ISSN:0975-9646

kNN-R:Building Confidential and Efficient Query Services in the Cloud Using RASP Data Perturbation

D. S. Shintre[1]& Dr. S. M. Jagade[2]
[1]*Department of M.E.(C.S.E.),* [2]*Principal*
T.P.C.T.'s College Of Engineering, Osmanabad, India.

Abstract: ***In this paper, the*** **outsourcing data-intensive services to service providers is increasingly popular with the great advantages of saving hardware and software maintenance cost. Range query and k nearest neighbors (kNN) search on large-scale databases are the important data services to many applications, from location-based services (LBS), machine learning, to similarity search in multimedia database. Once the kNN query service is outsourced, data confidentiality and query privacy become the important issues, because the data owner loses the control over the data. Adversaries, such as curious service providers, will try to breach the content of the database or intercept users queries to breach users privacy, especially when the queries are location queries. This security requirement dramatically increases the complexity of constructing a practical outsourced database services.**

Keywords: ***range query,kNN query,privacy,security***

I. INTRODUCTION

Hosting data-intensive query services in the cloud is increasingly popular because of the unique advantages in scalability and cost-saving. With the cloud infrastructures, the service owners can conveniently scale up or down the service and only pay for the hours of using the servers. Here, new approaches are needed to preserve data confidentiality and query privacy, the efficiency of query services and the benefits of using the clouds should also be preserved. It will not be meaningful to provide slow query services as a result of security and privacy assurance. It is also not practical for the data owner to use a significant amount of in-house resources, because the purpose of using cloud resources is to reduce the need of maintaining scalable in-house infrastructures. Therefore, there is an intricate relationship among the data confidentiality, query privacy, the quality of service, and the economics of using the cloud.[1]

We summarize these requirements for constructing a practical query service in the cloud as the CPEL criteria: data confidentiality, query privacy, efficient query processing, and low in-house processing cost. Satisfying these requirements will dramatically increase the complexity of constructing query services in the cloud. Some related approaches have been developed to address some aspects of the problem. However, they do not satisfactorily address all of these aspects. For example, the cryptoindex and order preserving encryption (OPE) are vulnerable to the attacks. The enhanced cryptoindex approach puts heavy burden on the in-house infrastructure to improve the security and privacy. The New Casper approach uses cloaking boxes to protect data objects and queries, which affects the efficiency of query processing and the inhouse workload.

We propose the random space perturbation (RASP) approach to constructing practical range query and k-nearest neighbor (kNN) query services in the cloud. The proposed approach will address all the four aspects of the CPEL criteria and aim to achieve a good balance on them.

To achieve better CPEL criteria for the kNN service, we propose RASP encryption that builds our outsourced databases to process range queries and k nearest neighbors search. The key components include (1) the server-side efficiency and secure query processing; and (2) the data confidentiality and query privacy guaranteed by the RASP encryption. This approach has a number of unique features.

- The outsourced data is encrypted with the RASP encryption showing resilience to various types of attacks on both the outsourced data and queries.
- The query services for range query and k nearest neighbors search can be implemented with indexed outsourced data - a unique benefit of the RASP encryption. Index-aided query processing guarantees the performance.
- The kNN-R query processing also guarantees 100% recall and high precision around 50%. In addition, this high precision is not subject to the change of the user defined level of confidentiality and privacy, which is a nice property that other approaches such as Casper do not offer. High precision guarantees the low workload of inhouse post-processing.

II. RELATED WORK

Distance-recoverable encryption is the most intuitive method for preserving the kNN relationship. However, this type of encryption is vulnerable to known plaintext-ciphertext attack, as discussed in. Wong et al. notices that comparing dot products instead of distances is sufficient to find kNN. However, this approach depends on the strong assumption that attackers cannot know plaintext-ciphertext query pairs, which is impossible in practice. Once the attacker knows one pair of plaintext-ciphertext query it is straightforward to reconstruct the key transformation matrix. In addition, the encrypted data cannot be indexed, which results in low server efficiency.

