Information Systems 40 (2014) 67–83

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

MSSQ: Manhattan Spatial Skyline Queries ☆

Wanbin Son, Seung-won Hwang, Hee-Kap Ahn *

Department of Computer Science and Engineering, Pohang University of Science and Technology, Republic of Korea

ARTICLE INFO

Article history:
Received 18 April 2012
Received in revised form
1 June 2013
Accepted 1 October 2013
Recommended by: Xifeng Yan
Available online 18 October 2013

Keywords:
Spatial skyline queries
Spatial databases
Manhattan distance
Query semantics

ABSTRACT

Skyline queries have gained attention lately for supporting effective retrieval over massive spatial data. While efficient algorithms have been studied for spatial skyline queries using the Euclidean distance, these algorithms are (1) still quite computationally intensive and (2) unaware of the road constraints. Our goal is to develop a more efficient algorithm for L_1 distance, also known as Manhattan distance, which closely reflects road network distance for metro areas. We present a simple and efficient algorithm which, given a set P of data points and a set Q of query points in the plane, returns the set of spatial skyline points in just $O(|P|\log|P|)$ time, assuming that $|Q| \le |P|$. This is significantly lower in complexity than the best known method. In addition to efficiency and applicability, our algorithm has another desirable property of independent computation and extensibility to L_∞ norm distance, which naturally invites parallelism and widens applicability. Our extensive empirical results suggest that our algorithm outperforms the state-of-the-art approaches by orders of magnitude. We also present efficient algorithms that report the changes of the skyline points when single or multiple query points move along the x- or y-axis.

1. Introduction

Skyline queries have gained attention [1–5] because of their ability to retrieve "desirable" objects that are not worse than any other object in the database. Recently, these queries have been applied to spatial data, as we illustrate with the example below.

Consider a hotel search scenario for a conference trip to Minneapolis, where the user marks two locations of interest, e.g., the conference venue and an airport, as Fig. 1(a) illustrates. Given these two query locations, one option is to identify hotels that are close to both locations. When considering the Euclidean distance, we can say that hotel H5, located in the middle of the two query points, is more desirable than H4, i.e., H5 "dominates" H4. The goal is to narrow down the choice of hotels to a few desirable hotels that are not dominated by any other objects, i.e., no other object is closer to all the given query points simultaneously.

However, as Fig. 1(b) shows, considering these query and data points on the map, the Euclidean distance, quantifying the length of the line segment between H5 and the query points, does not consider the road constraints and thus severely underestimates the actual distance.

Going back to Fig. 1(a), we can now assume that the dotted lines represent the underlying road network and revisit the problem to identify desirable objects with respect to L_1 distance. In this new problem, H4 and H5 are equally desirable, as both are three blocks away from the conference venue and two blocks from the airport.

In general, the Manhattan distance, or L_1 distance, reflects actual road network distances well for well-connected

☆ Work by Son and Ahn was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2011-0030044). Work by Hwang was supported by Microsoft Research Asia.

* Corresponding author. Tel.: +82 542792387.
E-mail addresses: mnbiny@postech.ac.kr (W. Son), swhwang@postech.ac.kr (S.-w. Hwang), heekap@postech.ac.kr (H.-K. Ahn).

