INTERNATIONAL JOURNAL OF INTELLIGENT COMPUTING AND TECHNOLOGY

ISSN: 2457 0249 **Website :** http://ijict.com/
Jan 2018-Vol.1, Iss.2 Page No: 1 -9

Skyline Computing for Uncertained Database: A Query Based Framework

S. Saravanapriya[1], Dr. V. Thiagarasu[2]

Research Scholar[1], Associate Professor[2], Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam, India

srsaravanapriya@gmail.com, profdravt@gmail.com

Article History : Received on : Nov 2017; Published on Jan 2018

Abstract

The skyline query, aiming at identifying a set of skyline tuples that are not dominated by any other tuple, is particularly useful for multicriteria data analysis and decision making. For uncertain, a probabilistic skyline query, called P-Skyline, has been developed to return skyline tuples by specifying a probability threshold. However, the answer obtained via a P-Skyline query usually includes skyline tuples undesirably dominating each other when a small threshold is specified; or it may contain much fewer skyline tuples if a larger threshold is employed. To address this concern, a new uncertain skyline location based query has been proposed, called U-Prefix Scan query. Instead of setting a probabilistic threshold to qualify each skyline tuple independently, the U-Prefix Scan query searches for a set of tuples that has the highest probability (aggregated from all possible scenarios) as the skyline answer. In order to answer U-Prefix scan queries efficiently, a number of optimization techniques for query processing, including 1) Computational simplification of K-NN for neighbour node prediction with U-Prefix Scan probability, 2) Pruning of unqualified candidate skylines and early termination of query processing, 3) Reduction of the input data set, and 4) Partition and conquest of the reduced data set. A comprehensive performance evaluation on their algorithm and an alternative approach that formulates the K-NN with U-Prefix scan Skyline processing problem by integer programming.

Keywords: Skyline Processing, Uncertain Data, Data Set, Big Data

INTRODUCTION

Data Mining is an analytic process designed to explore data (usually large amounts of data typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) The initial exploration, (2) Model building or pattern identification with validation/verification, and (3) Deployment (i.e., the application of the model to new data in order to generate predictions).

Location Based Services (LBS) are becoming increasingly important to the success and attractiveness of next generation information systems. However, a natural tension arises between the need for user privacy and the flexible use of location information. In this work presented a framework K-NN to support privacy enhanced location based services. Classified the services according to several basic criteria and a K-NN distribution method to support these services has been proposed. The main idea behind the system is to hierarchically maintain location like grid information under different routes and distribute the appropriate route information only to group members with the necessary permission. This method proposed to deliver hierarchical location information. Furthermore, present a practical LBS system implementation. Hierarchical location information coding offers flexible location information access which enables a rich set of location based services. A load test shows such a system is highly practical with good efficiency and scalability.

LITERATURE REVIEW

Chuan-Ming [2015] proposed to extend data systems by a Skyline operation. This operation filters out a set of interesting points from a potentially large set of data points. A point is interesting if it is not dominated by any other point. For example, a hotel might be interesting for somebody traveling to Nassau if no other hotel is both cheaper and closer to the beach. They show how SQL can be extended to pose Skyline queries, present and evaluate alternative algorithms to implement the Skyline operation, and shows how this operation can be combined with other data operations. Computing the Skyline is known as the maximum vector problem [KLP75, PS85]. They use the term Skyline because of its graphical representation. More formally, the Skyline is defined as those points which are not dominated by any other point. A point dominates another point if it is as good or better in all dimensions and better in at least one dimension. For example, a hotel with price = $50 and distance = 0.8 miles dominates a hotel with price = $100 and distance = 1.0 miles. Shows the Skyline of cheap hotels near the beach for a sample set of hotels. A travel agency is one application for which a Skyline operation would be useful. Clearly, many other applications in the area of decision support can be found finding good salespersons which have low salary. A Skyline operation can also be very useful for data visualization. The Skyline of

Manhattan, for instance, can be computed as the set of buildings which are high and close to the Hudson River.

