International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Survey of Searching Nearest Neighbor Based on Keywords using Spatial Inverted Index

Shilpa B. Patil
M.E. Computer-II,
Vidya Pratisthan`s College Of Engineering-Baramati.
patil.shilpa962@gmail.com

`

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 – Nearest Neighbor Search, IR2-tree, Nearest, Range search, Spatial inverted index.

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. 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. Spatial inverted index is the technique which will be the solution for this problem. Spatial database manages multidimensional data that is points, rectangles.

This paper gives importance spatial queries with keywords [5] [6] [9] [10]. Spatial queries with keywords take arguments like location and specified keywords and provide web objects that are arranged depending upon spatial proximity and text relevancy. Some other approaches take keywords as Boolean predicates [1] [2], searching out web objects that contain keywords and rearranging objects based on their spatial proximity. Some approaches use a linear ranking function [7] [8] to combine spatial proximity and textual relevance. Earlier study of keyword search in relational databases is gaining importance. Recently this attention is diverted to multidimensional data [3] [4] [11]. N. Rishe, V. Hristidis and D. Felipe [12] has proposed best method to develop neighbor search with keywords. For keyword-based retrieval, they have integrated R-tree [14] with spatial index and signature file [12]. By combining R-tree and signature they have developed a structure called the IR2-tree [12]. IR2-tree has merits of both R-trees and signature files. The IR2-tree preserves object's spatial proximity which important for solving spatial queries.

Literature Survey

Literature review is contains the points IR2 - Tree, Drawbacks of the IR2-tree, Previous methods.

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 (SC)[12].

Drawbacks of the IR2-Tree

IR2-Tree is first access method to answer nearest neighbour 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 [12].

Keyword search on spatial databases

This work, mainly focus on finding top-k Nearest Neighbors, in this method each node has to match the whole querying keywords. As this method match the whole query to each node, it does not consider the density of data objects in the spatial space. When number of queries increases then it leads to lower the efficiency and speed. They present an efficient method to answer top-k spatial keyword queries. This work has the following contributions: 1) the problem of top-k spatial keyword search is defined. 2) The IR2-Tree is proposed as an efficient indexing structure to store spatial and textual information for a set of objects. There are efficient algorithms are used to maintain the IR2-tree, that is, insert and delete objects. 3) An efficient incremental algorithm is presented to answer top-k spatial keyword queries using the IR2-Tree. Its performance is estimated and compared to the current approaches. Real datasets are used in our experiments that show the significant improvement in execution times.

Disadvantages: -

1. Each node has to match with querying keyword. So it affects on performance also it becomes time consuming and maximizing searching space.
2. IR2-tree has some drawbacks.

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

Location based information stored in GIS database. These information entities of such databases have both spatial and textual descriptions. This paper proposes a framework for GIR system and focus on indexing strategies that can process spatial keyword query. The following contributions in this paper: 1) It gives framework for query processing in Geo- graphic Information Retrieval (GIR) Systems. 2) Develop a novel indexing structure called KR*-tree that captures the joint distribution of keywords in space and significantly improves performance over existing index structures. 3) This method have conducted experiments on real GIS datasets showing the effectiveness of our techniques compared to the existing solutions. It introduces two index structures to store spatial and textual information.

A) Separate index for spatial and text attributes:

Advantages: -

1. Easy of maintaining two separate indices.
2. Performance bottleneck lies in the number of candidate object generated during the filtering stage.

Disadvantages: -

1. If spatial 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.

B) Hybrid index

Advantages and limitations: -

1. When query contains keywords that closely correlated in space, this approach suffer from paying extra disk cost accessing R*-tree and high overhead in subsequent merging process.

Hybrid Index Structures for Location-based Web Search.

There is more and more research interest in location-based web search, i.e. searching web content whose topic is related to a particular place or region. This type of search contains location information; it should be indexed as well as text information. text search engine is set-oriented where as location information is two-dimensional and in Euclidean space. In previous paper we see same two indexes for spatial as well as text information. This creates new problem, i.e. how to combine two types of indexes. This paper uses hybrid index structure, to handle textual and location based queries, with help of inverted files and R*-trees. It considered three strategies to combine these indexes namely: 1) inverted file and R*-tree double index.2) first inverted file then R*-tree.3) first R*-tree then inverted file. It implements search engine to check performance of hybrid structure, that contains four parts:(1) an extractor which detects geographical scopes of web pages and represents geographical scopes as multiple MBRs based on geographical coordinates. (2) The work of indexer is use to build hybrid index structures integrate text and location information. (3) The work of ranker is to ranks

the results by geographical relevance as well as non-geographical relevance. (4) an interface which is friendly for users to input location-based search queries and to obtain geographical and textual relevant results.

Advantages: -

1. Instead of using two indexes for textual and spatial information. this paper gives hybrid index structures that integrate text indexes and spatial indexes for location based web search.

Disadvantages: -

1. Indexer wants to build hybrid index structures to integrate text and location information of web pages. To textually index web pages, inverted files are a good. To spatially index web pages, two-dimensional spatial indexes are used, both include different approaches, this cause to degrading performance of indexer.

2. In ranking phase, it combine geographical ranking and non-geographical ranking, combination of two rankings and the computation of geographical relevance may affects on performance of ranking.

Conclusion

In this report, we have surveyed a Searching Nearest Neighbor based on Keywords using Spatial Inverted Index and evaluate the needs and challenges present in Nearest Neighbor Search. This report covers existing techniques for that and also covers upon new improvements in current technique. In this paper, we have surveyed topics like IR2 – Tree, Drawbacks of the IR2-Tree, Spatial keyword search, Solutions based on Inverted Indexes.

REFERENCES:

[1] I. De Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In ICDE, pp. 656–665, 2008.
[2] D. Zhang, Y. M. Chee, A. Mondal, A. K. H. Tung, and M. Kitsuregawa. Keyword search in spatial databases: Towards searching by document. In ICDE, pp. 688– 699, 2009
[3] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, "Processing Spatial- Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems," Proc. Scientific and Statistical Database Management (SSDBM), 2007.
[4] X. Cao, G. Cong, and C. S. Jensen. Retrieving top-k prestige-based relevant spatial web objects. PVLDB, 3(1):373–384, 2010.
[5] Y.-Y. Chen, T. Suel, and A. Markowetz. Efficient query processing in geographic web search engines. In SIGMOD, pp. 277–288, 2006.
[6] G. Cong, C. S. Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. PVLDB, 2(1):337–348, 2009.
[7] I. De Felipe, V. Hristidis, and N. Rishe. Keyword search on spatial databases. In ICDE, pp. 656–665, 2008.
[8] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, "Hybrid Index Structures for Location-Based Web Search," Proc. Conf. Information and Knowledge Management (CIKM), pp. 155-162, 2005.
[9] I.D. Felipe, V. Hristidis, and N. Rishe, "Keyword Search on Spatial Databases," Proc. Int'l Conf. Data Eng. (ICDE), pp. 656-665, 2008.
[10] C. Faloutsos and S. Christodoulakis, "Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation," ACM Trans. Information Systems, vol. 2, no. 4, pp. 267-288, 1984.
[11] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, "The R- tree: An Efficient and Robust Access Method for Points and Rectangles," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 322-331, 1990.
[12] G.R. Hjaltason and H. Samet, "Distance Browsing in Spatial Databases," ACM Trans. Database Systems, vol. 24, no. 2, pp. 265-318, 1999