ISSN:0975-9646

Categorize and Efficient: Top k keyword Search of Spatial-Textual data on the Road Network

[1] Santosh.I.Halpati, [2] D.M.Thakore

[1]M.Tech student, Department of computer BVUCOEP
[2]Professor, Department of computer BVUCOEP

***Abstract*: Spatio-Textual information nowadays becomes so popular and mandatory for each searching application. Many organization have shown their concern regarding importance of using spatial information ("latitude" & "longitude") along with Textual data (i.e. Keyword Query "Hospital"). In parallel study, of searching trend and getting Top k result out of Spatio-Textual search data by using two spatial-Textual indexing technique IR-Tree and IQuad-Tree. As previously prove that the IQuad-Tree having better performing result in terms of storage and retrieval of spatial query point and the content related to it spatial point. Consequently, Spatio-Textual retrieve data for top k result on road network have different result for the nearest neighbor search as compare to simple Spatio-Textual searching. While the information which is retrieve should be efficiently relevance to the query keyword and spatial constraint. In this paper, we discuss the technique to calculate distance for the nearest neighbor search on road network and efficient algorithm that can dynamically retrieve the relevant data for every keyword search query. Extracting actual data out of user unusual query keyword search many time prove to be bottle neck, so it is better trained data in that environment. In previous approaches, for extracting Top k result out of data on road network is not having efficiently relevance and faster result. In this paper, we are going to propose the method that will give boom to our Spatio-Textual relevance search on road network with quality of top k result as compare to previous methods.**

Keywords: Spatial-Textual, ILQuad-Tree, Information retrieve, machine learning.

I. INTRODUCTION

New searching trends have brought the revolutionary change in the field of spatial-Textual information search. As many organization want their information to be known spatially, this geo-location and geo-position technology have modernize the search capabilities. People nowadays find for the spatial information (latitude & longitude) and textual keyword information (i.e. "hospital", "hotel" etc.) more often in searching application. Now the social networking site also having GPS tracking concept, that whenever or whichever place somebody post photo or text content it is shows the spatial information attached to it. The smartphone intelligent technology are optimize to provide the geo-tag information for the photograph which are been capture by the phone camera.

Motivation: People are more addicted in spatial objects whose description is given through a set of query keywords. Now by giving a spatial-text query keywords (i.e. "orthopedic hospital"), it search for nearest neighbor keyword query returns the closet spatial information relevant for the keywords data. While, the point of interest is to know that the information for the spatial-textual data use the most popular indices structure which is embed with the inverted index and R-tree, named the IR-tree and the IR^2-tree. Methodology behind IR^2-tree index is that it maintained to get global spatial information and contain the keyword information with each spatial point, every node in the IR^2-tree is link with an inverted file index, which will can be refers to as a pseudo page document that represents all the data information whom spatial information embed into the node's Minimum Bounding Rectangle. Moreover, there is another type of technique indices used for storing the spatial data Polygon or point along with text data which shall discuss through this paper.

In this paper, the previous propose methodology called ILQuad-Tree the full form is Inverted linear Quad-Tree which is hierarchal placing the spatial point is also having good contribution towards the spatial-Textual searching experiment. But still finding the accurate keyword and less cost for query processing issues are not been sort out efficiently. So it should have technique that classified its information dynamically and give the faster result by lessen processing.

II. PRELIMINARY ABOUT SPATIAL TEXTUAL ON ROAD NETWORK

Now will stick to the concept of IR-Tree, which is having some revolution change in the idea of searching space. In analyzing the Rocha-Junior et al [1] work which have given the challenging problem for executing the top-k result for spatial keyword queries on following road networks providing a set of spatial point and its related information. The interesting things is that it is having a top-k spatial-Textual keyword query on road networks which perhaps returns the k best result in terms of finding nearest neighbor for the query location, and provide the accurate space information.

