Density-Based Top-K Spatial Textual Clusters Retrieval

Dingming Wu, Ilkcan Keles, Song Wu, Hao Zhou, Simonas Šaltenis, Christian S. Jensen, *Fellow, IEEE*, and Kezhong Lu

Abstract—So-called spatial web queries retrieve web content representing points of interest, such that the points of interest have descriptions that are relevant to query keywords and are located close to a query location. Two broad categories of such queries exist. The first encompasses queries that retrieve single spatial web objects that each satisfy the query arguments. Most proposals belong to this category. The second category, to which this paper's proposal belongs, encompasses queries that support exploratory user behavior and retrieve sets of objects that represent regions of space that may be of interest to the user. Specifically, the paper proposes a new type of query, the top-k spatial textual cluster retrieval (k-STC) query that returns the top-k clusters that (i) are located close to a query location, (ii) contain objects that are relevant with regard to given query keywords, and (iii) have an object density that exceeds a given threshold. To compute this query, we propose a DBSCAN-based approach and an OPTICS-based approach that rely on on-line density-based clustering and that exploit early stop conditions. Empirical studies on real data sets offer evidence that the paper's proposals can find good quality clusters and are capable of excellent performance.

Index Terms—Clustering, query processing, indexing, data management

1 Introduction

Spatial keyword query processing [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] allows users to receive answers to geographically constrained queries that take into account information about "what" the user is searching for as expressed by keywords. For instance, a query may request a "good microbrewery that serves pizza" that is close to the user's hotel. In general, a spatial keyword query retrieves a set of spatial web objects that are located close to the query location and whose text descriptions are relevant to the query keywords. Fig. 1a illustrates a query q (black dot) with keywords 'outdoor seating' that requests the top-5 restaurants (red squares) in London from TripAdvisor[1] based on a ranking function that is a weighted sum of spatial distance and text relevance. Several spatial keyword query variants have been studied. Proposals differ in terms of the query arguments and in how the objects matching the query arguments are found and ranked. A continuously moving top-k spatial keyword query [9], [13] requests the up-to-date result while the query location changes continuously. A location-aware top-k prestige-based text retrieval query [14] gives high rankings to the objects that have relevant surrounding objects. A collective spatial keyword query [15] retrieves a set of objects that taken together best match the query arguments. Spatial keyword queries are also investigated in road networks [16], [17], and Li *et al.* [18] study spatial keyword search constrained by a movement direction.

Most studies consider the retrieval of one or more objects, each of which satisfies the query. However, in some use cases, users may be interested in regions with many objects that satisfy query parameters rather than in a set of objects scattered in space. For instance, a user may prefer to visit one nearby shopping area to explore multiple outlets selling jeans, rather than visiting one jeans shop in zone A and another in zone B. Such functionality is also useful for a marketing manager, who wishes to get an overview of the locations of coffee shops in a central business district. In addition, similar businesses are often located close to each other and form small regions, such as shopping, dinning, and entertainment areas, to attract customers [19]. Some previous studies [20], [21], [22] consider co-location relationship between objects and retrieve regions so that the total weights of objects inside the regions are maximized. However, these studies are limited regarding the shapes of the regions retrieved, such as a fixed-size rectangle or a circle. A recent study [23] requests the length-constrained maximum-sum region of interest where the road network distance between objects is less than a query constraint and the sum of the ranking scores of the objects inside the region is maximized. This kind of query may still retrieve a region containing many objects with low ranking scores and may ignore promising regions with few objects with high ranking scores. Also a query range must be specified, which helps reduce the search space.

1. http://www.tripadvisor.com.

- *Dingming Wu, Song Wu, Hao Zhou, and Kezhong Lu are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. E-mail: {dingming, kzlu}@szu.edu.cn, {wusong2018, zhouhao2017}@email.szu.edu.cn.*
- *Ilkcan Keles is with Turkcell, Maltepe, Turkey, and also with the Department of Computer Science, Aalborg University, 9220 Aalborg, Denmark. E-mail: ilkcan.keles@turkcell.com.tr.*
- *Simonas Šaltenis and Christian S. Jensen are with the Department of Computer Science, Aalborg University, 9220 Aalborg, Denmark. E-mail: {simas, csj}@cs.aau.dk.*

Manuscript received 30 July 2020; revised 26 November 2020; accepted 3 January 2021. Date of publication 11 January 2021; date of current version 6 October 2022.
(Corresponding author: Kezhong Lu.)
Recommended for acceptance by C. Li.
Digital Object Identifier no. 10.1109/TKDE.2021.3049785