A k-Nearest Neighbor Search Algorithm for Enhancing Data Privacy in Outsourced Spatial Databases

Miyoung Jang, Min Yoon and Jae-Woo Chang

Department of Computer Engineering, Chonbuk National University
567 Baekje-daero, Deokjin-gu, Jeonju-si
Republic of Korea
{brilliant, myoon, jwchang}@jbnu.ac.kr

Abstract

With the advancement of cloud computing technologies and the propagation of location-based services, research on outsourced spatial databases has been spotlighted. Therefore, the traditional spatial databases owners want to outsource their resources to a service provider so that they can reduce cost for storage and management. However, the issue of privacy preservation is crucial in spatial database outsourcing since user location data is sensitive against unauthorized accesses. Existing privacy-preserving query processing algorithms encrypt spatial database and perform a query on encrypted data. Nevertheless, the existing algorithms may reveal the original database from encrypted database and the query processing algorithms fall short in offering query processing on road networks. In this paper, we propose a privacy-preserving query processing algorithm which performs on encrypted spatial database. A new node-anchor index is designed to reduce unnecessary network expansions for retrieving k-nearest neighbor (k-NN) objects from a query point. Performance analysis shows that our k-NN query processing algorithm outperforms the existing algorithm in terms of query processing time and the size of candidate result.

Keywords: *Outsourced spatial database, Location-based services, K-nearest neighbor search algorithm, Privacy, Query processing*

1. Introduction

Due to the advancement of cloud computing technologies, the research on outsourced databases has been spotlighted [1-2]. In the outsourced database environment, a data owner attempts to outsource his/her database to a service provider, in order to reduce cost for data storage and management. Only both authorized users and a data owner are allowed to access the outsourced data, but not the third parties. With the popularity of LBS, the traditional spatial databases owners want to outsource their databases to the service provider so that they can manage the spatial data efficiently. In this context, the issue of privacy preservation is very important in spatial database outsourcing because a user's location data is valuable and sensitive against unauthorized accesses.

In the literature, protecting outsourced spatial database has been actively studied [3-10]. The distance-oriented transformation technique [10] is proposed where the metric preserving transformation (MPT) converts an original spatial database into a numeric database by using a distance between POIs. Hence, the service provider cannot assume the original coordinate of POIs while guaranteeing the accuracy of the query processing result. However, because this technique only considers Euclidean distance between POIs, it cannot be directly applied to k-NN query processing in road networks. In real location-based applications, users move along

with the road network and so it is crucial to consider road network restrictions for k-NN query processing.

In this paper, we propose a spatial database encryption scheme that produces a transformed database from an original database by using network distances among POIs. We randomly select anchors for grouping POIs and devise an anchor-node index in order to store both network distances between POIs as well as those between anchors and network nodes. To generate the index, we encrypt both distances and anchor information by using an order preserving encryption scheme (OPES [12]). In addition, we propose a novel k-NN query processing algorithm performs on transformed data in road network by considering not only the spatial data privacy but also the query processing efficiency. Our anchor-node index can greatly reduce the network expansion cost when retrieving the k-NN POIs from the user's location. Moreover, we reduce the number of retrieved POIs by incrementally reducing a k-NN search range.

The rest of the paper is organized as follows. In Section 2, we present related work. Overall architecture of Section 3 describes proposed spatial data encryption scheme and k-NN query processing algorithm. To show the efficiency of proposed algorithm, extensive performance analysis is provided in Section 4. Finally, Section 5 concludes this research with future research directions.

2. Related Work

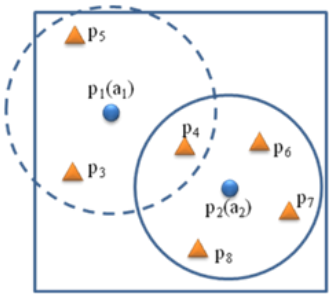
2.1. Spatial transformation

M. L. Yiu, *et al.*, [5] proposed Hierarchical Space Division (HSD), Error-Based Transformation (ERB) and HSD* by combing the advantage of HSD and ERB. The HSD protects data by using a spatial partitioning technique for redistributing the transformed data. This method is strong against the attack done by adversaries who have the subset of database information but weak against those who have a target distribution. The ERB introduces bounded errors into the data that are reversible with the help of the secure hash function [11]. The ERB transformation is weak against those who have background knowledge of data distribution. They also proposed HSD* by combining HSD and ERB. However, since the spatial transformation technique maintains coordinates of POIs, it is dangerous when the adversary has background knowledge of original POI distribution and some part of database.

