Sensitive Distance Estimates Technique Analysis for Continuously K-Nearest Neighbors Query in Multi-Stream Processing

Ling Wang[1], Tie Hua Zhou[1], Kyung Joo Cheoi[2], Kwang Deuk Kim[3] and Keun Ho Ryu[1*]

[1]*Database/Bioinformatics Laboratory, School of Electrical & Computer Engineering, Chungbuk National University, Chungbuk, Korea*
{smile2867, thzhou, khryu}@dblab.chungbuk.ac.kr
[2]*Department of Computer Science, School of Electrical & Computer Engineering, Chungbuk National University, Chungbuk, Korea*
(kjcheoi@chungbuk.ac.kr)
[3]*Korea Institute of Energy Research, Daejeon, Korea*
(kdkim@kier.re.kr)

Abstract

In many real-world applications, data streams are usually collected in a decentralized manner such like sensor network, ubiquities sensor network, internet traffic analysis, and so on. In particularly, requirements for continuous, fast, high-volumes, adaptability, costly streaming data, an approximated analysis is needful for fast response to users on forward predicates. Distance estimate for both of "continuously" queries and streams is still a more challenge area because of a smaller or larger threshold selected is very easily to lead to a wrong result for continuously k-nearest neighbor queries. Therefore, we proposed a required filtering method to help to choose a well threshold of distance estimate in order to control error rates of approximated answers.

Keywords: *filtering, continuously k-nearest neighbors, distributed streams, distance estimates.*

1. Introduction

Data stream process as an active research area is widely used for anywhere in real world, especially for people use smart phone almost cover the world. Smart phone could be considered as both a mobile sensor to generate information to the environment, and also could be a processor to do something using internet such as on-line movie, game, micro-blogs and so on. Because of streaming, people could use a very limited memory to gather information much, much larger than controlled on the requirement of on-line process and fast response. Then, we only need to retrieve useful contents to storage, and remove others is enough. This kind of mode is stream processing. Exactly, smart phone is not an only application and the motivation comes from a potentially large application domain, e.g., network monitoring, sensor networks, telecommunications, web applications, etc. [1, 2]. In a stream application, we need mechanisms to support continuous queries on data that are continuously updated from the environment. A stream scenario brings a number of unique queries processing [3], such as in order to achieve continuously high performance, the system needs to cope with

* Corresponding author

similarities among the many standing queries, adapt to the continuously changing environment and so on.

In the time series streaming environments, k-nearest neighbors' search, which aims at retrieving the similarity between two or more streams, is an important issue [4, 5]. Unlike a snapshot k-nearest neighbors' query, a continuously k-nearest neighbor query requires continuous evaluation as the query result becomes invalid with the change of information of the query or the database objects. In many real-world applications, data streams are usually collected in a decentralized manner such like sensor network, ubiquities sensor network, internet traffic analysis, and so on. Actually, readings from a sensor network are collected in a distributed fashion, and it is even impossible to do so when the available network bandwidth is limited. Especially for continuously k-nearest neighbors' search, which aims at retrieving the similarity between two or more streams, a bandwidth-efficient approach is needed to process continuously k-nearest neighbor queries among distributed streams. Focus on recent flooded data to evaluate and only store mining results to retrieve is much easier to do so cover entire data. Although this kind of approximated result has a challenge for degree of accuracy, a good mining method could help us to get a right answer rather than a result of exist data invalided. Therefore, in this paper, we have a discussion on this kind of distance sensitive estimate query and assume a filtering method to track a more safe distance threshold on the continuously process.

2. Framework Overview

Our regular process flows is shows in figure 1. Given a query stream, the goal of continuously k-nearest neighbor queries is to find the k streams among all input streams with the highest similarities to query stream than other streams in the user-defined time range. Multiple evolving streams as the input of data stream database, a common approach is to transform the input data cells after mining methods (FH-transforms) and then retain the most representative ones to store in memory and construct a synopsis by timestamp segmentation. Searching for a better solution of distance estimate, a filtering method (Q-filtering) could help us to choose a valid threshold of distance in order to guarantee low error for continuously k-nearest neighbor queries before data transforming. Finally, a continuous approximated answer is response to users.

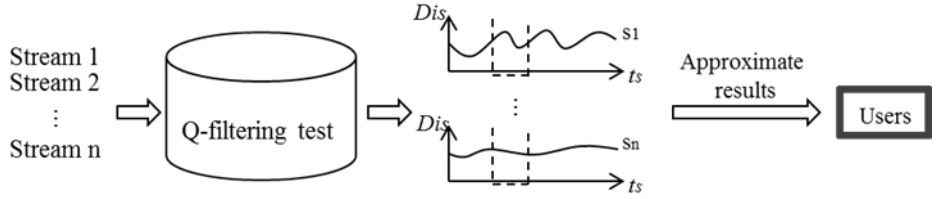

Figure 1. Work Flows

2.1. Q-Filtering Test Model

Generally, if we want to know how many nearest neighbors around us, the first thing has to know is how "nearest" could be considered as the neighbors, which always regards as the "distance" in k-nearest neighbors queries. Recent years, it's more challenges on both of "continuously" queries and objects discussion because of a smaller or larger threshold selected is very simply to lead to a wrong result for continuously k-nearest neighbor queries. For example, if you stand in the inside of a building, the query is how many peoples in this

