Cost-Aware and Distance-Constrained Collective Spatial Keyword Query (Extended Abstract)

Harry Kai-Ho Chan[*], Shengxin Liu[†], Cheng Long[‡], Raymond Chi-Wing Wong[§]
[*]*Roskilde University, Denmark*, kai-ho@ruc.dk
[†]*Harbin Institute of Technology (Shenzhen), Shenzhen, China*, sxliu@hit.edu.cn
[‡]*Nanyang Technological University, Singapore*, c.long@ntu.edu.sg
[§]*The Hong Kong University of Science and Technology, Hong Kong*, raywong@cse.ust.hk

2022 IEEE 38th International Conference on Data Engineering (ICDE) | 978-1-6654-0883-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDE53745.2022.00129

***Abstract*—With the proliferation of location-based services, geo-textual data is becoming ubiquitous. Objects involved in geo-textual data include geospatial locations, textual descriptions or keywords, and various attributes (e.g., a point-of-interest has its expenses and users' ratings). Many types of spatial keyword queries have been proposed on geo-textual data. Among them, one prominent type is to find, for a query consisting of a query location and some query keywords, a set of multiple objects such that the objects in the set collectively cover all the query keywords, and the object set is of good quality according to some criterion. Existing studies define the criterion either based on the geospatial information of the objects solely, or simply treat the geospatial information and the attribute information of the objects together without differentiation though they may have different semantics and scales. As a result, they cannot provide users flexibility to express finer grained preferences on the objects. In this paper, we propose a new criterion which is to find a set of objects where the distance (defined based on the geospatial information) is at most a threshold specified by users and the cost (defined based on the attribute information) is optimized. We develop a suite of two algorithms including an exact algorithm and an approximation algorithm with provable guarantees for the problem. We conducted extensive experiments on both real and synthetic datasets, which verified the efficiency and effectiveness of proposed algorithms.**

I. Introduction

Nowadays, geo-textual data which refers to data with both spatial and textual information is ubiquitous. Some examples of geo-textual data include the points-of-interests (POIs) in the physical world (e.g., restaurants, shops and hotels), geo-tagged web objects (e.g., webpages and photos at Flicker), and geo-social networking data (e.g., users of FourSquare have their check-in histories which are spatial and also have their profiles which are textual). Entities of geo-textual data, which we call geo-textual objects, are associated with various attributes as well. For example, a restaurant is usually associated with some expense attribute (e.g., in the Yelp APP, this information is shown by the number of "$" symbols).

Given a database of geo-textual objects, a popular query called *Collective Spatial Keyword Queries* have been proposed, which is to search for a set of multiple objects that collectively cover all query keywords and are desirable to the user based on some criterion [2], [6], [1], [5]. If an object set covers all query keywords, such an object set is said to be a *feasible set*. There are usually many feasible sets given a query - each combination of objects covering different query keywords would be a feasible set. Therefore, a key question that needs to be answered is that among all possible feasible sets, which one should be returned, i.e., what criterion should be used for picking a feasible set?

Most studies define the criterion based on the geospatial aspects of the objects solely [2], [6], [1], [5]. While this criterion would find a feasible set with some notation of distance optimized, it pays no attention to the attribute aspect of the objects and cannot guarantee the desirability of the returned feasible set in the attribute aspect. [4] defines the criterion such that it considers both the geospatial aspect and the attribute aspect of objects. While this criterion is superior over previous ones that ignore the cost part, it lacks of capability of providing users a finer grained control on their preferences on the cost part and the distance part. Moreover, a cost and a distance may have different semantics and scales, and combining them together using a product operator as adopted in [4] may cause problems such as one dimension dominates the other and essentially only one aspect is captured.