2.2. Distance-oriented transformation based methods

M. L. Yiu, *et al.*, [10] proposed the distance-oriented transformation techniques called MPT (Metric preservation technique). This method converts an original spatial database in a metric space into another metric space datasets by using distance between POIs. So, the data owner transforms his data objects into a distance-based metric dataset and outsources the converted database to a service provider. Therefore, the service provider cannot assume the original coordinate of POIs while the query processing algorithm guarantees the correct answer. The server builds an index structure on the encrypted dataset in order to facilitate efficient search. For query processing, the data owner informs every user of the transformation. At query time, a trusted user applies the transformation function (with a key) to the query and sends the transformed query to the server. Then, the server processes the query and reports the results back to the user. Eventually, the user decodes the retrieved results back into the actual results. Figure 1(left) shows the example of anchor nodes and its POIs. Assume that there are 2 anchor nodes and 8 POIs around them. The table in Figure 1 explains the transformed POI index by using the OPES for each POI in anchors. The p1, p3,

p4, p5 are assigned to anchor 1, and others assigned to anchor 2. The transformed database stores the order preserving encrypted distance between anchor and POI while preserve the order of distance.



ID	Anchor ID	Distance	ID	Anchor ID	Distance
1	1	OPE(0.00) = 0.00	5	1	OPE(0.32) = 0.80
2	2	OPE(0.00) = 0.00	6	2	OPE(0.10) = 0.20
3	1	OPE(0.35) = 1.00	7	2	OPE(0.18) = 0.40
4	1	OPE(0.20) = 0.60	8	2	OPE(0.34) = 0.90

Figure 1. Anchors and POI distribution and encrypted POI index

3. K-NN Query Processing Algorithm for Outsourced Databases

In this section, we present our k-NN query processing algorithm on the encrypted spatial database in road networks. The challenge here is to extend the existing MPT algorithm to process the k-NN query processing under road network restrictions. Our algorithm utilizes an anchor-node index in order to reduce unnecessary network expansions while processing k-NN queries in road networks.

3.1. Problem setting

This section introduces the problem of database encryption and query processing in the context of privacy-preserving publication of sensitive datasets. We focus on the most common type of two-dimensional point datasets, which are presented by (x,y) coordinates. After applying distance-oriented encryption, the two-dimensional point datasets become numeric values. We focus on a k-NN query processing, which is the most popular type of query in location-based applications. Since our spatial data encryption transforms the coordinate data into distance values, we first set a range for POI search and then retrieve a set of candidate POIs as a query result. Definition 1 explains the concept of range for POI search and Definition 2 defines distance-based k-NN query processing.

Definition 1 (POI search range) Given a query point and a set A of anchors, a POI search range is set to retrieve the POIs whose distances to their nearest anchors are closer than the distance between a query point and the k-NN POI from its nearest anchor. The range for POI search is represented as a search bound of each anchors, i.e., [min, max], where min and max are derived from the distance between each anchor and the query point.

Definition 2 (Distance-based k-NN query processing) Given a POI search range [min, max], a set A of anchors and a set P of POIs, the distance-based k-NN query processing retrieves each p□ P in A that has the distance within the POI search range.

Figure 2 depicts the overall architecture of our spatial data encryption and k-NN query processing. There are three main components in this architecture: query issuer, data owner and service provider. In spatial database outsourcing, a data owner and a service provider acts different roles. Because the data owner performs data encryption in the pre-processing phase for outsourcing his/her spatial data, the data owner transforms an original database (P) into a

transformed datasets (P') and outsources the data to the service provider(1-a). Also, the data owner generates an anchor-node index. At the same time, the data owner sends the encryption information and the anchor-node index to trusted users so that they can utilize the k-NN query processing with the service provider (1-b). At a query processing time, a trusted user transforms its query (Q) to a transformed query (Q') by applying the same transformation of outsourced database. Then, the trusted user issues a query to the service provider (2-a). The service provider performs the query (2-b) and returns a query result to the trusted user (2-c).