Hu et al. addresses the query privacy problem and requires the authorized query users, the data owner, and the cloud to

collaboratively process kNN queries. It makes implicit assumption that authorized query users do not collude with the cloud to breach the security. Collusion is possible when the curious cloud provider disguises as a client to submit queries. Furthermore, most computation depends on the query user's local processing and multiple rounds of interaction with the cloud server. The cloud server only aids query processing, which doesn't meet the principle of letting the cloud take most responsibility of query processing. Papadopoulos et al. also focuses on the query privacy problem, specifically, location privacy, with private information retrieval methods.

SpaceTwist proposed a method to query kNN with the encrypted users' positions for location privacy. But it didn't keep the data encrypted when it comes to sensitive data. The Casper approach considers both data confidentiality and query privacy. It cloaks each data point and query point with a rectangle. The point can be anywhere within the cloaking box. Therefore, the size of cloaking box determines the level of protection. An algorithm is designed to find the nearest neighbor of the cloaked query point from the cloaked data points. Because of cloaking the query result may contain irrelevant points. Depending on the size of the domain, an acceptable size of cloaking might be large for some applications. Our experiment shows that the query result precision drops dramatically with the slightly increased size of cloaking.

Yiu et al. uses a hierarchical space division method to encode spatial data points. It partitions the data space into blocks. In each block linear transformations are applied to *x*-axis and *y*-axis, respectively. Note that this transformation does not change the order ofthe *x* and *y* axis values in each block. Consequently, the dimensional order is preserved for the entire domain. The security weaknesses of order preserving encryption have been thoroughly discussed in several places.

Another line of research facilitates authorized users to access only the authorized portion of data, for example, a certain range, with a public key scheme. However, the underlying encryption schemes do not produce indexable encrypted data.

III. Problem Formulation

A. Query Services

Query is mainly used to search. Queries are constructed by using structured query language. It is mainly used to retrieving the needed information from the database. Query services are the method for services that are exposed through an implementation of service provider. Here by using RASP, range query and kNN query in cloud provide secure, fast storing and retrieving process of encryption and decryption of a data from database.

kNN query is to find the closest k records to the query point, where the euclidean distance is often used to measure the proximity. It is frequently used in locationbased services for searching the objects close to a query point, and also in machine learning algorithms such as hierarchical clustering and kNN classifier. A kNN query consists of the query point and the number of nearest neighbors, k.

B. System Architecture

Cloud computing infrastructures used to store large datasets and query services. The architecture shows two main parts in it. Fig. 1 shows how our model works to provide query services. The data owner encrypts the original data in its inhouse proxy server with the RASP encryption and uploads them to the service provider.

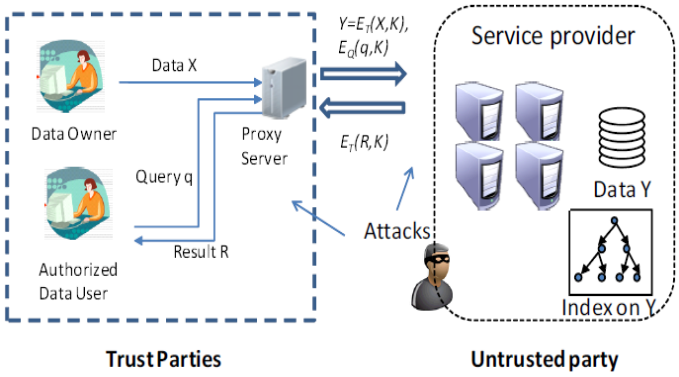

Fig. 1 The kNN-R System Architecture

The proxy server is the gateway for query processing. It encrypts queries, submits them to the service provider, and decrypts the query results returned from the service provider. The proxy server maintains the security keys, which are used in outsourcing data and encrypting queries. The traffic between the proxy server and the service provider contains only the encrypted data and queries. Although the proxy server does not handle the large dataset and process queries, it might still be a bottleneck for a large number of users and frequent query submissions. However, the cost to scale the proxy server should be much lower than that to host the entire query processing service and we will compare the time cost between the proxy server and the service provider.

