Information Systems 34 (2009) 328–352

Contents lists available at ScienceDirect

Information Systems

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

Node and edge selectivity estimation for range queries in spatial networks

E. Tiakas, A.N. Papadopoulos *, A. Nanopoulos, Y. Manolopoulos

Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 24 March 2008
Received in revised form
1 September 2008
Accepted 9 September 2008
Recommended by: F. Korn

Keywords:
Spatial networks
Selectivity estimation
Query optimization

ABSTRACT

Modern applications requiring spatial network processing pose several interesting query optimization challenges. Spatial networks are usually represented as graphs, and therefore, queries involving a spatial network can be executed by using the corresponding graph representation. This means that the cost for executing a query is determined by graph properties such as the graph order and size (i.e., number of nodes and edges) and other graph parameters. In this paper, we present novel methods to estimate the number of nodes and edges in regions of interest in spatial networks, towards predicting the space and time requirements for range queries. The methods are evaluated by using real-life and synthetic data sets. Experimental results show that the number of nodes and edges can be estimated efficiently and accurately, with relatively small space requirements, thus providing useful information to the query optimizer.

1. Introduction

Spatial networks can be represented as graphs, where road segments are represented by graph edges and crossroads (and other points of interest) are represented by graph vertices. Depending on the application, such a graph may be weighted, directed or undirected. This way, any spatial query on the original network can be executed on the underlying graph G. Evidently, the performance of such queries depends on the number of nodes and edges found in the region of interest, which defines a subgraph of G, as well as on the number of objects lying on the edges.

Several query processing techniques have been proposed for fundamental query types in spatial networks, such as range and k-nearest-neighbors (k-NN) [1–3]. However, when such queries are combined, appropriate query optimization techniques are necessary to increase efficiency. Therefore, estimations on factors affecting the performance of such queries are crucial for query optimization purposes. More specifically, the number of vertices and edges contained in a specific region is an indication of the required computational time required to store the corresponding subgraph, as well as the time required to execute queries.

This paper is a first effort towards a comprehensive study in estimating the number of vertices and edges contained in a region of a spatial network. This region of interest is defined by a starting vertex v_0 and a network-based distance e. The goal is to estimate the number of nodes and edges contained in the region of interest as accurately as possible. Such an estimation is useful in several cases such as:

- In location-based services, it is important to predict the trajectory of a moving object [4], taking into account that the motion is not predefined (e.g., a bus). The estimated number of junctions and street segments that a vehicle may visit provides significant help towards this direction, since they provide an indication regarding the size of uncertain region.

* Corresponding author. Tel.: +30 231 099 1918; fax: +30 231 099 1913.
E-mail addresses: tiakas@csd.auth.gr (E. Tiakas), papadopo@csd.auth.gr (A.N. Papadopoulos), ananopou@csd.auth.gr (A. Nanopoulos), manolopo@csd.auth.gr (Y. Manolopoulos).

0306-4379/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2008.09.004

- By using the estimated number of nodes and edges combined with information regarding the current positions of moving objects one can estimate the number of objects lying in a particular distance from a query object. This information can be used for query optimization purposes in spatiotemporal query processing, since it may be used for estimating the selectivity of range and k-NN queries as well as estimating the I/O cost of the query.
- In some spatiotemporal applications, there is a need for continuous evaluation of queries. As an example, consider a moving object for which we need to continuously monitor the set of objects residing within a specific distance from the query object. Our methods may be applied in such cases towards estimating the size of the result for each timestamp.

In this respect, several different directions are examined, each one with different requirements and estimation accuracy. The examined methods are: (i) a simple method based on multi-dimensional scaling (MDS), (ii) an estimation with global parameters, (iii) a local estimation using densities or kernels, and (iv) an estimation with binary encoding techniques.

The rest of this paper is organized as follows. In Section 2, we present the related work and specify our contributions. In Section 3, we formulate the problem and present the estimation methods in detail. Section 4 contains the performance evaluation and related experimental results. Section 5 presents a discussion for applications of the proposed methods, whereas Section 6 concludes our work.

2. Related work and contribution

Selectivity estimation has been examined in the past with respect to spatial or spatiotemporal queries. Below, we briefly present some fundamental work in the area.

In [5], the authors examine the performance of range queries in R-trees and variants. More specifically, estimation formulae have been proposed for the number of disk accesses using global parameters and local densities. Selectivity estimation for spatial joins has been studied in [6], where the authors provide efficient methods with relative errors below 30%. In [7] the authors propose two approaches for the selectivity estimation of spatiotemporal queries: a simple histogram approach and an index-based estimator. The Power-Method [8] provides accurate estimations for such queries by using a simple formula with minimal computational cost, small space requirements and average relative error rate below 20%. In [9] the authors propose a selectivity estimation method with low relative estimation error (about 10%) for spatial queries using specific global parameters formulae based on Hausdorff fractal dimension.

The concept of local density has been studied extensively in general and spatial data sets, but not in combination with spatial networks. In the bibliography, one can find a plethora of density estimation proposals. In this direction, an important method is the kernel density estimation method and its variants [10,11]. Kernel density estimators are used in many application domains such as clustering [12,13], outlier detection [14], and visualization [15]. Moreover, several variations of kernel density estimations and smoothing have been proposed in [16,17].

The basic limitation of the previous approaches is that they are restricted to Euclidean spaces only. Our contribution is the presentation of efficient methods for spatial query estimation for spatial networks assuming non-Euclidean spaces. More specifically, we introduce and evaluate three novel estimation methods:

Global parameters estimation method: Which is based on global parameters (see Section 3.3).

Local densities estimation method: Which extends the previous method by using local density factors (see Section 3.4). We present a new computational model of local node densities in Section 3.4.1, as well as an alternative approach by applying the well-known Gaussian kernel densities estimators in Section 3.4.2.

Binary encoding estimation method: Which uses specific graph transformations, a specific binary encoding technique and a formula with only binary and basic register operations for calculations (Section 3.5).

In addition to these three methods, a simple solution based on MDS is also evaluated. The advantage of this approach is that it can exploit previous results for the selectivity estimation of multi-dimensional objects using the Euclidean distance.

A preliminary version of this work appears in [18] where we have presented the basic estimation methods. The current version is more complete and in summary the new material is described as follows:

- the MDS method is included for comparison purposes,
- a more thorough theoretical analysis is performed and proofs of fundamental theoretical results are given,
- estimation of the number of edges contained in the region of interest is given (in addition to the estimation of the number of nodes),
- a discussion regarding the exploitation of the results by the query optimizer for range and k-NN processing is included, studying selectivity estimation issues and processing cost with respect to the number of I/O operations required,
- a more thorough experimental evaluation is carried out.

3. Estimation approaches

In this section, we define the problem and present the proposed methods aiming at effective estimation solutions. Table 1 contains the basic symbols used.

3.1. Problem definition

Let $G(V_G, E_G)$ be a connected weighted undirected graph where V_G and E_G is the set of nodes and edges, respectively. The distance measure $d(v,u)$ denotes the shortest path distance between nodes v and u. Given a specific starting node $v_0 \in V_G$, and a desired distance e, we