Using Skylines on Wavelet Synopses for CKNN Queries over Distributed Streams Processing

Ling Wang* TieHua Zhou** Kwang-Deuk Kim**** Yang-Koo Lee* Keun-Ho Ryu***

Abstract In this paper, we discuss the problem of continuous k nearest neighbors (CKNN) monitoring over distributed streams wavelet synopses, which also considered sliding window structure under stream based kNN query. We developed traditional skylines techniques and propose a new method which called DR skylines to process CKNN queries as a bandwidth efficient approach. It tries to process CKNN queries on synopses for optimized sliding window time and space computation.

Keywords : Sliding Window, Synopses, CKNN, Skylines

1. Introduction

Data stream applications such as network monitoring, on line transaction flow analysis, trading in financial markets, video surveillance, weather forecasting and sensor processing pose tremendous challenges for database systems. In data stream applications, data arrives very fast and the rate is so high that one may not wish to store all the data; yet, the need exists to query and analyze this data. There has been a concerted effort in recent years to build data stream management systems, either for general purpose or for a specific streaming application. Many of the DSMSs are motivated by monitoring applications. Example DSMSs are in [1, 2, 3, 4, 5, 6].

Since the data streams evolve continuously without limit, it is impractical to store complete details for each steam. Instead, the queries are usually processed from the limited memory in which the behaviors of the data streams are summarized. Among all kinds of sketching techniques, the wavelet based approaches [7, 8, 9] have been received the most research attention due to the property of dimensionality reduction and the simplicity of transforming the data cells.

In the time series streaming environments, similarity search, which aims at retrieving the similarity between two streams, is an important issue [10]. For a kNN query, the DSMS will find the k streams that have a more similar pattern than others to a given pattern contained in a reference stream. Compared to kNN query processing in traditional databases, streamkNN query processing is much more challenging. It must handle an endlessly growing amount of data with limited resources. Unlike a snapshot kNN query, a CKNN query requires continuous evaluation as the query result becomes invalid with the change of information 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. Read- ings from a sensor network are collected in a distributed fashion. It's inefficient to gather all of the distributed streams to a central site before doing any query processing. It is even impossible to do so when the available network bandwidth is limited. Hence, there is a need to develop a bandwidth efficient approach to processing kNN queries among distributed streams.

In this paper, we use a developed skylines method

‡This research was supported by a grant(#07KLSGC02) from Cutting edge Urban Development Korean Land Spatialization Research Project funded by Ministry of Construction & Transportation of Korean government and by a Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (R01 2009 000 10926 0), and by the Korea Institute of Energy Research (KIER).

* Ph. D Candidate, Database/Bioinformatics Laboratory, Chungbuk National University, {smile2867, leeyangkoo}@dblab.chungbuk.ac.kr
** Master Student, Database/Bioinformatics Laboratory, Chungbuk National University, thzhou@dblab.chungbuk.ac.kr
*** Professor, Database/Bioinformatics Laboratory, Chungbuk National University, khryu@dblab.chungbuk.ac.kr(corresponding author)
**** Principal Technologist, Korea Institute of Energy Research, kdkim@kier.re.kr

