Knowledge-Based Systems 93 (2016) 121–134

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Efficient reverse spatial and textual k nearest neighbor queries on road networks

Changyin Luo [a], Li Junlin [b], Guohui Li [a], Wei Wei [a,*], Yanhong Li [c], Jianjun Li [a]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[b] Wuhan Digital Engineering Institute, Wuhan, China
[c] Department of Computer Science, South-Central University for Nationalities, Wuhan, China

ARTICLE INFO

Article history:
Received 14 July 2015
Revised 30 October 2015
Accepted 6 November 2015
Available online 27 November 2015

Keywords:
Spatial keyword query
Reverse k nearest neighbor
Road network
Network voronoi diagram

ABSTRACT

The proliferation of geo-positioning technologies boosts the prevalence of GPS-enabled devices, and thus many spatial-textual objects that possess both text descriptions and geo-locations are extensively available in reality. Hence, how to efficiently exploit both spatial and textual description of objects to a spatial keyword query (SKQ) has increasingly become a challenging problem. Previous studies on SKQ problem usually focus on Euclidean space. In the real world, however, most of the spatial-textual objects lie on road networks. This paper takes the first step to investigate a novel problem, namely, reverse spatial and textual k nearest neighbor (RSTkNN) queries on road networks. We formalize the RSTkNN queries and present several spatial keyword pruning methods to accelerate the query processing. Then two effective verifying techniques are proposed, which can be seamlessly integrated into our RSTkNN query procedure. Finally, comprehensive experiments on real-world and synthetic data sets are conducted to demonstrate the performance of our approaches.

1. Introduction

With the rapid development of mobile portable devices and location positioning technologies, a large number of user locations are shared on various social platforms, such as Facebook, Twitter, Foursquare, Flickr and Gowalla. Meanwhile, increasing volumes of geo-textual objects that represent Point-of-interests (POIs, e.g., shopping mall, hotel or restaurant) are gaining in prevalence. Generally, a geo-textual object contains a geographical location (i.e., longitude, latitude) and a textual description (e.g., features, reviews, facilities). The massive amount of available geo-textual data enables users to retrieve a set of objects that best matches the user's submitted spatial keyword query (i.e., SKQ, which includes a geographical location and a set of keywords), in terms of both spatial proximity to query location and textual relevance to query keywords.

Reverse k Nearest Neighbor (RkNN) [1] query, which aims to find a set of objects that take the query as one of their kNN based on the spatial distance, has been studied extensively (e.g., [1–12]) over the past decade, due to its importance in a wide range of applications, such as location based service, resource allocation, marketing and decision support, profile-based management, etc. These traditional studies on the retrieval of RkNN only consider *spatial distance* as a unique influence factor. However, in real-world applications, both *spatial distance* and *textual relevance* should be taken into account. For example, if one plans to select a location from a given set of potential locations for establishing a new facility (e.g., restaurant, hospital, supermarket), a better choice might be choosing a location that could minimize the average distance among customers, and meanwhile have less textual relevance with their competitors. As another example, assume the customers specify their procurement plans via a set of keywords (e.g., computer, printer, fax) and their locations, a shopping mall can pose an RSTkNN query to find the potential buyers (customers) whose keywords are relevant to that of the shopping mall and meanwhile have the shopping mall as one of their k nearest neighbor.

In recent years, SKQ has become an active topic in database community. Most of the existing studies on SKQ are restricted to Euclidean space [13–22]. According to [23], previous works mainly focus on three types of SKQ in Euclidean space, i.e., Boolean range queries (BRQ) [24,25], Boolean kNN queries (BkQ) [17,26] and Top-k kNN queries (TkQ) [13,14,21,22,27]. Nevertheless, in reality, the position and accessibility of spatial-textual objects are constrained by network connectivity, and spatial proximity should be determined by the shortest path distance rather than Euclidean distance. Recently, spatial keywords queries on road networks have drawn increasing attention. Rocha et al. [28] pioneer TkQ queries on road networks. Range-constrained spatial keyword queries on road networks have

* Corresponding author. Tel.: +8618827440253.
E-mail address: weiwei8329@gmail.com, weiw@hust.edu.cn (W. Wei).

http://dx.doi.org/10.1016/j.knosys.2015.11.009
0950-7051/

been described in [29]. Guo et al. [30] propose a safe segment to continuously monitor T*k*Q queries on road networks. In order to obtain a spatially diversified SKQ result, diversified spatial keywords search (DSKQ) on road networks is investigated in [31], in which a signature-based inverted indexing and an incremental network expansion method are developed for DSKQ search. Gao et al. [32] design an innovative *count-tree* to study reverse top-*k* Boolean spatial keyword (R*k*BSK) retrieval on road networks. In their work, several novel pruning heuristic methods are developed to facilitate R*k*BSK queries processing, but they can only process Boolean spatial keyword query.

Although the traditional R*k*NN queries have been particularly well studied, they only focus on spatial location but ignore text (keywords) relevance. Recently, Lu et al. [33] first take the textual relevance into consideration for R*k*NN queries in Euclidean space. Albeit they design a branch-and-bound search algorithm based on an innovative index called IUR-tree (Intersection-Union R-tree) in their work, their approach cannot be employed to handle RST*k*NN queries on road networks. The key reason is that, IUR-tree is a combination of textual vectors and R-tree that is constructed in Euclidean space, while spatial distance between two objects on road networks should be evaluated by the shortest path distance rather than Euclidean distance. Hence, the pruning methods designed based on IUR-tree cannot work on road networks. As a result, their branch-and-bound search framework cannot be adopted to solve RST*k*NN queries on road networks.

