A Probabilistic Convex Hull Query Tool

Zhou Zhao, Da Yan and Wilfred Ng
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong, China
{zhaozhou, yanda, wilfred}@cse.ust.hk

ABSTRACT

Uncertain data is inherently important in a lot of real-world applications, such as environmental surveillance and mobile tracking. Probabilistic convex hull is very useful for discovering the territory of imprecise data in such applications with a high confidence. In order to deal with this, we propose and study probabilistic convex hull queries based on the possible world semantics, which are able to retrieve the objects whose probability of being on the convex hull is at least α. The demonstration is based on animal tracking whose GPS coordinate is no longer considered to be precise due to device limitation or privacy issues. We demonstrate two interesting results from studying the migration habit of one specific species and the correlation between species through probabilistic convex hull queries.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications

General Terms

Algorithms

Keywords

Probabilistic Convex Hull, Query Processing, Uncertain

1. INTRODUCTION

The problem of defining convex hull on a set of data points has attracted a lot of attention due to its wide spectrum of real life applications such as pattern recognition [1], cluster analysis [7] and linear optimization [2]. A large number of algorithms have been proposed to compute convex hull. Among them Andrew's Monotone Chain algorithm [6] finds the convex hull of a set of n 2D points in $O(n \log n)$ time and it is probably the best known one.

All the existing algorithms assume that the each data point must be certain. However, in real life applications, the collected data may be imprecise due to environment factors, device limitations and privacy issues. To the best of our knowledge, no previous work has studied the concept of probabilistic convex hull query.

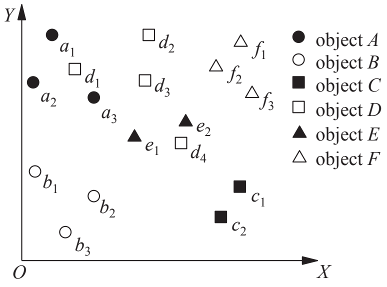

Figure 1: A Set of Uncertain Objects

Consider an animal tracking system, which continuously collects the sample location of animals. The sample locations of an animal are inherently noisy due to the instability of the tag and environmental factors. Figure 1 shows an example of the sample locations of six animals (i.e. objects A to F). Each animal has several sample locations that represent its possible positions. Suppose that the system sampled 3 times to localize these animals. Animal A was observed in three places, a_1, a_2 and a_3, each of which has a probability of $\frac{1}{3}$. There are two sample locations of C, since two of three samples have the same sample location. In this case, the probability of one sample location is $\frac{2}{3}$ while the probability of the other is $\frac{1}{3}$. Sample locations can be used to evaluate the uncertainty in the animal tracking system.

In this paper, we demonstrate our probabilistic convex hull query tool, which provides an efficient way for retrieving objects whose probability of being on the convex hull is at least α. Our system not only supports a single probabilistic convex hull query, but is also capable of handling dual probabilistic convex hull queries, which find the convex hulls for two pieces of data. Following our animal tracking example, we are able to learn the activity region of some animal through probabilistic convex hull queries. Dual convex hull queries are useful for discovering the correlation between two pieces of data. If the convex hull of the location of two species overlap, these two species possibly interact for some reason.

The probabilistic convex hull model is presented in Section 2. In Section 3, we describe our probabilistic convex hull query algorithm. We propose three types of convex hull queries supported by our system and demonstrate them in Sections 4 and 5. We give the conclusion in Section 6.

EDBT 2012, March 26–30, 2012, Berlin, Germany.
Copyright 2012 ACM 978-1-4503-0790-1/12/03 ...$10.00

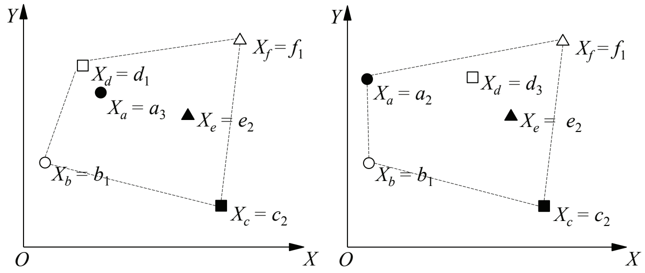

Figure 2: Two Possible Worlds of the Uncertain Object Set Presented in Figure 1

2. PROBABILISTIC CONVEX HULL

In this section, we present our uncertain data model and formally define the concept of *probabilistic convex hull* (PCH) based on the popular *possible world semantics* [3].

