Querying and Mining Geo-textual Data for Exploration: Challenges and Opportunities

Gao Cong Kaiyu Feng Kaiqi Zhao
Nanyang Technological University, Singapore
{gaocong@, kfeng002@e., kzhao002@e.}ntu.edu.sg

***Abstract*—Geo-textual data (e.g., geo-tagged tweets) is becoming increasingly available on the Web. This paper reviews recent studies on searching and mining geo-textual data for exploration, and discusses future directions along with open problems.**

I. Introduction

Massive volumes of data that contain both text content and geographical location information is being generated at an unprecedented scale on the Web. For example, tweets can be associated with locations, which may be coordinates (latitude and longitude) or semantic locations; Social photo sharing websites (e.g., Flickr and Instagram) often host photos with both descriptive tags and geographical information; Check-ins in location based social networks (e.g., Foursquare) or reviews on local business websites (e.g., Yelp and TripAdvisor) contain both text content and locations of points of interest (POIs). As another example, online news and some web pages can often be geo-tagged, and thus are associated with locations. We refer to such data as geo-textual, or spatio-textual, data.

Geo-textual data from various sources is often characterized by big volume, and some geo-textual data is continuously generated and arrives in a stream manner. For example, around 10 million geo-tagged tweets are generated every day in Twitter [1] and 7 million check-ins were submitted on Oct 3rd 2015 in Foursquare [2]. Geo-textual data often contains multi-dimensions. Apart from the rich textual dimension and geospatial dimension, geo-textual data is often featured with other dimensions, such as the user who created the content and the creation time for user generated geo-textual content, as well as the category information for POIs. Compared with traditional spatial data, the textual dimension greatly enriches the data. Meanwhile, the spatial dimension of geo-textual data also adds a semantically rich new aspect to textual data. Furthermore, the user information will enable us to explore the personal aspect and social aspect, and the time information adds the temporal dimension, which can greatly enrich geo-textual data.

With the aforementioned characteristics of geo-textual data, which can be readily presented on a map, naturally this calls for the requirement for user exploration. This opens interesting research topics on searching and mining on some dimensions of geo-textual data for exploration. First, it has many applications to retrieve a region for exploration that satisfies user-specified conditions (e.g., the size and shape of the region) while maximizing some other conditions (e.g., the relevance to the query keywords of the objects in the region). We use an example from the work [6] to explain the problem. Consider a tourist who is planning for a trip to Paris. The tourist may wish to explore a region of a certain size that has the most diverse collection of services and attractions (e.g., cinemas, shopping malls, restaurants, and museums). This will enable him/her to experience many different attractions and services in one place without the need to travel to different regions to experience them all together. In the example, the user specifies a size for the region and we want to maximize the diversity of the retrieved region. Second, it is useful to mine and explore different properties, e.g., topics [22] and top-k frequent words [17] of the geo-textual data, within a region and perhaps a timespan. For example, one may want to know what was happening in Singapore in September 2015 (Singapore election) and how was the supporting rate of the candidates in a region of Singapore. Exemplified by these studies, some progress has been made for querying and mining geo-textual data for exploration. However, the research has just begun and there are many opportunities for continued research to achieve exploratory search and mining on geo-textual data.

The rest of the paper is organized as follows. We first introduce geo-textual data in Section II. Then, we review existing research on searching and mining geo-textual data for exploration in Section III. Finally, we discuss interesting future work as well as the corresponding challenges in Section IV.