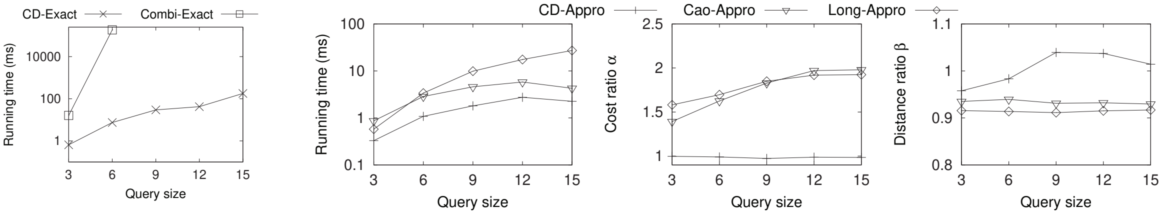
The main contribution of this paper is summarized as follows. (1) We propose a new type of query, namely *Cost-Aware and Distance-Constrained Collective Spatial Keyword Query* (CD-CoSKQ), which aims to find an object set with the smallest *cost* subject to a constraint on the *distance*. The cost is based on the attribute aspect of the objects and the distance on the geospatial aspect. This query provides users a finer grained interface to express their preferences on both the geospatial aspect and the attribute aspect of the geo-textual objects. (2) We prove the inapproximability result of the CD-CoSKQ and develop two algorithms, namely an exact algorithm *CD-Exact* and an (α, β)-approximation algorithm *CD-Appro*. (3) We conducted extensive experiments, which verified the efficiency and effectiveness of our algorithms.

II. Problem Definition

Let $\mathcal{O}$ be a set of geo-textual objects. Each object $o \in \mathcal{O}$ is associated with a location denoted by $o.\lambda$, a set of keywords denoted by $o.\psi$, and some attributes which we convert to a form of cost denoted by $o.w$ such that a lower cost is preferred. Given two objects o and o', we denote by $d(o, o')$ the Euclidean distance between $o.\lambda$ and $o'.\lambda$.

[†] Corresponding Author

2375-026X/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDE53745.2022.00129



(a) Exact Algorithms (b) Approximation Algorithms

Fig. 1. Effect of Query Size (i.e., $|q.\psi|$) ($cost_{Sum}$, $dist_{MaxSum}$, Yelp)

Problem 1 (CD-CoSKQ [3]): Given a query q which consists of a query location $q.\lambda$, a set of query keywords $q.\psi$, and a distance threshold $q.B$, the CD-CoSKQ problem is to find a set $G \subseteq \mathcal{O}$ of objects such that (1) G covers $q.\psi$, (2) the distance of G wrt q, denoted by $dist(G, q)$, is at most $q.B$, and (3) the cost of G, denoted by $cost(G)$, is minimized. □

Cost Functions. We consider two cost functions $cost(G)$.

$$cost_{Max}(G) = \max_{o \in G} o.w \qquad cost_{Sum}(G) = \sum_{o \in G} o.w$$

Distance Functions. We consider two commonly used functions as the distance function in this paper.

$$\begin{aligned} dist_{MaxSum}(G) &= \max_{o \in G} d(o, q) + \max_{o_1, o_2 \in G} d(o_1, o_2) \\ dist_{Dia}(G) &= \max_{o_1, o_2 \in G \cup \{q\}} d(o_1, o_2) \end{aligned}$$

Intractability. We have the following result.

Theorem 1: The CD-CoSKQ problem is NP-hard to approximate with any constant factor c ($c \geq 1$) [3]. □

III. Algorithms

Given a query q, an object o is said to be *relevant* if $o.\psi \cap q.\psi \neq \emptyset$, and a set S of objects is said to be a *feasible set* if S covers $q.\psi$ (i.e., $q.\psi \subseteq \cup_{o \in S} o.\psi$).

CD-Exact. Consider the optimal solution G^* for a given CD-CoSKQ q. By definition, G^* is a feasible set. Let t be a query keyword and $\mathcal{O}(t)$ be the set of objects, each of which contains t. We know that G^* must include an object in $\mathcal{O}(t)$. Motivated by this, we find G^* by searching around each object $o \in \mathcal{O}(t)$, which we call a **seed object**. To search around an object $o \in \mathcal{O}(t)$, we restrict our attention to a set S of relevant objects that are located close enough to o wrt the query distance threshold, which we call a **candidate set**, and then within S, we find the feasible set G with the smallest cost among all feasible sets involving o and having the distance at most B, which we call a **local optimal set**. At the end, we return the local optimal set which has the smallest cost among all local optimal sets found, and it is deemed to be the optimal solution.

