# Scalability and Performance Improvement in Graph Search Query Using FEM Framework

**Priyanka Chumbhale, Dr. M U Kharat**

Student of M.E., Computer Department,MET BKC Adgaon, Nasik, SavitribaiPhule Pune University, Maharashtra, India

Professor, Computer Department, MET BKC Adgaon, Nasik, SavitribaiPhule Pune University, Maharashtra,India

## Abstract

In graph or tree based search, shortest path computing can be defined as finding the distance between two vertices. Graph data can use in many domains like, in social networks or in knowledge graph. This graph search contains sub-graph. The problem of finding shortest path between two nodes can be solved using minimal spanning (Prime's or kruskal's algorithm) tree, the salesman traveling path, and the likewise. Problem is occurred with the graph based searching when graph is too big to fit in memory and hence it uses the external memory Disk-based method has some limitations when graph exceeds its size. In this paper, we are analyzing the shortest path for efficient relational approaches to graph search queries. For that, we describes FEM framework in this paper. It is used to bridge the gap between relational operations and graph operations. To improve the performance of FEM framework, we use window function and merge statement. Also to improve scalability without indexing we are proposing an edge weight aware graph partitioning scheme and design a bi-directional restrictive BFS.

## Introduction

Graph-based search is rapidly used in social networks or in knowledge graph. It is majorly required when graph search is over the graph. Graph may contain sub-graph [3]. Sometimes graph is too big to fit in memory and it exceeds the limit of memory. In previous disk-based system have limited memory therefore, they have limitations on graph search query if graph exceeds the main memory. Neo4j [2] is a technique which supports large size graph as well as it supports to primitive operations like, traversal or finding shortest path. This technique has some difficulty for supporting general graph search query. We observe that MapReduce framework and its implementation Hadoop [2] have capability of processing large graph in the distributed system. Relational Database (RDB) also supports graph search. According to the study of literature RDB and graph search have some overlapped functionalities, i.e. storage, data buffer, index, optimizations, etc. RDB can supports and manage complex data types, i.e. XML data [10], [7]. In this paper, we are focusing on shortest path search. There are two main reasons behind it, first is shortest path search plays key role in many large applications and second, it have similar evaluation pattern like other search query. In this paper, we are analyzing various shortest path discovery techniques.

## Literature Survey

**Jun Gao, Jiashuai Zhou, Jeffrey Xu Yu, and Tengjiao Wang[1]** proposed solution, to improve scalability without indexing we are proposing an edge weight aware graph partitioning scheme and design a bi-directional restrictive BFS. We also are representing weight aware edge partitioning scheme. We observe that MapReduce framework



ISSN: 2278 – 7798 

and its implementation Hadoop have capability of processing large graph in the distributed system. Relational Database (RDB) also supports graph search. According to the study of literature RDB and graph search have some overlapped functionalities, i.e. storage, data buffer, index, optimizations, etc. RDB can supports and manage complex data types, i.e. XML data.

**E. W. DrrKsrR[2]** provides the solution for constructing the tree of minimum total length between the n nodes. Authors were also focusing on solving the problem of finding minimal path between given node p and Q. In this approach nodes are divided into three sub-categories.

**Thomas Stutzle[3]** focused on contributions to TSP solving. In this Lin-Kernighan implementations are used. This paper reviewed symmetric and asymmetric TSP. heuristic solutions are constructed using SLS methods and also the Lin-Kernighan (LK) Algorithm, Population-based ILS Algorithms, The Memetic Algorithm (MA-MF), ACO Algorithms, Nearest-Neighbour Heuristic (NNH) for the TSP are proposed.

**Michalis Potamias, Francesco Bonchi, Carlos Castillo, Aristides Gionis[4]** describes methods that are landmark-based for point-to-point distance estimation in very large networks. It outperforms the current approximate standard Random and the state-of-the-art exact techniques. To provide fast estimates of the actual distance in very short time this paper use pre-computed information. This paper also uses social graphs with explicit or implicit links. For example, Flickr, Yahoo! Which having a graph based on the communication network and Instant Messenger service, etc. In this paper, five real-world datasets are used to show their experimental results. They are also researching about dynamic data structure.

**R. Prim[5]** described various algorithms for finding shortest path. i. e. Minimal spanning tree, traveling salesman problem, kruskal algorithm etc. In this paper, longest spanning sub-tree of a connected graph aims to maximize the symmentric functions. This paper, discussed about the arithimetizing of metric factors of the basic problem. This approach is used to handle the quite large-scale problems. This mainly focused to check closed cycles. Author kruskal discussed about the shortest spanning subtree in graph. Traveling salesman problem find the closed path minimum length.

**S. Trißl and U. Leser[6]** present the GRIPP index structure (GRaph Indexing based on Pre- and Postorder numbering). It is used to index graphs with 5-million and more nodes. It is fastest method for indexing typical and large biological networks. In this reachability queries are important in graph, one can recursively traverse the graph at query time, begins from v and then performing BFS or DFS search until no more edges remain. Authors adopt notations from Cormen. In this system, graphs are stored as a collection of nodes and edges in an RDBMS. Nodes in the graph includes unique identifier whereas binary relationship between two nodes. To answer the reachability pre-compute the transitive closure (TC). In this paper, proposed approach GRIPP extends the pre and post order labeling scheme to work on graphs. GRIPP indexing and sear h algorithm is also implemented.