Trimponias [2013] tackled the problem of skyline analysis on uncertain data. A novel probabilistic skyline model where an uncertain object may take a probability to be in the skyline has been proposed, and a p-skyline contains all objects whose skyline probabilities are at least p ($0 < p \leq 1$). Computing probabilistic skylines on large uncertain data sets is challenging. They developed a bounding-pruning-refining framework and three algorithms systematically. The bottom-up algorithm computes the skyline probabilities of a part of selected instances of uncertain objects, and uses those instances to prune other instances and uncertain objects effectively. The top-down algorithm recursively partitions the instances of uncertain objects into subsets, and prunes subsets and objects aggressively. Combining the advantages of the bottom-up algorithm and the top down algorithm, they developed a hybrid algorithm to further improve the performance. Their experimental results on both the real NBA player data set and the benchmark synthetic data sets shows that probabilistic skylines are interesting and useful, and their algorithms are efficient on large data sets.

Sunitha [2013] proposed the Probabilistic threshold k-Nearest-Neighbour Query (T-k-PNN), which returns sets of k objects that satisfy the query with probabilities higher than a few threshold T. Two steps are proposed to handle this query efficiently. In the first step, objects that cannot constitute an answer are filtered with the aid of a spatial index. The second step, called probabilistic candidate selection, significantly prunes a number of candidate sets to be examined. The remaining sets are sent for verification, which derives the lower and upper bounds of answer probabilities, so that a candidate set can be quickly decided on whether it should be included in the answer. They also examine spatially-efficient data structures that support these methods. Their solution can be applied to uncertain data with arbitrary probability density functions. They have also performed extensive experiments to examine the effectiveness of their methods. Uncertainty is inherent in many emerging applications. In the Global-Positioning System (GPS), for example, the location values collected from the mobile devices have measurement errors and it is difficult to remove them due to the lacking of domain knowledge as another example, consider a habitat monitoring system where data like temperature, humidity, and light intensity are acquired from sensors. Due to the impreciseness of sensing devices, the data obtained are often noisy.

III. PROPOSED ALGORITHM

The proposed system is an efficient method for skyline computation. In this work, a new skyline query for uncertain data. It focuses on meeting the non-dominance, incomparability and coverage properties simultaneously for uncertain skyline query. A search algorithm based on dynamic programming (DP) to find U-Prefix Scan is been used here. The algorithm is improved with pruning and early termination

(P&ET) techniques. Input data set reduction (SR) and partition (SP) techniques to reduce the input data set size in order to further expedite the U-Prefix Scan processing time.

Introduce a set of query evaluation techniques:

1. Show that exact query evaluation is expensive for part of their proposed queries.

2. Give branch-and-bound search algorithms to compute exact query answers based on A∗ search. The search algorithms lazily explore the space of possible answers, and early-prune partial answers that do not lead to final query answers.

3. Novel sampling techniques based on a U-Prefix Scan method to compute approximate query answer.

Address the challenges associated with dealing with uncertain scores and incorporating probabilistic score quantifications in both the semantics and processing of ranking queries. Summarize such challenges as follows:

Ranking Model: The conventional total order model cannot capture score uncertainty. While partial orders can represent incomparable objects, incorporating probabilistic score information in such model requires new probabilistic modelling of partial orders.

Query Semantics: Conventional ranking semantics assume that each record has a single score and a distinct rank (by resolving ties using a deterministic tie breaker).Query semantics allowing a score range, and hence different possible ranks per record needs to be adopted.

Query Processing: Adopting a probabilistic partial order model yields a probability distribution over a huge space of possible rankings that is exponential in the data size. Hence need efficient algorithms to process such space in order to compute query answers.

When person desire to know destination information based on consumer's requirement say for illustration user needs to reach nearest ATM or hospital. He can get ATM or hospital information using internet service provider. However the person wishes effective result with respect to travel time and fee (i.e. nearest route).

