Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 4, April 2016, pg.257 – 260

Survey on Fast Nearest Neighbor Search with a Keyword

Sayyed Muzammil Ali[1], Dr. A. S. Alvi[2]
[1]M.E student, Department computer Science & Engineering & S.G.B.A.U, India
[2]Professor, Department computer Science & Engineering & S.G.B.A.U, India
[1] smali.muzammil@gmail.com

Abstract— ***Conventional spatial queries, such as vary search and nearest neighbour retrieval, involve solely conditions on objects' geometric properties. Today, many modern applications support new form of queries that aim to find objects that satisfies both spatial data and their associated text. For example, rather than considering all the restaurants, a nearest neighbour query would instead invite the restaurant that is the nearest among those whose menus contain "steak, spaghetti, brandy" all at identical time. Presently the simplest solution to such queries relies on the IR2 -tree, which, as shown in this paper, includes a few deficiencies that seriously impact its efficiency. Impelled by this, we have a tendency to develop a brand new access methodology called the spatial inverted index that extends the standard inverted index to address multidimensional information, and comes with algorithms which will answer nearest neighbour queries with keywords in real time***

Keywords— Data Information Retrieval Tree, Keyword Search, Spatial Inverted Index.

I. INTRODUCTION

Spatial data mining is a special kind of data mining. The main difference between data mining and spatial data mining is that in spatial data mining tasks we use not only non-spatial attributes, but also spatial attributes. Spatial data mining is the application of data mining methods to spatial data. A spatial database is used to store large amount of space related data such as maps, medical imaging data etc. and manages multidimensional objects (such as points, rectangles, etc.), and provides quick access to those objects based on different choice criteria. Today, the widespread use of search engines has created it realistic to write down spatial queries in an exceedingly novel approach. Conventionally, queries focus on objects‟ geometric properties solely, like whether or not a point is in a rectangle, or however close two points are from one another. We have seen some trendy applications that have an ability to pick objects supported each of their geometric coordinates and their associated texts. As an example, Search engine will be fairly useful if it finds nearest restaurant that will offers the demanded food. Note that this is often not the "globally" nearest building (which would are came back by a standard nearest neighbor query), however the closest restaurant among solely those providing all the demanded foods and drinks. During this paper, we tend to design a variant of inverted index that's optimized for multidimensional points, and is therefore named the spatial inverted index (SI-index). This access technique with success incorporates point coordinates into a standard inverted index with little additional space, attributable to a delicate compact storage theme. Meanwhile, an SI-index preserves the spatial locality of Information points, and comes with an R-tree designed on each inverted list at space overhead.

II. RELATED WORK

Many applications need finding objects that are nearest to a given location that contains a group of keywords. An increasing variety of applications need the economical execution of nearest neighbour (NN) queries affected by the properties of the spatial objects A spatial keyword query consists of a query space and a group of keywords. The solution could be a list of objects hierarchical in line with a mix of their distance to the query space and also the connection of their text description to the query keywords. Spatial queries with keywords have not been extensively explored. In the past years, the community has sparked enthusiasm in studying keyword search in relational databases. It is until recently that attention was diverted to multidimensional data [12], [13]. The best method to date for nearest neighbor search with keywords is due to Felipe et al. [12]. They nicely integrate two well-known concepts: *R-tree* [2], a popular spatial index, and *signature file* [11], an effective method for keyword-based document retrieval. By doing so they develop a structure called the *IR*2-*tree* [12], which has the strengths of both R-trees and signature files. Like R-trees, the IR2-tree preserves objects' spatial proximity, which is the key to solving spatial queries efficiently. On the other hand, like signature files, the IR2-tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined.

A. R-TREE:

R-Tree makes use of solely Associate in Nursing R-Tree organization [16]. Given a distance-first top-k spatial keyword query, the algorithmic rule initial finds the top-1 nearest neighbor object to the query purpose Q.p. Then it retrieves that object (since the R-tree solely contains object pointers) and compares that object's matter description with the keywords of the query. If the comparison fails then that object is discarded, and therefore the next nearest object is retrieved. The progressive NN algorithmic rule is employed. Once a satisfying object is found it's came back and therefore the method repeats till k objects are came back. The drawback of this algorithmic rule is that it's to retrieve each object came back by the NN algorithmic rule till the top-k result objects are found. This doubtless will result in the retrieval of the many 'useless' objects. Within the worst case (when none of the objects satisfies the query's keywords) the whole tree must be traversed and each object must be inspected.

