ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 7, July 2018

Efficient Keyword Search on Large Scale Graph in a Distributed System Using Different Searching Techniques

Rutuja Jagtap[1], Deepak Uplaonkar[2]

P.G. Student, Department of Computer Engineering, JSPM Narhe Technical Campus, Pune, Maharashtra, India[1]

Associate Professor, Department of Computer Engineering, JSPM Narhe Technical Campus, Pune, Maharashtra, India[2]

ABSTRACT: Graph Keyword Search has derived interest of number of research scientists, since we can represent graph models in both forms - structured and unstructured database. Keyword Search on the other hand can extract valuable information for users without knowledge of fundamental schema and query language. However, there are number of applications where graph can be much large than the available memory. If user tries to search for a location then first admin should enter a particular domain in the large data set. A web scale graph contains billions of vertices. The State of art approach employs centralized algorithms to process graph Keyword searches which are infeasible for such large graph due to limited computational algorithms to process and storage space of Centralized Server. To address this problem, we have worked and investigated keyword search for Web-scale graphs deployed in a distributed environment. Firstly, we have used a naive search algorithm to answer the query efficiently. Later, we encoded the shortest- path distance from a vertex to any given keyword in the graph. Performing these algorithms, we can find query answers by exploring fewer paths, so that the time and communication costs can be reduced.

KEYWORDS*:* Keyword Search, Distributed Graph, Scale Graph, and graph models.

I. INTRODUCTION

Graph Keyword Search has derived interest of number of research scientists, since we can represent graph models in both forms - structured and unstructured database. Keyword Search on the other hand can extract valuable information for users without knowledge of fundamental schema and query language. However, there are number of applications where graph can be much large than the available memory. If user tries to search for a location, then first admin should enter a particular domain in the large data set. A web scale graph contains billions of vertices. The State of art approach employs centralized algorithms to process graph Keyword searches which are infeasible for such large graph due to limited computational algorithms to process and storage space of Centralized Server. To address this problem, we have worked and investigated keyword search for Web-scale graphs deployed in a distributed environment. Firstly, we have used a naive search algorithm to answer the query efficiently. Later, we encoded the shortest-path distance from a vertex to any given keyword in the graph. Performing these algorithms, we can find query answers by exploring fewer paths, so that the time and communication costs can be reduced.

II. RELATED SURVEY

B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin found centralized keyword search over graphs, that is, given a set of keywords, we required to return top-k subtrees, subgraphs or h-cliques that contain these keywords. [1]

G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan found top-k subtrees by using backward search in BANKS-I. [3]

DOI: 10.15680/IJIRCCE.2018.0607007

M. R. Garey and D. S. Johnson found the exact top-k subtrees is an instance of the group Steiner tree problem, which is NP-hard. [5]

H. Wang and C. C. Aggarwalcalculates a reachability index which is used to answer if two given nodes are reachable in a directed graph. In an answer tree, every leaf node is reachable from the root. However, given leaf nodes, we do not know which vertex of the graph is the root in advance. Therefore, we cannot simply use a reachability index to answer the keyword query. [6]

J. B. Rocha-Junior and K. Nrvag studied top-k spatial keyword queries on road networks, but with different query aims from this paper. The reason that the query in this paper returns the k best objects ranked in terms of both spatial distance to the query location and textual relevance to the query keywords, whereas our studied query returns k answer trees with the smallest total distances. [7]

Y. Yuan, G. Wang, L. Chen, and H. Wang aimed at finding top-k sub trees from a large uncertain graph. The filtering-and verification framework is employed to answer the query efficiently. There are some works to study the variants of graph keyword search. The problem definition and solutions in this paper focus on the probability computation of sub trees over uncertain graphs in a centralized manner, whereas those in this paper concentrate on the parallel and distributed computation of sub trees over deterministic graphs. [8]

M. Qiao, L. Qin, H. Cheng, J. X. Yu, and W. Tian studied the top-k nearest keyword (k-NK) query over a graph. A k-NK query searches for k nearest answer nodes, each of which contains all the query keywords. In contrast, the query in this paper looks for subtrees, in which leaf nodes jointly contain all the query keywords. [9]

Q. Lu, X. Y. Jeffrey, and C. Lijun utilized the power of database and backward search to find answer trees. [20]

III. PROPOSED SYSTEM

In the system, there are three parameters such as the user, provider and admin. User first registers with all details and login the user searches the location with Domain of keyword, keyword and Location in which he is interested to search. Search location is stored on the graph dataset or normal data set. Admin store the location along with the location name, location keyword and domain of location. When user searches for a location, first he needs the access permission of particular user to search the location. Provider sends the encryption key to user to search the location in application and user Search the location with route.
Explanation:-
User/ Client: User searches the location after valid login and Registration and get the location and later, he finds the route from current location and Search location.
Provider: Provider sends the location and generates a randomly encryption key and checks for user details.
Admin: Admin adds the location with longitude and latitudes

