GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

SURVEY ON EFFICIENT NEAREST NEIGHBOR SEARCH IN DATA MINING

A.Sri Vidhya M.E.(PG Scholar)*1 and Prof. M. Ashwin2

*1,2Dept of Computer science and engineering, Adhiyamaan College of engineering Hosur,India

ABSTRACT

Spatial query which focus only on the geometrics properties of an object like points, rectangle etc. Now a day's many new applications which involve the queries that completely aim to return an object which satisfy equally on spatial predicate and their associated text. Spatial query takes the given location and a keyword as the input and finds the object that matches the both spatial predicate and the text related to the given query. Some of the spatial queries are range search and nearest neighbor retrieval which includes only geometric properties of an object. This paper is about the related work involved for the efficient nearest neighbor search.

Keywords- *Spatial Query, Nearest Neighbor Search, IR2-Tree, Range Search, Spatial Predicate.*

I. INTRODUCTION

Nearest neighbor search (NNS), also known as closest point search. [16]It deals with an optimization drawback for returning closest (or most similar) points. Nearest neighbor search which locates the closest neighbor of a query point in a set of points, which is a crucial and wide studied drawback in several fields, and it has a wide selection of applications. Here we can search for the nearest point by giving keywords as input; it can be spatial or textual. A spatial database is to manage multidimensional objects i.e. points, rectangles, etc. Some spatial databases handle more complicated structures like 3D objects, topological coverage's.

Nearest Neighbor could be a technique applicable for classification models. Not like alternative algorithms, the work data is not scanned or processed to make the model. Instead, the work data is that the model. Once a innovative case or instance is presented to the model, the algorithmic rule look within the slightest points the information to go looking out a collection of cases that are most almost like it and uses them to predict the result. There are two principal drivers inside the k-NN algorithm: the amount of nearest cases to be used (k) and a metric to measure what's meant by nearest.

Every use of the k-NN algorithmic rule desires that we tend to specify a positive range value for k.[16] This determines what number existing cases are verified once predicting a innovative case. k-NN refers to a family of algorithms that we tend to could denote as 1-NN, 2-NN, and 3- NN, so forth. As an example, 4-NN indicates that the algorithmic rule will use the four nearest cases to predict the result of a innovative case. K-NN is based on a plan of distance, and this desires a metric to figure out distances.

II. NEAREST NEIGHBOR TECHNIQUES

Nearest neighbor techniques are classified into two types.

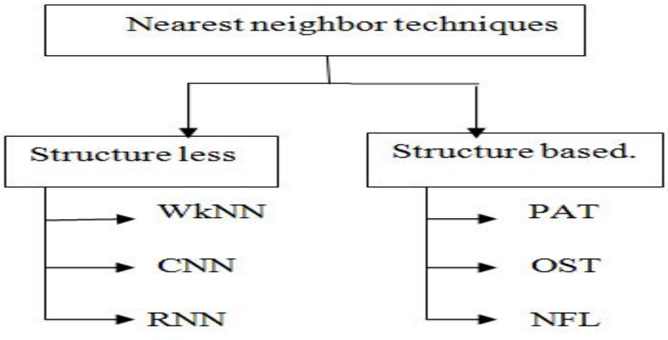

Fig.1 Types of NN techniques

Structure less NN techniques

The first type of NN technique is structure less. This structure less NN techniques includes much technique in that they are WkNN, CNN, RNN. [16]The k-nearest neighbor is the first category in which the complete data is classified into training data and sample data point. Distance is computed from all training points to sample point and the point with short distance is called nearest neighbor. This technique is very easy to implement but value of k affects the output in some cases. Bailey names his algorithm as weighted kNN because it uses weights with classical kNN and gives algorithm named weighted kNN (WkNN)[11]. The Condensed Nearest Neighbor (CNN) [6],[2],[1]algorithm saves the patterns one by one and removes the repeated ones. So, CNN removes the data points which do not add more information and show resemblance with other training data set. The Reduced Nearest Neighbor (RNN) [10] is better over CNN because it includes an additional step that is removing the patterns which are not affecting the training data set output.