II. Geo-textual Data

Geo-textual data can be divided into (i) streaming geo-textual data that arrives at a high rate, exemplified by geo-tagged tweets, and (ii) static geo-textual data that is relatively stable, exemplified by collections of POIs. A host of indexing techniques for static geo-textual data have been proposed in open sources/commercial systems. These techniques usually combine a spatial index and a text index structure. The existing indices can be categorized according to the spatial index they utilize: (i) R-tree based indices (e.g., [5], [9], [16], [23]);(ii) grid based indices (e.g., [19]); and (iii) space filling curve based indices (e.g., [4]). Using the text index employed, indexing techniques can also be classified as inverted file based (e.g., [23]) and signature-file based [9]. In addition, some techniques (e.g., [23]) loosely combine a spatial and a text index, while others integrate them tightly, resulting in hybrid index structures (e.g., [5], [9]). These indices can be used to support different types of queries containing both keyword component and spatial component, where the keyword component may be a boolean expression or a function that is used to compute the relevance of geo-textual objects; the spatial component often

[1] https://www.mapbox.com/blog/twitter-map-every-tweet/
[2] http://blog.foursquare.com/post/130625318273/7-million-check-ins

corresponds to a range query or a k nearest neighbor query. A survey and experimental study on these indices can be found in the paper [2]. These indexing and query processing techniques are proposed mainly for static geo-textual data. However, no indexing technique is dedicatedly designed for streaming geo-textual data arriving at a high rate.

III. Research on Exploring Geo-textual Data

We proceed to outline recent studies related to exploring geo-textual data, which are divided into two types, namely *region search* and *region exploration*. To give a flavor of what each type of studies is like we choose one work for each type to give a bit more detail.

A. Region Search

Given a collection of geo-textual objects, denoted by $\mathcal{O}$, the region search problem is to find a region (or a set of regions) for user exploration such that user specified conditions (e.g., size and shape of a region) are satisfied while some objective function is optimized (e.g., the aggregation score of the returned region is maximized). Based on the returned result, users may want to modify the specified condition, and issue a new region search query interactively. The existing proposals on region search have offered different definitions on the user specified conditions and objective functions.

We next use our recent work on *best region search* [6] to illustrate the region search problem, and then review other studies on region search. Given a set of spatial objects $\mathcal{O}$, a submodular monotone aggregate score function, which has a "diminishing return" property, and the size $a \times b$ of a query rectangle, the *best region search* (*BRS*) problem aims to find an $a \times b$ rectangular region such that the *aggregate score* of the spatial objects inside the region is maximized. This problem is fundamental for several real-world applications such as *most influential region* search (e.g., find the best location for a signage to influence as many users as possible to adopt a product, where the users may belong to a social network) and *most diversified region* search (e.g., find a region with most diverse collection of services and attractions). To efficiently solve the *BRS* problem, we propose a novel algorithm called *SliceBRS* that finds the exact answer to the *best region search* problem by reducing it to the submodular weighted rectangle intersection problem, which greatly reduces the search space. Several other pruning strategies are developed to further reduce the search space. The exact algorithm is still slow on very large datasets. Since slight imprecision is acceptable in many real-world applications, to further improve the efficiency we propose the *CoverBRS* algorithm, in which we select a set of spatial points T from the space so that each point in T can represent some spatial objects in $\mathcal{O}$. The points together preserve some properties of $\mathcal{O}$ such that the rectangular region found on the set of points T can be an approximation of the result on $\mathcal{O}$ with performance guarantees.

We proceed to review the other studies on region search. In the *Maximizing Range Sum* (*MaxRS*) problem, the user specified condition (i.e., the size of a rectangle) is the same as that in *BRS*; however, its objective function is different, which is the sum of the ranking scores of all the objects covered by the retrieved rectangle region. The *MaxRS* problem can be regarded as a specialization of the *BRS* problem since SUM is a special submodular monotone aggregate score function. To solve the *MaxRS* problem, Imai et al. [10] propose an $O(n \log n)$ optimal algorithm, where n is the number of objects in $\mathcal{O}$, for finding the position of a rectangle of the given size enclosing the maximum number of spatial objects. Nandy and Bhattacharya [14] propose another line-sweeping-based algorithm with the same complexity. An external memory algorithm for the *MaxRS* problem is proposed [3] and an approximate algorithm for the problem is also proposed [18]. Instead of finding the top-1 region, the problem of finding subject-oriented top-k hot regions in spatial databases is also studied [13].

