ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

A Review on Spatial Approximate String Search in Road Networks using RSASSOL Algorithm

Sunil Shelke[1], G. P. Chakote[2]

Scholar, Department of Computer Science Engineering, M.S.S. College, Jalna, India[1]

Head, Department of Computer Science Engineering, M.S.S. College, Jalna, India[2]

Abstract: The related work deals with the fast nearest string search in large spatial databases. Specifically, it investigates spatial range queries extent with a string similarity nearest search predicate in both Euclidean distance and road networks.. The spatial approximate string (SAS) query search problem is common in searching the approximate string in large spatial database. To apply the range queries augmented with a string similarity search predicate in both Euclidean space and road networks. In Euclidean space, we propose an approximate solution, the MHR-tree, which fix min-wise signatures into an R-tree. The min-wise signature for an index node keeps a pithy representation of the union of q -grams from strings under the sub tree of u. We analyze the cut away functionality of such signatures based on the set resemblance between the query string and the q -grams from the sub trees of index nodes. We also discuss how to estimate the selectivity of a SAS query in Euclidean space, for which we present a novel adaptive algorithm to find balanced partitions using both the spatial and string information stored in the tree. For queries on road networks, we propose a more exact method, The RSASSOL algorithm partitions the road network, adaptively searches relevant sub graphs, and prune candidate points using two attributes reference nodes and index of string matching. Lastly, an adapted Multipoint algorithm (MPALT) is applied, together with the exact edit distances, to verify the final set of candidates.

Keywords: Approximate string, Spatial database, RSASSOL, MPALT, Road network.

I. INTRODUCTION

Since some time, data mining has attracted a great deal of attention in the information industry. It is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses or other informationrepositories.

Today, an increasing number of usages of data set havebecomeavailable. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining is also known as Knowledge Discovery in Data.Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends knowledge-driven decisions. Data mining is accomplishedby building models. A model performs some actions on data based on some algorithm. The notion of automatic discovery refers to the execution of data mining models. Approximate string matching is the technique of findingstrings that match a pattern approximately rather thanexactly. The closeness of a match is measured in termsof the number of primitive operations necessary to convert the string into an exact match. This number is called the edit distance between the string and the pattern. Query processing on road networks has many in this work; we focus on range queries in road networks known as spatial approximate string (SAS) queries in road networks (RSAS queries) [2].Spatial range queries inquire about certain spatial objects related to other spatial objects within a certain distance .Given a query, this algorithm on road network returns the best objects with shortest path to the query location and textual relevance to the query keywords. For example given a query point q and a network distance r on a road network, we want to retrieve all objects within distance r to q and with the description similar to some keywords. Applications such as online map services and mobile services .In a road network, each node represents a location, the edge between two nodes represents the path between them. Each node can be tagged with textual information such as school or hospital.

In reality, keyword search for retrieving approximate string matches is required. Because exact match is a special case of approximate string match, it is clear that keyword search by approximate string matches has a much larger pool of purposes. Approximate string search is necessary when users have a fuzzy search condition, or a spelling error when submitting the query, or the strings in the database contain some degree of uncertainty or error. In the context of spatial databases, approximate string search could be combined with any type of spatial queries. In this survey we focus on range queries and dub such queries as Spatial Approximate String (SAS) queries [3].

II. LITERATURE SURVEY

Keyword search over a large amount of data is an important operation in a wide range of domains. Felipe et al. has recently extended its study to spatial databases, where keyword search becomes a fundamental building

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

block for an increasing number of real-world applications, and proposed the IR -Tree. The LBAK-tree was proposed after study on SAS queries in the Euclidean space and it has been compared against the MHR -tree. Their result s have shown that the LBAK -tree has achieved better query time than the MHR - tree, but using more space. Note that the LBAK - tree returns exact answers for the ESAS queries, and the MHR -tree returns approximate answers. The comparison of the LBAK - tree with the MHR - tree said, for ESAS queries, the LBAK - tree should be adopted when exact answers are required; hen space consumption must be small and approximate solutions are acceptable, the MHR - tree is the candidate. To the best of our knowledge, RSAS queries a D selectivity estimation of SAS queries have not been explored before. Approximate string search alone has been extensively studied in the literature. These works generally assume a similarity function to quantify the closeness between two strings. There are a variety of these functions such as edit distance and Jaccard [5] [6]. The Researchers are always being conducted to analyses the works associated with spatial approximate string queries. Some of the innovative approaches to road networks are:

A. Dijikstras Algorithm

Dijikstras algorithm is a graph search algorithm that solves the single-source shortest path problem for a graph with non-negative edge path costs, producing a shortest path tree. This algorithm is often used in routing and as a subroutine in other graph algorithms [2].For a given source vertex (node) in the graph, the algorithm finds the lowest cost (i.e. the shortest path) between that vertex andevery other vertex. It can also be used for finding costs of shortest paths from a single vertex to a single destination vertex by stopping the algorithm once the shortest path to the destination vertex has been determined. For example, if the vertices of the graph represent cities and edge path costs represent driving distances between pairs of cities connected by a direct road, Dijkstra's algorithm can be used to find the shortest route between one city and all other cities. For RSAS queries, the baseline spatial solution is based on the Dijikstras algorithm. Its performance degrades quickly when the query range enlarges and/or the data on the network increases [8][9].

