Efficient Query Retrevial Using Group Nearest Neighbour Algorithm

V.S.Madhumathi, T.Sounder rajan
P.G. Scholar, Assistant Professor
Department of Computer Science and Engineering,
K.S.R Institute of Engineering &Technology
K.S.R kalvi nagar,
Tiruchengode, Namakkal, Tamil Nadu, India

Abstract: **The Location-aware keyword query returns ranked objects that are near a query location and that have textual descriptions that match query keywords.IR^2-tree is used for finding nearest neighbor. Compression scheme and object aggregation method based on spatial inverted index. Spatial Inverted Index has collections of spatial represent points are conferred with the collections of key word .Compression scheme reducing the space cost. Object aggregation is done by grouping the aggregative objects to retrieve the placement and textual dependent data. Priority level search is used to search the objects based on the priority given for the keywords depends on the user.This method has few drawbacks. To overcome this technique, Using Group nearest neighbor search algorithm group the object .In GNN algorithm it has two techniques such as Location based service and Euclidean Distance .Location based service used to retrieve the location using algorithm. Euclidean Distance are using the formula find the distance. When using these algorithms, its fully efficient and more accurate .The optimized groups of objects are retrieved.**

Keywords: **GNN, Spatial database, nearest neighbor search, spatial index, keyword search**

I. INTRODUCTION

Conservative spatial queries, such as range search and nearest neighbor retrieval, occupy only conditions on matter geometric property. Nowadays, a lot of modern applications name for new forms of queries that aim to find objects satisfying both a spatial predicate, and a predicate on their associated text. For example, in its place of taking into reflection all the restaurants, a nearest neighbor query would in its place ask for the bistro that is the closest among those whose menus contain "chicken, spaghetti, mutton" all at the same time. At present the best solution to such queries is based on the IR^2-tree and compression scheme and it has few deficiencies that seriously impact its efficiency. Compression scheme is motivated by develop a access technique called the spatial wrong way up index that extends the conventional overturned index to cope with multidimensional data, and comes with algorithms that can come back with nearest neighbor queries with keywords in real time.

Object aggregation is introduced and it is done by grouping the aggregative objects to retrieve the placement and textual dependent data. Priority level search is used to search the objects based on the priority given for the keywords depends on the user. After the aggregated group of objects is collected weight will be calculated for each group. Based on that the height priority group of objects are returned .Then consider the partially matching groups based on weight will be retrieved. The optimized group of objects is retrieved as a result. If the fully satisfied group of objects is not retrieved next the partially satisfied result will be retrieved as a optimized result. In GNN algorithm it has two techniques such as Location based service and Euclidean Distance .Location based service used to retrieve the location using algorithm. Euclidean Distance are using the formula find the distance. When using these algorithms, its fully efficient and more accurate .The optimized groups of objects are retrieved.

A.K Nearest neighbor

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor

kNN which is based on weights .The training points are assigned weights according to their distances from sample data point. But still, the computational complexity and memory requirements remain the main concern always. To overcome memory limitation, size of data set is reduced. For this, the repeated patterns, which do not add extra information, are eliminated from training samples. To further improve, the data points which do not affect the result are also eliminated from training data set . Besides the time and memory limitation, another point which should be taken care of, is the value of k, on the basis of which category of the unknown sample is determined. Gongde Guo selects the value of k using model based approach . The model proposed automatically selects the value of k. Similarly, many improvements are proposed to improve speed of classical kNN using concept of ranking , false neighbor information ,clustering .

B .Rank Nearest neighbor

Assign ranks to training data for each category. Performs better when there are too much variations between features ,.Robust as based on rank. It used to group the object then it aggregate the object .it assign rank to the each object ,Less computation complexity as compare to kNN .It used to find the location using rank.

C.Condensed Nearest Neighbor

The Condensed Nearest Neighbor (CNN) algorithm stores the patterns one by one and eliminates the duplicate ones. Hence, CNN removes the data points which do not add more information and show similarity with other training data set. It is improvement includes one more step that is elimination of the
patterns which are not affecting the training data set result. The another technique called Model Based kNN selects similarity measures and create a 'similarity matrix' from given training set. Then, in the same category, largest local neighbor is found that covers large number of neighbors and a data tuple is located with largest global neighborhood.

D. .Voronoi Network Nearest Neighbor

Voronoi Network Nearest Neighbor (VN3) approach, can handle sparse datasets but is inappropriate for medium and dense datasets due to its high precomputation and storage overhead. A new approach is proposed that indexes the network topology based on a novel network reduction technique. To reduce index complexity and hence avoid unnecessary network expansion, a novel technique called network reduction on road networks are proposed. This is achieved by replacing the network topology with a set of interconnected tree-based structures (called SPIE's) while preserving all the network distances. By building a lightweight and index on each SPIE, the (k)NN search on these structures simply follows a predetermined path, i.e., the tree path, and network expansion only occurs when the search crosses SPIE boundaries.

