5M-2

An evaluation of m-Closest Keyword Search on spatial data using Flickr data

†Dang Hoang Anh †QIU Yuan †OHMORI Tadashi †FUJITA Hideyuki
†Graduate school of Information System, The University of Electro-Communications

1 Introduction

The mCK search (m-closest keyword search) [1] is a search of spatial objects for the spatial closest set of objects which match m keywords as a query. We developed a new algorithm to improve the efficiency of mCK search [2]. The aim of this paper is to evaluate the quality and practicality of mCK search on real spatial data. We use photograph data of Flickr which have geographic coordinates and tags for the evaluation.

2 Problems

Given a spatial data set $T = \{t_1, t_2, ..., t_n\}$ and a query keyword set $Q = \{k_1, k_2, ..., k_n\}$, mCK search is to find the best area that contains at least one object for every search keyword in query Q. A *diameter* of an object set is the value of maximum distance among all of distance value between any two objects in T. The less maximum distance value is, the closer objects in that set are to each other. Thus, mCK search is to find the group of objects that has the smallest diameter in the original object set. Our team has developed a new algorithm to improve the original mCK Search method [2]. The purpose of this paper is using real data from Flickr to evaluate the practical usage of our new algorithm.

Flickr is a photo sharing service where images are stored with a wide range of information including several semantic tags related to either georeference about the scene being photographed and the photo's location. We build a data collection system to retrieve photo metadata from Flickr and a data search system applying mCK method and visualizing search results on Google Maps in order to evaluate m-Closest keyword search over spatial web objects.

3 Flickr dataset

Our real data set is created from public API service provided by Flickr where photo metadata are collected by specifying a bounding box corresponding to a region. In the experiment, we extracted all photo metadata taken in Tokyo region that are semantically tagged and have geographically marked. Table 1 describes an overview of the collected data. The dataset of photos retrieved from Flickr is taken from 2013-01-01 to 2013-04-30 including

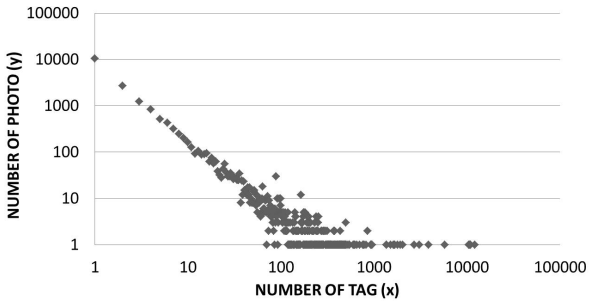

Figure 1: The number of tags appearing in y photos

95,306 tagged or non-tagged photos. After excluding non-tagged photos, we stored photo's metadata into a database. The database consists of 46,303 tagged geographical photos and the number of unique tags is 19,425. By investigating the properties of tags distribution as well as tag's occurrence, we can further understand tagging behavior. Most of the photos consist of about from 6 to 7 different kinds of tags. The characteristics of user's tagging behavior show that tags are not only referring to names representing places, landmark or geographic features such as "*shinagawa*", "*yokohama*" or "*cafe*", "*bookstore*" but also describing the content of the images and relating to user's interests such as "*flowers*", "*travel*", "*sky*". Fig. 1 describes there are x number of tags (x axis) which appear y times (number of photo) in the data collection. The graph has three distinguishing regions: 10,502 tags appear only one time and 2,683 tags appear two times, while 162 tags appear ten times and only 10 tags appear more than 900 times. Tags that appear about 28 to 300 times in the collection are considered to appeared the most frequently in all the tags. Fig. 2 describes the number of photos for each tag ID appearing in the collection. ID is given in the decreasing order of its occurrence. The y axis of the graph shows the number of photos that includes the tag ID. There

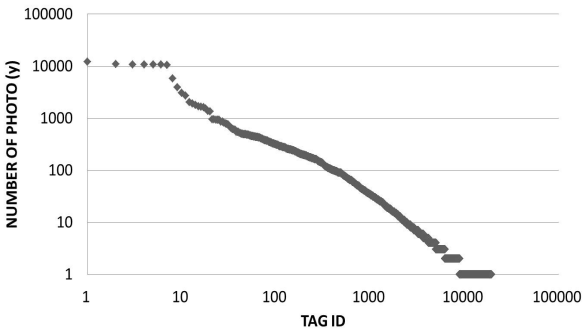

Figure 2: Occurrences of Tag ID

An evaluation of m-Closest Keyword Search on spatial data using Flickr data
†Dang Hoang Anh, Yuan Qiu, Tadashi Omori, Hideyuki Fujita
†Graduate School of Information System, The University of Electro-Communications

Table 1. Overview of collected data

Period taken time of collected data	From 2013-01-01 to 2013-04-30
Bound of collected data	Tokyo area
Number of Flickr photos (tag and non-tagged photos)	95,306
Number of collected photos (tagged photos)	46,303
Number of tag's varieties in collected data	19,425

are about over 10,000 photos which include such as these popular terms: "*tokyo*", "*japan*", *etc*. Many other tags are associated with locations.

