International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Review on Relevant Top-k Neighbor Search with Keywords

Sonali B. Gosavi[1], Shyamrao V. Gumaste[2]

[1]M.E. Second Year Student, Department of Computer Engineering, SPCOE, Otur, Pune, India

[2]Professor, Department of Computer Engineering, SPCOE, Otur, Pune, India

Abstract: ***Internet has became most integral part of today's modern era. Need for having required information on fingertip creating new challenges in engineering and technology world. As data size is increasing providing required information which can satisfy user's criteria is tedious job. Data mining help to process such huge data. Using these techniques people can access information like stores, hospitals, technical information, books etc. but today's trend is changing an along with just information people need geographical information too so they can reach particular shop or hospital. Today's applications also required to process such queries which need to satisfy both textual as well as spatial information need of user. For example instead of searching the hospital only, one can search the hospital having particular specialist doctors, no. of words, etc. To satisfy such queries the solution that is been used is the IR2-tree; but IR2-tree is having some pitfalls hence an alternative to this is to use WIBR- Tree which is capable to handle multidimensional data and process nearest neighbor queries with keywords.***

Keywords: Nearest neighbor search, keyword search, spatial inverted index, WIBR-tree, IR2-tree.

1. Introduction

Spatial database contains multidimensional objects like geographical data (such as points, rectangles, etc.). But because of the wide spread use of search engine, it is essential to search spatial data as well as information of that spatial data. let us focus on the basic notation of the spatial database. In the spatial database the locations of small entities are represented in the form of points in map e.g. restaurants, hospitals, hostels etc. and to represent larger entities such as lakes, gardens, grounds, etc. rectangle is used. GIS (Geographical Information System) gives all the restaurant in the address given in query. But the nearest neighbor methodology will give nearest restaurants of the given address. Now a days many search engines tries to answer spatial queries. But traditional spatial queries focus on the geometric properties of objects, such as whether point is in rectangle, or how close two points are from each other, etc. In modern era, it is quite essential to process queries that need both spatial properties as well as information about the spatial data. For example, it would be more useful if a search engine can be used to find the nearest restaurant that offers "chicken kolhapuri, veg manchurian and beer" all at the same time. The nearest neighbor algorithm gives the nearest restaurant among only those providing all the demanded drinks and foods. There are simple ways to process queries that contains both spatial as well as text features. For example to find restaurants for above query one can first retrieve the text features i.e. restaurant those menu is having given foods and drinks and then from resultant restaurants sort with spatial data i.e. nearest first. Alternative way to do it reversely by first finding spatial data of the query and then find the keywords. The major drawback of the above typical approach is that they are unable to process both spatial data and keywords at same time, so it is difficult to handle real time query using these approaches. This constraint may yields in missing of nearest neighbor. Previous study reflects that, authors used different approaches for retrieving data from server.[1]. Those concepts are R-tree for the spatial index and signature file for keyword based document retrieval. For that purpose they developed a structure named it as IR2 - tree which has advantages of both signature files and R-trees [3]. IR2-tree not only preserves the object of the spatial data like R-tree but also filter the exact portion of the query like signature file. Along with this IR2-tree also inherits pitfalls of signature files: false hits. In case of signature files, due to its conservative nature, sometimes it may still direct the search to some objects, even though they do not have all the keywords. Thus it needs cost to check an object which satisfying a query or not cannot be resolved using only its signature, but requires loading its whole text. It is very costly method for the Information retrieval. But the false hit problem is not only concern with signature file but also it may carry in other methodologies. Previous study reflects that, authors used different approaches for retrieving data from server. This method is very useful for the accessing the point co-ordinates in to an inverted bitmaps with small space. WIBR stores the spatial locality of points in R-tree and textual part in inverted bitmaps at very small space.

2. Literature Review

A. Overview of signature based file:

Signature file generally refers to a hashing-based framework known as superimposed coding (SC). Basically it is designed to perform membership tests. For example, to determine whether a query word w exists in a set W of words. If SC says no, then w is definitely not in W. On the other hand, SC returns yes, can be either way, in which case the whole W must be scanned to avoid a false hit. This is further explained using following example.