Problem Defining: Judging the concept of efficient top k spatial keyword search on road network and simple search of spatial-Textual data, both are having the different result for the nearest neighbor search. Meanwhile studying the challenging problem it can be visualize if the spatial point of interest are within your given range and sometime the point seem to be textual correct but it doesn't fulfil the spatial constraint which seem on the road. Even though the object in the map is near to the query point and fulfil the text information but the road at which it situated doesn't have meaningful way to reach it then it can be say that it is not one of the relevance result for the top k result. As figure 1 which is represent the spatial-textual concept on road

network at this time forget about the ILQuad-Tree on road network structure, visualize it is a simple street view and it is having information about number of hospital within the range of some kilometer. According to the fig 1 the query point ql (query location) of user is somewhere near S2 (spatial point) and it qk (query keyword) is "orthopedic hospital", so there are two keyword t1 and t2 which should be match correctly with data specified. Now it's clearly seen that from the query not it seem S2, S1 and S4 are nearest point but they are not text relevance by the qk and S4 is near but still not connected to the road that it can be follow. As result, S6 is within the range and also fulfilling the spatial constraint and the textual containing both the keyword so S6 will be the one of the top k result out of various information.

Problem Statement: Many time it seem that the cost of query processing for some keyword like hospital is increase because of it occurrence more in the document, it will analysis each and every node containing the keyword and put it into some signature file.

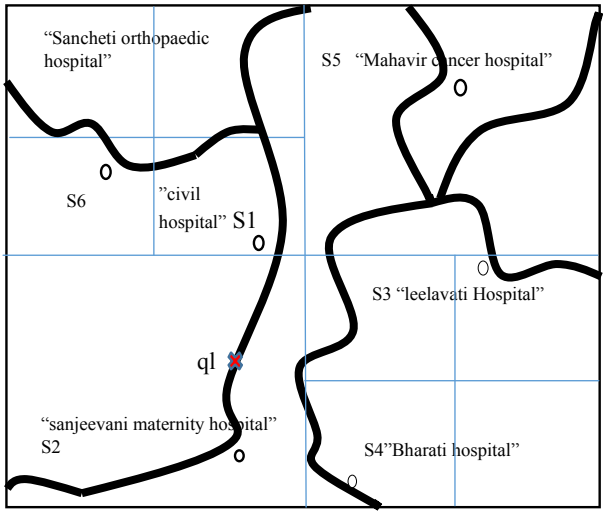

Figure 1 ILQuad-Tree on the road network

Signature file usually store the position of word where it occur and how many time it occur basically filter the information. The problem is that it require many filter and it should be work efficient every time still processing is more time consuming. This mechanism used by the previous propose system so that the faster rate of information should be retrieve with minimum processing by travels through various node's IR^2-Tree. Another is using And Semantics that will degrades the performance of IR-Tree when the number of keywords in the query increase because it doesn't exploit it semantics technique and extract all query keywords. So the problem which are been analysis, having great impact on the performance and the efficiency of retrieve accurate result on dynamically accessing the data.

In our proposed concept will going to give an efficient technique that used ILQuad-Tree hybrid index structure, which is having optimal capability for storing and retrieve the Spatial-Textual information. Moreover this hybrid index structure data are process through the machine learning methodology which will trained the data and whenever there is search for the query keyword it will classified data minutely and provide the high quality result for top k Spatial-Textual on the street.

III. RELATED WORK

In this section, we first present existing techniques, inverted R-tree and IR^2-tree information R-tree ILQuad-Tree, for the problem of top k result query as well as some other top *k* queries which dealing with both spatial-textual information on the road network.

Inverted R-tree: The usual concept that remove the stop word out of document and sort the meaningful keyword and put it into one linked list file that each keyword link with some document. Now for each keyword *t*, assign with the spatial point which the MBR has structure it in form of node graph tree, given the fact that the number of keyword are usually small in practice, the efficiently technique support keyword search since only a few lists need to be loaded and the target objects can be identified by merging the lists.

IR^2-tree: Initially, the inverted R-tree concept was good enough when there is only one query keyword but we only need to issue a top k nearest neighbour search result. Unfortunately, the performance of the inverted R-tree is low when the number of keyword are increasing for query. Even though the search region of the top k query will traverse through the most of the node from the location point to find the exact match and enlarge against the number of keywords. The author use Felipe *et al.* proposed system which the information retrieval R-tree (IR^2-tree) structure [2] which can solve the above problem by using *signature* technique.

