Research Article

Measuring similarity for query in geometric data

Bahram Sadeghi Bigham[1] · Raheleh Abyar Langroudi[2]

Abstract

One of the most important steps which is used in every data mining projects is searching an object or some similar objects in a data set. For geometric data, there are some methods that measure the difference between two objects. In recent years, researchers have focused on these types of metrics and used them in different applications (e.g., shape matching, machine vision, map generation, etc.). The query problem in these kinds of applications is more complicated when we have big data. In this paper, a new metric is presented which works efficiently when the geometric objects are in discrete form (e.g., polygon or chain). The presented method is important from a theoretical point of view, and its differences with other similar metrics are discussed in this paper.

Keywords Query · Metric · Similarity · Geometry · Big data

1 Introduction

Nowadays, with the advent and development of devices such as PDA smartphones and car navigation systems equipped with positioning systems such as GPS, the possibility of collecting position data generated by moving objects has increased dramatically. In addition, sensor tracing techniques, such as satellites and radars, collect and process large volumes of motion data. Generally, these tools generate raw data of motion with an identifier of the object and its position at a moment's time. These data are mostly stored as a string of spatial and temporal points, and called the trajectory. Tracing a moving object takes place continuously in a geographic space, so that a path only contains an example of the locations of the moving objects. One of the most common ways is the use of a variety of GPS-equipped vehicles. Moreover, sampling from other paths is likely to come from smartphones, online registration data. As a result, moving objects can be individuals, animals, vehicles, and even natural phenomena (e.g., storms). There is a wide range of applications that can be retrieved and improved by routing data.

In the first place, storing a large amount of rapidly growing data is considered as a primary task for processing path data, and then, a metric of similarity for comparing paths (which is an essential function in extracting routing data) should be specified, because the paths are probably produced by different sampling strategies and with different sampling rates. Finally, query processing is done on track data, which is difficult in terms of space and time complexity. To address these issues, a wide range of approaches and ideas has been proposed, and we classify them according to the main method of data mining.

In Sect. 2, we review some of the most commonly used meters in this regard, and then we introduce the metric that has been used previously only in statistical issues. This meter is the same as the Bhattacharya meter, which we will develop in Sect. 3, and we will propose a new meter that can be used to compare geometric shapes. Section 4 includes conclusions and suggestions that can be made in the future.

✉ Bahram Sadeghi Bigham, b_sadeghi_b@iasbs.ac.ir; Raheleh Abyar Langroudi, saghi.abyar20@gmail.com | [1]Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences, Zanjan, Iran. [2]Department of Computer, Abrar University, Tehran, Iran.

SN Applied Sciences (2019) 1:1193 | https://doi.org/10.1007/s42452-019-1184-y

Received: 22 May 2019 / Accepted: 28 August 2019 / Published online: 10 September 2019

2 Some common metrics

In many applications, data are recorded as a set of the length and latitude and/or geometric shapes and visible scenes [1]. These geometric data lead to the production of knowledge and the extraction of new concepts in which various data mining tools are used in this process. An important part of these processes is query operation that requires searching and comparing geometric objects and data. Due to the storage space constraints, different paths of moving objects are sampled at different rates. Vehicle routing data usually have a higher sampling rate than mobile devices because vehicles can provide proper battery and storage. In many applied scenarios [2, 3], a path is divided into sub-routes, each of which is often referred to as a section, a partition, or a frame. In as segmented approach, it attempts to simulate a person's description of reading the path data. It also divides the path into several partitions that move according to the behavior of the objects. In order to store the sampling points of a moving object that are aligned at intervals, the paths are divided into frames. Data recovery and processing [4] are critical in a storage system. The purpose of processing is to obtain the suitable data successfully. A location-based query tries to find paths that are close to all query locations in which a query is a small set of locations or without special order constraints. A typical program can recommend a route to travel to multiple locations. For example, suppose a passenger at a specified location wants to find a suitable way to transfer a request from a location to the destination. The route query range is essential for routing data mining applications [5]. Finding the nearest neighborhood is another major issue in spatial–temporal data mining. Several types of query of the nearest possible neighboring with inputs of a given path and a time interval are investigated based on a profile of unspecified paths as random processes. The main focus is mostly on the most similar paths for a path given in uncertain routes. The logical solution is to properly determine the similarity of two unknown paths. And this is the main topic of this paper that reviews the similarity criteria and introduces a new meter.

Template Extracting [6] is an analysis of the moving model of a motion object or movement objects. There are several patterns, such as collecting patterns, sequential patterns and periodic patterns. There is also a branch of research that addresses the issue of clustering the route. Clustering is the same as dividing the paths into groups with similar movement patterns. Groups of moving objects are identified based on information about the route (e.g., spatial dispersion, duration, velocity) as well as meaning, location meanings. Compared to the density-based clustering [7], there is also the use of motion-based clustering, whose logic is a simple observation, for example, a vehicle with high mobility (velocity). It probably indicates the low population of that area. Mobility-based clustering is based on less density-based clustering of the size of the path data set. Calculating the difference and the amount of difference between the two geometric shapes has many uses in other sciences. For example, in the discussion of artificial intelligence and machine vision, one of the most fundamental questions is finding the most similar form of data available to the requested data.