**A. Goldberg and C. Harrelson[7]** developed bidirectional variants of search and also investigate several variants of the new algorithms. These algorithms are used to compute optimal shortest path for graph. This paper proposed a new lower-building scheme based on landmarks and pre-processing technique for evaluating distance bounds. This approach considered a square grid with integral arc lengths which are selected randomly. Lower-building scheme is important for quality of the bounds. This works on general graphs by taking the advantage of additional information. ALT- algorithm is used for best practices. P2P algorithm notice the least vertices in the graph.

**B. Bahmani, K. Chakrabarti, and D. Xin[8],** proposed fast MapReduce algorithm. It belongs to Monte Carlo approximation of personalized



ISSN: 2278 – 7798 

PageRank vectors of all the nodes in a graph. This algorithm is used to design our PPR approximation algorithm. In this paper, scalability is applied to existing algorithms. To implement new algorithms without the need to be database experts machine learning researchers. This approach couples Weka and the database. To retrieve records basic model uses the DBMS. Advanced SQL statements are used to implement advanced SQL statements and also some popular libraries are implemented. WekaDb uses standard JDBC API.

**Beibei Zou , Xuesong Ma , Bettina Kemme , Glen Newton , and Doina Precup[9]** proposed an idea in which a relational database is described as secondary storage which aims to eliminate limitations of previous relational database system. Weka is added as back-tier to relational database system. System refers as WekaDB. The proposed algorithm in this paper uses MapReduce iterations for the problem of optimal among a broad family of algorithms. Algorithm used to personalize PageRank vectors. For MapReduce this paper compute single random walks of a given length for all nodes in a graph. In terms of efficiency and approximation error, outperforms the state of the art FPPR approximation algorithms.

**C. Aggarwal, Y. Xie, and P. Yu[10]** developed a connectivity index for massive-disk resident graphs. In this an edge sampling based approach to create compressed representations of the underlying graphs. The study about the minimum connectivity problem for massive disk-resident graphs is discussed. A disk-based query index for connectivity queries is designed to achieve the proposed goal. Rather than one graph of very large size, it decomposed into multiple problems of smaller size. This paper, firstly, creates the index for connectivity queries. Number of nodes in this paper are assumed by authors are very large. Also, indexed Representation from Compressed Graphs is created.

# Conclusion

Relational database is used in multiple techniques.We first abstract a relational generic graph search framework FEM with three new operators, and employ the new features of SQL such as window function and merge statement to improve the performance of the FEM framework. Second, we optimize the basic method via the bi-directional restrictive BFS over weight aware partitioned edge tables, which can improve both performance and scalability significantly without extra overheads.

FEM technique is used previously but no table partitioning is used. FEM requires multiple complex SQL functions such as SELECT, INSERT, UPDATE at a time. No updated techniques such as WINDOW and MERGE function are used. If we distribute the database over multiple systems, efficiency of system can be improved. There is need of such system that provides efficiency in shortest path discovery using RDB structure.

# References

1. Jun Gao, Jiashuai Zhou, Jeffrey Xu Yu, and Tengjiao Wang,"Shortest Path Computing in Relational DBMSs", IEEE Transaction on knowledge and data engineering, Volume-26, Issue No-04,Page No-997-1011,April-2014.

2. E. Dijkstra, "A Note on Two Problems in Connexion with Graphs," Numerische Mathematik, vol. 1, pp. 269-271, 1959.

3. Thomas Stutzle, "The Traveling Salesman Problem: State of the Art", Darmstadt University of Technology Department of Computer Science Intellectics Group, July 10, 2003.

4. M. Potamias, F. Bonchi, C. Castillo, and A. Gionis, "Fast Shortest Path Distance Estimation in Large Networks," Proc. Int'l Conf. Information and Knowledge Management (CIKM'09), pp. 453-470, 2009.

5. R. Prim, "Shortest Connection Networks and



ISSN: 2278 – 7798 

Some Generalizations," Bell System Technical J., vol. 36, pp. 1389-1401, 1957.

6. S. Trißl and U. Leser, "Fast and Practical Indexing and Querying of Very Large Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD'07), pp. 845-856, 2007.

7. Goldberg and C. Harrelson, "Computing the Shortest Path: Search Meets Graph Theory," Proc. 16th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA '05), pp. 156-165, 2005.

8. Bahmani, K. Chakrabarti, and D. Xin, "Fast Personalized Pagerank on Mapreduce," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '11), pp. 973-984, 2011.

9. Zou, X. Ma, B. Kemme, G. Newton, and D. Precup, "Data Mining Using Relational Database Management Systems," Proc. 10th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining ( '06), pp. 657-667, 2006.

10. Aggarwal, Y. Xie, and P. Yu, "GConnect: A Connectivity Index for Massive Disk-Resident Graphs," Proc. VLDB Endowment, vol. 2, no. 1, pp. 862-873, 2009.



ISSN: 2278 – 7798 