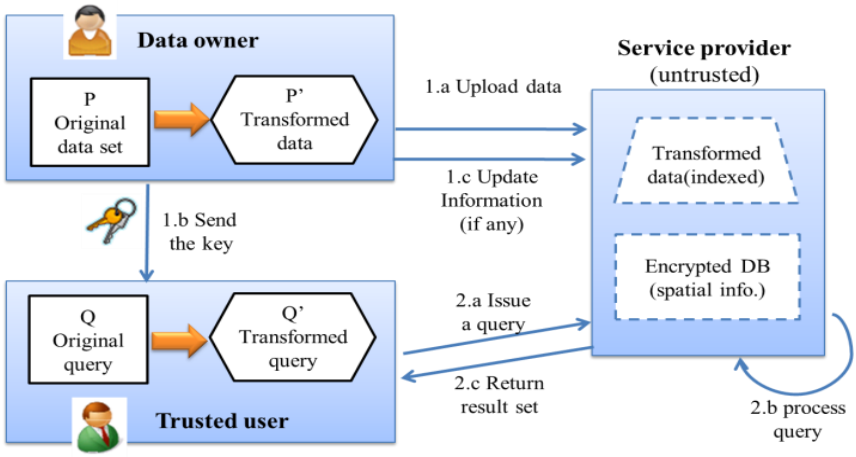


Figure 2. Overall architecture

3.2. Spatial data transformation and network distance index

In this section, we explain the spatial data transformation scheme for privacy-preserving database outsourcing and network distance pre-computation method for efficient query processing. First, the data owner performs spatial data transformation before outsourcing his private spatial data and sends the encrypted database to the service provider. M. L. Yiu, *et al.*, [10] proposed MPT protocol which generates the POI index using the Euclidean distance between POIs and anchor. Then, they encrypt the POI index by utilizing Order preserving encryption scheme (OPES) [12]. In this paper, we improve MPT algorithm to support k-NN query processing in the road network. The basic idea behind our method is to randomly choose a subset of the database a set of anchors and then assign each POIs in the original database to its nearest anchor. For each POI, we compute network distance from its anchor and then apply OPE on the distance value. These distance values are sorted and stored in the service provider in order to provide the k-NN query processing. The benefit of using OPE is that it hides the original distance values and while allowing comparisons to be correctly evaluated at the server side. In the query processing phase, we set the search range based on the value of anchor's coverage, distance between a query and an anchor, and the distance between a query and k-NN POI.

After encrypting the original database, a data owner generates an anchor-node index which stores distances between anchors and network nodes that lay on their anchor ranges. Hence each node stores its nearest anchor id and the distance from the anchor. By using the anchor-node index, a query issuer can retrieve the nearest anchor from its location by simply extending a network edge where he/she stands.

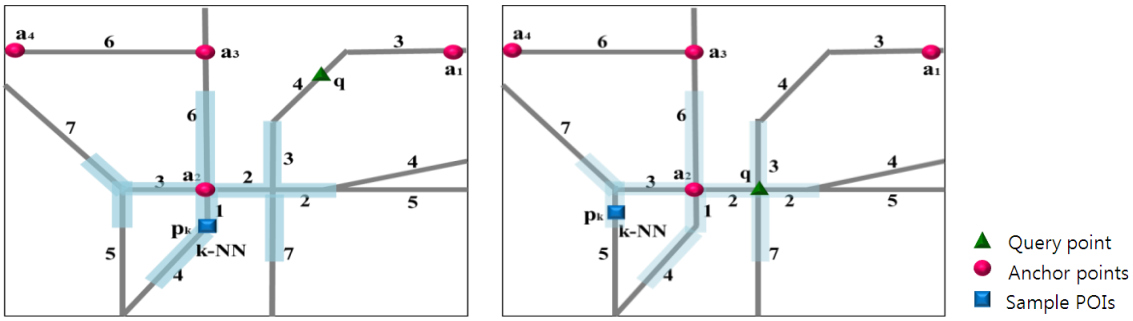
3.3. Spatial data transformation and network distance index

At query time, a trusted user applies the encryption method to its location data, and then sends the transformed query to a service provider. After receiving the query, the service provider performs k-NN query processing through 2 round communications with the query issuer. In the first round, the trusted user finds the nearest anchor from its location and set the search bound by using the distance to the anchor. Then, in the second round of communication, the server and query user retrieves all anchors lying in the query range and finds following POIs of each found anchors

3.3.1 The nearest anchor retrieval and the search range

When requesting a query, the trusted user transforms its location coordinate to the distance from the nearest anchor by using the same metric described in Section 3.2. The user can easily retrieve the nearest anchor from its location by using anchor-node index. After finding the nearest anchor, the trusted user requests r(%) of sample POIs from the nearest anchor to the service provider. Then, the user sets the tight search range of anchors to reduce the retrieved number of POIs. The search range can be defined by using the distance of the sampled k-NN POI from its anchor, i.e., dist(pk, ai), the distance between the query point and the anchor, *i.e.*, dist(ai, q), and the anchor range (ai, ri). There are two cases of search range setting for anchors considering the position of q.