A. Proposed System Architecture

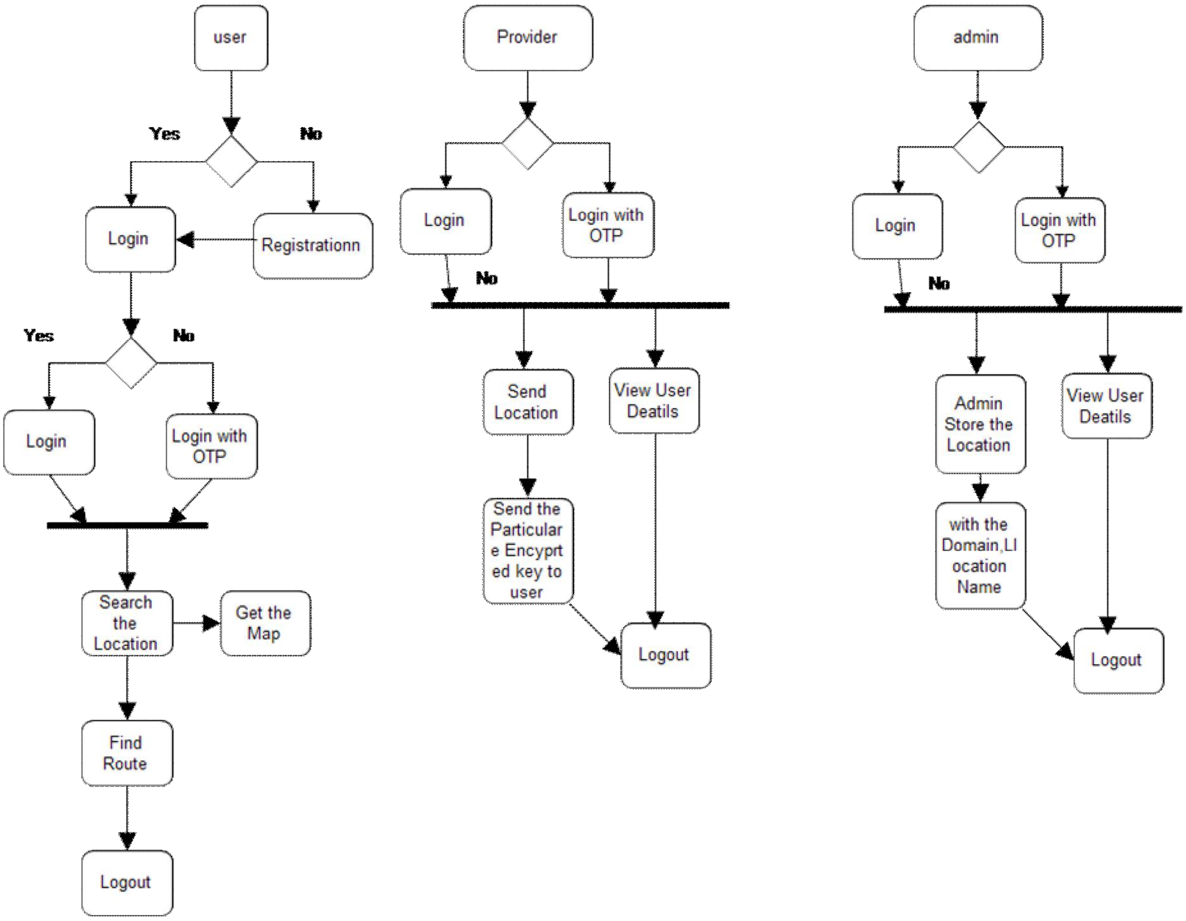

Fig. 1. Proposed Architecture

B. Mathematical Formulae:-

Input ni Belong V(G) , maximum Path desitances dc, NI index

Output if there is a path from ni to nj with distance no greater than dc return true; otherwise return false;
Let S, be a system such that,
S = s,Subi,Pubi,Si,E
Where,
S- Proposed System
s- Initial state at
Subi = the Registered Subscriber with attributes Sid,pwd,Li
where
Sid = subscriberid
Li = Locationid
P ubi = theRegisteredSubscriberwithattributesP id, pwd, Li
where

Li = Locationid
EEvent
N N ode
SSubscription
U BUpperBoundSidIdentifieransubscription
X – Input of the system
− L(Location), S(Searching)andRtreenode
Y OutputofSystemTopK Subscription matchingand show in Map

IV. EXPERIMENTAL RESULTS

Thus, we examined keyword search over distributed graph using signature based and partition algorithm. We studied how baseline algorithm is more efficient than signature based algorithm. Signature based algorithm searches for data from the original data set while baseline algorithm searches the data from partitioned database. Given graph shows how partition, baseline and AES algorithm works. It also shows the improved time and communication cost in baseline algorithm than the signature based algorithm.