We adopt the multi-instance mode to interpret the uncertain objects, which is also used in previous studies such as [5]. We assume that a database is composed of a set of objects $= \{o_1, o_2, \ldots, o_n\}$, and each object o_i is represented by a set of ℓ_i instances, which we denote as $o_i = \{s_i^1, s_i^2, \ldots, s_i^{\ell_i}\}$. For each uncertain object o_i, we define a random variable x_{o_i} describing the instance that o_i occurs to be.

To keep our model simple, we assume that (1) the uncertain objects are independent of each other, and (2) for each uncertain object, its instances are exclusive to each other.

Figure 2 illustrates two possible worlds of the uncertain objects presented in Figure 1. In the possible world on the left, object A is not on the convex hull while object D is on the convex hull and vice versa on the right. The probability of each possible world is the product of the existing probability of all its instances.

Definition 1. *Given a set of objects O, the probabilistic convex hull query with probabilistic threshold α returns the set of objects $PCH_\alpha = \{o \in O | Pr^h(o) \geq \alpha\}$, where $Pr^h(o) = \sum_{pw} Pr^h(o|pw) \cdot Pr(pw)$. The parameter pw is a possible world of O and $Pr^h(o|pw)$ is the probability of o on the convex hull in pw.*

3. ALGORITHM

In this section, we consider how to compute $Pr^h(o)$ for an object $o \in O$. In our model, each object o_i is represented by a set of instances s_j. Conceptually, in order to compute the $Pr^h(o_i)$, we have to consider all the $Pr^h(s_j)$ which is the probability of s_j being on the convex hull. According to the *law of total probability*, the expression of $Pr^h(o_i)$ is given by:

$$\begin{aligned} Pr^h(o_i) &= \sum_{pw} Pr^h(o_i, pw) \\ &= \sum_{pw} \sum_{s_j \in o_i} Pr^h(o_i, pw | s_j) Pr(s_j) \\ &= \sum_{s_j \in o_i} \sum_{pw} Pr^h(s_j, pw | s_j) Pr(s_j) \\ &= \sum_{s_j \in o_i} Pr^h(s_j) Pr(s_j) \qquad (1) \end{aligned}$$

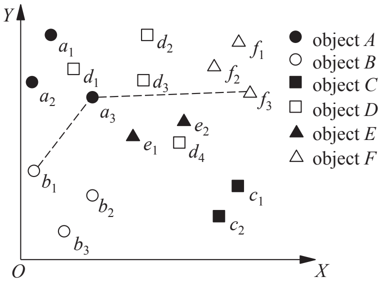

Figure 3: Illustration of $Pr^h(\widehat{b_1 a_3 f_3})$ Computation

As we know that $Pr^h(s_j)$ can be queried from the databases, then the remaining problem is how to compute $Pr^h(o_i)$. Before presenting the algorithm for computing $Pr^h(s_j)$, we consider the scenario in Figure 3 first.

Suppose that we already know that the instances b_1 and f_3 are the consecutive instances of a_3, denoted as $\widehat{f_3 a_3 b_1}$ and we want to compute the probability of the instance a_3 being on the convex hull. If a_3 is on the convex hull, all the instances of C, D and E must below lines $\overline{a_3 f_3}$ and $\overline{b_1 a_3}$. In Figure 3, it is observed that all the instances of C and E are below these two lines. For another object D, only one instance d_4 is below these two lines. Therefore, $Pr^h(\widehat{f_3 a_3 b_1}) = 1 \times \frac{1}{4} \times 1 = 0.25$.