II. COMPARATIVE ANALYSIS

Algorithms used	Functions	Drawbacks
K Nearest Neighbor search	Uses nearest neighbor rule.	1. Biased by value of k 2.Computation Complexity 3.Memory limitation 4. Being a supervised learning lazy algorithm i.e. runs slowly 5. Easily fooled by irrelevant attribute.
Rank Nearest Neighbor Search	Assign ranks to training data for each category	1.Multivariate kRNN depends on distribution of the data
Condensed Nearest Neighbor	Eliminate data sets which show similarity and do not add extra Information	1. CNN is order dependent; it is unlikely to pick up points on boundary. 2. Computation Complexity
Voronoi Network Nearest Neighbor	Focus on nearest neighbor prototype of query point	1.Large number of computations
Group Nearest neighbor	GNN algorithm used to group the object. It give accurate object and time consuming. It used for large number of data	1. 1.More computation

III. EXISTING SYSTEM

The existing system reviews the Information Retrieval R-tree (IR^2-tree) and compression scheme, Object Aggregation which is the state of the art for answering the nearest neighbor queries. IR^2 trees return the query points even if it does not contain all the query keywords. Compression scheme is widely used to reduce the size of an inverted index with gap keeping approach. Object aggregation is done by grouping the aggregative objects to retrieve the placement and textual dependent data. Priority level search is used to search the objects based on the priority given for the keywords depends on the user**,** When using objeaggregation, it based on the priority it take more time.

A. IR^2-TREE

The grown-up system IR^2-Tree follow the two kinds of strategies

- R trees
- Signature files

The R tree strategy wants the more number of keywords to search the user specification. The autograph files are loading the extra number of text to match the object for user specification. Here the drawback of IR^2 trees where discussed and it has the advantages of both R trees and signature files. The IR^2 trees does not contain all the query keywords. It will direct the search to some objects those does not contain all keywords.

Autograph file in general refers to a hashing-based structure, whose instantiation is known as superimposed coding (SC), which is shown to be more effective than other instantiations. It is considered to make membership tests: establish whether a query word w exists in a set W of words. SC is traditional, in the wisdom that if it says "no", then w is definitely not in W. If, on the other hand, SC returns "yes", the true answer can be either method, in which container the whole W must be scanned to avoid a false hit.

B. COMPRESSION SCHEME

Compression eliminates the defect of a conventional index such that an SI-index consumes much less space. Compression is already widely used to reduce the dimension of an inverted index in the conservative context where each inverted list contains only ids. In that case, an efficient approach is to record the gaps between consecutive ids, as opposed to the precise ids. Gap-keeping will be much less beneficial if the integers of set S are not in a sorted order. This is for the reason that the space saving comes from the hope that gaps would be much smaller (than the original values) and hence could be represented with smaller amount bits. This would not be true had S not been sorted. Compressing an SI-index is less uncomplicated. The differentiation here is that each element of a list, a.k.a. a point p, is a triplet (idp, xp, yp), including both the id. As gap-keeping requires a sorted arrange, it can be valuable on only single attribute of the triplet.

C. OBJECT AGGREGATIONS

The object collection module is used to collect the relevant points as object. Here systems get the specified keyword from user and system assumes the user as the origin of spatial. It gets the four regions points simultaneously. And it collects each region nearest points and forms the groups. The weight is calculated by distance between the each point by finding the adjacent point. Based on the highest priority the group of points will be retrieved. It produce optimized group of result to the user. Retrieving a group of spatial web objects, where the keywords of the objects are aggregated and the result must match the query keywords. The group must be nearest to the query location and also the objects in a group are interrelated. calculates the partial keyword of the known keyword list. The keywords which matches less than the given keyword are displayed.

DISADVANTAGE

- Initial search depends only on object geometric properties
- Not efficient for searching large number of data
- When Grouping the objects based on priority it takes more time and low performance
- It has less accuracy

IV. PROPOSED SYSTEM

In Proposed system, using Group nearest neighbor search algorithm group the object .In GNN algorithm it has two techniques such as Location based service and Euclidean Distance .Location based service used to retrieve the location using algorithm. Euclidean Distance are using formula find the distance.

Euclidean distance $\text{dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$

A.Group nearest neighbour

GNN query as fit as its three detachment functions (sum, max, min) were earliest introduced .Sum is used to play down the total distance traveled by a group of users, while max (min) can warranty the most up-to-date (earliest) incoming time for a group of users. GNN can be determined in many LBS applications, such as storm monitor, forest fire suppression. Group Nearest Neighbor (GNN) query has recently gained much thought. A typical scenario of GNN query is to find a capability which minimizes the maximum (minimum or total) journey coldness for a group of users. This, in revolve, leads to the latest (earliest or total) time that a user (users) will arrive at the facility.