First, when a query point resides outside the nearest anchor range, the query range can be defined as $[\min, \max] = [dist(a_i, p_k), dist(a_i, r_j)]$. In this case, it is required to find only POIs whose distances are greater than that of the k-NN POI from the anchor. Second, if a query point is located within the nearest anchor range, we need to consider the distance between the query point and k-NN POI as well as the distance between the anchor and the query point. In this case, the search range for k-NN POIs can be defined as $[dist(a_i, p_k), dist(a_i, q)]$. Figure 3 depicts two example cases.



(a) query point resides outside of anchor range

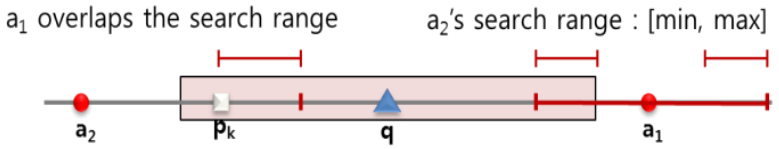
(b) query point resides within the anchor range

Figure 3. Query point within an anchor range

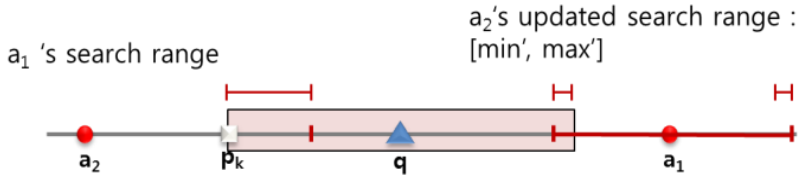
3.3.2 Retrieval of anchors within the search range and update of the search range

In order to guarantee a correct query result, the query issuer should retrieve all anchors that overlap the query range. When an additional anchor is found, the user repeats the first round for the anchor found. If the anchor returns POI that is closer to the query point than p_k in a_q, user should update the query range by using the newly retrieved k-NN POI. Figure 4 describes the basic idea behind the query range update. The search range of the nearest anchor a1 from query point q can be defined as [min, max]. a_2 is found within the search range so

that we request sampling POIs from the a_2. It is found that the distance between the k-NN POI of a_2 (p_k) and q is closer than the maximum distance from q to the search range of the a_1. Therefore, the search range of a_1 is updated with the p_2 (Figure 4(b)). Since the service provider only stores the encrypted value, it retrieves the set of potential candidates that can be the k-NN of query point by using OPE values. The user decrypts the result and computes the real network distance from candidate POIs and prunes out unnecessary POIs. When computing the real network distances from the query to the POIs, we make use of a heuristic for network expansion in order to reduce the number of unnecessary network expansion.



(a) k-NN search range with the nearest anchor a_1



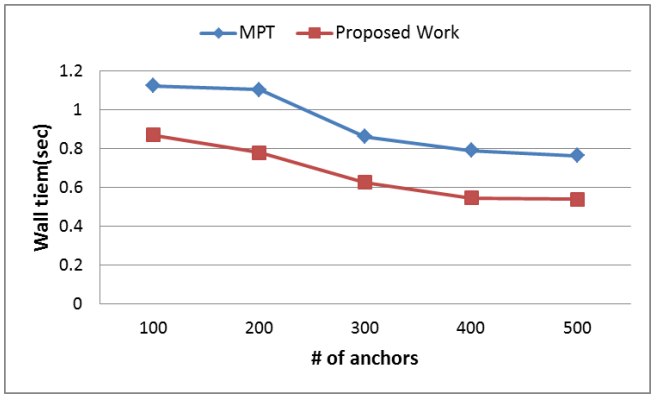
(b) update of the k-NN search range with a_2