In this work, we investigate RST*k*NN queries on road networks, which pose significant challenges to the existing approaches for processing both conventional RkNN queries (without taking textual relevance into account) and RST*k*NN queries in Euclidean space (its computation cost for the spatial proximity is much lower than that in road networks). Furthermore, RST*k*NN queries in our work belong to *score based* spatial keywords queries. Therefore, the techniques concerning the Boolean SKQ cannot be employed to solve our problem directly.

The contributions of this paper can be summarized as follows:

- We formalize reverse spatial and textual *k* nearest neighbor (RST*k*NN) queries on road networks, and identify the problem of RST*k*NN retrieval. To the best of our knowledge, this is the first work on RST*k*NN queries on road networks.
- We describe several pruning methods to prune non-promising objects at a low cost. The first verifying algorithm in our solutions is based on the network-expansion. In order to avoid expanding road networks multiple times as the first method does, we take advantage of Network Voronoi Diagram (NVD) to develop the second algorithm to obtain RST*k*NN results in an efficient way.
- Comprehensive experiments on real-world and synthetic datasets demonstrate the effectiveness and efficiency of our approach.

The rest of this paper is organized as follows. Section 2 reviews related work and Section 3 gives preliminaries and describes index structure. In Section 4, a basic approach is described. Two efficient algorithms for RST*k*NN queries are developed in Section 5. The experimental results are demonstrated in Section 6. Section 7 makes the conclusion.

2. Related work and background

2.1. RkNN queries on road networks

Korn et al. [1] are the pioneers who first research on RNN queries. They answer RNN query by pre-calculating and adopt three phases, namely pruning, containment and verification, to obtain the final results. After that, numerous literatures concerning the variants of RNN queries in Euclidean space have been particularly well studied [1–10,12]. The following will make an overview of RNN queries on road networks. The snapshot RNN queries in spatial networks are first discussed by Safar et al. [34], in which NVD is utilized to efficiently process RNN queries. In a following work [35], they extend their approach to answer RkNN queries in spatial networks. Sun et al. [36] study the continuous monitoring of RNN queries on road networks. Li et al. [37] design a novel DLM-tree that represents the whole monitoring area of a continuous RkNN (CRkNN) queries to explore CRkNN queries on road networks. Cheema et al. [11] employ a filter and refinement technique to first study CRkNN retrieval (monochromatic and bichromatic) in spatial networks where both the objects and queries continuously change their locations.

2.2. Spatial keyword queries

Retrieving geo-textual objects with query location and keywords has gained increasing attention recently for the popularity of location-based services. There are two types of SKQ, namely, Boolean SKQ and score-based SKQ. The Boolean SKQ is to find the *k* objects nearest to the query *q* among a set of objects whose keyword set covers the query keywords. While score-based SKQ is to obtain the results according to score evaluated by a ranking function that takes into account the spatial proximity and text relevancy (e.g. Eq. (1)). Comparing with Boolean SKQ, it is much more expensive to obtain a score-based SKQ result. A comprehensive experimental evaluation of different SKQ indexing and query processing techniques have been surveyed in [23]. Several geo-textual indices have been developed to efficiently answer T*k*Q, such as IR^2-tree [17], IR-tree [13], S2I [38], I^3 [39] and IL-Quadtree [19]. Top-*k* spatial keyword queries on trajectories are first investigated in [40], in which *k* trajectories whose text descriptions cover the keywords given by the user and that have the shortest match distance are found out. In order to preserve user privacy in text-based search, Wang et al. [41] propose a new dummy query generation method (called HDGA) to deal with various attacks discussed in their work. Literatures [42–44] study *closet keywords search* (Keyword Cover), which retrieves objects that should cover a set of query keywords and have the minimum inter-objects distance. Motivated by the observation of increasing availability and importance of keyword rating in decision marking, Deng et al. [45] investigate a generic version of closet keyword search (called Best Keyword Cover) which considers inter-objects distance as well as the keyword rating of objects. Sometimes, users may wonder why some known object is unexpectedly missing from a result when a SKQ is issued, [46] takes the lead in exploring how to answer why-not questions on spatial keyword top-*k* queries using query refinement. Wang et al. [47] propose a novel adaptive spatial textual partition index (AP-Tree) to support continuous spatial keyword queries over stream. Moreover, many variants of SKQ have been developed such as direction-aware SKQ [48], interactive Top-*k* spatial keyword (IT*k*SK) query [49], temporal spatial-keyword Top-*k* publish/subscribe (TaSK) query [50], approximate keyword query of sematic trajectory [51] and so on. However, all the methods mentioned above cannot be employed to support RST*k*NN retrieval.

3. Preliminaries

In this section, the problem of RST*k*NN queries on road networks as well as the necessary definitions is formally given in Section 3.1, followed by the indexing architecture in Section 3.2.

3.1. Problem definition

Road networks. We model a road network as a weighted graph $G = (V, E, W)$, where V is the set of vertices (i.e., road conjunctions or road borders), E is the set of edges, and W is the set of weights (network distance) that are associated with each edge. Without loss of generality, we assume bidirectional traffic which is pervasive in real life. Unidirectional traffic is also supported by our approach.