www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 9 Sep 2015, Page No. 14436-14438

Approximate Processing of Queries in P2P Networks

Suraj N. Arya[1], Rajesh V. Argiddi[2]

[1]Solapur University, Walchand Institute of Technology,
Seth Walchand Hirachand Marg, Ashok Chowk, Solapur-413006 India
surajarya21@gmail.com

[2]Solapur University, Walchand Institute of Technology,
Seth Walchand Hirachand Marg, Ashok Chowk, Solapur-413006 India
argiddi@gmail.com

Abstract*: Peer-to-peer (P2P) databases are mostly used on the Internet for distribution and sharing of documents, applications, and other data. Finding answers to large-scale ad hoc queries like aggregation queries on these databases gives rise to many new challenges. Finding the exact solutions can consume a large amount of time and is also difficult to implement since the P2P databases are distributed and dynamic. In this paper, an approach for approximately answering of ad hoc queries in such databases is presented. Generally, the data is distributed across many peers in a distributed environment, and most of the times, within each peer, the data is highly correlated. This fact is taken advantage of and an approach to process the queries in such an environment is proposed in this work.*

Keywords: distributed query processing, ad hoc queries, peer-to-peer databases, aggregation queries**.**

1. Introduction

Peer-to-Peer (P2P) databases are mostly used nowadays for sharing files and partitioning and distributing data over the network. A P2P network is composed of many peer nodes that share data and resources with other peers. Also, it is possible to establish a hierarchy amongst the nodes, if required. In typical client-server models, the server plays a major role in providing services to its clients and also processing queries initiated from clients. In a P2P network, there is no central authority for doing such administrative tasks. Hence, there is lack of coordination amongst the peer entities. Also, there is absence of performance bottlenecks due to absence of failures. Such networks can be considered to be scalable, and dynamic, and due to this, it makes no huge difference in performance if a few nodes join and/or depart from the network.

The important applications of P2P databases involve those that include tasks like file sharing and data retrieval. An important type of applications that involve such databases includes queries like aggregation queries. Processing aggregation queries has a wide scope in applications of the areas of decision support, data analysis and mining, etc. Aggregation queries may be widely used in sensor networks for temperature and anomaly detection. Similarly, they may also be used in Intrusion Detection Systems.

The system proposed in this paper intends to make it easy to keep updated information about the different nodes in the peer network and also to retrieve and analyze the results for the aggregation queries at the target nodes. It is desired that the results to the aggregation queries must be generated with minimum latency in order to support real-time systems. If the information about the different active peers in the network and the information about databases on each peer are present, then it becomes easier to generate approximate results with minimum risks. Aggregation queries could be of the following simple form:

```
SELECT  aggreg_oper (col_name)
FROM    T
WHERE  selection_condition;
```

In the above query, the table T may be distributed over the P2P system, most preferably using horizontal partitioning. The aggreg_oper is any aggregation operation like SUM, AVG, COUNT, MIN, MAX, etc. while col_name denotes the name of the column on which the aggregation operation is to be applied. Such type of query has to be transformed into multiple queries each operating on different partition of the original database.

2. Literature Review

Adaptive sampling-based techniques were presented for the answering of ad hoc aggregation queries in P2P database systems. Minimum numbers of messages were required to be sent over the network and accordingly tunable parameters were needed to be provided to maximize the performance for different network topologies [1]. A system was efficiently built that facilitated efficient searches of large numbers of data providers on the internet. Each data provider or data source could become an autonomous node in a very large peer-to-peer network [2]. Indices were used on each node and accordingly queries were directed to the respective relevant sources from any node where the query was submitted. Such an approach was found to be feasible especially in those applications that involve a large peer-to-peer network. A system was developed that made use of an adaptive two-phase sampling approach based on random walks of the P2P graph [3]. It also made use of the block-level sampling techniques.

Query routing and processing are the main problems that arise due to the absence of a global catalog in a P2P network.

Solutions were proposed to efficiently handle these problems and accordingly generate the results for the query initiated by the user [4]. A protocol is proposed for participants to build P2P networks in a distributed fashion that resulted in connected networks of constant degree [5]. This resulted in efficient search and data exchange. Also, global knowledge about all the peer nodes in the network was not a compulsion to be known previously.

A framework was developed to classify current and future P2P network technologies. The main task included in this work was identifying the basic characteristics of the P2P network applications [6]. The infrastructure that may be developed on this idea may focus on P2P computing. A Decision Support System named Aqua was designed to provide fast approximate answers to aggregate queries [7]. Such queries are very common in OLAP applications. It precomputes special statistical summaries of the original data which are then stored in database. Approximate answers were given by rewriting the queries so that they may be run on the synopses computed. Also, the system keeps the synopses consistent i.e. up-to-date as the contents of the original database undergo changes due to transactions executing and updating the contents.

An efficient range of query processing scheme supporting range queries was proposed [8]. Two query processing algorithms were proposed respectively for single-attribute and multiple-attribute range queries within a bounded delay. The ability to answer approximately aggregation queries efficiently proves to be of large benefit for decision support and data mining tools. The techniques recognize the importance of considering variance in the data distribution [9]. This work may be implemented on a database system and may turn out to be of superior quality in generating approximate results.