Figure 4. Update of k-NN search range

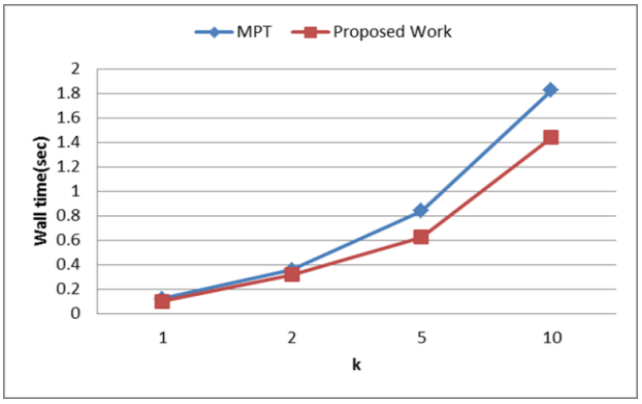
4. Experimental Evaluation

We compare our algorithm with the existing MPT algorithm in terms of preprocessing time for query processing time and returned candidate set size. For fair comparison, we extend the existing MPT method to support k-NN query processing in road network environment in a naïve way. The naïve extended MPT indexes the road network and POIs by R-tree which is commonly used in the traditional spatial network query processing. For the performance of k-NN query processing, we use the number of anchors ranged from 100 to 500 and the required k-NN as 1, 2, 5, 10 and 20. Both algorithms are implemented in Visual studio C++ 2005 on Window XP professional sp3 operating system with Intel Core(TM) i3 CPU 530 @ 2.93GHz and 2GB RAM. The San Francisco bay area map data consisting of 220,000 edges and 170,000 nodes are used. In addition, we made use of 22,025 POI data which is generated by using network-based generator of moving objects [13]. Every result reported in this paper is the average value of 100 k-NN query processing.

Figure 5 plots query processing time for both algorithms with different k and # of anchors. Our algorithm greatly reduced the query processing time since we utilize the anchor-node index for both cases. This is because when retrieving the nearest anchor from the user's location, our algorithm does not need to expand the network to find the nearest anchor but only the anchor-node that stores the network distance between each node and its nearest anchor. On the other hand, the extended MPT performs direct expansion of network edges from the user's location to the nearest anchor.



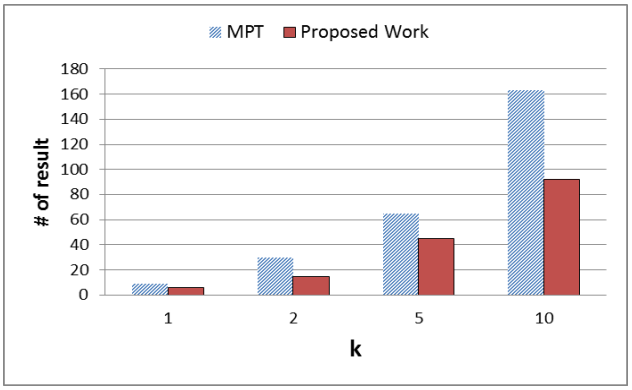
(a) varying k



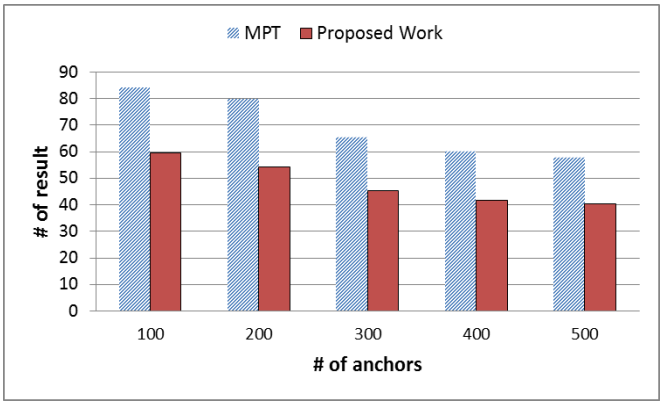
(b) varying # of anchors

Figure 5. Query processing time

Figure 6 presents the query result size (*i.e.*, number of candidate POIs) for both algorithms with different k and # of anchors. In all cases, our algorithm retrieved fewer candidates then the exiting algorithm. This is because our algorithm gradually reduces the anchor search range by updating the distance between a k-NN POI and the user.