Structure based NN techniques

The second type of nearest neighbor techniques is structures based technique. This structure based NN techniques includes many technique in that they are Ball Tree, k-d Tree, principal axis Tree (PAT), orthogonal structure Tree (OST), Nearest feature line (NFL), etc[16]. Ting Liu introduces the new concept called Ball Tree. Principal Axis Tree nearest Neighbor (PAT) [15] takes the advantages like Good performance, Fast Search. It suffers at the Computational time. Orthogonal Search Tree Nearest Neighbor (OST) [14] takes the advantages like Less Computation time, Effective for large data sets. The major drawback is Query time is large. Nearest feature Line Neighbor (NFL)[13] it takes the advantages like Improve classification accuracy, Highly effective for small size,utilises information ignored in nearest neighbor i.e. templates per class. Its takes the disadvantages like Fail when prototype in NFL is far away from query point, Computations Complexity, To describe features points by straight line is hard task.

III. LITERATURE REVIEW

IR^2-TREE

Information Retrieval R-tree which is the state of answering the NN Queries.IR^{2-}tree is the combination of the R-tree with signature files. First we will see about Signature file. Signature file refers to a hash-based framework, whose instantiation in [9] is called as superimposed coding (SC), which is shown to be more effective than other instantiation [5]. It is charted to perform membership tests that is to determine whether a query word q exists in a set of Words W.If it returns "no", then q is surely not in W. If SC returns "yes", then q is in W, to avoid a false hit W is scanned as a whole if it returns "yes". The IR^2-tree is an R-tree where each leaf or non leaf entry E is augmented with a signature which summarizes the union of the texts in the sub tree. On traditional R-trees, the best-first algorithm [12] is a well-known solution to NN search. It is now directly to adapt it to IR^2-trees.

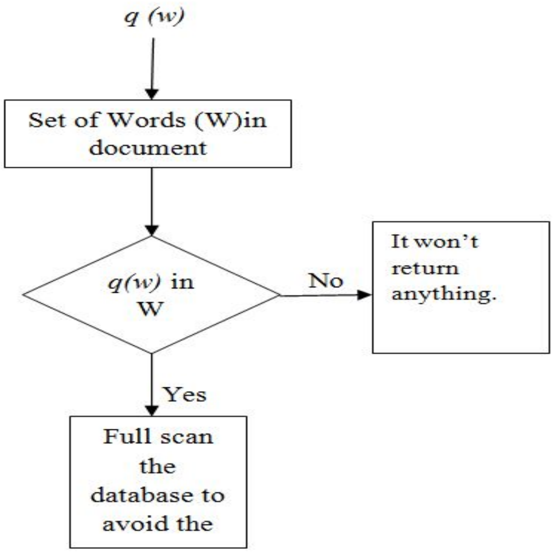
q (w)
Set of Words (W)in document
q(w) in W
No
It won't return anything.
Yes
Full scan the database to avoid the

Fig.2 Working of SC

G. Cong, C.S. Jensen, and D. Wu . [7], Keyword based nearest neighbor queries which relates the concept incense that how object text plays a role in finding the query result. This paper aims at IR flavor i.e. Information Retrieval. It computes the relevance between the documents of an object 'O' and a query 'q'. In order to calculate the similarity between the object 'O' and a query 'q', the relevance result is added with the Euclidean distance between the object 'O' and a query 'q'. As the result, the object with highest similarity is returned. The drawback of this relevance computing is sometimes it may return the objects that are not in the query keyword but still the returned object has the high similarity. The below diagram is about the query processing of reference [7].