C. Threat Model

We aim to protect the confidentiality of the outsourced data and the privacy of query. While query integrity is also an important issue, it is orthogonal to our study. Authentication techniques can be integrated into our framework to address the integrity problem. Thus, the integrity problem will be excluded from the paper. We can assume the curious service provider is interested in the data and queries, but it will follow the protocol to provide the service. Also, we assume that the attacker knows the algorithms to encrypt data and queries. Active attackers will also try to obtain as much prior knowledge as possible to break the encryption. According to the level of prior knowledge the attackermay have, we categorize the attacks into three categories.

- Level 1: The attacker observes only the encrypted database and encrypted queries. She/he is interested in the distributions (e.g., dimensional distributions) of the original database and the original data records. This corresponds to ciphertext only attack(COA) in cryptography.
- Level 2: Apart from the encrypted database, the attacker knows the original data distributions and

wants to recover the original records.

- Level 3: In addition to the above knowledge, the attacker manages to obtain a number of plaintext records/ queries and their ciphertext images. This corresponds to the chosen plaintext attack (CPA) in the cryptography.

The three levels also correspond to the difficulty level of obtaining the required prior know-ledge. We will analyze the security of our approach based on the three levels of attacker models.

D. Modules

Three modules are used. They are RASP, range query and kNN query.

RASP :

RASP denotes Random Space Perturbation. It also combines OPE, random projection and random noise injection. Here OPE denotes Order Preserving Encryption is used for data that allows any comparison. And that comparison will be applied for the encrypted data; this will be done without decryption. Random projection is mainly used to process the high dimensional data into low dimensional data representations. It contains features like good scaling potential and good performances.

Random noise injection is mainly used to adding noise to the input to get proper output when we compare it to the estimated power. The RASP method and its combination provide confidentiality of data and this approach is mainly used to protect the multidimensional range of queries in secure manner and also with indexing and efficient query processing will be done. RASP has some important features. In RASP the use of matrix multiplication does not protect the dimensional values so no need to suffer from the distribution based attack.

RASP prevents the data that are perturbed from distance based attacks; it does not protect the distances that are occurred between the records. And also it won't protect more difficult structures it may be a matrix and other components. The range queries can be send to the RASP perturbed data and this range query describes open bounds in the multidimensional space.

In random space perturbation, the word perturbation is used to do collapsing this process will happen according to the key value that is given by the owner. In this module the data owner have to register as owner and have to give owner name and key value. And then the user have register and get the key value and data owner name from the owner to do access in the cloud. Here user can submit their query as range query or kNN query and get their answer. We analyze and show the result with encrypted and also in decrypted format of the data for the query construct by the user.

Range Query:

Range query is the query used to retrieve the data from the database. It will retrieve the data value that is between the upper bound and lower bound. The range query is not usual because user won't know in advance about the result for the query, how much entries will come as result for the query.

For example

```
FROM table name
WHERE id (
SELECT top 10*
FROM United States
WHERE age >50
);
```

The above example shows the sample query for range query. Here the example query is to retrieve the entries from United States it will retrieve the persons who are above 50 years in the top 10 list from the record of United States.

kNN Query :

kNN query represents k-Nearest Neighbor query. This query is mainly used to retrieve the nearest neighbor values of k. here k used to denote positive integer value. kNN algorithm is mainly used for classification and regression. In this it uses kNN-R algorithm to process the range query to kNN query. This algorithm consists of two methods. That is used to make interaction between the client and the server. The client will send the query to the server with initial upper bound and lower bound. This upper bound range has to be more than the k points and the lower bound range have to be less than the k points.

E. Range Query Processing with RASP

Based on the RASP encryption we proposed, we target to provide two kinds of query services: range query and k nearest neighbors. Since RASP encryption is convexity preserving and a range query can represented as a convex set query, in the encrypted space there is a unique convex set that is the answer to the query.

