A Spatial Insight for UGC Apps: Fast Similarity Search on Keyword-Induced Point Groups

Zhe Li
Dept. of Computing
Hong Kong Polytechnic Univ.
richie.li@connect.polyu.hk

Yu Li
Dept. of Computer Science and Technology
Hangzhou Dianzi Univ.
flyzeroliyu@gmail.com

Man Lung Yiu
Dept. of Computing
Hong Kong Polytechnic Univ.
csmlyiu@comp.polyu.edu.hk

***Abstract*—In the era of smartphones, massive data are generated with geo-related info. A large portion of them come from UGC applications (e.g., Twitter, Instagram), where the content provider are users themselves. Such applications are highly attractive for targeted marketing and recommendation, which have been well studied in recommendation system. In this paper, we consider this from a brand new spatial aspect using UGC contents only. To do this we first representing each message as a point with its geo info as its location and then grouping all the points by their keywords to form multiple point groups. We form a similarity search problem that given a query keyword, our problem aims to find *k* keywords with the most similar distribution of locations. Our case study shows that with similar distribution, the keywords are highly likely to have semantic connections. However, the performance of existing solutions degrades when different point groups have significant overlapping, which frequently happens in UGC contents. We propose efficient techniques to process similarity search on this kind of point groups. Experimental results on Twitter data demonstrate that our solution is faster than the state-of-the-art by up to 6 times.**

***Index Terms*—Similarity Searching, Spatio-Textual Searching**

I. Introduction

Location-based social media have been generating massive amount of geo-related data. For instance, each tweet message can be combined with 2 parts, the textual part and the location part. Social photo sharing websites (e.g. Flickr) contain photos with both descriptive tags and locations. Foursquare, a location based social network, provides the "check-in" function for end-user to share a message tagged with a location.

With the above geo-tagged messages, we propose the concept of *keyword-induced point group*. Given a keyword key, we form a point group G_{key} as the set of locations such that their messages contain the keyword key. To illustrate this concept, we extract all tweets located in Washington, and then visualize the point groups of different keywords. For the keyword "flower", we obtain the point group G_{flower} and then plot its distribution in Figure 1 marked with black square. Similarly, the keyword "love", its point group G_{love} is shown with gray cross.

We are interested in comparing the spatial distributions between two keyword-induced point groups. By displaying both G_{flower} and G_{love} in the same map (in Figure 1), we observe that most of the tweets containing "flower" are close to some tweets containing "love". That would reveal

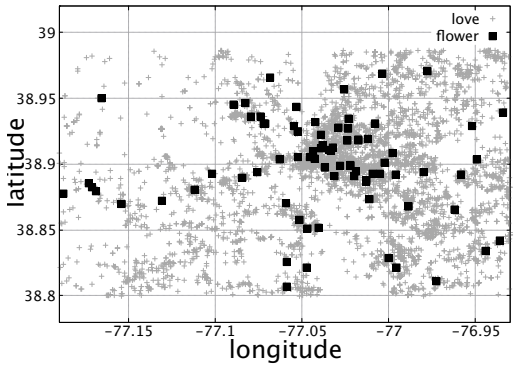

Fig. 1. Keyword-induced point groups in Washington

certain connection between the keywords "flower" and "love". Such information can be exploited in applications like targeted marketing and recommendation. For instance, a flower shop may wish to show advertisements (e.g., Twitter Ads) to nearby users who have just posted tweets about "love". Alternatively, when a user posts a tweet about "flower", the system may recommend a nearby tweet about "love". According to a survey [2], some works have considered using location information to recommend geo-tagged messages. We are the first to consider *the similarity of location distribution* in tweet recommendation.

Following the literature [3], we consider both the Hausdorff distance $dist_H(Q, G)$ and the symmetric Hausdorff distance $dist_{SH}(Q, G)$ as distance measures between two point groups Q and G. As a baseline for comparison, we also consider the Euclidean distance between the centroids of point groups $dist_{cen}(Q, G)$. The equations for these distance measures are given below:

Hausdorff distance:

$$dist_H(Q, G) = \max_{q_i \in Q} \min_{p_j \in G} dist(q_i, p_j)$$

Symmetric Hausdorff distance:

$$dist_{SH}(Q, G) = \max\{dist_H(Q, G), dist_H(G, Q)\}$$

Euclidean distance between centroids:

$$dist_{cen}(Q, G) = dist(q_c, p_c)$$

2375-0324/19/$31.00 ©2019 IEEE
DOI 10.1109/MDM.2019.00-26

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on December 01,2021 at 15:29:21 UTC from IEEE Xplore. Restrictions apply.

