iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** V **Month of publication:** May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5248

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Implementation Paper on Top-K Nearest Keyword Search on Cloud using Graph Encryption

Varsha Laxman Waje[1], V.S. Narayana Tinnaluri[2],
[1]*Computer Science and Engineering, Sandip University, Nashik, India*
[2]*Computer Science and Engineering, Sandip University, Nashik, India Narayana.*

Abstract: Now a days the security demands of data outsourcing applications is increasing and it is becoming an important issue in sustainable smart cities. Client's data which is encrypted has been widely accepted by industry. As the clouds and edges are not far trusted encryption of data should be done at client side and then it should be outsource. Therefore, it is challenging issue that how to correctly encrypt the data so that encrypted and remotely stored data can be queried to get it back. We noticed that not so much people have worked on approaches for graph –structured data encryption and support for graph queries answering so it is still lacking in studies. In this paper, we investigate one graph encryption method called top-K Nearest Keyword(kNk) searches .it is an important graph query type, several indexes are design to store information which is necessary to answer queries and to maintain the privacy or security about the graph .It may be vertex identifiers, keywords and edges. Our graph encryption methods are secure or not are demonstrated by theoretical proofs and experiments which is done on real-world datasets.

Keywords: Graph Encryption, Clouds, Edges, Top-K nearest Keyword Search, KnK, Encryption.

I. INTRODUCTION

Information re-appropriating has turned into a significant use of distributed computing and edge figuring, as information proprietors free themselves from looking after IT framework and information management. It has been recognized that security issues have turned into the greatest difficulties towards distributed computing. The pattern that calculation and capacity administrations are moving from mists to edges further features information protection issues, since security dangers at the edge side may be much more noteworthy than the cloud side. To secure information protection, information proprietors ought to encode information before redistributing. Be that as it may, customary encryption strategies make re-appropriated information never again inquiry capable, which would severely effect on information usability. Though a great deal endeavors have been made to empower watchwords seek on scrambled literary information how to perform different inquiries on encoded chart organized information is as yet a difficult issue. Distributed computing and edge figuring comprises of various applications, information redistributing is additionally significant of distributed computing. Information redistributing is only putting away customer's information remotely on cloud and getting to it at whatever point required. While information is redistributed the security of information ought to be kept up so customer will store his/her information with trust, so encryption of information is done before it is redistributed. Information encryption is nothing be that as it may, procedure of changing over plain content information to figure content information, this is finished with the help of one secure key. Be that as it may, the conventional encryption strategy does not bolster more information ease of use on the grounds that such information is never again query able. At that point likewise heaps of works have been made for watchword seek on scrambled textual information, yet at the same time different questions on encoded chart organized information is testing issue. Top-K nearest keyword(kNk) look is mulled over due to its significant applications in chart. kNk comprise of a diagram G= (V,E) where V is set of all vertices of chart. (for example v) and E is set of all edges of diagram.

In this technique each vertex named with scope of watchwords (w). An info given to the kNk seek is (k,v,w), kNk seek give result to such an extent that k Vertices in chart which are marked with w catchphrase and are closest to vertex v. In this task we contemplate, kNk seek gives secure information re-appropriating setting, for example the most effective method to appropriately scramble chart and safely answer kNk look inquiries. It is seen that while performing kNk inquiries on chart there was heaps of data spillage from inquiries and diagram too. In this way, for inquiry we ought to at any rate ready to shroud substance and identifiers of w and v separately. What's more, for chart we should ready to shroud all vertex identifiers. For instance, in genuine life the vertex identifiers could be Email-Id, portable number, Name, Address and so forth. To spread every single above need, the uncommon encryption strategy for diagram ought to be planned. By utilizing the AES encryption strategy the data spillage because of chart can be maintained a strategic distance from however it doesn't bolster the inquiry on such diagram, it encodes entirety chart. While encoding fractional diagram leeks a lot of data and creates high hazard in certifiable utilization. That implies in the event that entire chart organized is released, at that point it is anything but difficult to perform different assaults on such a chart. For instance, vertex re- ID assaults. We will contribute following things:

1) On encoded diagram explore kNk questions.
2) Defining a run of the mill diagram encryption conspire which will bolster kNk questions and its security model is offer.
3) Conducting execution

In this project we have utilized named diagram. We characterize chart as G=(V,E) yet in our venture we are utilizing the catchphrases which are marked to the vertices along these lines, here diagram can be characterized as G=(V,W) where V is set or word reference if identifiers of versus neighbor indicates as V[v], each vertex of chart has this sort of word reference. Each w watchword marked in the chart has dictionary (i.e. W). It is indicated as W[w]. It stores identifiers of vertices which are marked with w. The kNk questions ought to be replied on scrambled diagram appropriately and for that reason the chart ought to be accurately encoded. At that point and afterward the best possible information redistributing will happen, at that point customer will ready to store his information on cloud and access it at whatever point required. For concentrating on kNk questions noting plan, a chart encryption plot comprise of 5 calculations, for example, 1.KeyGen for example key age calculation,2. Encrypt for encryption, 3.TokenGen for example token age calculation, 4.Answer 5.Decrypt for decoding. Our re-appropriating framework includes customer and capacity supplier.