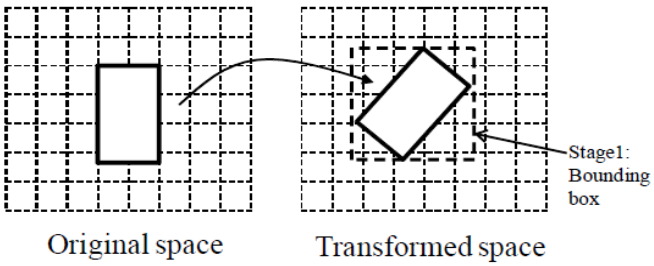

Fig. 2 Illustration of the two-stage processing algorithm

To efficiently process range query, we could use multidimensional index trees, such as R-Tree that handles axis-aligned minimum bounding boxes (MBR). However, if we still depend on the multidimensional indexing to process the transformed queries, the processing algorithm should be slightly modified to handle arbitrary convex areas because the boundaries of transformed queries are in arbitrary convex shape, not necessarily axis-aligned as Fig. 2 shows.

So we use a two-stage query processing strategy to process the encrypted query. In the first stage, the proxy transforms the original range space to a polyhedron and finds the MBR of polyhedron usingvertex-based algorithm. To submit a range query for the MBR of this polyhedron, we couldget the initial result set. In the second stage, we could linear scan the data in the initialresult set and filter out the data which are not in the polyhedron.

F. KNN-R: Using Range Queries to Process kNN Queries

We have mentioned, the quality of secure outsourced kNN query service can be summarized as the CPEL criteria: data Confidentiality, query Privacy, Efficient query processing, and Low inhouse workload. The proposed kNN-R approach aims to achieve a balanced overall CPEL criteria. In the following discussion, we use the "client" to represent the data owner's inhouse proxy server.

The confidentiality of data and the query privacy are guaranteed by using the RASP encryption. We will discuss the security issues in next section. The server-side efficiency is achieved by the index-aided query processing and an efficient secure binary range query algorithm. Low client-side cost is achieved by fast query preprocessing and high precision query results.

Overview of the kNN-R Algorithm:

The kNN-R algorithm aims to improve the confidentiality guarantee while preserving the efficiency of query processing. The basic idea is to use the RASP encryption to protect the confidentiality of data, and to use secure range query to protect the privacy of kNN query. The key is to develop an efficient kNN query algorithm based on the RASP encrypted data and queries. The design of kNN-R algorithm keeps the following problems in mind. (1) While the RASP protects the data confidentiality, it does not preserve distances or distance ranks. Therefore, the traditional distance-based kNN search algorithm does not work with the RASP encrypted data. Can we design a kNN search algorithm based on existing RASP range query algorithm? (2)Because of the limited computing capacity of the client side, the new algorithm should minimize the client's responsibility in query processing, which includes pre-processing, post-processing, and in-processing aid. Thus, the second question is how we design the algorithms to minimize the client's costs.

The kNN-R algorithm consists of five steps involving both the client and the server. Fig. 3 demonstrates the whole procedure of the algorithm and the Algorithm 1 explains the algorithm step by step.

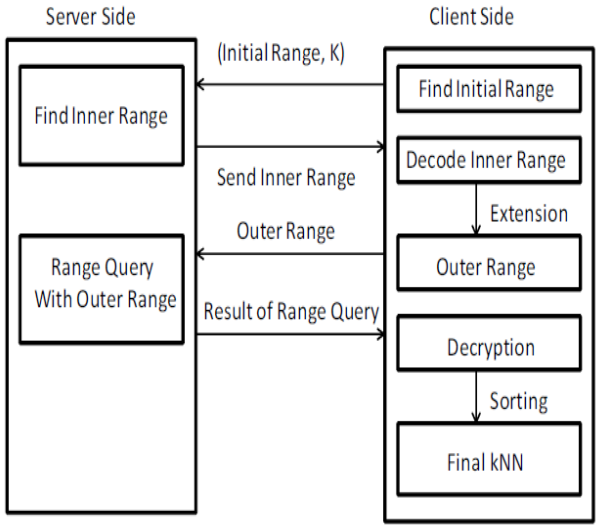

Fig. 3 Procedure of the KNN-R algorithm

The client will generate the initial upper bound range (that contains more than k points) and the lower bound range (that contains less than k points) and send them to the server.