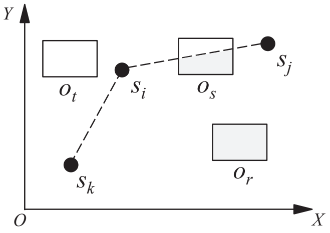

Figure 4: Illustration of $Pr^h(\widehat{s_j s_i s_k})$ Computation

After explaining the computation of $Pr^h(\widehat{b_1 a_3 f_3})$ in Figure 3, we now give the general computing formula. Suppose that s_j and s_k are the consecutive instances of s_i, then we have the following three cases:

Case 1 $\exists o_t \in O - \{o_i, o_j, o_k\}$, such that $\forall s_j \in o_t$, s_j is above the line $\overline{s_j s_i}$ or $\overline{s_i s_k}$, then $Pr^h(\widehat{s_j s_i s_k}) = 0$.

Case 2 If $o_r \in O - \{o_i, o_j, o_k\}$, satisfies that $\forall s_j \in o_r$, s_j is above the line $s_j s_i$ or $s_i s_k$, then o_r does not have effect on $Pr^h(\widehat{s_j s_i s_k})$.

Case 3 If $o_s \in O - \{o_j, o_i, o_k\}$, the instances of o_s across the lines $\overline{s_k s_i}$ or $\overline{s_i s_k}$, then only the instances above the lines have positive effect on $Pr^h(\widehat{s_j s_i s_k})$.

Given $\widehat{s_j s_i s_k}$, if we find an object that conforms to Case 1, then $Pr^h(\widehat{s_j s_i s_k}) = 0$ without any further computation. If an object satisfies Case 2, we skip that object. For the objects in Case 3, we need to compute the probability of that object being below the two lines by considering the probability of its instances.

Once we have computed $Pr^h(\widehat{s_j s_i s_k})$, we can enumerate all the possible consecutive instances of s_i in order to get $Pr^h(s_i)$. The algorithm for computing $Pr^h(o_i)$ is summarized as follows:

1. For all consecutive instances s_j and $s_k \notin o_i$ compute $Pr^h(\widehat{s_j s_i s_k})$

2. $Pr^h(s_i) \leftarrow \sum_{s_j, s_k} Pr^h(\widehat{s_j s_i s_k})$ for all s_j, s_k consecutive to s_i

3. $Pr^h(o_i) \leftarrow \sum_{s_j \in o_i} Pr^h(s_j) Pr^h(s_j)$

If $Pr^h(o_i) \geq \alpha$, o_i is considered to be an object on the convex hull, otherwise it is regarded to be an interior point.

4. PCH QUERY PROCESSING

We implemented a prototype that supports three types of probabilistic convex hull queries: single, dual and regional PCH queries. For each type of PCH query, our system returns both the information of objects on and inside the convex hull separately.

4.1 Single PCH query

Single PCH is useful for describing the territory of uncertain data, since the number of objects on the convex hull is much smaller than the size of data. Given a set of uncertain data objects, our system bulk-loads the uncertain data to R-tree and performs the computation of PCH as illustrated in Section 3.

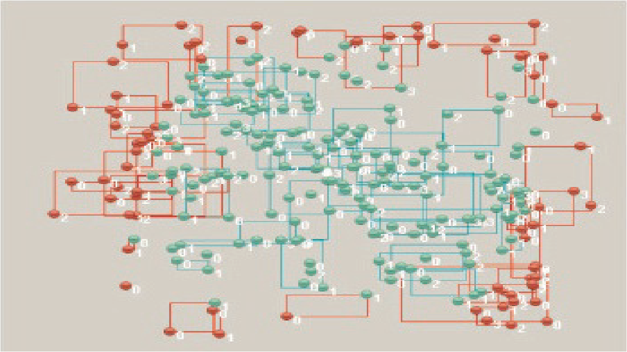

Figure 5: Single Probabilistic Convex Hull

Figure 5 demonstrates the single PCH query result of one sample uncertain data. In Figure 5, the red points are the objects on the convex hull with probability greater than probability threshold α.

4.2 Dual PCH query

Dual probabilistic convex hull is able to discover the common region and spatial relationship between two uncertain data. Similar to the PCH query, our system carries out the computation of probabilistic dual convex hull by using the input of two given pieces of data. Figure 6 shows the result of dual PCH query on two sample uncertain data. The purple and red points show the convex hull of both data.