In mapping topics, drawing, comparing city maps, different applications of robot locator or unmanned car can be seen in the abundance of applications of the problem of determining the difference between two geometric shapes. Hausdorff, Frechet and the turning function meters are commonly used meters, which are described below. In addition, other meters are also used for this application, which, however, are not very efficient. An important point to be taken into account in these meters is that some of these meters (such as the turning function) measure only the similarity of the two shapes, and in either case they first measure the two shapes and then the similarity measures between them. This is while some other meters (such as the Frechet distance) without measuring the two shapes obtain the distance or difference between them. One of the methods of measuring the similarity is to approximate each curve with a set of points and then use the Hausdorff distance defined as:

$$\delta_H(A \cdot B) = \max\left\{ \max_{a \epsilon A} \left\{ \min_{b \epsilon B} d(a \cdot b) \right\} \cdot \max_{b \epsilon B} \left\{ \min_{a \epsilon A} d(a \cdot b) \right\} \right\}$$

Here d is the Euclidian distance between two points. A and B are two sets of points that describe the two curves we want to compare. In this method, first we convert each curve to a set of points: then, we obtain the minimum distance of each point of a curve with another curve. We do the same for the other curve, and eventually we report the largest value as the Hausdorff distance. While Hausdorff distance is a good measure of measurement in many applications, it is not always the case. The reason for this difference is that the Hausdorff distance only looks at a set of points in both curves and does not pay attention to the direction of the curve. However, curve direction is important in many applications. Suppose a man and his dog are walking and both should move on separate curves, and both can independently control their speed but are not allowed to return back. The Frechet distance of the two curves is equal to the minimum length of the dog's leash, which is necessary for both the man and the dog to move on their curves. Since it is difficult to perform mathematical operations on curves of arbitrary shape in some cases, the curves are approximated by a polygon curve, and in this case, the polygon curve.

The man's position is represented by a function based on t with $P(\alpha(t))$ and the position of the dog with $Q(\beta(t))$. The distance between the two curves is defined as follows:

$$\delta_F(P \cdot Q) = \min\{\max d(p(\alpha(t)) \cdot Q(\beta(t)))\}$$

Another common method for comparison is the use of a turning function. In this way, the comparison of the forms with each other is done on a scale, so the difference in size is not taken into account and only the degree of similarity in the structure of the forms is examined. For this, first, a diagram, whose axis x is the length of the shape and the y-axis is the angle in radians, is considered. Then on the shape of a rib, we start drawing by angle size toward the vertical axis and edge length toward the horizon. In the next step, the next side and the amount changing its degree are included in the chart. This will continue to return to the starting point. Similarly, the same steps for another were performed. To calculate the difference between two shapes using the turning function, it is enough to calculate the area between the two graphs. But it is worth noting that changing the starting point in the bug gives you different answers. In order to overcome this problem, separate diagrams should be drawn from each vertex at the beginning of the figure, and among all of them, the acceptable answer, is that it shows the least difference.

We now consider two series of data, each of which is divided into N categories. According to the distribution, each of the categories has a probability that the total probability of occurrence of all data in a series will be equal to one.

If the probability of each category is represented by a rectangular diagram, the total height of all rectangles in a group is equal to one. Now, if the probability of each category of the first series of data is represented by $\boldsymbol{P}$, the probability of each category is represented by $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{\boldsymbol{n}}$, respectively, and the probability of each category of the second series of data is represented by $\boldsymbol{P}'$. We give the probability of each group being $\boldsymbol{p}'_1, \boldsymbol{p}'_2, \ldots, \boldsymbol{p}'_{\boldsymbol{n}}$, respectively.

The formula for the Bhattacharya coefficient indicated by ρ is as follows:

$$\rho(P \cdot P') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}$$

The maximum Bhattacharya coefficient occurs when the one-by-one rectangles (the probabilities of the classes) are the same, and its value is equal to one, and if there are many differences, this value is zero or close to zero. The Bhattacharya meter is defined by the Bhattacharya coefficient as follows:

$$d(p \cdot p') = \sqrt{1 - \rho(p \cdot p')}$$

Contrary to the Bhattacharya coefficient, more similarity means less Bhattacharya coefficient for two same shapes, the Bhattacharya meter approaches zero. This formula is defined for Bhattacharya meter in interval [0,1], for the solution of this problem this meter is also introduced as $d = -\ln(\rho)$. In this case, if the similarity is high, the Bhattacharya coefficient is equal to 1, and if $\ln(1) = 0$, then the value of the Bhattacharya is zero, and if the similarity is low, the Bhattacharya coefficient is equal to zero and ln(0) is equal to the infinite negative; therefore, the value of the Bhattacharya distance positive is infinitely reported.