KNN-Route analysis: Consequently person needs application that supplies all of the expertise he desires. The proposed procedure entails almost always three predominant modules, user module, LBS module and Route-Saver module. In user module user receives a location map includes locations, user location and route map from user place (source) and possible destination. Proposed work, the users require accurate results that are computed with appreciate to live traffic information. The entire works require the LBS to know the weights (travel times) of all road segments .Considering that the LBS lack the Infrastructure for monitoring road traffic, the above works are inapplicable to the problem. Few works try and model the entire works require the LBS to know the weights (travel times) of all road segments.

Considering that the LBS lack the infrastructure for monitoring road traffic, the above works are inapplicable to the problem. Approximately works try and model the travel occasions of street segments as time-various features, which may also be extracted from historical traffic patterns. These services may just capture the consequences of periodic events (e.g. rush hours, weekdays). Nevertheless, they nonetheless cannot reflect traffic information, which can be effected by sudden events, e.g. congestions, accidents and road maintenance.

The LBS module is responsible for accumulating the specified data from consumer and LBS generate optimized information which includes consumer's present area and route log to the destinations. Then this information is transferred to the Route-saver. Route-saver utilizes the contemporary traffic understanding bought from traffic provider and calculates the journey time and most beneficial path to source and destinations by using Nearest Neighbour queries.

To reduce the number of route requests while providing efficient results, combine information throughout a couple of routes within the log to derive tight lessen/higher bounding journey times. An effective strategy is proposed to compute bounds effectively. Additionally, compare the influence of exclusive orderings for issuing route requests on saving route requests. And learn the best way to parallelize route requests in order to reduce the query present Route-Saver algorithm for processing a range query. It applies the travel time bounds discussed above to reduce the number of route requests. To guarantee the accuracy of returned results, it removes all expired routes. The algorithm first conducts a distance range search to obtain set of candidate points. It also consists of two phases to process the candidate points in the query results in the set of exact results for user query.

Skyline Route API: Examples are: Google/Bing route APIs. Such API computes the shortest route between two points on a road network, based on live traffic. It has the latest road network G with live travel time information. Mobile User using a mobile device (Smartphone), the user can acquire his current geo-location q and then issue queries to a location-based server. In this project, consider range and K-NN queries based on live traffic.

Location-Based Service/Server: It provides mobile users with query services on a data set P, whose POIs (e.g., restaurants, cafes) are specific to the LBS's application. The LBS may store a road network G with edge weights as spatial distances, however G cannot provide live travel times. In case P and G do not fit in main memory, the LBS may store P as an R-tree and store the G as a disk-based adjacency list.

RESULTS & DISCUSSION

Assign each region an ordered region number, in which each bit encodes a data space, split whose detail is kept in the region split history. In such way, route a query by computing the target region number based on local split histories. Second, define the skyline search space to limit the number of involved nodes and partition the search space into subspaces adaptively at each hop to parallelize processing and control the

number of search messages. Third, balance the query loads in load balanced partitioning of data space during the joining/leaving of nodes.

Furthermore, dynamically sample load from both linked and random nodes and migrate data in case of load imbalance. A novel approach called KNN with Prefix scan that partitions and numbers the data space among the peer nodes such that the target subspace (region) number can be derived with good accuracy in order to control the peers accessed and search messages during skyline query processing. Formally prove the necessity and completeness of visiting non-dominated nodes within the delimited skyline search space.

Balance the query loads among peers through both load balanced data space partition and dynamic load migration. A novel load sampling mechanism that attends the quality of sampled load distribution and the efficiency of sampling process by combining direct and random sampling. Have an addressed the efficient processing of traditional centralized skyline querying on system to system networks. Based on a tree structured network, have a proposed skyline processing U-Prefix Scanning algorithm to partition the skyline space adaptively to control query forwarding behavior effectively. Consequently, have been able to significantly reduce the number of visited nodes and search messages. Also have devised approaches for effective query load balancing. The correctness and effectiveness of proposed algorithm were formally proved and validated by experiments. As for future research, investigate efficient approximate algorithm for high dimensional data querying in the system to system settings.