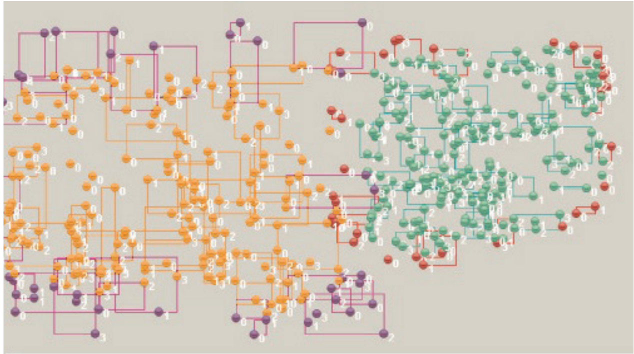

Figure 6: Dual Probabilistic Convex Hull

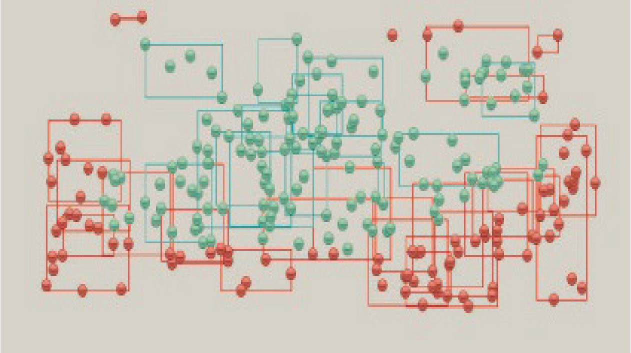

Figure 7: Regional Probabilistic Convex Hull

4.3 Regional PCH query

Regional probabilistic convex hull query is used to retrieve the probabilistic convex hull of a specified region of original uncertain data. This query makes our system flexible and useful for more real-world applications. Since the distribution of the data is not always strongly connected, regional PCH query is able to offer finer results. Figure 7 shows a PCH of a specified region of one sample uncertain data.

5. DEMONSTRATION PLAN

Our system and web interface are implemented in Java programming language on top of R-tree[4]. In the demo, we present application scenarios for all three types of queries: single single, dual and regional PCH queries using real dataset that we collected from GTOPP[1].

GTOPP dataset contains the sample locations of the tagged animals sorted by timestamp, providing the geographical information of these animals. Since the sample locations are not reliable and the consecutive sample locations may represent the same location, we use the sliding window method to group the consecutive samples into one object. The samples in the objects have the same weight and the probability of each sample location is assigned by voting of the samples.

Figure 8 demonstrates the result of probabilistic single convex hull query on the movement area of the marine species, Atlantic bluefin tuna. Through querying convex hull of the locations of this specie, we are able to know its movement

[1] `http://gtopp.org`

range and learn about the territory it inhabits.

However, the sample locations of Atlantic bluefin tuna are imprecise and uncertain, as shown in Figure 8. Notice the three sample locations with lowest latitude in the south, it is clear that these three consecutive sample locations are not certain. They are too far away and the Atlantic bluefin cannot pass through these locations within a short time. The deterministic convex hull algorithms are not able to solve this problem, since it returns a convex hull with a lot of false positive regions because it considers all the possible sample locations. However, Our system produces a much more concise convex hull which is illustrated in Figure 8.