IR-tree: In [3], [4], the full form is information *R*-tree (IR-tree) structure is proposed to bring support environment for the spatial keyword ranking query, which is similar to the IR^2-tree method. As compare to IR^2-tree instead of using the signature file at each keyword it is using inverted file at each and every node. Further optimizations are applied to improve the information retrieve performance by taking advantage of the tree Structure.

IL-Quad-Tree Structure: A Quad-Tree is a hierarchal space partitioning tree data structure in which a d-dimensional space subdivided into 2d regions. Assuming that Quad-Trees resulting from a split are numbered in the order SW(southwest), SE(southeast), NW(northwest) and NE(northeast), which are represented by 00, 01, 10 and 11 respectively 3. According to the index structure of quad-tree we can predict the point in given figure 1 that S2 is in SW and S5 is in NE another SE and NW block are recursively again split sequentially into four block. In figure 2 it can be seen that a leaf node is set black if it is not empty else it is white leaf node. Now searching query maximal depth of the Quad-Tree is 2, the split sequence of the node 1 is "SE, NE" and its code is represented by 0111. While author [7] also state that the Quad-Tree structure is kept log of space partition based signature of the objects, that ways the level of a node in the Quad-Tree is available. However, the correct node (region) information on the code and the level information are visualize.

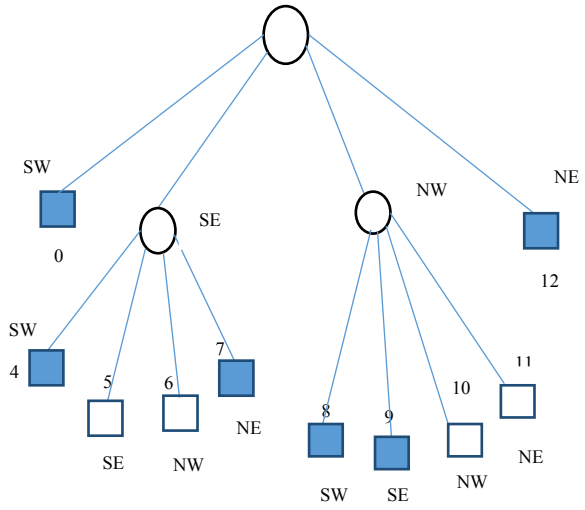

Figure 2 quad-Tree structure on road network

Top k spatial-textual search for each keyword t1 and t2 (i.e. "orthopedic" and "hospital") which will belong to the vocabulary from the document for that we build a linear Quad-Tree, which is denoted by L1, which shows the occurrence of keyword t1.

While the blue color filled leaf nodes, which is explicitly keep the Quad-Tree structure, and also serves as the signature of the objects in Li. This information can be easily fit into the main memory just require a 1 bit for each node of the Quad-Tree and signature is set to 1 for blue leaf nodes and non-leaf nodes would be 0.

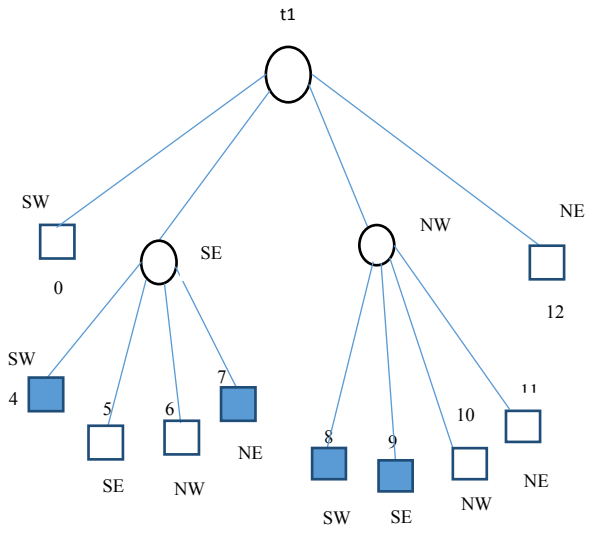

Figure 3 t1 keyword on quad tree

In given figure 3 and 4 it can be seen that for each keyword t1 and t2 it constructed a Quad-Tree separately. Merging this concept in figure 5 with B^+ tree for linking with the textual information and get the accurately ranked query data for k result to show.

