Searching of Nearest Neighbor Based on Keywords using Spatial Inverted Index

B. SATYA MOUNIKA[1], J. VENKATA KRISHNA[2]

[1]M-Tech Dept. of CSE SreeVahini Institute of Science and Technology TiruvuruAndhra Pradesh

[2]Assoc.Professor Dept. of CSE SreeVahini Institute of Science and Technology TiruvuruAndhra Pradesh

Abstract:

Many search engines are used to search anything from anywhere; this system is used to fast nearest neighbor search using keyword. Existing works mainly focus on finding top-k Nearest Neighbors, where each node has to match the whole querying keywords .It does not consider the density of data objects in the spatial space. Also these methods are low efficient for incremental query. But in intended system, for example when there is search for nearest restaurant, instead of considering all the restaurants, a nearest neighbor query would ask for the restaurant that is, closest among those whose menus contain spicy, brandy all at the same time, solution to such queries is based on the IR2-tree, but IR2-tree having some drawbacks. Efficiency of IR2-tree badly is impacted because of some drawbacks in it. The solution for overcoming this problem should be searched. The spatial inverted index is the technique which will be the solution for this problem.

Keywords: - Access control, e -health, privacy preserving, cloud computing.

1. INTRODUCTION

Nearest neighbor search (NNS), also known as closest point search, similarity search. It is an optimization problem for finding closest (or most similar) points. Nearest neighbor search which returns the nearest neighbor of a query point in a set of points, is an important and widely studied problem in many fields, and it has wide range of applications. We can search closest point by giving keywords as input; it can be spatial or textual. A spatial database use to manage multidimensional objects i.e. points, rectangles, etc. Some spatial databases handle more complex structures such as 3D objects, topological coverage's, linear networks. While typical databases are designed to manage various NUMERIC'S and character types of data, additional functionality needs to be added for databases to process spatial data type's efficiently and it provides fast access to those objects based on Different selection criteria. Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. The search engines available today provide keyword search on top of sets of documents. When a set of query keywords is provided by the user, the search engine returns all documents that are associated with these query keywords.

2. LITERATURE SURVEY

1. Keyword search on spatial databases:

This method focuses on searching top-k nearest neighbor query. They present an efficient method to answer top-k spatial keyword queries, it introduce an indexing structure called IR2-Tree (Information Retrieval R-Tree). IR2-tree is the most related work which is similar to our work. Here incremental algorithm is used to answer Top-k spatial keyword queries using the IR2 -Tree. The IR2-tree combines the features of R-tree with signature files. Signature file is nothing but a hashing-based framework that is based on the concept of superimposed coding as explored in this method. The problem with this technique is that as the number of words grows in size, scanning the entire list become tedious. When the list is not scanned it may result in false hits. As a solution to this problem inverted indexes were introduced in 5.

2. Spatial Keyword Query:

This hybrid index structure is used to search m-closest keywords. This technique finds the closest tuples that matches the keywords provided by the user. This structure combines the R*-tree and bitmap indexing to process the m closest keyword query that returns the spatially closest objects matching m keywords. To reduce the search space a priori based search strategy is used. Two monotone constraints are used as a priori properties to facilitate efficient pruning which is called as distance mutex and keyword mutex. But this approach is not suitable for handling ranking queries and in this number of false hits is large.

3. Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems:

Location based information is stored in GIS database. These information entities of such databases have both spatial and textual descriptions. This method introduces a framework for GIR system and focus is on indexing strategies that can process spatial keyword query. It introduces two index structures to store spatial and textual information.1) Separate index for spatial and text attributes 2) Hybrid index. But by using first structure that is separate index for spatial and text attributes, if filtering is done first, many objects may lie within a query is spatial extent, but very few of them are relevant to query keywords. This increases the disk access cost by generating a large number of candidate objects. The subsequent stage of keyword filtering becomes expensive. And by using second structure that is hybrid index there are high overhead in subsequent merging process. Idea of geographical web search was illustrated in [2], [4] [5].

3. RELATED WORK

1. IR2 Tree:

The IR2 – Tree [12] combines the R-Tree and signature file. First we will review Signature files. Then IR2-trees are discussed. Consider the knowledge of R-trees and the best- first algorithm [12] for Near Neighbor Search. Signature file is known as a hashing-based framework and hashing -based framework is which is known as superimposed coding.

2. Drawbacks of the IR2-Tree:

IR2-Tree is first access method to answer nearest neighbor queries. IR2-tree is popular technique for indexing data but it having some drawbacks, which impacted on its efficiency. The disadvantage called as false hit affecting it seriously. The number of false positive ratio is large when the aim of the final result is far away from the query point and also when the result is simply empty. In these cases, the query algorithm will load the documents of many objects; as each loading necessitates a random access, it acquires costly overhead.

4. PROPOSED SYSTEM

We design a variant of inverted index that is optimized for multidimensional points and is thus named the spatial inverted index (SI-index). This access method successfully incorporates point coordinates into a conventional inverted index with small extra space. Mean while an SI-index preserves the spatial locality of data points.