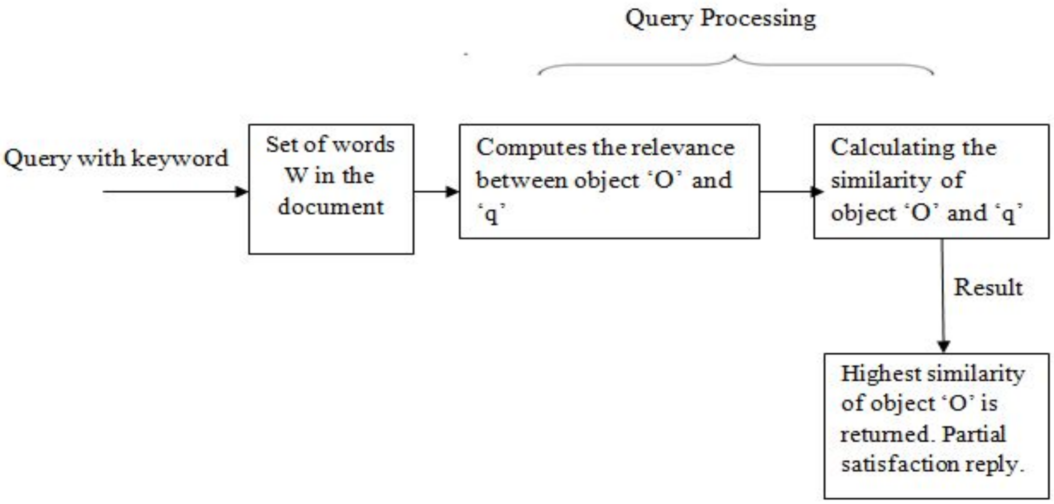
Query Processing
Query with keyword
Set of words W in the document
Computes the relevance between object 'O' and 'q'
Calculating the similarity of object 'O' and 'q'
Result
Highest similarity of object 'O' is returned. Partial satisfaction reply.

Fig.3 Query Processing diagram

I.D. Felipe, V. Hristidis, and N. Rishe, [9] In this paper they evaluates in the basics of Boolean predicate. It overcomes the drawback that is described in[7] paper that is if any of query keyword is missing in the document, it must not return it. In this paper it overcomes the drawback by not allowing partial satisfaction reply. The result must be if it exists with all specified 'q' keyword, it must return or else it must not return The below diagram is about the query processing of reference [9].

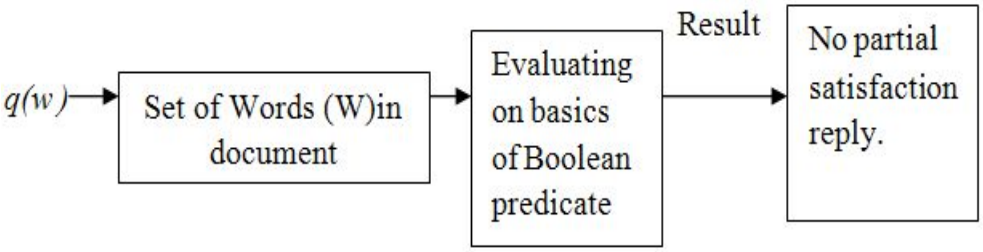
q(w)
Set of Words (W)in document
Evaluating on basics of Boolean predicate
Result
No partial satisfaction reply.

Fig.4 Query Processing diagram

Y.-Y. Chen, T. Suel, and A. Markowetz, [8] This author specify Geographic web search. Each web page is assigned with geographic region that is enabled to the web page contents. Usually in web search, each web page is returned to the user query on the basics of ranking each webpage. The higher ranked web page is return first on the basics of user query. Here there is the drawback that is underpinning problem which is overcome by combination of keyword search and range queries.