Definition 1 (**Dominance)** Assume that minimum value is preferred. Given two objects, p and q with the same number of dimensions; object p is said to dominate object q, formally written as $p \prec q$, if and only if object p is as good as object q in all dimensions and better than object q in at least one dimension.

Definition 2 (Skyline Query) Given a set of objects O, an object p of O is said to be a skyline object if and only if there does not exist any other objects k in O which dominates p. Then the skyline on O is the set of all skyline objects in O. Applying the skyline definition on objects presented in Fig. 1, with the assumption that minimum value is preferred for both dimensions, then the set of skyline objects is {a, b, c}.

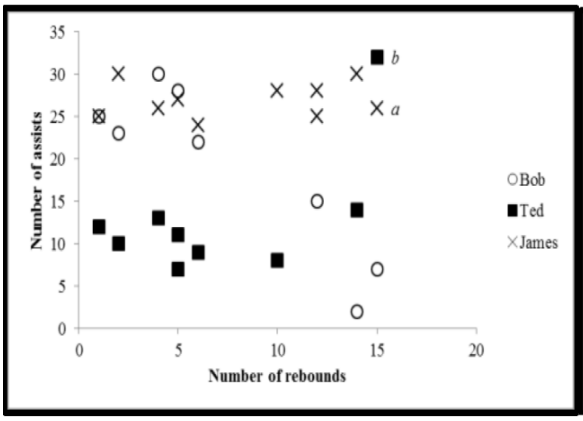

Fig.1Skyline query

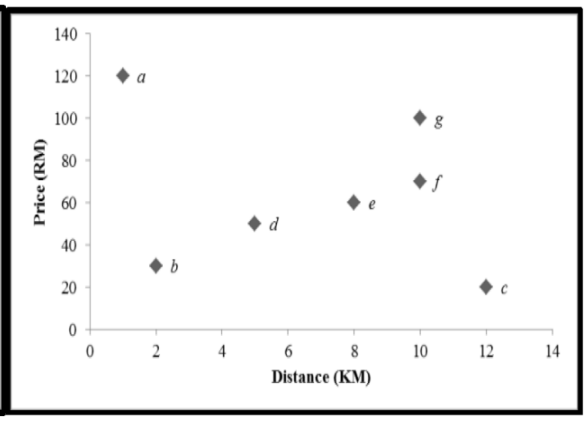

Fig.2 Uncertain data at the object level

Definition 3 (**Uncertain Data**) Let D is an n-dimensional data. The data D can either contain uncertainty at the object level if each object of D has several instances or at the dimension level if any of its objects can have different forms of data values for a dimension. Fig. 2 is an example of a data with uncertainty at the object level, while Fig. 3 and Fig. 4 depict examples of uncertain data at the dimension level.

Definition 4 (Uncertain Dimension) given a data D with n-dimensions, a dimension Aj is said to be an uncertain dimension if it contains different data value representations (i.e. points and range values).