In the region search problem defined by Cao et al. [1], the user specified condition is that the total length of the retrieved road network is smaller than a specified threshold, and the objective function is the sum of the relevance score of the geo-textual objects in the retrieved road network to the specified keywords. This region search problem is to find a road network that satisfies the user specified conditions while the objective function is maximized. Alexander et al. propose Semantic Window [12] to study the region search problem for interactive data exploration on multidimensional data, in which a user explores a data space by posing a number of queries that find rectangular regions the user is interested in. The space is divided into cells and a window is a combination of cells.

The region search problem is also studied in the context of continuously moving objects. Jensen et al. propose the density query on continuously moving objects [11]. Given certain shape and an area range at time t as the user-specified condition, the density query is to find all regions such that they satisfy the user-specified condition and their density exceeds a given threshold. Ni et al. propose the pointwise-dense region query [15] in spatio-temporal databases. The query is to find a region such that any square of given size centered in the region has a density higher than a given threshold.

B. Region Exploration

Unlike region search, the problem of region exploration is to explore a specified region. Based on the returned results, users may interactively specify a different region and pose a different region exploration query.

We illustrate the region exploration problem with our work on mining topics for geo-textual objects within a user specified region and a time interval [22]. The user can submit a query with a rectangular spatial region and a time interval, and our system returns the top-k (e.g., 100) topics mined from the spatial-temporal objects (e.g., geo-tagged tweets) in the user specified region and time interval. Since learning a topic model every time for each query is time-consuming, we propose the following two phase framework for the exploratory topic mining task. (1) **Indexing and Pre-Computation:** To support efficient online learning, we first partition the geographical space and time space into small cells and construct an Octree index to organize the documents. Then, some cells in the Octree are selected by considering the memory constraint and accuracy guarantee (because the online combining algorithm is an approximate algorithm). Finally, an Latent Dirichlet Allocation (LDA) model is learnt for each selected cell.

(2) **Online Topic Learning:** We propose an algorithm to efficiently combine the pre-trained topic models on the cells that are covered by the user specified spatial-temporal space.

The combining algorithm for two cells in the online topic learning phase uses one pre-trained topic models as prior knowledge to resample the topics for the other one. The sampling process follows the idea that the pre-trained models have already captured the correlation between words in the form of topics and assigned each word token a topic label. We make use of this information to group the word tokens by topic and sample topics for each word token group instead of sampling individual word tokens. This sampling algorithm is at least an order of magnitude faster than the one that samples topics for individual word tokens. We prove that the approximation error of our sampling algorithm is proportional to the number of word tokens that appear in both cells, and thus we optimize the partition in the pre-computation phase by considering the overlap of word tokens of the resulting subcells.

We proceed to review the other work on region exploration. Given a user-specified region, one study [17] considers the problem of retrieving the top-k frequent words over the geo-textual data stream for the region. Another study considers the problem of selectivity estimation [8] for a user-specified region over the geo-textual data stream. The study [7] is to discover events for a user specified region and a time interval by clustering hashtags contained in the tweets falling in the specified region and time interval.

IV. Future Directions

Users often do not have a clear idea about how to specify parameters to perform a region search. Exploratory search [20] helps users to search, navigate, and discover new facts. Typically, users combine querying and browsing strategies to foster learning and investigation. Thus, exploratory search is a great help for users to explore the geo-textual data and find interesting information by interactively revising their queries. Most of existing work on exploring geo-textual data does not investigate the problem from an exploratory way, and only considers some dimensions of geo-textual data. This raises many opportunities for continued research. We proceed to discuss some factors to be considered to help find potential opportunities and issues that may direct future research efforts.

A. Interactive Region Search

It is common that users revise their queries and issue new queries based on their previous results. Ideally, we can use the results of previous queries to tune the results of the current query. We would need to investigate the relationships between different queries. For example, when the current region search query is similar to the previous query, the pruned space of the previous query is less likely to be the result for the current query. We also need indices to organize the search space and results for previous queries.