Customer will claim the diagram say G which is to be redistributed, and capacity supplier will store encoded from of G. Presently customer will ready to allude kNk question on scrambled chart to get to it again from capacity supplier and capacity supplier will answer customers kNk questions. The framework contains 2 conventions Setup convention ans Query convention. In Setup convention, customer claims a chart and allots the data in diagram and after that scramble with the assistance of diagram encryption plot, and re-appropriate it to capacity supplier. What's more, amid Query convention, customer issues kNk inquiry with the assistance of token which is likewise scrambled by token age calculation. The capacity supplier will return rundown of k diagram vertices to the customer.

II. LITERATURE REVIEW

I.Abraham, D.Delling,A.V.Goldberg, and R.F. Werneck. We study a class of robust network design problems motivated by the need to scale core networks to fulfill increasingly dynamic capacity needs. Past work has focused on designing the network to support all those matrices (all matrices not exceeding marginal traffic patterns is available. Another extreme is the fixed demand model, where one designs the network to support peak point-to-point demands. We introduce a capped hose model to explore a broader range of traffic matrices which includes the above two as special cases. It is realized that optimal designs for the hose model are always determined by single-hub routing, and for the fixed-demand model are based on shortest-path routing. We shed light on the wider space of capped hose matrices in order to see which traffic models are more shortest path-like as opposed to hub-like. To address the space in the middle of, we utilize hierarchical multi-hub routing templates, a generalization of hub and tree routing. In particular, we demonstrate that by adding peak capacities into the hose model, the single-hub tree-routing tem-plate is never again practical. This initiates the study of a class of robust network design (RND) problems restricted to these templates. Our empirical analysis is based on a heuristic for this new hierarchical RND problem. We also propose that it is possible to define a routing indicator that accounts for the strengths of the marginal and peak demands and utilize this information to pick the appropriate routing template. We benchmark our approach against other surely understood routing templates, using representative carrier networks and a variety of different capped hose traffic demands, parameterized by the relative importance of their marginal as opposed to their point-to-point peak demands. This study also reveals conditions under which multi-hub routing gives improvements over single-hub and shortest-path routings.

T. Akiba, Y. Iwata, and Y. Yoshida. This paper considers the task of noting shortest path queries in large certifiable charts, for example, interpersonal organizations, communication networks and web charts. The traditional Breadth First Search (BFS) approach for comprehending this issue is too time-consuming when networks with millions of hubs and possibly billions of edges are considered. A common technique to address these complexity issues utilizes a little set of landmark hubs from which the distance to every single other hub is precomputed so as to then answer arbitrary distance queries by navigating through one of the selected landmarks. Although many strate-gies to select landmarks have been introduced in past work, the issue of finding an optimal set that covers the entire chart remains NP-hard. Our contribution starts with a study of characteristics that determine the achievement fullness of a landmark selection strategy. We propose another adaptive heuristic for selecting landmarks that does not only pick central hubs, but additionally guarantees that these landmarks properly spread different regions of the diagram. Experiments on a different set of large diagrams demonstrate that the proposed selection strategy and assisting hub handling technique can efficiently estimate the hub to-hub distance in charts with millions of hubs with very high accuracy, while utilizing a similar amount of precipitation time as previously proposed strategies.

J. Blocki, A. Blum, A. Datta, and O. Sheffet, Phenomenal volumes of location-based information have been produced because of the widespread adoption of social network applications and GPS-enabled gadgets and sensors. Publication of such location data can provide valuable resources for researchers and government agencies in applications ranging from near real-time population-wide health monitoring to planning for future urban communities. However, such data hold personally distinguishing information, which offers ascend to many privacy issues. There is along these lines a squeezing requirement for ways to confine this characteristically distinguishing location-related information; how-ever ideally we would get a kick out of the chance to safeguard the utility of the data. Importantly, any such solution has to be scalable to large population-wide data scenarios. To tackle this, We introduce a novel differentially private hierarchical location sanitization (DPHLS) approach based on the concept (, r)- dataset actualized through a Variable Order Mobility Markov Model (VO3M). We show how this framework allows individual locations in personal trajectories to be protected using selection and recurrence perturbation mechanisms utilizing the (, r)- dataset, leveraging past (distributed) location histories to obfuscate the client location in an adaptable and controllable manner. The adequacy and effectiveness of the proposed solution is evaluated through the huge data tries that have been carried out utilizing an Open Stack-based Cloud and Apache Spark-based platform using large-scale social media trajectories. The experimental results recommend that the privacy publication algorithm can effectively scale to enormous data scenarios while retaining the utility of the datasets (trajectories) and saving individual client privacy.