Manuscript received : 2009.05.12 Revised : 2009.06.19 Accepted : 2009.06.20

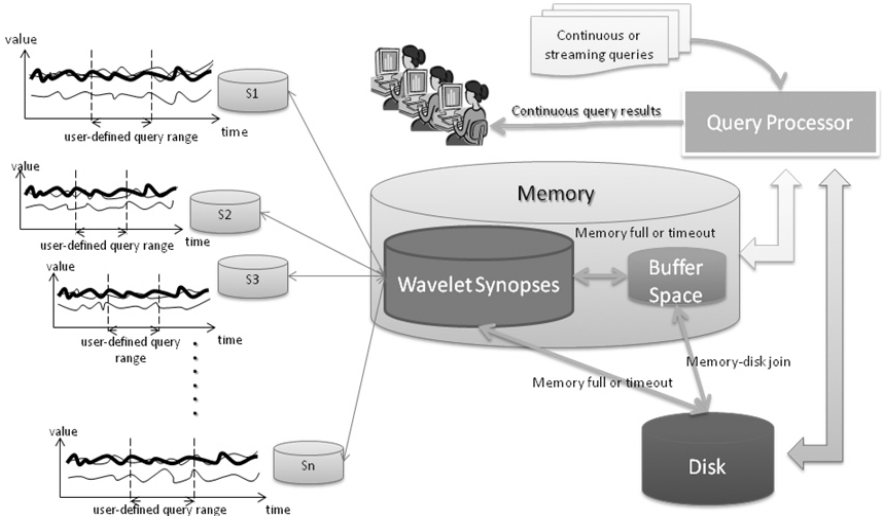


Fig. 1. System Design

to study the problem of processing CKNN queries over distributed streams on wavelet synopses. The system model as shown in figure 1, consider multiple evolving streams as the input of the DSMS. The length of each stream is beyond the storage capacity of the main memory. To approximate the behavior each stream, a common approach is to transform the input data cells as wavelet coefficients and then retain the most representative ones. At the client side, we consider the fact that different users may be interested in different ranges, and allow the DSMS to process the queries according to the range of interest. Given a query stream, the goal 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. Searching for a better solution, we notice that summary sketches(wavelet synopses), instead of complete details, of the streams are usually maintained in a data streaming environment. In this paper, we developed skylines as a bandwidth efficient approach to combine with our useful sliding window to discuss CKNN query problem over distributed streams environment by wavelet coefficients.

2. Related Work

K nearest neighbors query is an important research topic in a streaming environment [10, 11, 12, and 13]. In [11], the authors proposed to continuously retrieve the latest L points of a stream as a query pattern and then find its nearest neighbors form a time series database. Base on traditional indexing methods, the proposed scheme achieves efficient query response via perfecting. Searching nearest neighbors efficiently is also an important issue in a high dimensional database, which applies a spatial data structure (such as R tree) in searching exact nearest neighbors. And some other methods which the nearest neighbors are approximated within an error bound. In our proposed system, we not only consider the stream environment, and also support the spatial -temporal cases on moving objects which both queries and objects are changing online. And another useful design is that we consider the continuous queries from "now" or until "future" which always processes in main memory. Even for the "past" which gets the history data from disk and combine with real processing data of memory for processing the continuously queries from "past" to "future" in extended buffer space as shown in figure 1.

Skyline computation has received considerable attention in relational databases [14] and web information system [15, 16]. The skyline maintenance is performed by an in memory incremental algorithm, which discards records that cannot participate in the skyline until their expiration. However, recently, the database community witnessed a paradigm shift to query processing over continuous streams. The goal is to continuously report the qualifying records for long

standing queries in a real time manner. [17] Proposes methods for skyline monitoring over sliding windows. Generally, sliding window may be too large to fit in the main memory, so, we need algorithms that can summarize the underlying streams in concise, but reasonably accurate, synopses that can be stored in the allotted amount of memory and can be used to provide approximate answer to user queries along with some reasonable guarantees on the quality of the approximation. So, in this paper, we supply the skylines on the wavelet synopses to process CKNN queries over distributed streams as a time efficient approach.