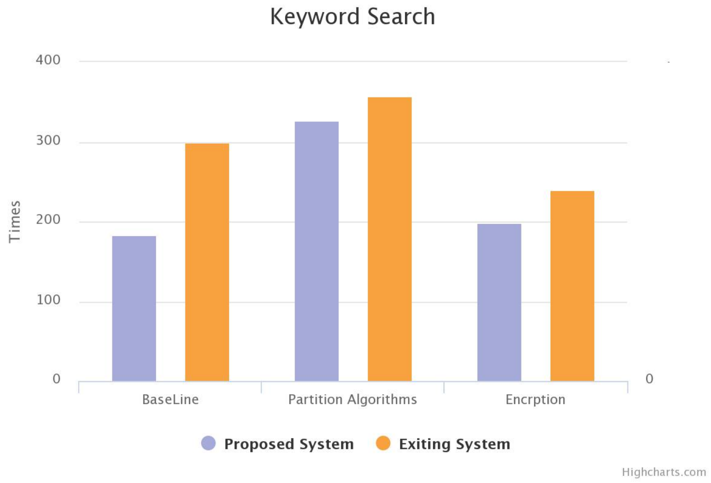

Fig. Existing System VS Proposed System

V. CONCLUSION

Applications on big graph usually suffer from extremely high time and space complexities, and thus require solutions of the distributed and parallel computing. In this paper, we have studied the problem of the keyword search over a distributed graph, which returns top-k answer trees. To efficiently find query answers, we give a distributed backward search algorithm and synchronize it over the cluster to terminate the search as early as possible. However, this naive solution incurs massive time and communication costs, due to the flooding nature of the search strategy. To solve this issue, we design a signature-based search mechanism, which explores the approach of spreading summary information about the vertex of hosted keywords in its neighbourhood, and then uses this information for forwarding the distributed backward search. Therefore, the signature-based search algorithm determines query answers by traversing only few paths of the graph, so that it costs very low time and network overhead. In addition, we re-organize the distributed graph to optimize the processes of the Decay and Diffusion Rules. Experiments show that our approach can improve the search efficiency by many times faster than the naive algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 7, July 2018

REFERENCES

[1] B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin, Finding top-k min-cost connected trees in databases, in ICDE, 2007.

[2] H. He, H. Wang, J. Yang, and P. S. Yu, Blinks: ranked keyword searches on graphs, in Proc. of SIGMOD, 2007, pp. 305316.

[3] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan, Keyword searching and browsing in databases using banks, in Proc. of ICDE, 2002, pp. 431440.

[4] J. Li, C. Liu, and J. X. Yu, Context-based diversification for keyword queries over xml data, Knowledge and Data Engineering, IEEE Transactions on, vol. 27, no. 3, pp. 660672, 2015.

[5] M. R. Garey and D. S. Johnson, Computers and intractability: a guide to the theory of NP- completeness. W.H.Freeman, 1979.

[6] H. Wang and C. C. Aggarwal, A survey of algorithms for keyword search on graph data, in Managing and Mining Graph Data. Springer, 2010, pp. 249273.

[7] J. B. Rocha-Junior and K. Nrvag, Top-k spatial keyword queries on road networks, in Proc. of EDBT, 2012, pp. 168179.

[8] Y. Yuan, G. Wang, L. Chen, and H. Wang, Efficient keyword search on uncertain graph data, TKDE, vol. 25, no. 12, pp. 27672779, 2013.

[9] M. Qiao, L. Qin, H. Cheng, J. X. Yu, and W. Tian, Top-k nearest keyword search on large graphs, Proceedings of the VLDB Endowment, vol. 6, no. 10, pp. 901912, 2013.

[10] L. Qin, J. X. Yu, L. Chang, H. Cheng, C. Zhang, and X. Lin, Scalable big graph processing in mapreduce, in Proc. of SIGMOD, 2014, pp. 827838.

[11] S. Luo, Y. Luo, S. Zhou, G. Cong, J. Guan, and Z. Yong, Distributed spatial keyword querying on road networks. in EDBT, 2014, pp. 235246.

[12] R. De Virgilio and A. Maccioni, Distributed keyword search over rdf via mapreduce, in European Semantic Web Conference, 2014, pp. 208223.

[13] C. Liu, L. Yao, J. Li, R. Zhou, and Z. He, Finding smallest kcompact tree set for keyword queries on graphs using mapreduce, World Wide Web, pp. 120, 2015.

[14] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, Pregel: a system for large-scale graph processing, in SIGMOD. ACM, 2010, pp. 135146.

[15] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, and A. Kyrola, Distributed graphlab: a framework for machine learning and data mining in the cloud, PVLDB, vol. 5, no. 8, pp. 716727, 2012.

[16] G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou, Ease: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data, in Proc. of SIGMOD, 2008, pp. 903914.

[17] D. H. (ed.), Approximation algorithms for NP-Hard problems. PWS, 1997.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to algorithms. MIT press, 2001.

[19] D. Guo, J. Wu, H. Chen, Y. Yuan, and X. Luo, The dynamic bloom filters, IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 1, pp. 120133, 2010.

DOI: 10.15680/IJIRCCE.2018.0607007