A. Algorithm Used

1) *DES Algorithm:* The Data Encryption Standard is a block cipher, meaning a cryptographic key and algorithm are applied to a block of data simultaneously rather than one bit at a time. To encrypt a plain text message, DES groups it into 64-bit blocks. Each block is enciphered using the secret key into a 64-bit cipher text by means of permutation and substitution. The process involves 16 rounds and can run in four different modes, encrypting blocks individually or making each cipher block dependent on all the previous blocks. Decryption is simply the inverse of encryption, following the same steps but reversing the order in which the keys are applied.

2) *DES Encryption*

a) Plainext is broken into blocks of length 64 bits. Encryption is blockwise.

b) A message block is first gone through an initial permutation IP,then divided into two parts L_0,where L_0 is the left part of 32 bits and R_0 is the right part of the 32 bits

c) Round i has input L_{i-1},R_{i-1} and output L_i,R_i

$L_i = R_{i-1}, R_i = L_{i-1} \oplus f(R_{i-1},K_i)$

and K_i is the subkey for the 'i'th where $1 \leq i \leq 16$

$L_1 = R_0, \quad R_1 = L_0 \oplus f(R_0,K_1)$

$L_2 = R_1, \quad R_2 = L_1 \oplus f(R_1,K_2)$

$L_3 = R_2, \quad R_3 = L_2 \oplus f(R_2,K_3)$

....

....

....

$L_{16} = R_{15}, \quad R_{16} = L_{15} \oplus f(R_{15},K_{16})$

d) After round 16,L_{16} and R_{16} are swapped,so that the decryption algorithm has the same structure as the encrption algorithm.

e) Finally,the block is gone through the inverse the permutation IP^{-1} and then output

f) One round of DES in very simple way during encryption

3) *DES Decryption*

a) Observation:In encryption,we have

$L_i = R_{i-1}, R_i = R_i = L_{i-1} \oplus f(R_{i-1},K_i)$

b) and K_i is the subkey for the 'i'th round.Hence

$R_{i-1} = L_i, L_{i-1} = R_i \oplus f(L_i,K_i)$ for each 'i'

c) Due to swap operation after the 16th round encryption,the output of encryption is $IP^{-1}(R_{16},L_{16})$

d) Equation(1) as follows:

$R_{15} = L_{16}$, $L_{15} = R_{16} \oplus f(L_{16},K_{16})$
$R_{14} = L_{15}$, $L_{14} = R_{15} \oplus f(L_{15},K_{15})$
$R_{13} = L_{14}$, $L_{13} = R_{14} \oplus f(L_{14},K_{14})$
....
....
....
$R_1 = L_2$, $L_1 = R_2 \oplus f(L_2,K_2)$

e) If we give $IP^{-1}(R_{16},L_{16})$ as the input for the same algorithm with round subkeys($K_{16},K_{15},......K_1$),then the output is $IP^{-1}(L_0,R_0)$,the original message block

f) Decryption is performed using the same algorithm, except the K_{16} is used as the first round,K_{15} in the second,and so on,with K_1 used in the 16th round

B. KNN Algorithm

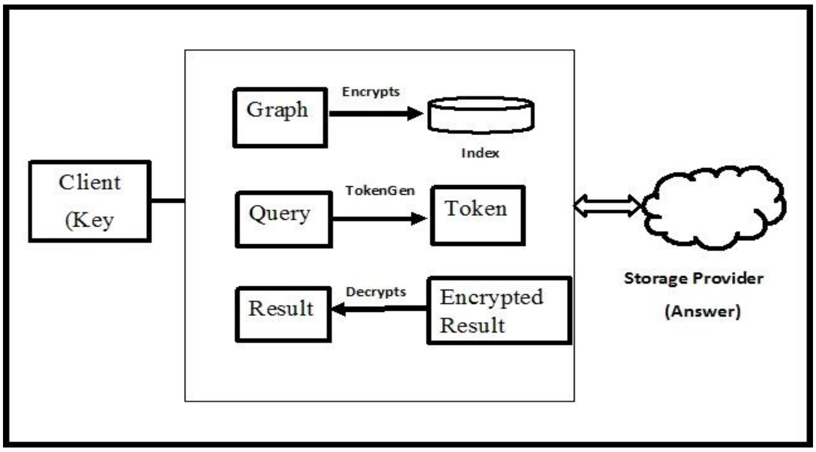
K-nearest neighbor algorithm (KNN) is part of supervised learning that has been used in many applications in the field of data mining, statistical pattern recognition and many others. KNN is a method for classifying objects based on closest training examples in the feature space. An object is classified by a majority vote of its neighbors. K is always a positive integer. The neighbors are taken from a set of objects for which the correct classification is known. It is usual to use the Euclidean distance, though other distance measures such as the Manhattan distance could in principle be used instead.