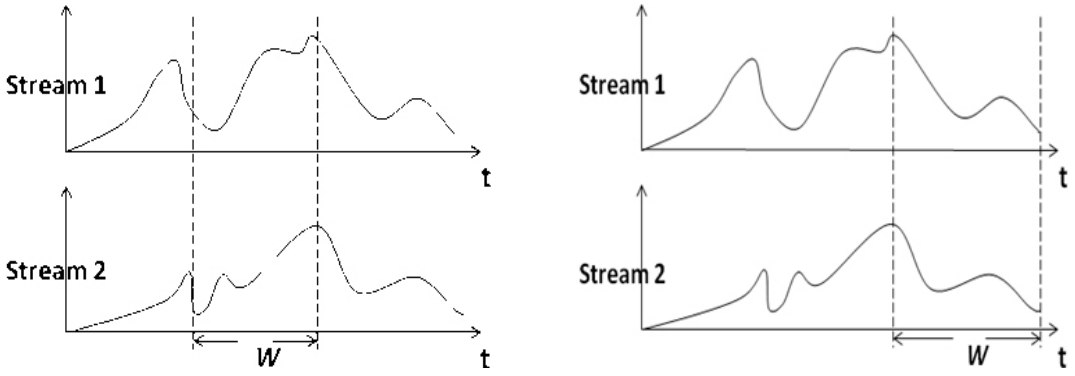
3. CKNN Query over Distributed Streams

The continuous k nearest neighbor (CKNN) query is an important type of query that finds continuously the k nearest objects to a query point. As shown in the figure 1, here CKNN query is to find the k streams among all distributed streams with the highest similarities to query stream than other streams in the user defined time range. A class of algorithms for stream processing focuses on the recent past of data streams by applying a sliding window on the data stream. In this way, only the last W values of each streaming time series is considered for query processing, whereas older values are considered obsolete and they are not taken into account. As it is shown in figure 2, streams that are non similar for a window of length W left, may be similar if the window is shifted in the time axis right.

3.1. Skylines for Wavelet Synopses

In this paper, we study skyline computation in stream systems that consider only the tuples that arrived in a sliding window covering the W most recent timestamps, where W is a system parameter called the window length. But for the wavelet decomposition, we divided the sliding window into some small sub which most for aggregate process as a synopsis. And we only count the tuples which stored in each sub window over synopses, and find the most similarity streams by using skylines compare with all sub windows in synopses over distributed streams. An example as shows in figure 3, (s2, s3……) considered as some distributed streams, and we use the sum function in each sub window as aggregated process over synopses, and the x axis shows the time series by timestamps.

By the general skylines used in sliding window, the steams always begin with the first slide until them expire, but exactly for base on wavelet synopses processing, the example aggregate sum always be changed by time series going (here sum means the count distance with query count). So, how we proposed a developed skylines method which we called DR Skylines to fit in this kind of data structure. For example CKNN query “Continuously to find the most 2 similar ports with port1 by count of coming tourists in Inchon by ships.” Here, first we can always get the query stream from port1. At the same time, some other streams data can be compared with query stream by real time, and make a coordinate axis such as figure 3. The y axis means the distance counts compare with test query, and x axis means the timestamp for each sub window over wavelet synopses. The 2NN



(a) Non similar streams (b) Similar streams

Fig. 2. Similarity using a Sliding Window of Length W and Time Series t

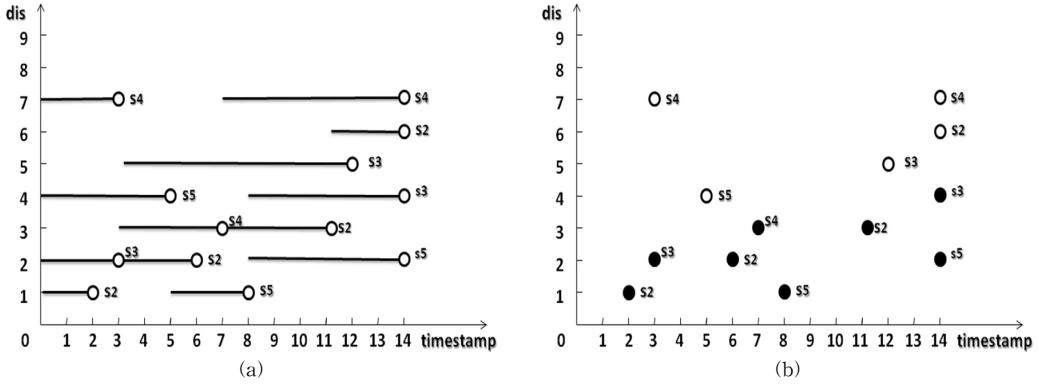


Fig. 3. Multiple-streams over Sub-windows (a) and 2-skyband Results (b)

set at time 0 is {s2,s3}. When s3 expires at time 3, it is replaced by s4. At time 5, s4 is replaced by s5 because of low distance contrast. By this approach, we can get several time series tuples by each timestamp to calculate the final results by the probability table which is most time to retain in the 2NN sets. (The 2 skyband results as shown in figure 3.)