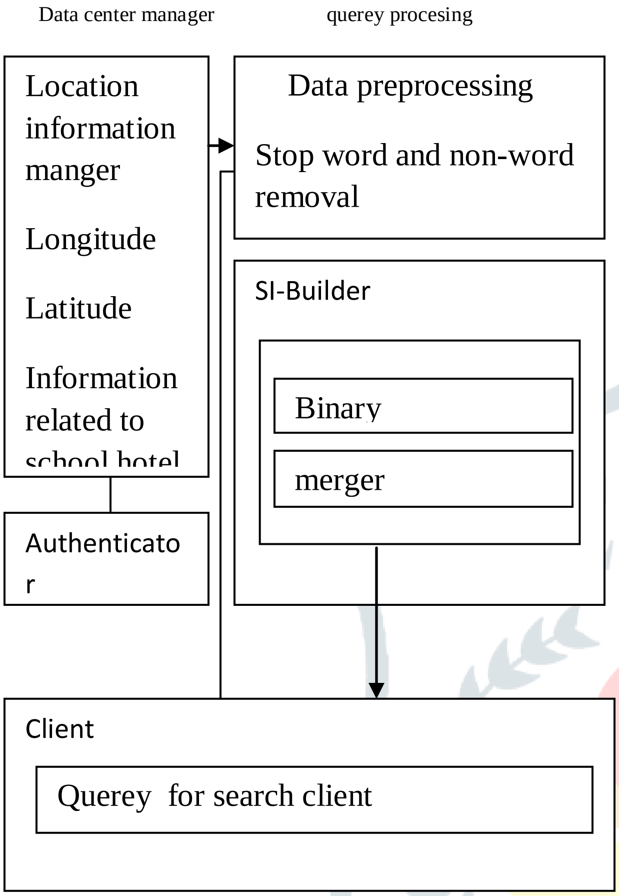
Advantages:

1. It an gives the nearest neighbor queries with keywords in real time.

2. Computational cost is less.

3. It is straightforward to extend our compression scheme to any dimensional space.

5. ARCHITECTURE



The architecture of keyword search using spatial inverted index

6. IMPLEMENTATION

1. Data Center Manager:

In first module location information manager use “Google Map” of particular area, by selecting any location on that map it gives longitude and latitude. The retrieved longitude and latitude saved into the database.It also includes information related to school, bank, hotel, restaurant, college etc. Data preprocessing is applied on the data that is managed by information manager.

2. Querey Processing:

a) Data preprocessing:

When user enter the query to search then it is passed to the data preprocessing, first preprocessing of the query is stop-words removal from that query. Stop-words are frequently occurring and unimportant words in a language that helps to construct sentences but do not represent any content of the documents. Stop-words in include: a, about, an, are, as, at, be, by, for, from, how, in, is, of, on, or, that, the, these, this, to, was, what, when, where, who, will, with. Such words should be removed before documents are indexed and stored. For example if user query is “search for nearest school” then remove “for” word from that query. Next step is stemming which is the process of reducing words to their stems or roots. A stem is the portion of a word that is left after removing its prefixes and suffixes. For example if user query is “search nearest engineering colleges” then remove “s” from colleges it remains stem that is “college”.

b) SI-Indexer :

The proposed system is based on the spatial inverted index. SI-index is an compressed version of I-index with compressed coordinates. Compression eliminates the defect of I-index such that SI-index consumes much less space. Compression methods are used in order to reduce size of index where each inverted index contains only IDs.

7. MODULES

1. Customer Registration:

In this module, the user will have to register first. Once the user does the registration then he/she can access the application. For registration user have to enter the basic information about himself. User also have to set the username and password. This all registration information is get stored into database. The IMEI number is automatically get stored into database once user do the registration.

2. Customer Login:

In this module, after the registration customer can login through mentioned username and password.

3. Hotel Registration:

In this module, Admin register the hotel with its famous dish. Hotel owner have to do the registration then only the hotel get search through application. Also hotel owner have to add the menu which is available in the hotel so that user can search the hotel through keyword. Only registered hotels will be displayed in the application. These hotel's location will be seen in the map with distance. Each hotel owner will get the separate username and password for login.

4. Hotel Login / Admin:

In this model once Hotel Owner login into application then he can insert the menu or update the menu.

8. ALGORITHM

Output: Semantic nearest neighbor.

Description:

```
1: Read three values as (id, x, y
        Where, id - id of place or word
         x - Position on x-axis of place
        y - Position on 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}
         Convert x to binary values as,
         b1= binary(x);
        c. Repeat step a and b for y and
create b2.
        d. Merge b1and b2bit by bit and
store in b3
        For each bit in b1and b2,
        B = b1&b2;
        e. Convert B 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.
```

9. CONCLUSION AND FUTURE WORK

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. In this paper we used two level gap-keeping to save space cost.

10. ACKNOWLEDGEMENT

I express great many thanks to Prof. S. S. Nandgaonkar for his great effort of supervising and leading me, to accomplish this fine work. Also to college and department staff, they were a great source of support and encouragement. To my friends and family, for their warm, kind encourages and loves. To every person gave us something too light my pathway, I thanks for believing in me.

11. 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] 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 277–288, 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.

Authors Profiles

B. SATYA MOUNIKA
M-Tech Dept. of CSE SreeVahini Institute
of Science and Technology Tiruvuru
Andhra Pradesh.

J. Venkata Krishna
Assoc.ProfessorSreeVahini Institute of
Science and Technology Tiruvuru
Andhra Pradesh
"M.TECH, Ph. D"