Consider a string of length L bits. Repeated bit M is 2. SC works in the same manner as of Bloom Filter. Let us consider

Table 1: Hash Format of string computation with L=5, M=2

word	Hashed bit string
A	00101
B	01001
C	00011
D	00110
E	10010

Table 2: Spatial points with text

P	**W**p
P1	a, b
P2	b, d
P3	D
P4	a ,e
P5	c, e
P6	c, d, e
P7	b, e
P8	c,d

a word / string of length L = 5 and repeated bit M = 2 as shown in Table I. For example: In the above table we have to assume l=5,m=2 means if we have h(a) of a word, third and fifth bit set to 1.from left. In the hashing function we will OR the bit string of all the bit. Suppose I want to calculate signature of a,b so 01101.For finding the query word wq from W the SC perform the membership test in that test it will check whether all 1's of W appear at same location. If not it is sure that wq is not in the W. But sometimes false hit may occurs like assume that we want to check 'c' is member of the a,b using the set of signature h(c)=00011.and h(a,b)=01101 in the h(c) the fourth bit is 1,and h(a,b) fourth bit is 0 so the 'c' is not the member of a,b.

False hit example: Consider the membership test of SC in which 'c' will be test in (b,d) in h(b,d) the fourth bit is 0 and h(c) fourth bit is 0 and SC report that 'c' is member of (b.d) and that is false hit.

B. IR2-tree

IR2-tree is based on the R-tree in fig.1. Each leaf, non-leaf entry is E which is summary of the text object. In the fig.1 we will illustrate the example based on data set of fig.1and hash value which is represented in the Table 1. The string value 01111 in Fig 1 is the leap entry, which is P2.The signature of the wP2 = b, d which is the document of the P2.

The string 11111 is the non-leaf entry E3 is the signature of wp2+ wp6 means the signature of the non-leaf entry is the combination of the signature of leaf node. Normally R tree, best first search algorithm is the better option of the NN nearest neighbor search.[1]For IR2-tree we have to fire the query point 'q' with associated text wq. The IR2-tree generate the ascending order of the distance of MBR to 'q' (MBR is the leaf entry).Pruning the entry whose signature which is absent the any one word of Wq.so in the Fig.1 for the verification the algorithm read all node of the tree and fetch the entry of p1,p4,p6 for the word c,d because the Wq is c,d and the final answer is p6 while p2,p4 are the false hit. So in the IR2-tree avoid the false hit which was occurred in signature file.

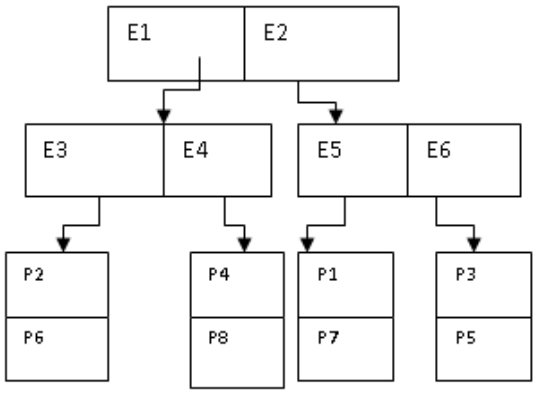

Figure 1: Signature of the entry

Table 3: Example of an inverted Index

word	Inverted list
a	p1,p4
b	p1,p2,p7
c	p5,p6,p8
d	p2,p3,p6,p8
e	p4,p5,p6,p7

1) Review study of Spatial invert index: Spatial invert index is the best method for accessing the keyword based retrieval .in the following list we will see the how to arrange the inverted index of points and the associated text of that point[5]. According to above list we have to create the list of inverted index which is having query word and associated point which having the same word [1]. One more point is that the list of the word is sorted order with regards point ID. So at the time of the query processing merge step will be performed on list. For example suppose we want find the point which is having words c, d because of that we will compute insertion of the inverted list. In the NN algorithm NN processing is with the IR2-tree. In that the points are retrieved from the index .Specifically NN query q with keyword set Wq the query method of I-index first determine the set of pq of all the points that will contains all the query word and then do —pq— randomly for finding the distance of pq from q.
2) Overview of Dbxplorer: The above paper is related to Keyword-Based Search over Relational Databases [2]. In day today internet is very user friendly for accessing the data. In this paper they have to give us the powerful question language. It will find the keyword from the server and retrieve the related web pages for the user.
3) Query processing on geographic data: In the geographic search the search engine allow the user to fire the query or find the result based geographic region [7]. It's also called local search, it is also useful for the extracting the knowledge of any location. It is also useful in the GIS. For the geographical search engine we need association of text as well as spatial data.