TABLE I
Q: POINT GROUP OF "PRESIDENT"

Distance Rank	$dist_H(Q,G)$	$dist_{SH}(Q,G)$	$dist_{cen}(Q,G)$
1	time	pain	place
2	love	loose	infinitely
3	people	send	seewhy

Table I gives an example demonstrating the nearest 3 similarity search results for point group tagged with keyword 'President' under the above mentioned distance metrics. We a conduct more complex case study using the point groups from Washington DC, New York, Los Angeles, and Hong Kong. The results of $dist_H(Q,G)$ and $dist_{SH}(Q,G)$ are viewed to be meaningful for human users. While the results found by $dist_{cen}(Q,G)$ are not meaningful.

II. Searching Acceleration

The state-of-the-art solution for our problem is [3]. Assume that all data point groups have been indexed by R-trees. At query time, it builds an R-tree for the query point group Q, then utilizes minimum bounding rectangles (MBRs) to derive lower bound distance for $dist_H(Q,G)$ and attempt pruning unpromising data point groups. Nevertheless, the solution in [3] has not taken the characteristics of keyword-induced point groups into account. Notice in most cases, the regions covered by two keyword-induced point groups overlap heavily, thus rendering MBR-based lower bound distances loose.

Our solution [1] tackles this problem by designing a much tighter lower bound along with an optimization and filtering technique for acceleration.

The Hausdorff distance $dist_H(Q,G_{key})$ is expensive to compute, incurring $O(|Q|\cdot|G_{key}|)$ time. To skip such expensive computation, we will develop a fast lower bound function $LB(Q,G_{key})$ so that $LB(Q,G_{key}) \leq dist_H(Q,G_{key})$. During similarity search, we maintain the threshold $dist_{kBest}$ for the best k Hausdorff distance of point groups examined so far. If a point group satisfies $LB(Q,G_{key}) \geq dist_{kBest}$, then G_{key} can be safely pruned without computing $dist_H(Q,G_{key})$.

The key idea is how to select the representatives. There are two principles we conclude, the first one is obvious, that the representatives should have a similar distribution, the general shape should be similar. And the second one is they should be spatial separate, this is because when we calculate the Hausdorff distance, if a region can not contribute the growth of max, the local region is highly not chance either, therefore the next representative we try should be somehow far away from the previous one.

We further verify the effectiveness of our representatives by comparing with randomly selected ones. Figure 2 compare the tightness of our representative-based lower bound with random selected representative based lower bound.

With our representative based lower bound, we could achieve an extremely tight bound by only calculating with about 5% of the query point group. And even if the point group can not be pruned, the calculation will not be wasted, as in our algorithm, the lower bound calculation is also a part of the entire Hausdorff calculation process.

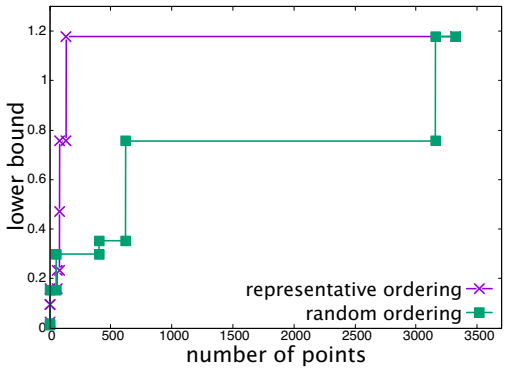

Fig. 2. Incremental lower bound on Twitter data

This technique brings the most significant acceleration with regard to the state-of-the-art. The optimization in lower bound calculation will bring an extra maximum 17% improvement and the filtering techniques will bring an extra 40% improvement in maximum.

III. Discussion

The Hausdorff distance and Symmetric Hausdorff distance are highly likely to provide results with semantic similarities. While whether other spatial distance metrics could also provide valuable semantic connection between point groups for UGC contents? Whether these metrics are efficient enough to derive analytical results and how to accelerate them if they are not. All these questions remain a research problem and are worthy for exploring.

IV. Conclusion

In this paper, we discuss the spatial property of keyword-induced point groups, which provide a brand new aspect for recommendation and targeted marketing that are especially suitable for UGC applications such as Twitter and Instagram. We formally form this as a similarity search problem under the Hausdorff distance metric. With the acceleration techniques we proposed, our solution on real data could be faster than the state-of-the-art by up to 6 times.

Acknowledgment

The author would like to thank his advisor, Dr. Ken Yiu for the insightful discussion and suggestions.

References

[1] Li Z, Li Y, Yiu M L. Fast similarity search on keyword-induced point groups[C]//Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2018: 109-118.

[2] Bao J, Zheng Y, Wilkie D, et al. Recommendations in location-based social networks: a survey[J]. GeoInformatica, 2015, 19(3): 525-565.

[3] Adelfio M D, Nutanong S, Samet H. Similarity search on a large collection of point sets[C] Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2011: 132-141.

Authorized licensed use limited to: Hong Kong Polytechnic University. Downloaded on December 01,2021 at 15:29:21 UTC from IEEE Xplore. Restrictions apply.