Efficient Query Service Provider using Clustering K-Nearest Neighborhood Algorithm

A. Roslin Deepa, Dr. Ramalingam Sugumar

1Ph.D Scholar, Professor
Christhu Raj College, Bharathidasan University, Trichy, Tamilnadu - India.
Christhu Raj College, Trichy, Tamilnadu - India.

Abstract -
Data mining has wide variety of real time application in many fields such as financial, telecommunication, biological, and among government agencies. So the query processing system is also an important thing to access and search the database. Once the KNN query service is outsourced, data confidentiality and query privacy become the important issues, because the data owner loses the control over the data. This type of queries are very useful in many applications namely decision making, data mining and pattern recognition. In this paper we studies a KNN based search which was worked on very efficiently. This method utilize a conventional data-partitioning index on the dataset, employ the state-of-the-art database techniques including k nearest neighbor (KNN) retrieval and reverse KNN search technique using Clustering find the minimum value and calculate the average using k-means algorithm. The empirical study of this paper is also providing the efficiency of the KNN based query processing on spatial databases.

Keywords -- Data mining, K-Nearest Neighborhood, Query processing, Range Query, Spatial Data base.

1. INTRODUCTION
The problem of K nearest neighbor (KNN) queries in spatial databases is the main type of query which is frequently used in Geographical Information Systems. It is defined as: given a set of spatial objects and a query point, find the K closest objects to the query (2). An example of KNN query is a query initiated by a GPS device in a vehicle to find the five near service center to the vehicle. The final result of this query is a set of intervals, or split points, and their associated KNNs. The split points also specify where on the path the KNNs of a moving object will change, and the intervals specify the appropriate locations that the KNNs of a moving object remains the same. The challenge in this type of query is to efficiently specify the location and the KNNs of the split points (2).

The existing work on KNN queries and its variations are intended at Euclidian spaces, where the bath between two objects is the straight line connecting them. These approaches are basically depended on utilizing index structures. In Spatial data mining it differs from regular data mining in parallel with the differences between non-spatial data and spatial data (7). The attributes of a spatial object stored in a database may be affected by the attributes of the spatial neighbors of that object. In addition, spatial location, and implicit information about the location of an object, may be exactly the information that can be extracted through spatial data mining.
In efficient query service provider area, clustering is as applied to large datasets, is the process of creating a group of objects organized on some similarity among the members. In spatial data sets, clustering permits a generalization of the spatial component that allows for successful data mining (3). Basically, a number of various methods are available for performing clustering, in particular it can be classified as (3), partitional clustering, Hierarchical clustering and locality based clustering. Partitional clustering develops a partition of the data such that objects in a cluster are more similar to each other than they are to objects in other clusters. The k-means and k-medoid methods are forms of partitional clustering. Hierarchical clustering performs a sequence of partitioning operations. Hierarchical clustering is frequently used in document and text analysis. Locality-based clustering algorithms group objects based on local relationships, and therefore the entire database can be scanned in one pass. Some locality based algorithms are density based, while others assume a random distribution.

The rest of the paper is structure as follows, section 1 provide the basic information and importance of data mining and clustering methods. Section 2 embraces the various existing papers which are based on clustering and query processing solutions. It is followed by section 3 includes the KNN algorithm details with its simple example and spatial data processing methods. The next section 4 contains the location based query processing techniques and its enhanced clustering schemes. Finally section 5 brings to a close with conclusion of the efficient query service provider mechanism using KNN clustering process.

2. RELATED WORK
In 2013, Ali Khoshgozaran, Houtan Shirani-Mehr et al. (1) propose a fundamental approach to perform the class of Range and Nearest Neighbor (NN) queries, the core class of spatial queries used in location-based services, without revealing any location information about the query in order to preserve users’ private location information. The authors propose a new approach is to utilize the power of one-way transformations to map the space of all objects and queries to another space and resolve spatial queries blindly in the transformed space. They also discussed the problem of location privacy and introduced a new way of blindly evaluating spatial queries by using one way space transformations to map objects and query points into an unknown space and addressing the query in this new space.

In 2013, Chang-Sup Park and Sungchae Lim (3) aims to propose an effective method to process keyword-based queries over graph-structured databases which are widely used in various applications such as XML, semantic web, and social network services. They proposes an extended answer structure for keyword queries, inverted list indexes on keywords and nodes, and query processing algorithms exploiting the inverted lists. The study of this paper aims to provide more effective and relevant answers to a given query than the previous approaches in an efficient way. The authors also proposed a query processing method for keyword search on graph databases based on a new answer structure. It selects keyword nodes using a new relevance metric for the root node to the keyword nodes containing a query keyword.