3.2. DR Skylines Algorithm

Our proposed DR Skylines algorithm as follows:

Algorithm DR-skylines (*W(TS), k*)
Input: S_{in}: set of arriving points; S_{out} :set of *k*-NN points; TimeSeries: TS(t0,t1, t2……)
Output: S_{out} is the most CKNN answers.
For each S_{in} in *sub_window (TS)*
  if the count(*q_in_set*)<k , insert S_{in} into *q_in_set* and *W(TS)_list*
  else *count_dist(s', q)* < *count_dist(q_in_set(s),q)*, insert s' into *q_in_set* and *W(TS)_list*, remove *q_in_set(s)*
For each query *q* and size (S_{out})=*k*
  if count *W(TS)_list(s)* / count *W(TS)_list* is larger than count *(TS)_list(s')*/count *W(TS)_list* , insert *s* into S_{out} until S_{out} is full.
Report S_{out} .

4. Conclusions

For a kNN query, the DSMS will the k most streams that have more similar patterns than others to a given pattern contained in a query stream. Compared to kNN query processing in traditional databases, stream based kNN query processing is much more challenging. So, in this paper, we discussed the problem of continuous kNN monitoring over distributed streams wavelet synopses, which also considered sliding window structure under stream based kNN query. We developed traditional skylines techniques which called DR skylines method to process CKNN queries as a bandwidth efficient approach. It tries to process CKNN queries on synopses for optimized sliding window time and space computation.

In future work, we intend to extend our research on discussing about moving objects based on sensor networks, also for spatial -temporal applications. And then, we will compare with proposed DR Skyline method to give a detailed discussion by evaluation.

References

[1] Arasu, A., Babcock, B., Babu, S., Datar, M., Ito, K., Motwani, R., Nishizawa, I., Srivastava, U., Thomas, D., Varma, R., Widom, J.: "STREAM: The Stanford stream data manager" IEEE Data Engineering Bulletin, 2003, pp. 19-26.

[2] Don, C., Uqur, C., Mitch, C., Christian, C., Sangdon, L., Greg, S., Michael, S., Nesime, T., Stan, Z.: "Monitoring streams a new class of data management applications." VLDB, 2002, pp. 215-226.

[3] Sirish, C., Owen, C., Amol, D., Michael, J.F., Joseph, M.H., Wei, H., Sailesh, K., Samuel, R.M., Fred, R., Mehul, A.S.: "TelegraphCQ: Continuous dataflow processing for an uncertain world." CIDR, ACM Press, 2003, pp. 668-668.

[4] Chuck, C., Theodore, J., Oliver, S., Vladislav, S.: "Gigascope: A stream database for network applications." SIGMOD, 2003, pp. 647-651.

[5] Ki Hyun Yoo, Kwang Woo Nam, "Strategies and Cost Model for Spatial Stream Join," Journal of Korea Spatial Information System Society, Vol.10, No.4, 2008.

[6] Yang Koo Lee, Keun Ho Ryu, "Historical Sensor

Data Management using Temporal Information," Journal of Korea Spatial Information System Society, Vol.10, No.4, 2008, pp. 143-813.

[7] Anna, C.G., Yannis, K., Muthukrishnan, S., Martin, J.S.: "One Pass Wavelet Decompositions of Data Streams." IEEE Transactions on Knowledge and Data Engineering, 2003, pp. 541-554.

[8] Panagiotis, K., Nikos, M.: "One pass wavelet synopses for maximum error metrics." VLDB, 2005 , pp. 421-432.

