Neurocomputing 173 (2016) 118–123

Contents lists available at ScienceDirect

Neurocomputing

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

Finding regions of interest using location based social media

Shuo Shang [a], Danhuai Guo [b], Jiajun Liu [c], Kai Zheng [d], Ji-Rong Wen [e]

[a] China University of Petroleum, Beijing, China
[b] CNIC, Chinese Academy of Sciences, Beijing, China
[c] CSIRO, Brisbane, Australia
[d] The University of Queensland, Brisbane, Australia
[e] Renmin University of China, Beijing, China

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form
18 May 2015
Accepted 3 June 2015
Available online 20 August 2015

Keywords:
Region of interest
Distance
Density
Location based social media

ABSTRACT

The discovery of regions of interest in city groups is increasingly important in recent years. In this light, we propose and investigate a novel problem called Region Discovery query (RD query) that finds regions of interest with respect to a user's current geographic location. Given a set of spatial objects O and a query location q, if a circular region ω is with high spatial-object density and is spatially close to q, it is returned by the query and is recommended to users. This type of query can bring significant benefit to users in many useful applications such as trip planning and region recommendation. The RD query faces a big challenge: how to prune the search space in the spatial and density domains. To overcome the challenge and process the RD query efficiently, we propose a novel collaboration search method and we define a pair of bounds to prune the search space effectively. The performance of the RD query is studied by extensive experiments on real and synthetic spatial data.

1. Introduction

With the rapid development of GPS-equipped mobile device (e.g., smart phones, car navigation systems, and PDAs) and online map services (e.g., Google Maps[1] and Bing Maps[2]), people can easily acquire their current geographic location in real time and can retrieve spatial information relevant to their trips [6]. In this light, we propose and study a novel query called Region Discovery query (RD query) that identifies regions of interest with respect to a user's current geographic location. Spatial objects can be geo-tagged tweets and micro-blog posts from location-based social media, such as Twitter,[3] Weibo,[4] and Foursquare[5][3,27].

A region ω is a circular area defined by a center point $\omega.c$ and a radius $\omega.r$. We define two thresholds to identify qualified regions: (1) a size threshold $\tau.s$, i.e., a region should contains at least a threshold number of spatial objects; (2) a radius threshold $\tau.r$, i.e., the radius of the region should not exceed a radius threshold. Given a query location q, if a region ω is with dense spatial-object distribution and is spatially close to q, the region ω is returned by the query and is recommended to users. The region discovery query (RD query) is useful in many popular mobile applications such as trip planning and location recommendation. For example, when traveling overseas, travelers may wish to know about regions of interest (e.g., commercial districts and dining areas) around him/her. Intuitively, regions with high spatial-object density (e.g., geo-tagged tweets, micro-blog posts, and points of interest) are assumed to be more attractive to users. Also, a region located close to a user's current location is more attractive than a far-away region.

To the best of our knowledge, this is the first work that study the region recommendation problem while taking both spatial distance and spatial-object density into account. Previous studies (e.g., nearest neighbor query [10,9,11,23]) only use spatial distance as the sole factor when computing the query results. In contrast, the RD query takes both spatial distance and density distribution into account. A linear combination method [18,17] is adopted to combine the spatial and density domains.

An example of the RD query is shown in Fig. 1. Here, q is a query point, and vertices p_1, p_2, and p_3 are the center points of regions ω_1, ω_2, and ω_3, respectively. The distance between a region and a query point is defined by the shortest network distance between the region center point and the query point (e.g., $dist(\omega_2, q) = dist(p_2, q)$). If considering the spatial distance only (e.g., the same as the nearest neighbor query [9]), ω_2 is the region closest to q. However, when considering both spatial distance and the density of spatial objects, ω_2 is less attractive

E-mail address: jedi.shang@gmail.com (S. Shang).

[1] http://maps.google.com/
[2] http://www.bing.com/maps/
[3] https://twitter.com/
[4] http://weibo.com/
[5] https://foursquare.com/

http://dx.doi.org/10.1016/j.neucom.2015.06.086
0925-2312/

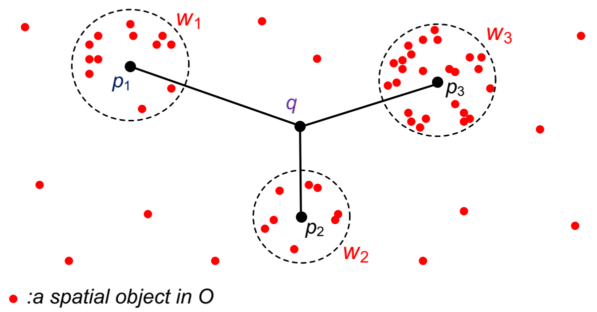

Fig. 1. An example of region discovery query.

than ω_3 because of its sparser spatial-object distribution. Although ω_3 is not as good as ω_2 according to spatial distance, we still consider ω_3 as the best choice for region recommendation when taking both spatial distance and spatial-object density into account.