Furthermore, it is useful to suggest how to revise a query in the interactive exploration. For example, given a set of geo-textual objects, a user may want to find a region of a specified size such that the number of objects inside the region is maximized. However, it is very likely that a region of a slightly larger size contains much more objects. This larger region may also be meaningful to the user. We may want to suggest such regions to users, which may reduce exploratory iterations that a user needs to before finding satisfactory result. Here, research issues are:

- How to reuse the result of the previous query to speed up the processing for the current query?
- How to index the search space of previous queries?
- How to suggest meaningful revision for user exploration query?

B. Personalized Region Search

Users can specify different conditions for region search, such as keywords and region size. However, users may not always know how to specify. It helps if we can automatically learn user's preferences to assist users to specify and revise their query conditions in the scenario of exploratory search and mining. On the other hand, it is also very helpful if we can personalize the region search result by incorporating the user's preference in ranking the regions. One possibility of learning users' preferences is from the users' geo-textual posts [21]. Here, research problems could be:

- How to utilize users' preferences to assist users to specify and revise queries?
- How to make use of the user preferences to compute an overall personalized score for a region search query, and how to combine the personalized score with existing search models?
- How to search efficiently after incorporating the personalized score?

C. Region Search on Geo-textual Stream

All the existing studies on region search consider the static geo-textual data. If we take into account the streaming geo-textual data, the region search problem will need to be revisited. Potential research problems could be:

- How to index geo-textual data stream to support region search and region exploration?
- What user conditions are specified and how to define the objective function?
- If we specify a spatio-temporal cube as the user condition for region search, how to efficiently handle this over streams?
- If the region search query is a standing query over geo-textual data stream, how to efficiently handle it?
- How to conduct exploratory search and personalized search over geo-textual streams?

D. What to Explore and How?

Region exploration is a new topic for both static and streaming geo-textual data. Many fundamental problems remain open. First, what else do we explore for a region apart from those tasks studied in several existing proposals? Many

potential answers exist for the problem. For example, one answer is to discover the evolution of topics and their corresponding proportion in a user specified region and timespan to perform business analysis (e.g., it is interesting to know what is the popularity of different brands, Apple v.s. Samsung, of smart watches in Singapore in early 2015 and what is the trend in 2016). As another example of answer, how to generate spatial-aware and time-aware summaries over geo-textual data streams?

Second, it is desirable if we can perform Datacube like region exploration. In addition to spatial and temporal dimensions, other dimensions of geo-textual data can be considered in Datacube like exploration, such as users and keywords.

Third, similar to region search, in region exploration users also need to interactively revise their regions and pose new queries. How do we support the exploratory mining on the geo-textual data interactively? Specifically, we may face the following problems:

- What are interesting mining tasks for supporting region exploration?
- If we explore the evolution of topics for a specified query region and timespan, is it possible to combine data management principles and machine learning techniques to mine topic evolution efficiently as it is done for mining topics [22]?
- If we generate spatial-aware and time-aware summaries, what is the best way to do it, and how to present the timeline of spatial-aware summaries?
- How to build "Datacube" to support topic exploration on geo-textual data, e.g., supporting roll-up and drill-down along different dimensions?
- What is the appropriate way to present location-aware and time-aware topics on a map?
- How to support interactive region exploration?

E. Combining Region Search and Region Exploration

On the one hand, users may be interested to explore the topics of the region returned by region search. For example, when a user searches for regions with keywords "bar" and "cinema", he may also want to know the topics which are related to the two keywords in the region. On the other hand, users may want to search region based on the results of region exploration. For example, a user may want to search similar regions based on the topics of a specified region. Here, research problems are:

- What is the best way to combine region search and region exploration?
- How can we find similar regions for a given region in terms of the results of region exploration (e.g., topics)?
- How to perform region exploration for the result of region search?

