spatial Databases", *12th IEEE International Conference on Mobile Data Managemen*t, 2011.

[15] Christos Faloutsos, Shari Roseman, "Fractals for Secondary Key Retrieval", In *Proceedings of the Eighth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems ,PODS '89*, 247-252.

[16] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, "Introduction To Algroithm", Third Edition, pg 484-499

[17] M. T. Khine, S. N. Aung and M. M. Sein, "Geo-textual Index Structure for Spatial Keyword Query with Myanmar Language", in Proceedings of the 14th International Conference

on Computer Applications (ICCA2016), Yangon, Myanmar, pp. 47-51, February 2016.

[18] X. Chen, C. Zhang, B. Ge, W. Xiao, " Temporal Social Network: Storage, Indexing and Query Processing", in the Workshop Proceedings of the EDBT/ICDT 2016 Joint Conference, Bordeaux, France, March 15, 2016.

[19] Z. Wei , W. Wanzhen, Y. Xingguang, X. Gang, "An optimized query index method based on R-tree," in *Proc Fourth International Joint Conference on Computational Sciences and Optimization, pages:1007-1010,* 2011.

[20] Y. Lifang; L. Rui; H. Xianglin; L. Yueping, "Performance of Rtree with slim-down & Reinsertion Algorithm," in *Proc International Conference on Signal Acquisition and Processing*, pages:291-294, 2010.