C. Basic technology for the geographic search:

1) Geo coding: In the geo coding technique three steps are necessary that are geo extraction, geo matching, geo propagation. Geo extraction: All the elements from a page which indicate query location. That is city name, contact number distance and generate the footprint. For the second step that is geo matching that foot print of same page will

be considered and in the third step that is geo propagation increase the quality scope of the geocoding by analyzing the link structure and the web pages topology [9]. And from that site map they will generate tree result.

2) Geographic query processing: Each query is having text term and query footprint means geographic ranking regarding user request. Thus in the above technique geographic ranking assign the score to each document footprint. Thus our overall ranking function in the form of

$$f(D, q) = g(fd, fq) \sum pr(D) \sum f(D, q) \quad (1)$$

In the above expression f(D,q)=Term based ranking, pr(D)=Global ranking , g(fd,fq)=geographic score And the expression will be calculated from footprint.

D. Concept of Bloom filter

Bloom filter [8] is one of the data structures that are useful for the membership queries to a set. The bloom filter needs very less space. Bloom filter avoids the false hit. It is normally used in the network. It is also used in the distributed database. As per above section in the signature file also use the bloom filter for the membership testing. Also it used in the password data structure and spell checking. Let F be a function of D = 0, 1, 2, n-1 to R= ⊥, 1, 2, 2r-. According to above expression $f(x) = \perp$ forallx $\perp$. $\perp$ this symbol is used to represent 0.

E. View of Spatial Keyword Index

Above topic name suggest the retrieving concept that is one will retrieve the geographic location as well as associated text with the query. For the spatial ,keyword retrieval one need to first of all collective answer of the spatial query that full fill the user requirement for that one have to assume the database of spatial multidimensional object and after that one will find the set of keyword[6].

In that case q = (λ, ψ) where λ = location and ψ = keywords. This type of query is called spatial group keyword. In the above paper IR-tree and approximation algorithms are used.

3. Mathematical Model

Let S is a system of relevant Top-k neighbor search with Keywords which can be represented as,

S = I, P, O

Where,

I = Set of all inputs given to the system represented as,

I = wi1, wi2,. . . win where i1,i2,..in are input keywords

P = Set of all possible processes required to get expected

Output it includes

1. Searching iteratively
2. Pruning
3. Word Partitioning

O = Set of all the required output can be represented as,

O = Do1, Do2,. . . where O1, O2 . . . are all output Documents which has required keywords.

This method gives output in non-polynomial time. So, problem becomes NP-Complete.

4. Implementation Details

The goal of the implementation is to make an all-around evaluation on important aspects of the geo-textual indices and compare their performance. All indices and algorithms were implemented in Java running the Windows 8 OS. Machine is equipped with Intel(R) Core(TM)i3 @ 1.70GHz, 8GB RAM, and 1TB SATA disk. For ensuring a comparable evaluation results, the same server is used for conducting experiments on the same dataset. The Java Virtual Machine Heap is set to 2GB.

A. Datasets

For this study we are considering Hotel dataset which is prepared containing Hotel's details of location and types of Services they are providing. We concentrate on Hotels in Pune region.

B. Working Details

The architecture of proposed system is shown in Fig.2. It works in following steps:

1. Merging and distance browsing: In the spatial retrieving the basic problem is the bottleneck so need to avoid it. But in the I-index is having the simple way to recover it. In the I-index we have to preserve co-ordinates of the point in one group in the inverted list and that co-ordinates of the list are used to generate the R-tree. Now discuss how to perform keyword based NN. In this technique NN queries are processed with I index. For answering the query first of all we will access all the points which is having all the query keywords in Wq. It is very useful if we find the p very early in all the relevant inverted list. In that case we can access the list of element which are having less distance with q. so the p will be discovered all points of the list.

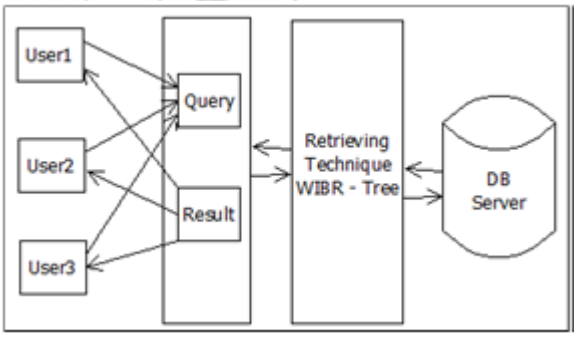

Figure 2: Proposed System

By using that we can count the number of copies of same point that will be relevant data. Consider an example if we want to NN search whose query point q and associated text is (c,d). for that search we have to use the list of word(c) and (d).from list

1) And now the new access order is depend on the distance of the given q. If we use the kNN then it will reported k nearest neighbor point and finish. Distance browsing is simple in the R-trees because R-tree uses best first algorithm which will give the exactly point with ascending order of the distance to q and R-trees are also the global access of the tree. For example at each step taking the point with the next point and return it. This algorithm is

normally work in the condition when the Wq small. But if the Wq is large then the out performance of sequential algorithm will be merged.