In 2009, Gao Cong Christian and S. Jensen et al. (5) proposes a new indexing framework for location aware top-k text retrieval. The framework leverages the inverted file for text retrieval and the R-tree for spatial proximity querying. Several indexing approaches are explored within the framework. The
framework encompasses algorithms that utilize the proposed indexes for computing the top-\(k\) query, thus taking into accounts both text relevancy and location proximity to prune the search space. The authors also provide the results of empirical studies with an implementation of the framework demonstrate that the paper’s proposal offers scalability and is capable of excellent performance.

In 2007, Panos Kalnis and Gabriel Ghinita et al. (8) present a framework for preventing location-based identity inference of users who issue spatial queries to Location-Based Services. The authors propose transformations based on the well-established K-anonymity concept to compute exact answers for range and nearest neighbor search, without revealing the query source. In this method they optimize the entire process of anonymizing the requests and processing the transformed spatial queries. Extensive experimental studies in this paper suggest that the proposed techniques are applicable to real-life scenarios with numerous mobile users.

3. KNN ALGORITHM

K nearest neighbor classifier algorithm, which stores all training tuples and classifies new tuples based on a distance metric (e.g., Euclidean distance functions). It is mainly used in a transformer fault database in which the data items are grouped into different similar classes to classify the new similarity metric data point (9). It predicts to determine the tuples with missing values and redundant values in accurately. Classes that are redundant that must be discarded for improved performance. The KNN algorithm (5) is used to fills missing values from stored observation tuples in same data set. This method searches the K nearest neighbor of the case with missing values and replaces by the mean or mode value if the value of a given fault symptoms value is missing in tuple t1, we assume previous stored values appeared in the dataset to be normalized. It can easily deal with multiple missing values which mainly used to produce complete dataset for efficient diagnosis new faults (9).

k-Nearest neighbors is a classification algorithm (8). It doesn’t create any classification model instead it just stores the labeled data. It looks at the k-closest labeled points for example k-nearest neighbors. Using neighbor’s classes it classifies the new data. For continuous data, KNN uses Euclidean distance (4). For Discrete data it first transform into continuous data and uses hamming distance to find the closest neighbor. It is expensive when applying on large dataset. It cannot be applicable for noisy data. Finding distance between 2 neighbors is difficult for larger range of values. It requires greater storage. Here the computation of K-nearest neighbors KNN algorithm steps is as follows,

- **Step 1:** Determine parameter K = number of nearest neighbors
- **Step 2:** Calculate the distance between the query-instance and all the training samples
- **Step 3:** Sort the distance and determine nearest neighbors based on the K-th minimum distance
- **Step 4:** Gather the category Y of the nearest neighbors
- **Step 5:** Use simple majority of the category of nearest neighbors as the prediction value of the query instance.
Example;
The example data is from a survey which has two attributes named by acid durability and strength to classify whether a special paper tissue is good or not. Here are the four training samples,

<table>
<thead>
<tr>
<th>A (Acid Durability)</th>
<th>B (Strength)</th>
<th>Z (Classification)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seconds</td>
<td>Kg/square meter</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>Bad</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>Bad</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>Good</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>Good</td>
</tr>
</tbody>
</table>

Apply the computation steps of K-nearest neighbors KNN algorithm,
Apply steps 1 to 5,
1. Determine parameter K = number of nearest neighbors
Suppose we use K = 3,
2. Calculate the distance between the query-instance and all the training samples
Coordinate of query instance is (3, 7) instead of calculating the distance we compute square distance which is faster to calculate (without square root)

<table>
<thead>
<tr>
<th>A (Acid Durability)</th>
<th>B (Strength)</th>
<th>Square Distance to query instance (3, 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seconds</td>
<td>Kg/square meter</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>(7-3)^2 + (7-7)^2 = 16</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>(7-3)^2 + (4-7)^2 = 25</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>(3-3)^2 + (4-7)^2 = 9</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>(1-3)^2 + (4-7)^2 = 13</td>
</tr>
</tbody>
</table>

3. Sort the distance and determine nearest neighbors based on the K-th minimum distance

<table>
<thead>
<tr>
<th>A (Acid Durability)</th>
<th>B (Strength)</th>
<th>Square Distance to query instance (3, 7)</th>
<th>Rank Minimum Distance</th>
<th>Is it included in 3-Nearest neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seconds</td>
<td>Kg/square meter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>(7-3)^2 + (7-7)^2 = 16</td>
<td>3</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>(7-3)^2 + (4-7)^2 = 25</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>(3-3)^2 + (4-7)^2 = 9</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>(1-3)^2 + (4-7)^2 = 13</td>
<td>2</td>
<td>Yes</td>
</tr>
</tbody>
</table>
4. Gather the category \( Y \) of the nearest neighbors. Notice in the second row last column that the category of nearest neighbor \( (Y) \) is not included because the rank of this data is more than 3 (=K)