B. OPRN Algorithm

We model a road network as a graph G = (V, E) where V (E) denotes the set of nodes (edges) in G. We denote index nodes in G by unique ids and specify an edge by it two end nodes, placing the nodes with the smaller id first. A spatial approximate string query Q consists of two parts: the spatial predicate Q_r and the string predicate Q_s. In road networks, Q_r is specified by a query point q and a radius r and the string predicate Q_s is defined by a set of strings and an edit distance threshold. The RSAS query framework consists of seven steps .Given a query, we first find all regions that intersect with the query range. Next, we use the similarity functions to retrieve the points whose strings are potentially similar to the query string. In the third step, we prune away some of these candidate points by calculating the lower and upper bounds of their distances to the query point. The fourth step is to further prune away some candidate points using the exact edit distance between the query strings and strings remaining candidates. After this step, the string predicate has been fully explored. In the next step, for the remaining candidate points, we calculate distances to the query point and return those with shortest distances within r .Then we find the points with better cost values. In the last step display the optimal path using graph [9].

C. Nearest Neighbor and Top-k Queries

The processing of k-nearest neighbor queries (kNNs) in spatial databases is a classical subject. Most proposals use index structures to assist in the kNN processing. Perhaps the most influential kNN algorithm is due to Roussopoulos et al. In this solution, R-tree indexes the points, potential nearest neighbors are maintained in a priority queue, and the tree is traversed according to a number of heuristics. Other branch-and-bound methods modify the index structures to better suit the particular problem addressed. Hjaltason and Samet propose an incremental nearest neighbor algorithm based on an R^*-tree. We propose a novel exact method, RSASSOL, which significantly outperforms the baseline algorithm in practice. The RSASSOL combines the q -gram based inverted lists and the reference nodes based pruning. Extensive experiments on large real data sets demonstrate the efficiency and effectiveness of our approaches [4].

D. THE MHR-TREE

The query algorithms for the MHR-tree generally follow the same principles as the corresponding algorithms for the spatial query component. However, we would like to incorporate the pruning method based on q-grams and Lemma 2 without the explicit knowledge of g_u for a given R-tree node u. We need to achieve this with the help of $s(g_u)$. Thus, the key issue boils down to estimating $|g_u \cap g_u|$ using $s(g_u)$ and $_\partial$. We can easily compute gσ and $s(g_\partial)$ from the query string once, using the same hash functions that were used for constructing the MHR-tree. When encountering a node u, let g refer to $g_u \cup g_\partial$ (g cannot be computed explicitly as g_u is not available) [1].

E. Range Query and Edit distance

In range query data base operation all records are retrieved within the range of upper and lower boundaries. The Data structures for range query are Range tree is the data structure used for organizing range queries. Range query operation involves in preprocessing input data onto the data structure and to retrieve the efficient answers for the queries by taking any subset as an input. Edit distance is a quantifying methodology of how two different strings are to one another and calculate minimum operations for transforming a string. Edit distances operation also find the automatic spell checks corrections and determine the candidate required corrections if there is any misspelled words which has low distance to the word in the tree.

III.FRAMEWORK

A. RSASSOL Algorithm

The paper on spatial approximate string search [1] presents a comprehensive study for spatial approximate

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

string queries in road networks. We use the edit distance, cosine similarity as the similarity measurement for the string predicate and focus on the range queries as the spatial predicate. Given a query, the RSASSOL algorithm on road network returns the best objects with shortest path to the querylocation and textual relevance to the query keyword.

TheRSASSOL method partitions the road network, adaptivelysearches relevant sub graphs and prunes candidate pointsusing both the string matching index and the spatialreference nodes .Lastly the MPALT algorithm is used toverify the final set of candidates .This works returns onlyone facility which matches the string predicate. Future work includes finding several facilities together with leastcost (shortest path).Given a query point q and a network distance r on a roadnetwork; we want to retrieve all objects within distance r to qand with the description similar to "theatre," where thedistance between two points is the length of their shortestpath. The MPALT algorithm minimizes the access to thenetwork by avoiding the nodes that will not be on anyshortest path between s and any destination ti. It alsoavoids repeatedly access to the explored part of thenetwork when calculating multiple shortest paths tomultiple destinations.The basic idea works as follow: We start the expansion ofthe network from s with the two nodes from the edgecontaining s, and always expand the network from anexplored node n (by adding adjacent nodes of n to apriority queue and checking points on correspondingedges) that has the shortest possible distance to any one ofthe destinations [2].