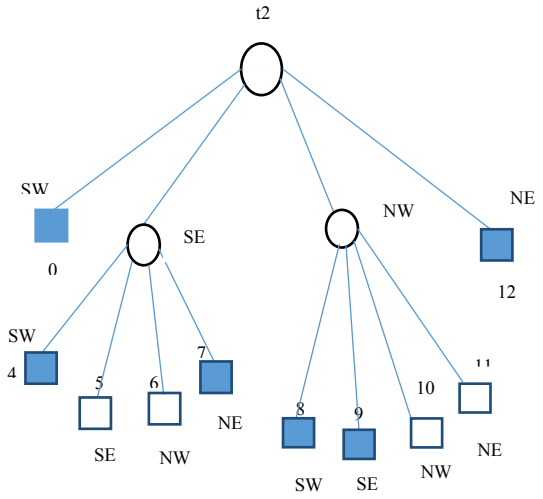

Figure 4 t 2 keyword on quad-Tree

Spatial-Textual query on the road network is also one challenging problem for that it is having the set of spatio-textual objects p € P on the edges E of the road network G. Each object p has a spatial location p.l and a textual description p.d. The ratio of a set of objects P is denoted as |P|.

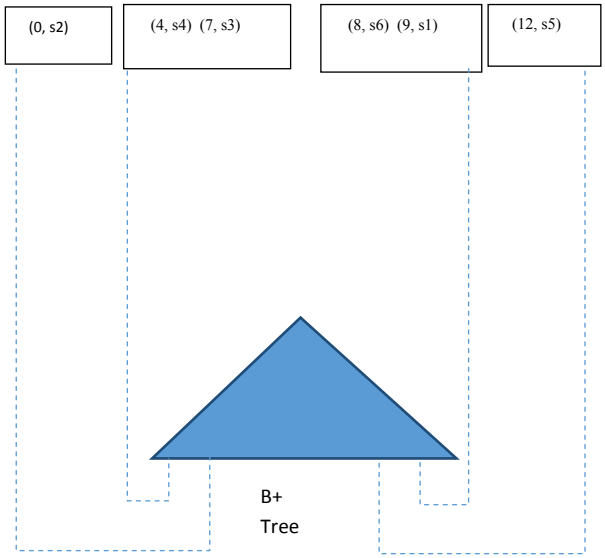

Figure 5 hybrid index

The road network distance between the object p and the ends of the edge in which it also define |v, p| or |v′, p|, where v and v′ are the endpoints of the edge (v, v′) where p located. The smallest path for the two objects p and p′ in the street network graph G is being ||p, p′||. Moreover a set of objects in the edge (v, v′) and (v′, v) are the same, and the distance |v, p| is equal to the distance |v, v′| − |v′, p|. So the distance can be known between p and one vertex is sufficient to obtain the distance between p and the other vertex v. Instead of storing information of vertex and edge we can also use another approach that used latitude & longitude information and get the distance out of it.

In parallel study about the proposed method is using the IR-Tree index structure, but as per the study it results are not very flexible compare to ILQuad-Tree. So it is worth discussing about it index maintained as the data are growing. ILQuad-Tree structure being choose by the author for future enhancement in the concept for search new capable of getting more efficiently.

IV. PROPOSED WORK

Basically Search enzymes are keyword-Oriented, many time it searching is talking about the semantic search that meaningful search. But always they rely on the keyword and it synonym searching which are the list of another synonym. In our study about search engine the point of analysis is it categorization technique which nowadays more used by various search engines. Categorize the query keyword which are used for search such hotel, hospital, temple, lodge etc. for better search experience. The aim is to achieve is that classified the document into some type of prefund category and each document either contain the categorize keyword in once or number of amount. Now here the challenge is for making the text classifier by hand is difficult and time consuming. So it more been advantageous of learn from the example. Make it index and tokenize it by applying the filter so it will produce faster result which will boom the spatial-textual search for quality most k result on the road network searching.