B. SIGNATURE FILE:

Signature files were introduced by Faloutsos and Christodoulakis [11] as a technique to with efficiency search a group of text documents. Signature files appear to be a promising access technique for text and attributes. Consistent with this technique, the documents (or records) are keep consecutive in one file ('text file'), whereas abstractions of the documents ('signatures') are keep consecutive in another file ('signature). So as to resolve a query, the signature file is scanned 1st, and plenty of no qualifying documents are at once rejected. In general signature file refers to a hashing-based framework, whose internal representation in keyword search on spatial information is thought as superimposed committal to writing (SC). It's designed to perform membership tests that confirm whether or not a query word w exists in a very set W of words. SC is conservative, within the sense that if it says "no" , then w is certainly not in W. on the opposite hand, if SC returns "yes", truth answer will be either manner, during which case the total W should be scanned to avoid a false hit.

Cao et al. [6] proposed collective spatial keyword query, they presented the new problem of retrieving a group of spatial objects, and each associated with a set of keywords. They develop approximation algorithm s with provable approximation bounds and exact algorithms to solve the two problems.

Lu et al. [18], combined the notion of keyword search with reverse nearest neighbor queries. They propose a hybrid index tree called IUR-tree (Intersection-Union RTree) to answer the Reverse Spatial Textual k Nearest Neighbor (RSTkNN) query that effectively combines location proximity with textual similarity. They design a branch-and-bound search algorithm which is based on the IUR-tree. To further increase the query processing, they proposed an improved variant of the IUR-tree called cluster IUR-tree and two corresponding optimization algorithm.

Zhang and Chee[19] introduced hybrid indexing structure bR*-tree, that combines the R*-tree and bitmap indexing to process the m-closest keyword query that returns the spatially closest objects matching m keywords. They utilized a priori based search strategy that successfully reduce the search space and also proposed two monotone constraints, distance mutex and keyword mutex to help effective pruning.

Ian De Flipe[10] presented an efficient method to answer top-K spatial keyword query. They proposed an index structure IR2-tree that combines signature files and R-tree to allow keyword search on spatial data objects that

each have limited number of keywords. Using the IR_2-tree an efficient incremental algorithm is presented to answer the spatial keyword queries.

G. Cong, C.S. Jensen, and D. Wu [12] proposed an approach that computes the relevance between the documents of an object and a query. This relevance is then incorporated with the Euclidean distance between object and query to calculate an overall similarity of object to query.

Yufie Tao and Cheng Sheng[17], developed a new access method which is called as spatial inverted index. It extends the conventional inverted index to lay hold on multidimensional data, and uses the algorithms that can answer nearest neighbor queries with keywords in real time. They designed a variant of inverted index called spatial inverted index that is optimized for multidimensional points. This access method successfully includes point coordinates into a conventional inverted index with small space.

III.EXISTING SYSTEM

The best method to date for nearest neighbour search with keywords is due to Felipe ET AL [12]. They nicely integrate two well-known concepts: R-tree [2], a popular spatial index, and signature file [11], an effective method for keyword based document retrieval. By doing so they develop a structure called the IR2-tree [12], which has the strengths of both R-trees and signature files. Like R-trees, the IR2-tree preserves objects‟ spatial proximity, which is the key to solving spatial queries efficiently. On the other hand, like signature files, the IR2-tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined. The IR2-tree, however, also inherits a drawback of signature files: false hits. That is, a signature file, due to its conservative nature, may still direct the search to some objects, even though they do not have all the keywords. The penalty thus caused is the need to verify an object whose satisfying a query or not cannot be resolved using only its signature, but requires loading its full text description, which is expensive due to the resulting random accesses. It is noteworthy that the false hit problem is not specific only to signature files, but also exists in other methods for approximate set membership tests with compact storage. Therefore, the problem cannot be remedied by simply replacing signature file with any of those methods.