building want to know. The general acknowledges could help us to make a primary judgment of this building, and give a probable "distance" to confine. However, computer itself cannot do anything without any further information supported. If the distance is too small such as 1 meter, maybe nobody could be scanned, else distance is too large such 1mile, it will lead too many noise data are avoided outside this building. Actually, this distance threshold is really not easy to control in k-nearest neighbors discussion by personal decided. Therefore, we using a lot of real data and methods to test and find a general regular direction of threshold which is related to distance obtained.

2.1.1. Q-Filtering phases: The entire Q-Filtering test for three steps (Table 1 shows the terminology that we use in the rest of this section):

Table 1. Terminology used in Distance Estimation

Parameter	Interpretation
S_{stream}	Number of each stream
D_{set}	The whole streams dataset
T_{type}	Number of completed data types
S_{type}	Number of streams
D_{is}	Distance
S_{item}	Frequency of each stream
S_{max}	The most frequency of each steam under selected
θ	Threshold

Step 1: select a sample real dataset to test, then initialized $T_{type} = S_{stream}/ S_{type}$ and $D_{is}=1$.

Step 2: for giving S_{max} and T_{type} to calculate a threshold $\theta = S_{max}/T_{type}$, until θ is equals or larger than 50%then stop processing. Otherwise, modify the D_{is} increasing until the threshold conditions satisfied.

Step 3: upon the satisfied threshold value, return the distance value and let it for the whole incoming datasets outside of sample.

Testing, this threshold is very safe critical point to get a valid distance further to retrieval a more accurate approximated answer to continuously k-nearest neighbors query. In this paper, we test this filtering method on two different kinds of transforms in the following parts, and finally experiments show detailed benefits.

2.2. FH-transform

A helpful tool for exploring and understanding the key properties of the Haar decomposition is the error tree structure [6]. In our FH-transform method, we average the values together by pairwise [*max*, *min*] to get a new "low-resolution" representation of the data. In the further study, we find two sensitive characters ($\overline{\varepsilon}$, δ_{diff}), $\overline{\varepsilon}$ is the mean among stream data and δ_{diff} is the average of max and min for the whole wavelet seem the most sensitive-append features over coefficients. Assume that we only pick out the pairs ($\overline{\varepsilon}$, δ_{diff}) as a typical features to present the data set stored in a subspace synopsis, whether also give us a right result on a approximated approach. Actually, these mining processes are not only give us low errors ensure for an approximated answer, but also greatly optimize space usage such that is more important in the streaming process.

3. Experiment and Evaluation

In this section, we describe a number of experiments we have run to evaluate the performance of Q-filtering with Haar and FH transforms. In the test, we used the real data here were the daily average temperature data of 19 cities random selected around Asia, and total of 106039 data for this evaluation. In addition, these streams were evenly distributed for all the experiments. We just want to know whose temperature changing regularity is most similar to an appointed stream as a query stream in this test.

3.1. Results for Q-filtering

Table 2. Distance Estimates Results

Methods	Distance	4NN results
Haar	1	5, 11, 13, **12**
FH-100	4	5, **2**, 11, 13
FH-200	3	5, **2**, 13, 11
FH-500	3	5

In order to know the sliding window size effect to these methods, three different valid ranges 100, 200 and 500 of sliding window are compared. There are 19 cities temperature information flows into memory as 19 different streams, and one of them consider as a query stream. Q-filtering is a filtering algorithm to evaluate a variety distance to the effect of approximated answers, and then return a more efficient valid distance value for the future processing. We test the approximated result accuracy rate effect on Haar, and FH methods by using Q-filtering test model to track a valid distance value on the continuously flooding datasets. The results as shows in Table 2. Here, the values of distance attributes are the results of using Q-filtering for each method. No sliding window processing is applied for Haar, so, we only consider 1 for distance in our test, and others are calculated by Q-filtering.

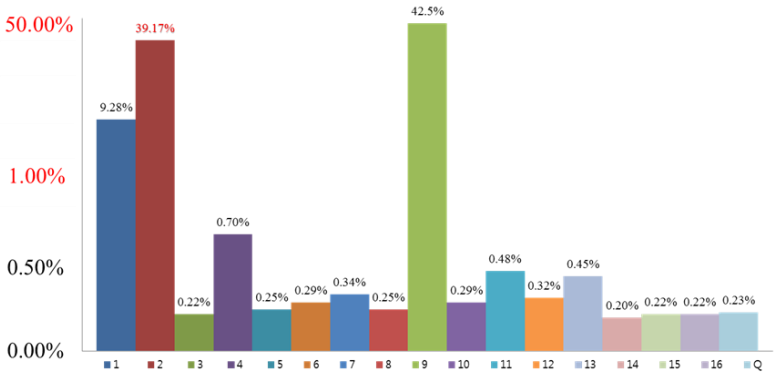

Figure 2. Error Rate of each Original Stream

In the result area, all of these methods give a similar result except two factors 2 and 12 (2, 12 means stream 2 and stream 12), because there are a lot of disabled data in the streams 2 and Haar is hard to identify. The error rate of original datasets is possibility generated from testing error by the machines, missing data or some other mistakes lead to these data lost.