The RD query is applied in spatial networks, since in a large number of practical scenarios, travelers move in spatial networks (e.g., road networks) rather than in a Euclidean space [25,26,19,21,20]. To enable efficient processing of the RD query, for each vertex p, we pre-compute the number of spatial objects that are covered by a circular region defined by $(p, \tau.r)$, where p is the center point and $\tau.r$ is the radius. These counts are useful in pruning the search space during query processing. Based on the pre-computation results, we develop an adaptive collaboration algorithm to compute the RD query efficiently. The search process is conducted in the spatial and density domains concurrently. A pair of upper and lower bounds are defined to prune the search space.

To the best of our knowledge, there is no existing approach that can compute the RD query efficiently. To sum up, the main contributions of this work are as follows:

- We define a novel region discovery (RD) query, and it is useful in many mobile applications such as trip planning and location recommendation.
- We propose a set of new metrics to evaluate the distance-and-density score of regions.
- We develop an adaptive collaboration algorithm to compute the RD query efficiently.
- We conduct extensive experiments on real and synthetic data to study the performance of RD query.

The rest of the paper is organized as follows. Related work is covered in Section 2. Section 3 introduces spatial networks and the distance metrics used in the paper; and it also gives problem definitions. The collaboration search method is covered in Section 4, which is followed by the experimental results in Section 5. Conclusions are drawn in Section 6.

2. Related work

Spatial queries in advanced traveler information system continue to proliferate in recent years. Nearest Neighbor (NN) query is considered as an important issue in such kind of applications. This kind of query aims to retrieve the closest neighbor to a query point from a set of given objects. Based on different constraint conditions, NN query processing can be classified into three categories, such that in Euclidean spaces (e.g. [14,7]), in spatial networks (e.g. [10,9,11,23,22]), and in higher dimensional spaces (e.g. [8,2]).

As a variant of NN queries, Continuous Nearest Neighbor queries (CNN) [1,13,24,15,16] report the kNN results continuously while the user is moving along a path. This type of queries aims to find the split points on the query path where an update of the kNN is required, and thus to avoid unnecessary re-computation. In [13], Mouratidis et al. investigate the CNN monitoring problem in a road network, in which the query point moves freely and the data objects' positions are also changing dynamically. The basic idea of [13] is to maintain a spanning tree originated from the query point and to grow or discard branches of the spanning tree according to the data objects and query point's movements.

To the best of our knowledge, there is no existing approach that can compute the RD query efficiently. Previous studies only use spatial distance as the sole factor when computing the query results, and the density of spatial objects is not taken into account. In contrast, the RD query takes both spatial distance and density distribution into account. A linear combination method [18,17] is adopted to combine the spatial and density domains.

3. Preliminaries

3.1. Network modeling and preprocessing

In this work, spatial networks are modeled by connected and undirected planar graphs $G(V, E)$, where V is the set of vertices and E is the set of edges. A weight can be assigned to each edge to represent its length or application specific factors such as traveling time obtained from historical traffic data [5]. Given two points a and b in road networks, the network distance between them is the length of their shortest network path (i.e., a sequence of edges linking a and b where the accumulated weight is minimal). The data points are distributed along roads and if a data point is not located at a road intersection, we treat the data point as a vertex and further divide the edge that it lies on into two edges. Thus, we assume that all data points are in vertices for the sake of clarity. We assume that each spatial object (e.g., geo-tagged tweets, geo-tagged photos) is attached to its nearest vertex. For each vertex $p \in G.V$, the number of spatial objects that are attached to p is maintained as an attribute of p, denoted by $p.g$. A vertex and its attached spatial objects make up the minimum unit in spatial-object density computations, and thus we do not need to access individual spatial objects during RD query processing.

3.2. Problem definition

A region ω is a circular area defined by a center point $\omega.c$ and a radius $\omega.r$. We define two thresholds to identify qualified regions: (1) a size threshold $\tau.s$, i.e., a region should contains at least a threshold number of spatial objects; (2) a radius threshold $\tau.r$, i.e., the radius of the region should not exceed a radius threshold. A region ω has an associated subgraph $\omega._G$, which contains the vertices $\omega._V$ and edges $\omega._E$ from G that are in the circular region.

The density $\omega.\rho$ of region ω is defined as

$$\omega.\rho = \frac{\sum_{p \in \omega._V} p.g}{\sum_{e \in \omega._E} W(e)}, \tag{1}$$

where p is a vertex in $\omega._V$ and $p.g$ is the number of spatial objects that are attached to p; and e is an edge in $\omega._E$ and $W(e)$ is its weight.

Given a region ω and a query point q, a distance function $E_s(\omega, q)$ and a density function $E_d(\omega)$ are defined in Eqs. (2) and (3), respectively. We use Sigmoid function [12] to normalize the values of $E_s(\omega, q)$ and $E_d(\omega)$ to [0, 1].

$$E_s(\omega, q) = \frac{2}{1+e^{-sd(\omega.c,q)}} - 1 \tag{2}$$