The IR2-tree is the first access method for answering nearest neighbor search with keywords. As with many pioneering solutions, the IR2-tree also has a some drawbacks that affect its efficiency. The most serious one of all is that the number of false hits can be really large when the object of the final result is far away from the query point, or the result is simply empty. In these cases, the query algorithm would need to load the documents of many objects, incurring expensive overhead as each loading necessitates a random access. *IR2-tree*, fails to give real time answers, and is often slower by a factor of orders of magnitude, the deficiency of *IR2-tree* is mainly caused by the need to verify a vast number of false hits.

IV.PROPOSED SYSTEM AND ITS ADVANTAGES

So, new access method spatial inverted access method is used to remove the drawbacks of previous methods such as false hits. In this paper, 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, owing to a delicate compact storage scheme. Meanwhile, an SI-index preserves the spatial locality of data points, and comes with an R-tree built on every inverted list at little space overhead.

As a result, it offers two competing ways for query processing.

• We can (sequentially) merge multiple lists very much like merging traditional inverted lists by ids.

• Alternatively, we can also leverage the R-trees to browse the points of all relevant lists in ascending order of their distances to the query point.

Functional requirements are as follows:

• The developed system should be able to perform keyword-augmented nearest neighbour search in time.

• It should be incorporable in a commercial search engine that applies massive parallelism, implying its immediate industrial merits.

• It is straight forward to extend our compression scheme to any dimensional space.

V. CONCLUSIONS

This paper presents the survey of various techniques for nearest neighbor search for spatial database. As in the previous methods there were many drawbacks. The existing solutions incur too expensive space consumption or they are unable to give real time answer. The planned system has remedied the situation by developing an access methodology referred to as the abstraction Inverted index (SI-index). Not solely that the SI-index is fairly space economical, however additionally it's the flexibility to perform keyword-augmented nearest neighbor search in time that's at the order of dozens of milliseconds. Moreover, because the SI- index relies on the standard technology of inverted index, it's readily incorporable in a business search engine that applies huge similarity, implying its immediate industrial merits.

REFERENCES

[1] 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 5–16, 2002.
[2] 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 322–331, 1990.
[3] 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 431–440, 2002.
[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, pages 16–29, 2012.
[5] X. Cao, G. Cong, and C. S. Jensen. Retrieving top-k prestige-based relevant spatial web objects. PVLDB, 3(1):373– 384, 2010.
[6] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In Proc. of ACM Management of Data (SIG- MOD), pages 373–384, 2011.
[7] 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 30–39, 2004.
[8] 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.
[9] E. Chu, A. Baid, X. Chai, A. Doan, and J. Naughton. Combining keyword search and forms for ad hoc querying of databases. In Proc. of ACM Management of Data (SIGMOD), 2009.
[10] 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.
[11] C. Faloutsos and S. Christodoulakis. Signature files: An access method for documents and its analytical performance evaluation. ACM Trans- actions on Information Systems (TOIS), 2(4):267–288, 1984.
[12] 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.
[13] 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.
[14] G. R. Hjaltason and H. Samet. Distance browsing in spatial databases. ACM Transactions on Database Systems (TODS), 24(2):265–318, 1999.
[15] V. Hristidis and Y. Papakonstantinou. Discover: Keyword search in relational databases. In Proc. of Very Large Data Bases (VLDB), pages 670–681, 2002.
[16] I. Kamel and C. Faloutsos. Hilbert R-tree: An improved r-tree using fractals. In Proc. of Very Large Data Bases (VLDB), pages 500–509, 1994.
[17] Yufei Tao and Cheng Sheng, "Fast Nearest Neighbor Search with Keywords", IEEE transactions on knowledge and data engineering, VOL. 26, NO. 4, APRIL 2014.
[18] J. Lu, Y. Lu, and G. Cong, "Reverse Spatial and Textual k Nearest Neighbor Search," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 349-360, 2011.
[19] 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.