4 mCK Searching results

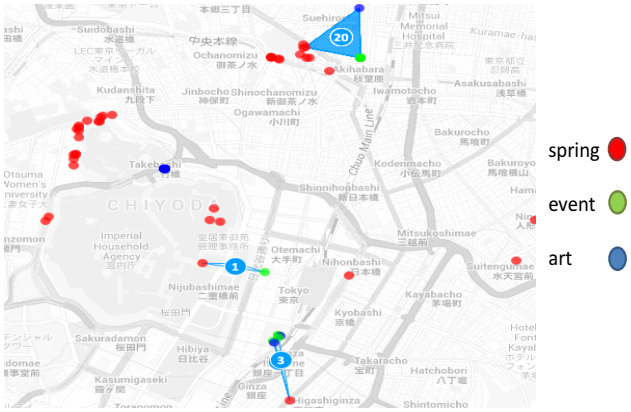

Figure 3: mck search result for spring-event-art

We build a search system using the mCK search method which allows user to input m-keywords ($m \geq 2$), the results are represented in polygon shapes depending on the number of input keywords. Our search result's ranking strategy is based on rating the spatial regions that have top-K smallest diameters. The smaller the diameter is, the nearer the spatial objects are. It indicates that place is more relevant to user's search intension. For meaningful searching experiments we make search with terms that related to people's interests instead of specific name of areas such as "*shinjuku*", "*yokohama*", etc. Fig. 3 and Fig. 4 illustrate results of example queries returned by mCK method. Assume that a user would like to find an area where holds art events in spring, with the three input keywords "*spring*", "*art*" ,"*event*", the search system returns a result shown in Fig. 3. It consists of 3 clusters representing 20 best matched areas among 442 points data for keyword "*spring*", 226 points data for "*art*" and 118 points data for "*event*". The top 2 answers are the 2 clusters in the lower half part of Fig. 3.They indicate places near Ginza, while the next best 17 answers are overlapped and represented by a triangle located at Akihaibara. Fig. 4 shows another example with three keywords "*sakura*", "*river*", "*temple*". We want to find an area that has temple and sakura scenes near a river. Our mCK search system returns a result which is a cluster of overlapped triangles sharing the same edge connecting 2 keywords "*river*" and "*temple*" near Ochanomizu (among 826 points data for "*sakura*", 92 points data for "*river*" and 309 points data for "*temple*"). It is considered that the cluster which has high density of candidate answers has good quality.

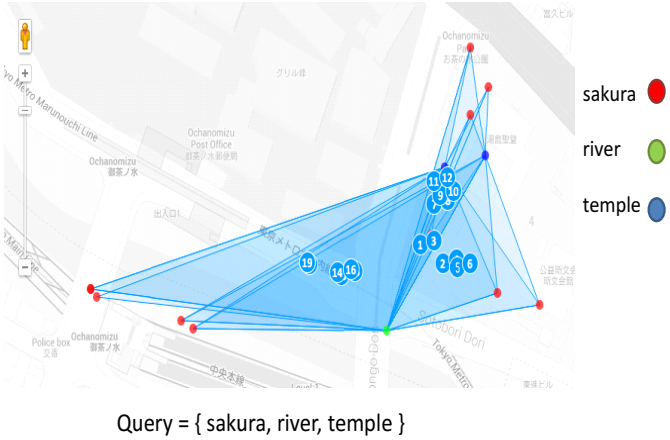

Figure 4: mck search result for sakura-river-temple

5 Conclusion

We developed a data collection and searching system using mCK search method to evaluate quality and practicality of mCK search algorithm. We are currently examining how to apply ranking for the density of area which includes candidate answers and how to measure the precision and recall of our mCK search.

References

[1] D.Z.Zhang, et al., "Keyword Search in Spatial Databases: Towards Searching by Document", IEEE ICDE, pp.688-699, 2009.

[2] Yuan Qiu, et al., "A new algorithm for m-Closest Keyword Search on spatial Data", IPSJ Annual Convention 5M-1 , 2014.