The algorithm on how to compute the K-nearest neighbors is as follows.

1) Determine the parameter K = number of nearest neighbors beforehand. This value is all up to you.
2) Calculate the distance between the query-instance and all the training samples. You can use any distance algorithm.
3) Sort the distances for all the training samples and determine the nearest neighbor based on the K-th minimum distance.
4) Since this is supervised learning, get all the Categories of your training data for the sorted value which fall under K.
5) Use the majority of nearest neighbors as the prediction value.

III.PROPOSED SYSTEM

When performing Top k queries on an encoded graph, privacy data would be leaked from both the diagram and the questions. For the diagram, we ought to at any rate conceal all vertex identifiers, as which may be email addresses, full names or on the other hand telephone numbers in genuine utilization. For this we proposed Index Based Top Value Matching Technique. For a question (k; v;w), we ought to in any event conceal the identifier of v and the substance of w. To accomplish every single above prerequisite, the encryption technique for the chart ought to be extraordinarily designed. Despite the fact that utilizing customary cryptographic encryption apparatuses, for example, DES to encode the whole diagram can maintain a strategic distance from any data leakage, the resultant scrambled chart appears to be difficult to be questioned. Though we halfway encode a diagram, for example, as it were scrambling vertex identifiers and keywords, the resultant encoded diagram leaks as well much data which would result in high risks in genuine use. For instance, the diagram structure is totally leaked with the goal that a foe is anything but difficult to perform different attacks.



Our redistributing framework includes customer and capacity supplier. Customer will possess the chart say G which is to be redistributed, and capacity supplier will store encrypted type of G. Presently customer will ready to flame kNk query on encrypted chart to get to it again from storage provider and storage provider will answer customers kNk questions.The framework contains 2 protocols Setup protocol and Query protocol. In Setup protocol, customer possesses a chart and doles out the data in diagram and afterward encrypt it with the assistance of chart encryption plot, and redistribute it to capacity supplier. Also, During Query protocol, customer issues kNk query with the assistance of token which is additionally encrypted by token age calculation. The capacity supplier will return list of k graph vertices to the customer.
Proposed system has following modules

A. Client Sign Up
B. Client Login
C. Upload
D. Graphical Password
E. Active User
F. Search File
G. Service Provider

IV. TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demand being placed on the client. The developed system must have the modest requirements, as only minimal or null changes are required for implementing this system. Technical feasibility assessment can be done through following ways:

A. NP-Complete

P Class: Class of all deterministic polynomial language problems.
NP Class: Class of all non-deterministic polynomial language problems. NP Complete problems are always solves within given time and space.

B. NP-Hard

These are problems for which there are no efficient solutions are found. Generally complexity of these problems is more than P, NP, NP-Complete. These may include higher multiplicative constants, exponent's terms or high order polynomial.

C. Satisfiability

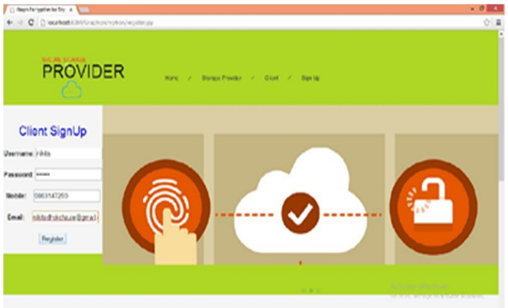
Boolean formula is satisfiable if there exists at least one way of assigning value to its variable so as to make it true and we denote it by using SAT. The problem of deciding whether given formula is satisfiable or not.

V. COST FEASIBILITY

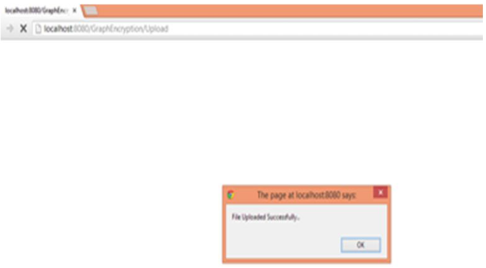
This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into research and development of system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized product had to be purchased.