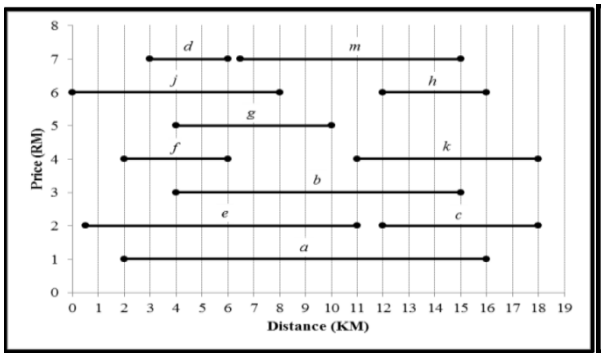

Fig.3Uncertain data with dimension

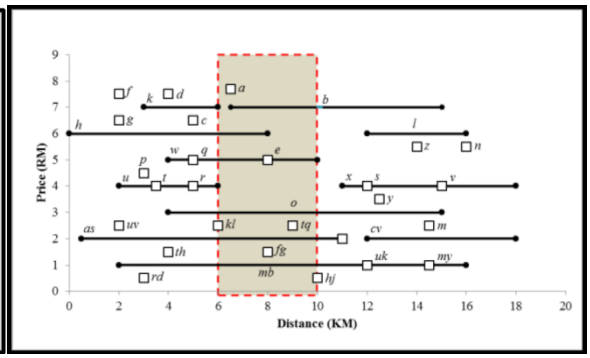

Fig.4 Uncertain DB with heterogeneous dimension

If the entire data values of any dimension in D of the same form (i.e. either all are points or all are range values), such data is said to be an uncertain data with homogeneous dimensions; otherwise, if it contains different data value representations (i.e. points and range values) the data is referred to as uncertain data with heterogeneous dimensions. An example of uncertain data with homogeneous and heterogeneous dimension is depicted in Figure 3 and Figure 4 respectively.

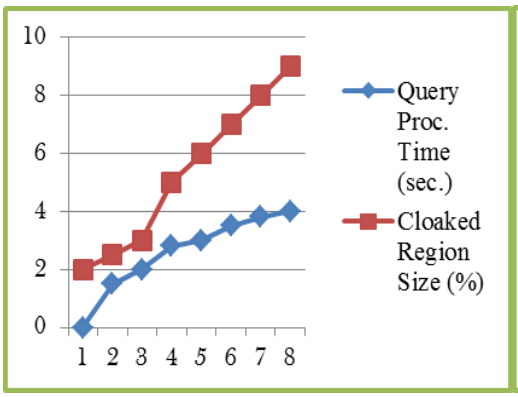

Fig.5 Result set size

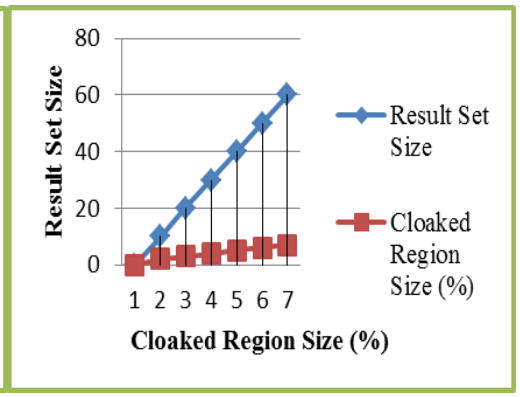

Fig.6 Query Processing Time

CONCLUSION

Presented a personalization framework for data queries based on information stored in structured user profiles that keep single, unconditional preferences. Formulated the main personalization step as a graph computation problem and presented and evaluated, through a number of experiments, algorithms for the personalization of a query. Defined skyline queries over continuous uncertain data, and proposed a novel, efficient framework to answer these queries. Query answers are probabilistic, where each object is associated with a probability value of being in skyline objects. Users can specify a probability threshold, that each object in the answer set must exceed, and a tolerance that defines the allowed error margin in probability calculation. Described framework in the context of skyline have proposed three methods to bound each object probability for being a preferred object, namely, proposed uncertainty reduction, pairwise comparison and bound tightening. Two-phase framework has been presented which encapsulates the three proposed methods along with filter-refine approach.