The MPALT algorithm is alsoused as the Multipoint Abbreviated List Table. This algorithm computes multiple shortest paths, within the query range, simultaneously at once between a single source point and multiple destination points. The distances computed and stored in storage model between a node to all reference nodes, which allows us to compute lower and upper distance bounds for any given node and any destination. The basic idea is that it starts the expansion of the network from source with the two nodes from the edge containing source node and always expand the network from an explored node that has the shortest possible distance to any one of the destinations. The algorithm terminates when the priority queue becomes empty. This algorithm minimizes the access to the network by avoiding the nodes that will not be on any shortest path distance between source and destination. It avoids repeatedly access to the explored part of the network when calculating multiple shortest paths to multiple destinations [10].

Conceptually, our RSAS query framework consists of five steps (refer to Figure 1).

Given a query, we first find all sub-graphs that intersect with the query range. Next, we use the Filter Trees of these sub-graphs to retrieve the points whose strings are potentially similar to the query string. In the third step, we prune away some of these candidate points by calculating the lower and upper bounds of their distances to the query point, using VR. The fourth step is to further prune away some candidate points using the exacted it distance between query string and strings of remaining candidates.

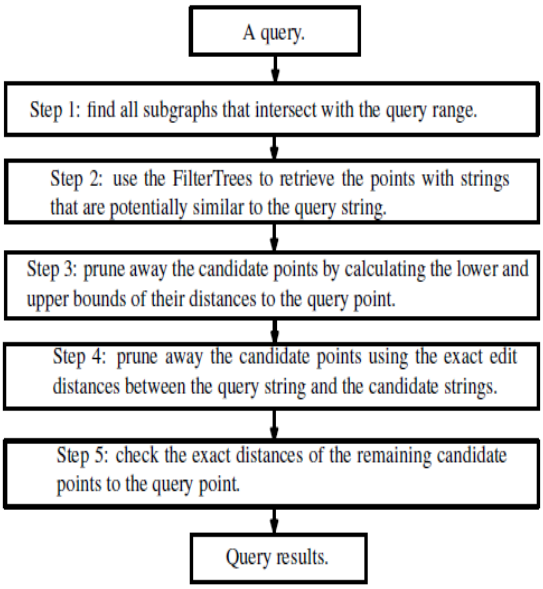

Fig.1. Overview of the RSASSOL algorithm.

After this step, the string predicate has been fullyexplored. In the final step, for the remaining candidate points, we check their exact distances to the query point and returnthose with distances within r [1].

IV. CONCLUSION

In this paper we have introduced a new approximate string search method which is RSASSOL algorithm. It is more efficient and faster than the previous Dijikstra's algorithm. It's performance does not degrades even though the range of query is vast.

REFERENCES

[1] Bin Yao, Mingwang Tang, MariosHadjieleftheriou. "Spatial Approximate String Search"; IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:25 NO:6 AYEAR 2013

[2] Tang and Mario Hadjieleftheriou "Spatial Approximate String Search" Proc.IEEE Int'l Conf. Knowledge and Data Engineering vol .5.No.6.June2013

[3] Joao B.Rocha-Junior and KjetilNorvag, "Top-k-Spatial Keyword Queries on Road Networks" EDBT26–30, Berlin, Germany 2012.

[4] XinCaoyGaoCongy Christian S. Jensen "Retrieving Top k Prestige Based Relevant Spatial Web Objects", Proceedings of the VLDB Endowment, Vol. 3, No.1,2010.

[5] D. Felipe, V. Hristidis, and N. Rishe. "Keyword Search on Spatial Databases" Proc.IEEE Int'l Conf.Data Eng.(ICDE), pp.656-665,2008.

[6] Oystein Egeland Carlsson "Keyword Search on Spatial Network Databases".This research wassupported in part by NSF grants CNS0320956, CNS-0220562, HRD-0317692, and IIS-0534530, 2004.

[7] S.Alsubaiee, A. Behm, and C. Li, "Supporting Location-Based Approximate-Keyword Queries, "Proc. SIGSPATIAL 18th Int'l Conf. Advances inGeographic Information Systems (GIS), pp. 61 - 70,2010.

[8] Ron Gutman "Reach-based Routing: A New Approach to Shortest Path Algorithms Optimized for RoadNetworks", proceedings of the sixth workshop onalgorithm engineering and experiment, New York,2010

[9] Sherlin Susan George, Remya R. "A Review on String Queries on Road Networks"; IJCAT International Journal of Computing and Technology, Volume 1, Issue 4, May 2014 ISSN: 2348 – 6090.

[10] S.Anandhi1, B. Anantharaj, R. Harriman. " Geographical Search with Approximate String in Spatial Databases";International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 4 938 – 941