(a) varying k



(b) varying # of anchors

Figure 6. Query result size

5. Conclusion

In this paper, we design a spatial database encryption scheme that produces a transformed database from the original database by using network distances among POIs. We generate an anchor-node index that reduces a network expansion cost. In addition, we propose a novel k-NN query processing algorithm that is efficiently performed on the encrypted database by using the anchor-node index. Through our performance analysis, we have shown that our algorithm outperforms the existing MPT method in terms of query processing time and candidate result size.

As a future work, we plan to study on a pruning technique to improve the performance of our method by reducing the size of the returned candidate set.

Acknowledgements

This research was supported by Basic Science Research program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(grant number 2010-0023800)

References

[1] X. Jiang, J. Gao, T. Wang and D. Yang, "Multiple sensitive association protection in the outsourced database", Proceedings of the 10th International Conference, Database Systems for Advanced Applications, **(2010)** April 1-4, Tsukuba, Japan.
[2] D. Sacharidis, K. Mouratidis and D. Papadias, "k-Anonymity in the Presence of External Databases", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, **(2010).**
[3] Y. Yang, D. Papadias, S. Papadopoulos and P. Kalnis, "Authenticated Join Processing in Outsourced Databases", Proceedings of ACM Special Interest Group on Management Of Data, **(2009)** June 29-July 2, Providence, Rhode island, USA.
[4] A. Singh, M. Srivatsa and L. Liu, "Search-as-a-Service: Outsourced Search over Outsourced Storage", ACM Transactions on the Web, vol. 3 and 4, no. 13, **(2009).**
[5] M. L. Yiu, G. Ghinita, C. S. Jensen and P. Kalnis, "Outsourcing Search Services on Private Spatial Data", Proceeding of the 25th IEEE International Conference on Data Engineering, **(2009)** March 29-April 2, Shanghai, China.
[6] W. K. Wong, D. W. Cheung, B. Kao and N. Mamoulis, "Secure k-NN computation on encrypted databases", Proceedings of ACM SIGMOD, **(2009)** June 29-July 2, Providence, Rhode island, USA.

[7] D. Sacharidis, K. Mouratidis and D. Papadias, "k-Anonymity in the Presence of External Databases", IEEE Transactions on Knowledge and Data, vol. 22, no. 3, **(2010).**
[8] Y. Yang, S. Papadopoulos, D. Papadias and G. Kollios, "Spatial Outsourcing for Location-based Services", Proceeding of the 25th IEEE International Conference on Data Engineering, **(2009)** March 29-April 2, Shanghai, China.
[9] M. L. Yiu, G. Ghinita, C. S. Jensen and P. Kalnis, "Enabling Search Services on Outsourced Private Spatial Data", The International Journal on Very Large Data Bases, vol. 19, no. 3, **(2010).**
[10] M. L. Yiu, I.Assent, C. S. Jensen and P. Kalnis, "Outsourced Similarity Search on Metric Data Assets", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 2, **(2010).**
[11] National Institute of Standards and Technology, Secure Hashing, http://csrc.nist.gov/groups/ST /toolkit/secure_hashing.html.
[12] R. Agrawal, J. Kiernan, R. Srikant and Y. Xu, "Order-Preserving Encryption for Numeric Data", Proceedings of ACM Special Interest Group on Management Of Data, **(2004)** June 13-18, Paris, France.
[13] T. Brinkhoff, "A Framework for Generating Network-Based Moving Objects", GeoInformatica, vol. 6, no. 2, **(2002).**

Authors

Miyoung Jang

She is a Ph.D candidate in the Chonbuk Nation University. She received the B.S and M.S degrees in Chonbuk National University in 2009 and 2011, respectively. Her research interests include security and privacy of database in cloud computing

Min Yoon

He is a Ph.D candidate in the Chonbuk National University. He received the B.S and M.S degrees in Chonbuk National University in 2009 and 2011, respectively. His research interests include security and privacy of sensor network and database outsourcing.

Jae-Woo Chang

He is a professor in the Department of Information and Technology, Chonbuk National University, Korea from 1991. He received the B.S. degrees in Computer Engineering from Seoul National University in 1984. He received the M. S. and Ph. D degrees in Computer Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1986 and 1991, respectively. During 1996–1997, he stayed in University of Minnesota for visiting scholar. And during 2003–2004, he worked for Penn State University (PSU) as a visiting professor. His research interests include sensor networks, spatial network database, context awareness and storage system.