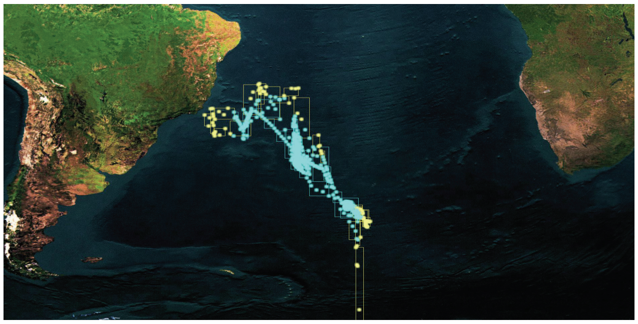

Figure 8: Atlantic Bluefin Tuna

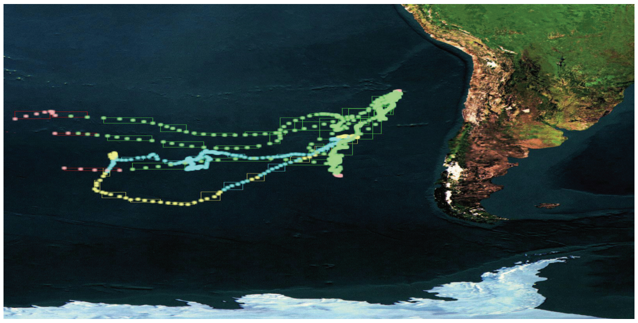

Figure 9: Pacific Bluefin Tuna and Northern Elephant Seal

The dual PCH query is useful for discovering the spatial correlation between two datasets. In Figure 9, the dual PCH query result shows that there exists an intersecting area between the sample locations on the convex hull of these two marine species: Pacific bluefin tuna and northern elephant seal. Through this query, we are able to know that these two marine species may have similar living habits because their territories overlap in our query result. This may be a signal that indicates a kind of correlation between these two species and this findings would be worthwhile for scientists to study further.

Regional PCH query is able to provide results with range constraints and it is very useful in real-life situations. For example, Pacific bluefin is widely spread over the South Pacific shown in the left map of Figure 10. However, the fishermen in Chile may only care about the living territory of Pacific bluefin near their habitation. In this case, Regional PCH query provides an area of fishing for the fishermen.

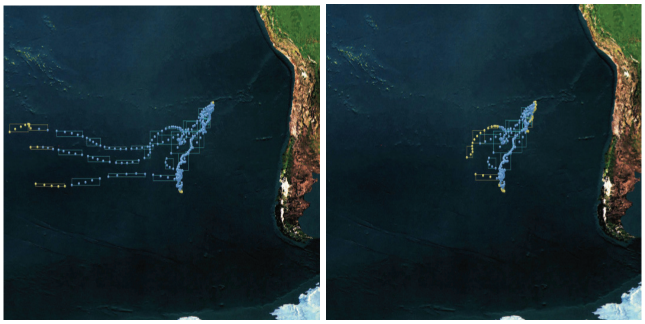

Figure 10: Pacific Bluefin Tuna Near South American

6. CONCLUSION

In this demo, first we aim to explain the problem of probabilistic convex hull query. Secondly, we show that our system is based on an uncertain model and an efficient algorithm, which returns objects on PCH of probability at least α. Based on the algorithm, we propose three types of PCH queries and demonstrate the results of them by synthetic and real datasets. The PCH query is able to show that our system is superior to the existing deterministic approach by providing more concise and significant results.

ACKNOWLEDGEMENTS: We are grateful to the anonymous reviewers for their insightful comments on this paper. This work is partially supported by RGC GRF under grant number HKUST 618509.

7. REFERENCES

[1] S. Akl and G. Toussaint. A fast convex hull algorithm* 1. *Information Processing Letters*, 7(5):219–222, 1978.

[2] Y. Chang, L. Bergman, V. Castelli, C. Li, M. Lo, and J. Smith. The onion technique: indexing for linear optimization queries. In *ACM SIGMOD Record*, volume 29, pages 391–402. ACM, 2000.

[3] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal, The International Journal on Very Large Data Bases*, 16(4):523–544, 2007.

[4] S. Leutenegger, M. Lopez, and J. Edgington. Str: A simple and efficient algorithm for r-tree packing. In *icde*, page 497. Published by the IEEE Computer Society, 1997.

[5] J. Pei, B. Jiang, X. Lin, and Y. Yuan. Probabilistic skylines on uncertain data. In *Proceedings of the 33rd international conference on Very large data bases*, pages 15–26. VLDB Endowment, 2007.

[6] F. Preparata and M. Shamos. *Computational geometry: an introduction*. Springer, 1985.

[7] J. Sander, M. Ester, H. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.