Given two sets of points P and Q, a group nearest neighbor (GNN) query retrieves the point(s) of P with the minimum sum of distances to all point in Q. Consider, for example, three users at locations q1, q2 and q3 that want to get a meeting point (e.g., a eating place); the corresponding query returns the data point p that minimizes the sum of Euclidean distances $|pqi|$ for $1 \le i \le 3$. Assuming that Q fits in memory and P is indexed by an R-tree, we advise several algorithms for finding the group nearest neighbors professionally. As a second step, we broaden our techniques for situations where Q cannot in shape in memory, layer both indexed and non-indexed query points. An investigational assessment identifies the best alternative based on the data and query properties.

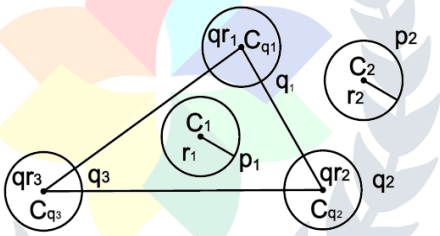
In many LBS scenarios, location information becomes uncertain, especially, when privacy concerns, sampling

precisions, and network transmission delays are taken into consideration. Previous works of GNN query mainly focuses on the scenarios when data objects (P) are uncertain. However, very small work has done to the scenario when query objects (Q) are also uncertain.

$$\alpha = \int_{r_{\min}}^{r_{\max}} (\Pr\{adist(o,Q) = r\} \cdot \prod_{\forall p \in P \setminus \{o\}} \Pr\{adist(p,Q) \geq r\}) dr$$

Symbols	Descriptions
P	The data object set with size $\|P\|$.
Q	The query object set with size n.
d	The dimensionality of the data object set.
UR(o)	The uncertain region of object o.
MBR(o)	The Minimum Bounding Rectangle (MBR) of object o.
dist(. , .)	The Euclidean distance between two objects.
adist(o, Q)	The aggregate distance from object o to query object set Q.
LB_adist(o, Q) (UB_adist(o, Q))	The lower (upper) bound of aggregate distance adist(o, Q).
α	The confidence of result.

EXAMPLE QUERY



ADVANTAGE

- It efficient for large amount for large amount data.
- Using GNN it easy to group the object
- Here there is no priority used
- It is time consuming and accuracy data

IV CONCLUSION

In existing work it follows compression scheme and it does not consider partial keyword match and priority level search. Object aggregation is capable of taking into explanation both text relevancy and place closeness used for finding group of objects based on the user's priority level. The retrieval of data is in the order. False hits will be reduced by the priority given for the keywords. The consumer specification not only depends on the object geometric properties but also the associated text and user's priority level. Partially satisfied results will be obtained if fully satisfied results are not present. Thus the nearest groups of objects are searched effectively.

References

[1]. Agrawal, Chaudhuri, S. and Das,'Dbxplorer: A system for keyword-based search over relational databases' , International Conference on information Engineering (ICDE), pages 5–16.

[2]. Anandhi R J, Natarajan and Subramanyam (2009) ' Efficient Consensus Function for Spatial Cluster Ensembles: An Heuristic encrusted Approach', International Symposium on compute, message, and Control (ISCCC).

[3]. Bhalotia,A. Nakhe, C. Chakrabarti, S. and Sudarshan, S. (2002) 'Keyword searching and browsing in databases using banks', International Conference on Data Engineering (ICDE), pages 431–440.

[4]. Cao, Chen, Cong, Jensen, Skovsgaard, A. and Wu, D. and Yiu, M. L (2012) 'Spatial keyword querying', In ER, pages 16–29. Cao,Cong, G. and Jensen, C. S. (2010) 'Retrieving top-k prestige-based relevant spatial mesh objects', PVLDB, 3(1):373–384

[5]. Cao,Cong, G. and Jensen, C. S. (2010) 'Retrieving top-k prestige-based relevant spatial net objects', PVLDB :373–384.

[6]. Cao,Cong, G. Jensen, C. S. and Ooi, B. C. (2011) 'Collective spatial keyword querying',ACM Management of Data (SIGMOD), pages 373–384.

[7]. Chazelle, Rubinfeld, R. and Tal, A.(2004) 'The bloomier cleaner: an professional data structure for static support find tables', Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 30–39.

[8]. Chen, and Markowetz, A.(2006) 'Efficient query processing in geographic web search engines', In Proc. of ACM Management of Data (SIGMOD), pages 277–288.

[9]. Chu, Baid A. and Naughton, J.(2009) 'Combining keyword search and forms for ad hoc querying of databases', In Proc. of ACM Management of Data (SIGMOD).Cong, Jensen, C. S. and Wu, D.(2009) 'Efficient retrieval of the top-k most relevant spatial network objects' , PVLDB, :337–348.