3 Determining similarity in queries

In this section, Bhattacharya meter is introduced to measure similarity in the calculation of the difference between two chains (trajectory). For the two chains (or part of the path) given by P_1 and P_2, we first divide them into pieces in Fig. 1. This division on the x-axis is $a_1 \ldots a_n$. The slope of each point on each of the paths gives the angles of that point.

$\theta_j^{P_i}$ is the angle corresponding to the point a_j on the path P_i. In this way, for the paths P_1 and P_2, the following angular vector is obtained:

$$\theta^{pi} = \left\langle \theta_1^{pi}, \theta_2^{pi}, \ldots, \theta_n^{pi} \right\rangle$$

The Bhattacharya coefficient between two paths P_1 and P_2 is defined in the general (continuous) case as follows:

$$Bh(P_1 \cdot P_2) = \frac{\int_x^{\infty} \sqrt{\left|\cos\left(\theta_x^{p1} - \theta_x^{p2}\right)\right|}}{||P_1||}$$

This coefficient is always a real number in the interval [0,1]. If the two paths P_1 and P_2 are exactly the same, the

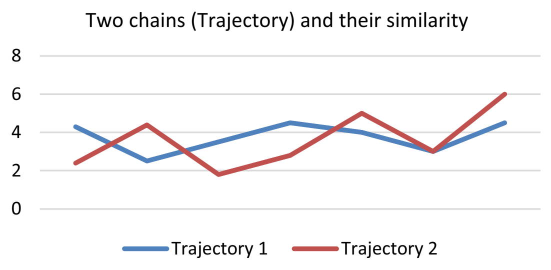

Fig. 1 Segmentation of two trajectories to calculate the Bhattacharya distance

slope difference at any desired point between the two paths is equal to zero and the face of the fraction is equal to the image of the path P_1 or P_2 on the X-axis. Means $||P_1||$ at the bottom of the fraction is the same amount. Therefore, for two completely identical paths, $Bh\left(P_1 \cdot P_2\right)$ is equal to 1.

The calculation of the above integral for a discrete state in which the entire image of the paths on the X-axis is converted to the intervals $x_1, x_2, \ldots, x_n$ is also similarly calculated as follows:

$$Bh\left(P_1 \cdot P_2\right) = \frac{\sum_{i=1}^{n} \sqrt{\left|\cos\left(\theta_{\alpha_i}^{p1} - \theta_{\alpha_i}^{p2}\right)\right|}}{|n-1|}$$

The above formula is not required; the lengths of the intervals are the same, and it is also obvious that whatever n is larger, more precision is measured.

As mentioned, $Bh\left(P_1 \cdot P_2\right)$ is always a number in the interval [0,1], and to convert it as a meter to measure the difference between two paths, perform the following transformation, and the distance between two paths P_1 and P_2 is as:

$$d_{Bh}\left(P_1 \cdot P_2\right) = -\ln\left(Bh\left(P_1 \cdot P_2\right)\right)$$

With the above mapping, we will have $0 \le d_{Bh} < \infty$ and the greater the difference between the two paths, the more d_{Bh} they are.

4 Conclusion and future works

Query is one of the key steps in most data mining processes. When the volume of data increases, this becomes much more complicated, and often, if the appropriate methods are not considered, it reduces system performance. This issue becomes more complicated when data themselves are given in non-numeric form. This paper introduced a new method that can be used in query processes for a data mining project. The basic application of this approach is to compare two geometric shapes. Although our definition for all input states is correct, for some apps and also to increase performance, we propose it for discrete geometric data. Introducing a new metric for measuring the difference between two chains is theoretically important which is presented in this paper. For subsequent research, one can focus on its various uses in machine vision and image processing, as well as data related to GIS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

References

1. Lee J-G et al (2010) Mining discriminative patterns for classifying trajectories on road networks. IEEE Trans Knowl Data Eng 23(5):713–726
2. Giannotti F, Nanni M, Pinelli F, Pedreschi D (2007). Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 330–339
3. Gidófalvi G, Pedersen TB (2009) Mining long, sharable patterns in trajectories of moving objects. Geoinformatica 13(1):27–55
4. Cheng H et al (2007) Discriminative frequent pattern analysis for effective classification. In: 2007 IEEE 23rd international conference on data engineering. IEEE
5. Han J, Kamber M, Pei J (2011) Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems 83–124
6. Chang J, Chowdhury NK, Lee H (2010) New travel time prediction algorithms for intelligent transportation systems. J Intell Fuzzy Syst 21(1, 2):5–7
7. Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World Wide Web. ACM, pp 791–800