TABLE I
SUMMARIES OF CD-APPRO

		Time Complexity	Appro. Ratio	
			α	β
$cost_{Max}$	$dist_{MaxSum}$	$O(\lvert\mathcal{O}''\rvert \cdot (\log\lvert\mathcal{O}\rvert +$	1	1.375
	$dist_{Dia}$	$\lvert S\rvert \log\lvert S\rvert + \sum_{o \in S}\lvert o.\psi\rvert))$		$\sqrt{3}$
$cost_{Sum}$	$dist_{MaxSum}$	$O(\lvert\mathcal{O}''\rvert \cdot (\log\lvert\mathcal{O}\rvert$	$O(\log\lvert q.\psi\rvert)$	1.375
	$dist_{Dia}$	$+\lvert\psi\rvert^2\lvert S\rvert))$		$\sqrt{3}$

CD-Appro. CD-Appro would output a solution G with $dist(G) \leq \beta \cdot B$ and the cost is at most α times that of the optimal solution G^*, where $dist(G^*) \leq B$. For both ratios, the smaller the value is, the better an algorithm performs. Compared to CD-Exact, CD-Appro uses a larger set of seed objects while constructing the candidate sets in a smaller region, and replaces the expensive enumeration procedure with a greedy procedure which finds an approximate feasible set that might violate the distance constraint but has both cost ratio α and distance ratio β bounded, and runs much faster. Table I shows the time complexities and the approximation ratios of CD-Appro ($|\mathcal{O}''| << |\mathcal{O}|$, $|S| << |\mathcal{O}|$ and $|\psi| < |q.\psi|$).

IV. Empirical Studies

Datasets. We used three real datasets, namely Yelp, Hotel and GN. Dataset Yelp (https://www.yelp.com/dataset/challenge) contains 192,609 POIs (i.e., objects), where each object has a location, a rating and belongs to a set of categories. It contains 788,841 words in total with 2,468 unique words. More details of the datasets could be found in [3].

Algorithms. We compare Combi-Exact (which is an exact algorithm baseline), and adaptions of Cao-Appro [1] and Long-Appro [6]. We measured the running time and the approximation ratios, i.e., the cost ratio α and the distance ratio β (for approximation algorithms only).

The results with $cost_{Sum}$ and $dist_{MaxSum}$ are presented in Figure 1. Our exact algorithm CD-Exact runs consistently faster than Combi-Exact. Our CD-Appro runs faster than the competitors, and achieves the cost ratio α very close to 1.

Acknowledgments: This research is supported by the Nanyang Technological University Start-Up Grant from the College of Engineering under Grant M4082302 and by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG20/19 (S)). This work was partially supported by the National Natural Science Foundation of China under Grant 62102117. The research of the HKUST side is supported by PRP/026/21FX.

References

[1] X. Cao, G. Cong, T. Guo, C. S. Jensen, and B. C. Ooi. Efficient processing of spatial group keyword queries. *TODS*, 40(2):13, 2015.

[2] X. Cao, G. Cong, C. S. Jensen, and B. C. Ooi. Collective spatial keyword querying. In *SIGMOD*, pages 373–384. ACM, 2011.

[3] H. K.-H. Chan, S. Liu, C. Long, and R. C.-W. Wong. Cost-aware and distance-constrained collective spatial keyword query. *TKDE*, 2021.

[4] H. K.-H. Chan, C. Long, and R. C.-W. Wong. Inherent-cost aware collective spatial keyword queries. In *SSTD*, 2017.

[5] H. K.-H. Chan, C. Long, and R. C.-W. Wong. On generalizing collective spatial keyword queries. *TKDE*, 30(9):1712–1726, 2018.

[6] C. Long, R. C.-W. Wong, K. Wang, and A. W.-C. Fu. Collective spatial keyword queries:a distance owner-driven approach. In *SIGMOD*, pages 689–700. ACM, 2013.