2) Weighted Independent Binary Representation (WIBR)tree: WIBR-tree [10] is a variant of IR-tree. It aims at partitioning objects into multiple groups such that each group shares as few keywords as possible. To achieve this goal, the objects in D is partitioned first into two groups using the most frequent word w1: one group whose objects contain w1 and the other group whose objects do not. Then it partition each of these two groups by the next frequent word w2. This process is repeated iteratively until each partition contains a certain number of objects. After partitioning, each group of objects becomes the leaf node of the WIBR-tree. Afterwards the tree is constructed following the structure of the IR-tree. When used for processing Boolean queries, the WIBR-tree [10] uses the inverted bitmap to replace the inverted file, which is denoted as the WIBR-tree, where a bitmap position corresponds to the relative position of an entry in its WIBR-tree node. The length of a bitmap is equal to the fan out of a node. Like the IR tree, the WIBR-tree can handle all the three types of query - Boolean kNN query, the top-k kNN query, and the Boolean range query.

C. Algorithm

Searching Algorithm
The search algorithm traverse the tree from the root in a way similar to B-tree. In the following we denote the rectangle part of an index entry E by EI, and the tuple-identifier or child pointer part by EP. Algorithm Search: Given an WIBR-tree whose root node is T, find all index records whose rectangles overlap a search rectangle S.

- Sl [Search subtrees] If T 1s not a leaf, check each entry E to determine whether EI overlaps S For all overlapping entries, invoke Search on the tree whose root node is pointed to by EP.
- S2 [Search Leaf node] If T 1s a leaf, check all entries E to determine whether EI overlaps S If so, E is a qualifying record.

5. Conclusion

This paper provides solution for problem of spatial keyword search and explained the performance limitations of current approaches. This system proposed a solution which is dramatically faster than current approaches and is based on a WIBR- tree. In particular we used the WIBR-Tree and showed how it is better than prior approaches. An efficient incremental algorithm was presented that uses the WIBR – Tree to answer spatial keyword queries. Here we are concentrating on application of searching of hotels at particular geographic area and which satisfy user requirement.

6. Acknowledgement

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We also thank the college authority for providing required infrastructure and support.

References

[1] Yufei Tao and Cheng Sheng, ”Fast Nearest Neighbor Search with Keywords,” National Research Foundation of Korea, GRF 4166/10,4165/11, and 4164/12 from HKRGC .
[2] S. Agrawal, S. Chaudhuri, and G. Das.Dbxplorer, “A system for keyword based search over relational databases”. In Proc. Of International Conference on Data Engineering (ICDE), pages 5 a 16, 2002.
[3] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger. The R*-tree, “An efficient and robust access method for points and rectangles”, In Proc. Of ACM Management of Data (SIGMOD), pages 322 a 331, 1990.
[4] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan, “Keyword searching and browsing in databases using banks”. In Proc. Of International Conference on Data Engineering (ICDE) , pages 431 a 440, 2002.
[5] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu, “Spatial keyword querying,” In ER, pages 16 a 29, 2012.
[6] X. Cao, G. Cong, and C. S. Jensen, “.Retrieving top-k prestige-based relevant spatial web objects”, PVLDB, 3(1):373 a 384, 2010.
[7] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi, “Collective spatial keyword querying”. In Proc. of ACM Management of Data (SIGMOD), pages 373 a 384, 2011.
[8] B. Chazelle, J. Kilian, R. Rubinfeld, and A. Tal. “The bloomier filter: an efficient data structure for static support lookup tables”. In Proc. of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 30 a 39, 2004.
[9] Y.-Y. Chen, T. Suel, and A. Markowetz, “Efficient query processing in geographic web search engines”. In Proc. of ACM Management of Data (SIGMOD), pages 277 a 288, 2006.
[10] D. Wu, M. L. Yiu, G. Cong, and C. S. Jensen, “Joint top-k spatial keyword query processing”. IEEE TKDE, 24(10):18891903, 2012.