TABLE 1 Content of Previous Work in Table

AUTHOR	PAPER TITLE	CONCEPT	APPROACHES
G. Cong, C.S. Jensen, and D. Wu[7]	Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects	Keyword based nearest neighbor Query	Computes the relevance between the object p and query q.It returns Partial satisfaction.
Y.-Y. Chen, T. Suel, and A. Markowetz[8] R. Hariharan, B. Hore, C. Li, and S. Mehrotra, [11]	Efficient Query Processing in Geographic Web Search Engines.[8] Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems[11]	Combines keyword search and range queries	Geographic web search returns based on ranking
I.D. Felipe, V. Hristidis, and N. Rishe,[9]	Keyword Search on Spatial Databases[9]	Nearest neighbor search using keywords by IR^2-Tree	IR^2-Tree preserves object's spatial proximity, which solve the spatial query efficiently. Here there is no partial satisfaction in returning the output.
D. Zhang, Y.M. Chee, A. Mondal, A.K.H. Tung, and M. Kitsuregawa,[17]	Keyword Search in Spatial Databases: Towards Searching by Document	Search for m-closest keyword	Collaborative in nature, resulting m points should cover the query keyword together.
X. Cao, G. Cong, C.S. Jensen, and B.C. Ooi,[5]	Collective Spatial Keyword Querying	Collective Spatial keyword querying	Similar idea as the above but it mainly focuses at optimizing different objective function.
X. Cao, G. Cong, and C.S. Jensen, [4]	Retrieving Top-k Prestige-Based Relevant Spatial Web Objects	Prestige based spatial keyword	Evaluates the similarity of an object p to a query q by considering the objects p neighbors.

IV. CONCLUSION

This paper describes nearest neighbor techniques, its classification, literature review and other related works involved. We have a table that describes the Content of Previous Work of the efficient search of nearest neighbor techniques that includes the concept and approaches for each paper. Some of nearest neighbor techniques are structure less and some are structured base. This nearest neighbor technique which improves over basic kNN techniques.

REFERENCES

1. *E.Alpaydin, "Voting Over Multiple Condensed Nearest Neighbors", Artificial Intelligent Review 11:115-132, 1997.*
2. *F. Angiulli, "Fast Condensed Nearest Neighbor", ACM International Conference Proceedings, Vol 119, pp 25-32.*
3. *T.Bailey and A. K. Jain, "A note on Distance weighted k-nearest neighbor rules", IEEE Trans. Systems, Man Cybernatics, Vol.8, pp 311- 313, 1978.*
4. *X.Cao, G. Cong, and C.S. Jensen, "Retrieving Top-k Prestige- Based Relevant Spatial Web Objects," Proc. VLDB Endowment, vol. 3, no. 1, pp. 373-384, 2010.*
5. *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.*

6. K.Chidananda and G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighbor", IEEE Trans. Information Theory, Vol IT- 25 pp. 488-490, 1979.
7. G.Cong, C.S. Jensen, and D. Wu, "Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects," PVLDB, vol. 2, no. 1, pp. 337- 348, 2009.
8. 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.
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. Geoffrey W. Gates, "Reduced Nearest Neighbor Rule", IEEE Trans Information Theory, Vol. 18 No. 3, pp 431-433.
11. 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.
12. G.R.Hjaltason and H. Samet, "Distance Browsing in Spatial Databases," ACM Trans. Database Systems, vol. 24, no. 2, pp. 265-318,1999.
13. S. Z Li, K. L. Chan, "Performance Evaluation of The NFL Method in Image Classification and Retrieval", IEEE Trans On Pattern Analysis and Machine Intelligence, Vol 22, 2000.
14. Y. C.Liaw, M. L. Leou, "Fast Exact k Nearest Neighbor Search using Orthogonal Search Tree", Pattern Recognition 43 No. 6, pp 2351-2358.
15. J.Mcname, "Fast Nearest Neighbor Algorithm based on Principal Axis Search Tree", IEEE Trans on Pattern Analysis and Machine Intelligence, Vol 23, pp 964-976.
16. Nitin Bhatia, Vandana" Survey of Nearest Neighbor Techniques" (IJCSIS) International Journal of Computer Science and Information Security,Vol. 8, No. 2, 2010
17. 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.