REFERENCES

1. Bin Jiang, Jian Pei, Xuemin Lin and Yidong Yuan, "Probabilistic skylines on uncertain data: model and bounding-pruning-refining methods" , Springer Science Business Media, LLC, 2010
2. Chuan-Ming Liu and Syuan-Wei Tang , "An Effective Probabilistic Skyline Query Process on Uncertain Data Streams" , Procedia Computer Science, 63, pp. 40 – 47, 2015
3. T.Sunitha and L.Indu,S.Anbu, **"**An Adaptive Algorithm for Computing Subspace SKYLINE Queries Over Distributed Uncertain Data ", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3(7), July 2013, ISSN: 2277 128X
4. K. Hose and A. Vlachou, "A survey of skyline processing in highly distributed environments," The VLDB Journal The International Journal on Very Large Data Bases, vol. 21(3), pp. 359–384, 2012.
5. K. Dongwon, I. Hyeonseung, and P. Sungwoo, "Computing exact skyline probabilities for uncertain databases", IEEE Transactions on Knowledge and Data Engineering, vol. 24(12), pp. 2113–2126, 2012.
6. M. J. Atallah and Y. Qi, "Computing all skyline probabilities for uncertain data", Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, pp. 279-287, 2009
7. W. Zhang, X. Lin, Y. Zhang, W. Wang, G. Zhu, and J. X. Yu, "Probabilistic skyline operator over sliding windows", Information Systems, vol. 38(8), pp. 1212-1233, 2013.

8. M. Balasubramanian, R. Anitha, L. Dhakshayani, V. Kavitha, "Adaptive Processing for Distributed Skyline Queries over Uncertain Data", International Research Journal of Engineering and Technology, Vol. 4(3), 2017, ISSN: 2395-0072
9. J. B. Rocha-Junior, A. Vlachou, C. Doulkeridis, and K. Norvag, "Agids: A grid-based strategy for distributed skyline query processing", Data Management in Grid and Peer-to-Peer Systems, Springer, pp. 12-23, 2009
10. L. Zhu, Y. Tao, and S. Zhou, "Distributed skyline retrieval with low bandwidth consumption", IEEE Transactions on Knowledge and Data Engineering, vol. 21(3), pp. 384-400, 2009.
11. T. Sunitha and L. Indu,S.Anbu, "An Adaptive Algorithm for Computing Subspace SKYLINE Queries Over Distributed Uncertain Data", International Journal of Advanced Research in Computer Science and Software Engineering , vol. 3(7), July 2013, ISSN: 2277 128X
12. G. Trimponias, I. Bartolini, D. Papadias, and Y. Yang, "Skyline processing on distributed vertical decompositions", IEEE Transactions on Knowledge and Data Engineering, vol. 25(4), pp. 850-862, 2013
13. B. Chen and W. Liang, "Progressive skyline query processing in wireless sensor networks", 5th International Conference on. IEEE in Mobile Adhoc and Sensor Networks, pp. 17-24, 2009
14. C. Doulkeridis and K.. Norvag, "A survey of large-scale analytical query processing in map reduce", The VLDB Journal, vol. 23(3), pp. 355-380, 2014.
15. B. Zhang, S. Zhou, and J. Guan, "Adapting skyline computation to the map reduce framework: Algorithms and experiments", Database Systems for Advanced Applications in Springer, pp. 403-414, 2011
16. Y. Tao, W. Lin, and X. Xiao, "Minimal map reduce algorithms," International Conference on Management of Data in Proceedings of ACM SIGMOD, pp. 529-540, 2013
17. Y. Park, J.-K. Min, and K. Shim, "Parallel computation of skyline and reverse skyline queries using map reduce", Proceedings of the VLDB Endowment, vol. 6(14), pp. 2002-2013, 2013.
18. Y. Tao, X. Xiao, and J. Pei, "Efficient skyline and top-k retrieval in subspaces", IEEE Transactions on Knowledge and Data Engineering , vol. 19(8), pp. 1072-1084

How to cite this article:

S. Saravanapriya & Dr. V. Thiagarasu, "Skyline Computing for Uncertained Database: A Query Based Framework", International Journal of Intelligent Computing and Technology, Vol. 1(2), 1-9, 2018