The server finds the records in the outer range and sends them to the client. The client decrypts the records and pick the top k candidates as the final result.

Algorithm 1 KNN-R algorithm

1: The client generates the initial range and sends its secure form to the server;

2: The server works on the secure range queries and finds the inner range covering at least k points;
3: The client decodes the secure inner range from the server and extends it to the outer range, which is sent back to the server;
4: The server returns the points in the outer range

5: The client decrypts the points and extracts the k nearest points;

IV. Experimental Results

The cloud computing contains two types of parties or people involved in data access in cloud. Customer and Cloud service provider, here customer represent as end user who store their data in cloud, as shown in Fig. 4. Cloud service provider has a responsibility to store customer data in secure format. Cloud service provider do encryption and decryption to ensure secure data processing. In customer party side we have data owner, end user, internal proxy server, and the users who can only submit queries. The data owners upload the perturbed data to the cloud. In the period in-between, the authorized users can submit range queries or kNN queries to find some records.

Fig. 4 Login Window for Admin and User

The approved customer can submit range queries or kNN queries to discover a some records. Here the data owner can store their information in cloud while those information will encrypted in cloud and stored in the cloud database furthermore the data owner will give encryption key by utilizing this key value just cloud will encode the data by utilizing random space perturbation method. The untrusted parties comprise of the inquisitive cloud service provider who hosts the query services and the ensured protected database. The RASP-perturbed data will be utilized to fabricate records to keep up query processing. Only Admin has authority to make changes in uploaded files. If user want to change the data in files, then the server blocks the user, as shown in Fig. 5.

Fig. 5 Blocked Data Content of different users

Recovery of the files can be done by the admin as shown in the following recoverey window Fig. 6.

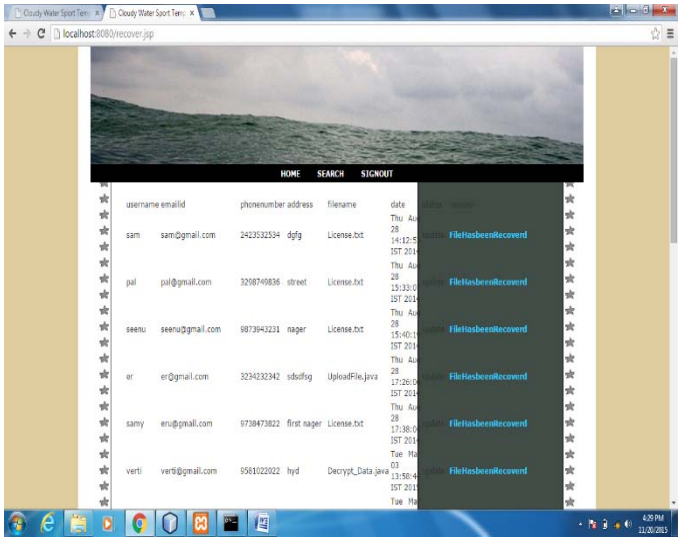

Fig. 6 Recovery Window of Users Blocked Files

V. CONCLUSION

We propose to study an outsourced service based on the CPEL criteria: data Confidentiality, query Privacy, Efficient query processing, and Low inhouse workload. With the CPEL criteria in mind, we develop the kNN-R approach for secure outsourced kNN query service. The kNN-R approach takes advantage of fast and secure RASP range query processing to implement kNN query processing. It can find high precision kNN results and also minimize the interactions between the cloud server and the inhouse client. High precision kNNresults and minimized interactions result in low inhouse workload. We have conducted a thorough security analysis on data confidentiality and query privacy. Compared to the related approaches, the kNN-R approach achieves a better balance over the CPEL criteria.

VI. ACKNOWLEDGMENT

I would like to express my deep gratitude to my research, Dr. S. M. Jagade has a research guide, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to extend my thanks to the technicians of the laboratory of the Computer department for their help in offering me the resources in running the program.

Finally, I wish to thank my parents and friends for their support and encouragement throughout my project tenure.