0306-4379/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2013.10.001

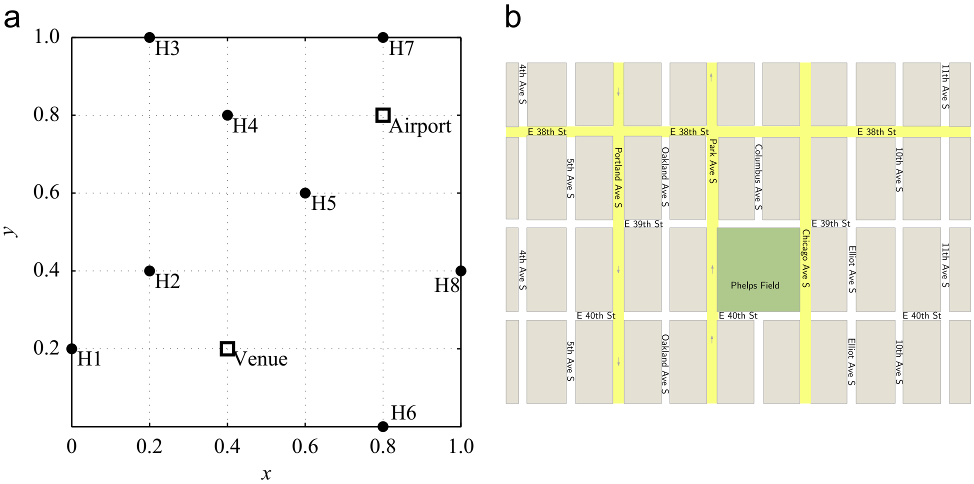

Fig. 1. Hotel search scenario: (a) abstract view and (b) actual Minneapolis map.

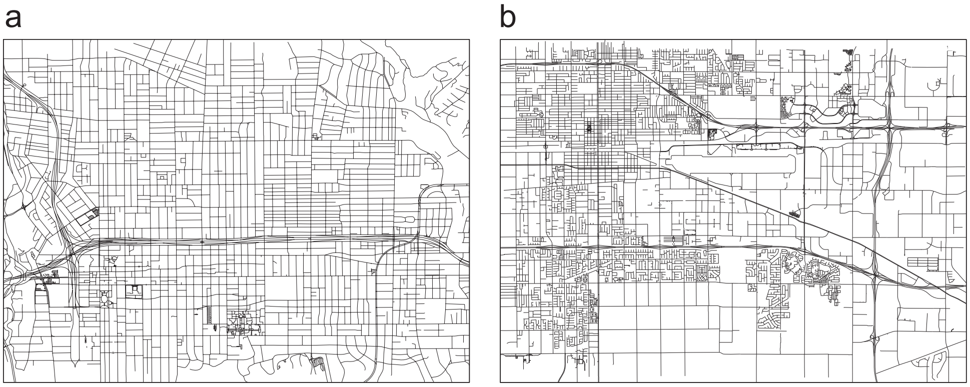

Fig. 2. Road network maps of (a) Pasadena and (b) Ontario, California.

Table 1
Inversion ratios of four real road networks.

Road network	Inversion ratio (%)
City of Pasadena	4.13
City of Ontario	4.61
City of San Joaquin County	6.13
California	6.78

metro areas such as Pasadena and Ontario (Fig. 2) in California. The experimental results for real road networks, summarized in Table 1, support this claim.[1] In the experiment, we repeated the following 1000 times for each network. We chose a node randomly and constructed two sorted lists of the nodes of the network, one in the ascending order of network distance and the other in the ascending order of L_1 distance from the chosen node. Then we counted the number of *inversions* between the two lists. Table 1 shows the average inversion ratio of each road network, which is less than 7%. For Pasadena and Ontario, the inversion ratios are even less than 5%.

Skyline queries have been actively studied for Euclidean distance [6–9]. Given a set P of data points and a set Q of query points in the plane, the most efficient algorithm known so far has the time complexity of $O(|P|(|S|\log|\mathcal{CH}(Q)|+\log|P|))$ [8,9]. Here S denotes the set of spatial skyline points, and $\mathcal{CH}(Q)$ denotes the standard *convex hull* of Q in the underlying metric. These algorithms are based on a geometric interpretation of spatial dominance of a point p over another point p': p is not spatially dominated by p' if and only if there is at least one query point in the side of the bisecting line of p and p' that contains p. From this observation, they showed that every data point p lying in $\mathcal{CH}(Q)$ is a skyline point, because there is at least one query point in the side of the line bisecting p and any other data point that contains p. They also showed, using a similar argument, that a site of the Voronoi diagram of P is a skyline point if its Voronoi cell makes nonempty intersection with $\mathcal{CH}(Q)$.

The geometric interpretation of spatial dominance also holds for L_1, because the bisecting line of two points p and p' in L_1 norm distance is the set of points at equidistance from p and p', and therefore there is at least one query

[1] First two road networks are from OpenStreetMap (http://www.openstreetmap.org/), and the other two road networks are available at http://www.cs.fsu.edu/~lifeifei/SpatialDataset.htm.