ACKNOWLEDGMENT

This research is supported in part by a Singapore MOE AcRF Tier 2 Grant (ARC30/12), and a Singapore MOE AcRF Tier 1 Grant (RG22/15).

REFERENCES

[1] X. Cao, G. Cong, C. S. Jensen, and M. L. Yiu. Retrieving regions of interest for user exploration. *PVLDB*, 7(9):733–744, 2014.

[2] L. Chen, G. Cong, C. S. Jensen, and D. Wu. Spatial keyword query processing: An experimental evaluation. *PVLDB*, 6(3):217–228, 2013.

[3] D.-W. Choi, C.-W. Chung, and Y. Tao. A scalable algorithm for maximizing range sum in spatial databases. *PVLDB*, 5(11):1088–1099, 2012.

[4] M. Christoforaki, J. He, C. Dimopoulos, A. Markowetz, and T. Suel. Text vs. space: efficient geo-search query processing. In *CIKM*, pages 423–432, 2011.

[5] G. Cong, C. S.Jensen, and D. Wu. Efficient retrieval of the top-k most relevant spatial web objects. In *PVLDB*, pages 337–348, 2009.

[6] K. Feng, G. Cong, S. S. Bhowmick, W.-C. Peng, and C. Miao. Towards best region search for data exploration. In *SIGMOD*. ACM, 2016.

[7] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang. STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream. In *ICDE*, pages 1561–1572, 2015.

[8] Y. Huang, F. Bastani, R. Jin, and X. S. Wang. Large scale real-time ridesharing with service guarantee on road networks. *PVLDB*, 7(14):2017–2028, 2014.

[9] I.D.Felipe, V.Hristidis, and N.Rishe. Keyword search on spatial databases. In *ICDE*, pages 656–665, 2008.

[10] H. Imai and T. Asano. Finding the connected components and a maximum clique of an intersection graph of rectangles in the plane. *Journal of algorithms*, 4(4):310–323, 1983.

[11] C. S. Jensen, D. Lin, B. C. Ooi, and R. Zhang. Effective density queries on continuouslymoving objects. In *ICDE*, pages 71–71. IEEE, 2006.

[12] A. Kalinin, U. Cetintemel, and S. Zdonik. Interactive data exploration using semantic windows. In *SIGMOD*, pages 505–516. ACM, 2014.

[13] J. Liu, G. Yu, and H. Sun. Subject-oriented top-k hot region queries in spatial dataset. In *CIKM*, pages 2409–2412, 2011.

[14] S. Nandy and B. Bhattacharya. A unified algorithm for finding maximum and minimum object enclosing rectangles and cuboids. *Computers & Mathematics with Applications*, 29(8):45–61, 1995.

[15] J. Ni and C. V. Ravishankar. Pointwise-dense region queries in spatio-temporal databases. In *ICDE*, pages 1066–1075. IEEE, 2007.

[16] J. B. Rocha-Junior, O. Gkorgkas, S. Jonassen, and K. Nørvåg. Efficient processing of top-k spatial keyword queries. In *SSTD*, pages 205–222, 2011.

[17] A. Skovsgaard, D. Sidlauskas, and C. S. Jensen. Scalable top-k spatio-temporal term querying. In *ICDE*, pages 148–159, 2014.

[18] Y. Tao, X. Hu, D.-W. Choi, and C.-W. Chung. Approximate maxrs in spatial databases. *PVLDB*, 6(13):1546–1557, 2013.

[19] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the web. In *SSTD*, pages 218–235, 2005.

[20] R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[21] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *SIGKDD*, pages 605–613, 2013.

[22] K. Zhao, L. Chen, and G. Cong. Topic exploration in spatio-temporal document collections. In *SIGMOD*, 2016.

[23] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma. Hybrid index structures for location-based web search. In *CIKM*, pages 155–162, 2005.