Before moving to toward the textual information for the hybrid structure of index, first will discuss how to calculate distance from the latitude & longitude information. For this purpose in our propose work we will uses haversine formula which is design specially to calculate distance from the latitude and longitude information. Now for any two points on a road network, the haversine of the central angle between them is given by

$$haversine\left(\frac{d}{r}\right) = haversin(\phi 2 - \phi 1) + \cos(\phi 1)\cos(\phi 2)\text{haversin}(\lambda 2 - \lambda 1)$$

$$haversin(\theta) = sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos(\theta)}{2}$$

Here d stand for the distance between the two latitudes longitude point (along a great circle of the sphere; see spherical distance), r is the radius of the sphere. $\phi 2 - \phi 1$: Latitude of point 1 and latitude of point 2. $\lambda 2 - \lambda 1$: Longitude of point 1 and longitude of point 2. To Solve d by applying the inverse haversine or by using the arcsine function:

$$d = 2r\ arcsin\left(\sqrt{\begin{matrix} haversin(\phi 1 - \phi 2) + \\ \cos(\phi 1)\cos(\phi 2)\text{haversin}(\lambda 2 - \lambda 1) \end{matrix}}\right)$$

$$= 2r\ arcsin\left(\sqrt{\begin{matrix} sin^2\left(\frac{\phi 2 - \phi 1}{2}\right) + \\ \cos(\phi 1)\cos(\phi 2) sin^2\left(\frac{\lambda 2 - \lambda 1}{2}\right) \end{matrix}}\right)$$

Hereby getting value d from the haversine formula we get accurate distance with is spatial true.

Now the problem sort out regarding the distance and create range in the spatial environment. The machine learning approach, as in the first phase will talking about the categorization of keyword or information related to the query object. Unfortunately, it won't help it if it been classified by hand and will take so long time. Else it remain will one solution that to trained data by machine learning approach and it is will compute many iteration and as a result it will provide with the more classified or categorization result. In this propose system given by the figure 6 in the first phase it is having the unknown function f: x->y which is ideal not know by the user. Second phase is training data which means that now the data have to be specified with the various parameter. Training data concept is usually specifications of various parameter by the user, if the data is fulfilling it parameter information then according to it vector fulfilling criteria it is been sort out or classified. But before that the third phase is learning algorithm and hypothesis set are been apply. Machine learning have many algorithm for classifier but will choose SVM for the efficient result for the spatial-Textual data classifying.

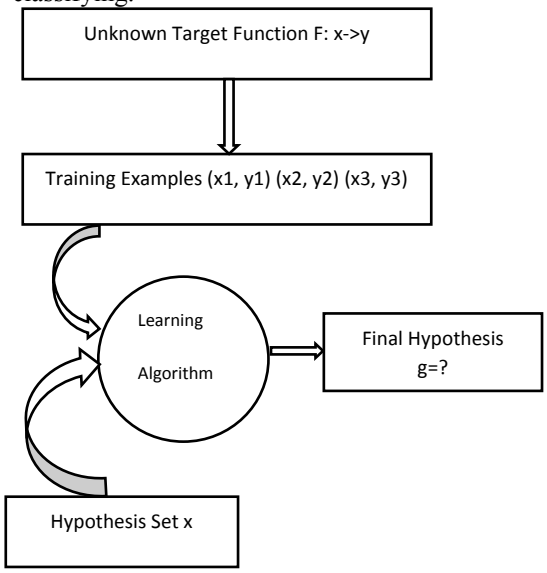

Figure 6 simple machine learning

Now for any query keyword, SVM will process through all the information literately and will provide the efficient top k result for the spatial-Textual data on the road network.

