ISSN : 2231-0711

V.S.Madhumathi et al | International Journal of Computer Science Engineering and Technology(IJCSET) | October 2014 | Vol 4, Issue 10,263-265

Review on: Nearest Neighbor Search with Keywords

V.S.Madhumathi[1], T.Sounder rajan[2]

[1]*P.G. Scholar,* [2]*Assistant Professor*
Department of Computer Science and Engineering,
K.S.R Institute of Engineering &Technology K.S.R kalvi nagar,
Tiruchengode, Namakkal Coimbatore-641035, Tamil Nadu, India
madhumathisowdamuthu@gmail.com

***Abstract* - In data mining, we have many nearest neighbor search algorithm .the nearest neighbor (NN) algorithm is very easy, highly well-organized and successful in the field of pattern recognition, text categorization, object recognition, location etc. Nearest Neighbor (NN) overcome the memory requirement and calculation difficulty. This paper discusses nearest neighbor (NN) methods for the analysis of spatial point patterns and the analysis of field experiments. The distance from a point to its NN, or correlated quantities, can be used to test whether the observed locations are a realization of a specific spatial point process. This paper provides the comparison of nearest neighbor search.**

Keywords - Nearest neighbor(NN), K nearest neighbor(KNN), Condensed nearest neighbor(CNN), Rank nearest neighbor(RNN)

I. INTRODUCTION

A spatial database manages multidimensional objects (such as points, rectangles, etc.), and provides fast access to those objects based on different selection criteria. The importance of spatial databases is reflected by the convenience of modeling entities of reality in a geometric manner. For example, locations of restaurants, hotels, hospitals and so on are often represented as points in a map, while larger extents such as parks, lakes, and landscapes often as a combination of rectangles. Many functionalities of a spatial database are useful in various ways in specific contexts. For instance, in a geography information system, range search can be deployed to find all restaurants in a certain area, while nearest neighbor retrieval can discover the restaurant closest to a given address. To find a spatial location using nearest neighbor algorithm. it use different algorithm such as KNN,RNN,CNN,VN3,GNN.

GIS databases contain vital location based information and play an important role in many applications such as disaster management, national infrastructure protection, crime analysis, etc. Such information has played a significant role during national disasters like 9/11 or more recent Hurricane Katrina, in saving lives, resources and properties. GIS databases accessible at one place underline the urgent need to make such location-based information readily available to a large number of organizations who deal with such data on a day-to-day basis.

Location-based information contained in publicly available GIS databases is invaluable for many applications such as disaster response, national infrastructure protection, crime analysis, and numerous others. The information entities of such databases have both spatial and textual descriptions. Likewise, queries issued to the databases also contain spatial and textual components, for example, “Find shelters with emergency medical facilities in Orange County,” or “Find earthquake-prone zones in Southern California.” Such queries are referred as spatial-keyword queries or SK queries for short. In recent times, a lot of interest has been generated in efficient processing of SK queries for a variety of applications from Web-search to GIS decision support systems. The systems built for enabling such applications as Geographic Information Retrieval (GIR) Systems. An example GIR system that address in this paper is a search engine built on top of hundreds of thousands of publicly available GIS databases.

II. RELATED WORK

Different studies are perform for finding the spatial database ,it based on text and location retrieval. It use different nearest neighbor search algorithm with keywords. It retrieve the location and text.

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.

E. Group Nearest Neighbor

Given two sets of points P and Q, a group nearest neighbor(GNN) query retrieves the point(s) of P with the smallest sum of distances to all points in Q. Consider, for instance, three users at locations q_1, q_2 and q_3 that want to find a meeting point (e.g., a restaurant); the corresponding query returns the data point p that minimizes the sum of Euclidean distances $|pq_i|$ for $1 \delta i \delta 3$. Assuming that Q fits in memory and P is indexed by an R-tree, we propose several algorithms for finding the group nearest neighbors efficiently. As a second step, we extend our techniques for situations where Q cannot fit in memory, covering both indexed and non-indexed query points. An experimental evaluation identifies the best alternative based on the data and query properties.

III. 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

IV. CONCLUSION

This paper presents an extensive survey on spatial database using nearest neighbor algorithm. ,many new approaches are proposed in the field of spatial database. many research issues have been highlighted and direction for future work has been suggested. many open issues have been highlighted by the researchers such s dealing with spatial database by different techniques future work will done

REFERENCES

[1]. Agrawal, Chaudhuri, S. and Das, G. (2002) 'Dbxplorer: A system for keyword-based search over relational databases' , In Proc. of International Conference on Data Engineering (ICDE), pages 5–16.

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

[3]. Bhalotia, Hulgeri, A. Nakhe, C. Chakrabarti, S. and Sudarshan, S. (2002) 'Keyword searching and browsing in databases using banks', In Proc. of 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 web objects', PVLDB, 3(1):373–384

[5]. Cao,Cong, G. and Jensen, C. S. (2010) 'Retrieving top-k prestige-based relevant spatial web objects', PVLDB, 3(1):373–384.

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

[7]. Chazelle, Kilian, Rubinfeld, R. and Tal, A.(2004) '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.

[8]. Chen, Suel, T. 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, Chai, Doan, A. and Naughton, J.(2009) 'Combining keyword search and forms for ad hoc querying of databases', In Proc. of ACM Management of Data (SIGMOD).
[10]. Cong, Jensen, C. S. and Wu, D.(2009) 'Efficient retrieval of the top-k most relevant spatial web objects' , PVLDB, 2(1):337–348.
[11]. Debing Zhang, Genmao Yang, Hu, Zhongming Jin ,Deng Cai , Xiaofei He (2013), 'A Unified Approximate Nearest Neighbor Search Scheme by Combining Data Structure and Hashing' , Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.
[12]. Hariharan, Hor, Li, C. and Mehrotra, S.(2007) 'Processing spatial keyword.(SK) queries in geographic information retrieval (GIR) systems', In Proc. of Scientific and Statistical Database Management (SSDBM).