[9] Sudipto, G., Boulos, H.: "Wavelet synopsis for data streams: minimizing non euclidean error." ACM SIGKDD, 2005, pp. 88-97.

[10] Hao, P.H., Ming, S.C.:" Efficient range constrained similarity search on wavelet synopses over multiple streams," 15th ACM international conference on Information and knowledge management, 2006, pp. 327-336.

[11] Like, G., Zheng, R.Y., Xiaoyang, S.W.: "Evaluating continuous nearest neighbor queries for streaming time series via pre fetching." Conference on Information and Knowledge Management, 2002, pp. 485-492.

[12] Nick, K., Beng, C.O., Kian, L.T., Rui, Z.: "Approximate NN queries on streams with guaranteed error/performance bounds." VLDB, 2004, pp. 804-815.

[13] Xiao, Y.L., Hakan, F.: "Efficient k NN search on streaming data series." SSTD, 2003, pp. 83-101.

[14] Kian, L.T., Pin, K.E., Beng, C.O.: "Efficient Progressive Skyline Computation." VLDB, 2001, pp. 301-310.

[15] Wolf, T.B., Ulrich, G, Jason, X.Z.: "Efficient Distributed Skylining for Web Information Systems.," EDBT, 2004, pp.256-273.

[16] Xue, M.L., Yi, D.Y., Wei, W., Hong, J.L.: "Stabbing the Sky: Efficient Skyline Computation over Sliding Windows," ICDE, 2005, pp. 502-513.

[17] Yu, F.T., Dimitris, P.: "Maintaining Sliding Window Skylines on Data Streams," IEEE TKDE, 2006, pp. 377-391.

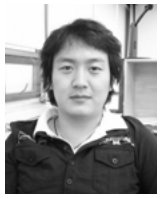
Ling Wang

2004 Dept. of Computer Science, Beihua University of China, Computer Science (B.S.)

2007 Dept. of Computer Science, Chungbuk National University, Computer Science(M.S.)

2007~Present Ph.D. Candidate, Dept. of Computer Science,, Chungbuk National University

Research Interests : GIS, Spatial-temporal Database, Data Stream Processing, Sensor Data Processing, and Data Mining



Tie Hua Zhou

2004 Dept. of Information and Computational Science, Beihua University of China(B.S.)

2007~Present Master Student, Dept. of Computer Science, Chungbuk National University

Research Interests : GIS, Spatial-temporal Database, Multimedia Image Processing, and Data Mining

Kwang Deck Kim

1987 Dept. of Computer Science, Hanbat National University of Korea(B.S.)

1989 Dept. of Computer and Statistics Science, Chonbuk National University(M.S.)

2000 Dept. of Computer Science, Chungbuk National University(Ph.D.)

1981~Present Principal Technologist, Korea Institute of Energy Research

Research Interests : GIS, Spatial-temporal Database, Computer Network Security, Data Mining

Yang Koo Lee

2002 Dept. of Computer Information Engineering, Chongju University of Korea(B.S.)
2004 Dept. of Computer Science, Chungbuk National University(M.S.)
2005~Present Ph.D. Candidate, Dept. of Computer
Science, Chungbuk National University
Research Interests : Spatial-temporal Database, Sensor Data Stream, u-GIS, and Multimedia Ddatabase

Keun Ho Ryu

1976 Dept. of Computer Science, Soongsil University(B.S.)
1980 Dept. of Computer Science, Yonsei University(M.S.)
1998 Dept. of Computer Science, Yonsei University(Ph.D.)
1976~1986 Post-doc of Arizona University, Research Scientist of Electronics & Telecommunications Research Institute of Korea
1989~1991 Research Staff of Arizona University
1986~Present Professor, Dept. of Computer Science,, Chungbuk National University
Research Interests : Temporal Database,
Spatial-temporal Database, Temporal GIS, Ubiquitous Computing and Stream Data Processing, Knowledge- base Information Retrieval, Database Security, Data Mining, and Bioinformatics