In figure 7 SVM classifier [9],[10] show as differentiate between relevance and non-relevance result, -1 is not relevance to the query keyword which is fired by the user and +1 is the relevance result for the query keyword. In our example from figure 1 upper +1 is a keyword data which are having both of the keyword t1 and t2 means it is fulfil the spatial constraints and the textual relevance information and -1 which will not having the t1 and t2. Still don't underestimate SVM since it iterative classification still able to classifier the data which is in -1 section, mean it is having some relation to the query keyword so it should not be leave. SVM classifier decided where to put it optimal hyper plane so that all the relevance keyword query should be cover up. But still it move +1 and -1 to find any object which can actively participate for relevant and non-relevant game or not.

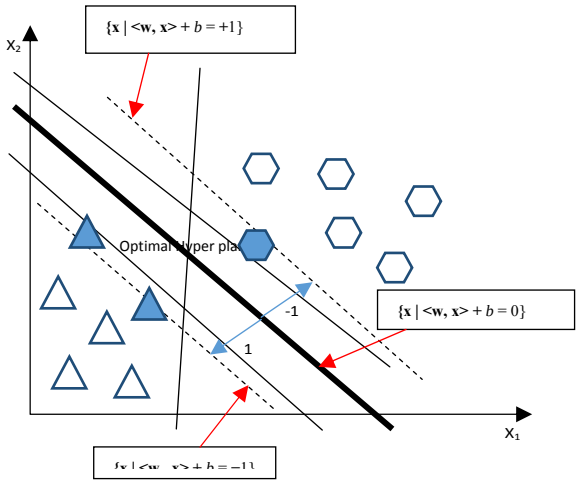

Figure 7. SVM structure

AT last the proposed algorithm will give the efficient Top k Result for spatial-Textual information retrieve on road network we can have the accurate result with using SVM classifier.

V. CONCLUSION

In overall concept of getting top k result which are having efficiently and accurately Spatial and textual data information. Our proposed system is having efficient result whenever the query is been dynamically asked. SVM classifier that work for many dimension provide and good classifier information it also increase the performance and also the cost of query processing is decreases.

REFERENCE

1. João B. Rocha-Junior_and Kjetil Nørvåg, "Top-k Spatial Keyword Queries on Road Networks", ACM, March 2012.
2. Ian De Felipe Vagelis Hristidis Naphtali Rishe, "Keyword Search on Spatial Databases*", ACM, March 2012.
3. G. Cong, C. S. Jensen, and D. Wu, "Efficient retrieval of the top-k most relevant spatial web objects," *PVLDB*, vol. 2, no. 1, 2009.
4. D. Wu, G. Cong, and C. S. Jensen, "A framework for efficient spatial web object retrieval," *VLDB J.*, 2012.
5. Ali Khodaei, Cyrus Shahabi, and Chen Li, "Hybrid Indexing and Seamless Ranking of Spatial and Textual Features of Web Documents", ACM, March 2013.
6. Ashish Kundu, ElisaBertino, "Structural Signatures for Tree Data Structures", ACM, August 23-28, 2008.
7. Chengyuan Zhang, Ying Zhang, Wenjie Zhang†, Xuemin Lin, "Inverted Linear Quad tree: Efficient Top Spatial Keyword Search", IEEE, 2013.
8. Geo Cong, Christian S. Jensen, Dingming Wu, "Efficient Retrieval of the Top-k Most Relevant Spatial web object", ACM,2009.
9. Mingjun Song and Daniel Civco,"Road Extraction Using SVM and Image Segmentation",Photogrammetric Engineering & Remote Sensing Vol. 70, No. 12, December 2004.
10. Shengyan Zhou, Jianwei Gong, Guangming Xiong, Huiyan Chen and Karl Iagnemma."Road Detection Using Support Vector Machine based on Online Learning and Evaluation",Manuscript received January 15, 2010. .This work was supported by the National Natural Science Foundation of China under grant No. 90920304.
11. Neelima Guduru,"Text Mining With Support Vector Machines And Non-Negative Matrix Factorization Algorithms",A Thesis Submitted Computer Science, University Of Rhode Island 2006
12. Dongxiang Zhang, Kian-Lee Tan, Anthony K. H. Tung, "Scalable Top-K Spatial Keyword Search", ACM, March 2013.