VI. RESULT ANALYSIS

In this module an owner has to upload its files in a cloud server, he/she should register first. Then only he/she can be able to do it. For that he needs to fill the details in the registration form. These details are maintained in a database. In this module, any of the above mentioned person have to login, they should login by giving their name and password. Following fig. shows the Client Sign Up form.

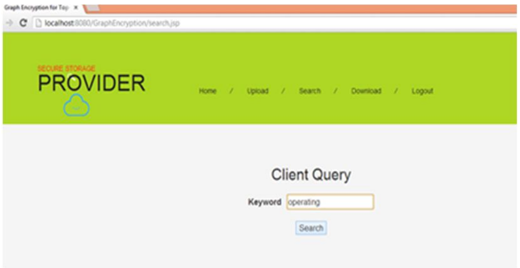


In this module user upload his file. The uploaded file is encrypted format. In this encryption process we are implementing DES (Data Encryption Standard). The uploaded file is not stored into the cloud server. In this project, we implemented the integrity for Uploaded data every time.


localhost:8080/GraphEncryption/Upload
The page at localhost:8080 says:
OK

File Uploaded Successfully..

In this module the Client can search for the files that he/she has uploaded after completing the process of sign up and after successfully uploading the file on the cloud server.


localhost:8080/GraphEncryption/search.jsp
SECURE STORAGE
PROVIDER
Home / Upload / Search / Download / Logout
Client Query
Search

Keyword: operating

VII. OBJECTIVES

1) *Flexibility:* It furnishes getting to data put away on cloud with chart encryption utilizing Top-k closest catchphrase seeks strategy.
2) *Low Cost of Storage:* Storage space of CSP is spared as just a single duplicate of similar data is put away.
3) *Big Data Support:* System supports huge data and gives low cost to data transfer and storage cost in productive manner

VIII. CONCLUSION

In this Project, we present a diagram encryption plot for Knk inquiries. The presented chart encryption conspire just utilizes lightweight cryptographic primitives, for example, pseudorandom work and symmetric key encryption, as opposed to moderate homomorphism encryptions. Along these lines, the proposed chart encryption conspire is benevolent to a wide arrangement of chart information based distributed computing and edge computing applications, for example, interpersonal organizations, e maps, criminal examinations, and so on. Contrast with chart anonymization comes closer from database network, our plot achieves higher security level as the chart itself is encoded and we do not make any suspicions on the kinds of assaults.

IX. ACKNOWLEDGMENT

We would like to acknowledge Sandip University, Computer department for their support, guidance and encouragement for writing this paper.

REFERENCES

[1] D. X. Song, D. Wagner, and A. Perrig. Practical techniques for searching on encrypted data. In IEEE Symposium on Security and Privacy, SP'00, pages 44–55, 2000.

[2] Babitha M.P, K.R. Ramesh Bab Secure Cloud storage using AES encryption. Department of IT, Government EngineeringCollege,2005

[3] C.R. Barde, Pooja Katkade, Deepali Shewale, Rohit Khatale. Secured multiple-keyword search over Encrypted Cloud Data. Department of computer, GESRHSCOE, Nasik, Maharashtra,2008

[4] Pengtao Xie and Eric Xing. CryptGraph: Privacy Graph Analytics on encrypted Graph. School of Computer Science, Carnegie Mellon University Pittsburg, 2008

[5] M. Chase and S. Kamara. Structured encryption and controlled disclosure. In ASIACRYPT, pages 577 -594. Springer,2010

[6] J. Cheng, A. W.-C. Fu, and J. Liu. K-isomorphism: privacy pre- serving network publication against structural attacks. In ACM SIGMOD, pages 459–470, 2010.

[7] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In Theory of Cryptography, pages 457–476. 2013.

[8] S. Chechik. Approximate distance oracles with constant query time. In ACM STOC, pages 654-663,2014

[9] D. Cash, J. Jager, S. Jareki, C. Jutla, H. Krawczyk, M. C Rosu, and M. Steiner. encryption in very large databases : Data structures and implementation. In NDSS, 2014

[10] J. He, M. Dong, K. Ota, M. Fan, and G. Wang. Netseccc: A scalable and fault-tolerant architecture for cloud computing security. Peer- to-Peer Networking and Applications, 9(1):67–81, 2016.

[11] A.P.Jaware, N.P. Borkar, "Implementation of a Secure and Dynamic Multi-keyword Ranked Search Scheme", IJCSMC, Vol. 6, Issues. 12, December 2017.