Most of these streams datasets are valid and error rate is below 0.5%, except 2 and 9 give a warning about almost 40% dataset are not available see figure 2. Therefore, we are more concerned whether it will give us a more dangerous approximate result. Actually, the real result for the real 4-NN (4-nearest neighbors) query is {5, 2, 11 and 13}, and these techniques except Haar give us a right response. 12 is a wrong value of the final result which has been supplied by Haar. Except that, there are almost right answers to k-nearest neighbors query after processing by Q-filtering. As a result, distance 3 is a valid distance estimate of FH-transform that gives a better approximated result.

4. Conclusion

In this paper, we assumed a required Q-filtering model to help to track a valid threshold of distance estimate for continuously k-nearest neighbor queries. Then, test this method on proposed FH-transform to compare with investigated Haar to discuss error rate. Although FH-transform as a mining method on approximated analysis, it gives a much better exact result compare with Haar. In future work, we intend to extend our research on discussing about more complex multi-resources, then compare with proposed Q-filtering and FH methods to give a deeper discussion.

Acknowledgements

This work was supported by the Korea Institute of Energy Research (KIER) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0000478).

References

[1] Y. J. Jung, Y. K. Lee, D. G. Lee, Y. M. Lee, S. Nittel, K. Beard, K. W. Nam and K. H. Ryu, J. Sensors. 11, 11235 **(2011)**.

[2] L. Wang, Y. K. Lee and K. H. Ryu, (Eds.), "Supporting Top-k Aggregate Queries over Unequal Synopsis on Internet Traffic Streams", Proceedings of the 10th Asia-Pacific web conference on Progress in WWW research and development, **(2008)** April 26-28; Shenyang, China.

[3] J. Li, K. Tufte, V. Shkapenyuk, V. Papadimos, T. Johnson and D. Maier, J. VLDB Endowment. 1, 274 **(2008)**.

[4] Y. F. Tao, K. Yi, C. Sheng and P. Kalnis, (Eds.), "Quality and Efficiency in High Dimensional Nearest Neighbor Search", Proceedings of the 35th SIGMOD international conference on Management of data, **(2009)** June 29 - July 2; Providence, USA.

[5] M. Sharifzadeh and C. Shahabi, J. VLDB Endowment. 3, 1231 **(2010)**.

[6] M. Garofalakis and P. B. Gibbons, (Eds.), "Wavelet Synopses with Error Guarantees", Proceedings of the ACM SIGMOD international conference on Management of data, **(2002)** June 3-6; Madison, USA.

Authors

Ling Wang is a Ph.D. student in Database/Bioinformatics Laboratory of Chungbuk University. She received the M.S. degree from Chungbuk University, Korea, in 2007.

Her research interests are mainly in the areas of multimedia database, image processing, data mining, data stream processing, sensor data processing, and spatial-temporal database.

Tie Hua Zhou is a Ph.D. student in Database/Bioinformatics Laboratory of Chungbuk University. He received the M.S. degree from Chungbuk University, Korea, in 2010.

His research interests mainly include multimedia image processing, data mining, and spatial-temporal database.

Kyung Joo Cheoi is currently an assistant professor in the Department of Electrical & Computer Engineering, Chungbuk University. She received the Ph.D degree from Yonsei University, Korea, in 2002.

She worked as a research engineer in TI-specialist-Tech. of LG CNS, Korea (2002 – 2005).

Her research interests mainly include computer vision, image processing, pattern recognition, brain science, cognitive science, and biometrics.

Kwang Deuk Kim is currently a principal researcher in the Korea Institute of Energy Research, Korea. He received the Ph.D degree in Computer Science from Chungbuk University, Korea, in 2000.

His research interests mainly include GIS, spatial-temporal database, computer network security, and data mining.

Keun Ho Ryu is currently a professor in the Department of Electrical & Computer Engineering, Chungbuk University, as well as a Leader of Database/Bioinformatics Laboratory. He also served a Director of Research Institute for Computer and Communication. He received the Ph.D degree from Yonsei University, Korea, in 1988.

Prof. Ryu worked not only at University of Arizona as Post-doc and research scientist but also at Electronics & Telecommunications Research Institute, Korea. He has served on numerous program committees including AINA, ICWE, , WAIM, APWeb, WISE, and FITAT, and so on. He is a member of the IEEE since 1982 and member of the ACM since 1983.

His research interests are mainly in the fields of temporal, spatial, and spatiotemporal databases and their related area included temporal GIS, ubiquitous computing and stream data processing, active database, data mining, database security, knowledge base information retrieval, and biomedical and bioinformatics.