<table>
<thead>
<tr>
<th>( A ) (Acid Durability)</th>
<th>( B ) (Strength)</th>
<th>Square Distance to query instance ((3, 7))</th>
<th>Rank Minimum Distance</th>
<th>Is it included in 3-Nearest neighbors</th>
<th>( Y = ) Category of nearest Neighbor</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
<td>((7-3)^2 + (7-7)^2 = 16)</td>
<td>3</td>
<td>Yes</td>
<td>Bad</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>((7-3)^2 + (4-7)^2 = 25)</td>
<td>4</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>((3-3)^2 + (4-7)^2 = 9)</td>
<td>1</td>
<td>Yes</td>
<td>Good</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>((1-3)^2 + (4-7)^2 = 13)</td>
<td>2</td>
<td>Yes</td>
<td>Good</td>
</tr>
</tbody>
</table>

5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance.
We have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with \( A = 3 \) and \( B = 7 \) is included in good category.

3.1 Spatial Clustering
Clustering, as applied to large datasets, is the process of creating a group of objects organized on some similarity among the members. Although there are similarities between spatial and non-spatial clustering, large databases, and spatial databases in particular, have unique requirements that create special needs for clustering algorithms (10).

- An obvious need considering the large amount of data to be handled is that algorithms be efficient and scalable.
- Algorithms need to be able to identify irregular shapes, including those with lacunae or concave sections and nested shapes.
- The clustering mechanism should be insensitive to large amounts of noise.
- Algorithms should not be sensitive to the order of input. That is, clustering results should be independent of data order.
- Algorithms should handle data with large numbers of features, that is, higher dimensionality.

Huge amounts of data have been collected through the advances in data collection, database technologies and data collection techniques (6). This explosive growth of data creates the necessity of automated knowledge/information discovery from data, which leads to a promising and emerging field, called data mining or knowledge discovery in databases. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial databases. KDD follows several stages including data selection, data preprocessing, information extraction or spatial data mining, interpretation and reporting (7). Data mining is a core component of the KDD process. Spatial data mining techniques are divided into four general groups: spatial association rules, spatial clustering, spatial trend detection and spatial classification.
4. LOCATION BASED QUERY PROCESSING

In this paper, we have discussed a query processing service provider using KNN algorithm. It consists of two types of location based query processing namely, K-anonymized relational databases and K-anonymized location based services (8).

K-anonymized relational databases

Anonymity was first discussed in relational databases, where published data (for example, census or medical) should not be linked to specific persons. Two techniques are used for transforming a relation to a K-anonymized one: suppression, where some of the attributes or tuples are removed, and generalization, which involves replacing specific values (for example, the phone number) with more general ones (for example, only the area code). Both methods lead to information loss. Algorithms for anonymizing an entire relation while preserving as much information as possible K-anonymity and diversity problems.

K-anonymized location based services

In K-anonymity location based services it using the framework the user sends his position, query, and K to the anonymizer, which removes the ID of the user and transforms his location through cloaking. The first cloaking technique, called Interval Cloak (IC) is based on quadtrees. A quadtree recursively partitions the space into quadrants until the points in each quadrant fit in a page/node. It completely shows the space partitioning and a simple quadtree, assuming that a node contains a single point. The anonymizer maintains a quadtree with the locations of all users.

4.1 Cluster enhancing Method

The cluster enhanced method is the hybrid indexing framework with clustering, which makes it possible to estimate tighter bounds at each tree node, thus improving query performance. Spatial web objects often belong to different categories. For example, the geo-referenced points of interest may belong to specific categories, such as retail, accommodations, restaurants, and tourist attractions. Points of interest from different such categories may appear in the same node of the hybrid index, and thus two objects in the same node can be very different in terms of their document similarities. The idea is to cluster objects into groups according to their corresponding documents. Each index node may then contain objects from different clusters. Instead of constructing a single pseudo document for each node, we construct a pseudo document for each cluster in each node (4).

Algorithm of clustering enhanced method

1. Retrieving initial samples of services
2. Eliminating irrelevant services to form a working dataset
3. Applying KNN algorithm to the dataset
4. Semantic matching query on spatial databases in related clustered group
5. if the results match the query then goto step 8
6. else choosing next cluster goto step 4
7. end if
8. end

5. CONCLUSION

The query processing system is the necessary aspect to access and search the database. So clustering method is important to get the appropriate results on the query searching. Whenever the user used query services it also contains some confidentiality and privacy issues due to some outsourcing of services. In this paper, we studies a KNN based search on spatial databases which were worked on very efficiently. This paper include the spatial clustering techniques utilize a conventional data-partitioning index on the dataset by using K-Nearest Neighborhood algorithm. In KNN query processing the location based services
are also discussed in this paper. The study of cluster enhancing methods in this paper is used to provide the simple and efficient way of clustering to provide the efficient query service provider using KNN algorithm.

References