REFERENCES

[1] Xu, H., Guo, S., and Chen, K. “Building confidential and efficient query services in the cloud with RASP dataperturbation”, IEEE Transactions on Knowledge andData Engineering 26, 2 (2014).

[2] K. Chen, R. Kavuluru, and S. Guo, “RASP: Efficient Multidimensional Range Query on Attack-Resilient Encrypted Databases,” Proc. ACM Conf. Data and Application Security and Privacy, pp. 249-260, 2011.

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, “Order Preserving Encryption for Numeric Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD), 2004.

[4] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.K. Andy Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “Above the Clouds: A Berkeley View of Cloud Computing,” technical report, Univ. of Berkerley, 2009.

[5] J. Bau and J.C. Mitchell, “Security Modeling and Analysis,” IEEE Security and Privacy, vol. 9, no. 3, pp. 18-25, May/June 2011.

[6] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, “Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data,” Proc. IEEE INFOCOMM, 2011.

[7] K. Chen and L. Liu, “Geometric Data Perturbation for Outsourced Data Mining,” Knowledge and Information Systems, vol. 29, pp. 657- 695, 2011.

[8] K. Chen, L. Liu, and G. Sun, “Towards Attack-Resilient Geometric Data Perturbation,” Proc. SIAM Int’l Conf. Data Mining, 2007.

[9] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, “Private Information Retrieval,” ACM Computer Survey, vol. 45, no. 6, pp. 965-981, 1998.

[10] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, “Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions,” Proc. 13th ACM Conf. Computer and Comm. Security, pp. 79-88, 2006.

[11] R. Marimont and M. Shapiro, “Nearest Neighbour Searches and the Curse of Dimensionality,” J. Inst. of Math. and Its Applications, vol. 24, pp. 59-70, 1979.

[12] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra, “Executing SQL over Encrypted Data in the Database-Service-Provider Model,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD), 2002.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer-Verlag, 2001.

[14] B. Hore, S. Mehrotra, and G. Tsudik, “A Privacy-Preserving Index for Range Queries,” Proc. Very Large Databases Conf. (VLDB), 2004.

[15] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2005.
[16] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis. Wiley, 2001.
[17] I.T. Jolliffe, Principal Component Analysis. Springer, 1986.
[18] F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin, "Dynamic Authenticated Index Structures for Outsourced Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
[19] Y. Manolopoulos, A. Nanopoulos, A. Papadopoulos, and Y. Theodoridis, R-Trees: Theory and Applications. Springer-Verlag, 2005.
[20] M.L. Liu, G. Ghinita, C.S. Jensen, and P. Kalnis, "Enabling Search Services on Outsourced Private Spatial Data," The Int'l J. Very Large Data Base, vol. 19, no. 3, pp. 363-384, 2010.
[21] H. Xu, S. Guo, and K. Chen, "Building Confidential and Efficient Query Services in the Cloud with Rasp Data Perturbation," Wright State Technical Report, http://arxiv.org/abs/1212.0610, 2012.
[35] M.L. L, C.S. S, X. Huang, and H. Lu, "SpaceTwist: Managing the Trade-Offs among Location Privacy, Query Performance, and Query Accuracy in Mobile Services," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 366-375, 2008.

AUTHOR PROFILE

Ms. Dipali Sambhaji Shintre, is a M.E student of Computer Science & Engineering from T.P.C.T.'s College of Engineering, Osmanabad, India. She graduated in Computer Science and Engineering from Dr. BAMU University, Aurangabad. Her current research work focuses on "kNN-R:Building Confidential and Efficient Query Services in **the Cloud using RASP Data Perturbation".**

Dr. S. M. Jagade received Ph.D. in (Electronics & Telecommunication) from SGGS IE & T Research center Nanded, ME (Ec) specialization in Computer Science from SGGS, College of Engineering & Technology, Nanded. Completed BE (Electronics and Telecommunication) Degree from Govt. Engineering College, Aurangabad. He is currently working as a Principal in T.P.C.T.'s College of Engineering, Osmanabad, India.