3. Methodology

3.1 Algorithm

The following is the algorithm for the proposed system:

1) Start.
2) An input file named ActiveNodes.txt containing the detailed information about active nodes present in the system is provided to the system.
3) With the input provided in previous step, peer-network would be generated for executing the aggregation queries of the user.
4) A set of rules generated as a result of efficient study and analysis is made use of for assisting user for business analysis and constraints.
5) After completing the business analysis, approximate results are generated along with final error estimate.
6) Stop.

3.2 Proposed System

The system is designed in order to make it simpler to keep the updated information about the different nodes in the network. Then, it is expected to retrieve and analyze the results for the initiated query at the query node. The details of the network, different node locations and the databases at each of the peer nodes in the network, execution of the query along with their results' collection, etc. could be maintained. The system is designed for the processing of aggregate queries over the peer nodes in the network. The following is the architecture for the proposed system:

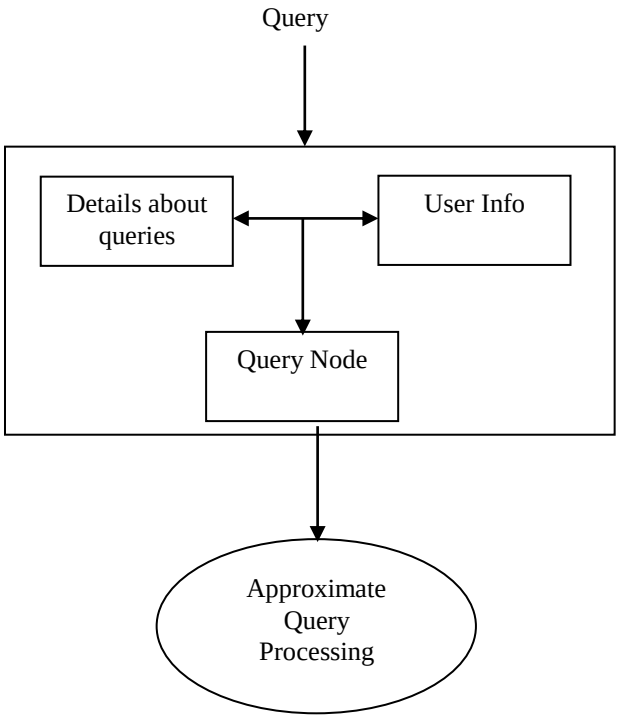

Figure 1: System Architecture

The proposed system includes connection of the nodes to form the peer network. As shown in the figure, the aggregation query would be taken for which the results are to be evaluated. The details about the query and the user information may be used for the further processing of the query. This may be desired as the approximate results of the queries need to be generated. Accordingly, the different peer nodes would be selected for the purpose of the query evaluation. The complete system includes four steps. The first step is the node construction where the peer-to-peer network is set up. The second step involves the selection of a node randomly for processing of the aggregate queries. Different algorithms or techniques may be used for the selection of such nodes. The third step includes selection of the records from the database that would be required for the further processing of the query. Also, the query would be run on multiple peer nodes and the results would be generated. Then the final step is to do the performance evaluation. This would involve comparison of the results generated at multiple peer nodes and then checking whether the results generated are valid or not. This could be called as approximate query processing. The aim is to increase the efficiency of such approximate query processing. The input given to the system at the initial stage would be a file that includes information about all the active nodes in the network that could work on real time databases.

3.3 Advantages

1) The proposed system can generate better results with large database size.

2) The proposed system can improvise the business of systems by introducing enhanced quality of service, efficient utilization of resources, etc.
3) The system could be even applied for those networks where the database is dynamic and constantly changing due to multiple transactions going on at different peer nodes.
4) The system may work efficiently even in those networks where the database is partitioned within the network.
5) The system may provide flexibility into ecommerce with the help of different techniques so as to achieve set of goals and to reduce the expected database latency.

4. Conclusion

The performance of aggregation queries in real-time database applications could be enhanced. Various techniques could be used for the approximate answering of ad hoc aggregation queries in P2P databases. The proposed system makes use of minimal number of messages to be sent over the network so as to achieve good quality service, increase in throughput, minimization of response time, and efficient utilization of resources.

REFERENCES

[1] Amol Bhagat, P. P. Pawade, V. T. Gaikwad, "Efficient Approximate Query Processing in P2P Network", National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2012) Proceedings published by International Journal of Computer Applications (IJCA), pp. 25-30, 2012.

[2] Leonidas Galanis, Yuan Wang, Shawn Jeffery, David DeWitt, "Processing Queries in a Large Peer-to-Peer System", Springer-Verlag Berlin Heidelberg, pp. 273-288, 2003

[3] Benjamin Arai, Gautam Das, Dimitrios Gunopulos, Vana Kalogeraki, "Efficient Approximate Query Processing in Peer-to-Peer Networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 7, pp. 919-933, July 2007.

[4] Raddad Al King, Abdelkader Hameurlain, Franck Morvan, "Query Routing and Processing in Peer-to-Peer Data Sharing Systems", International Journal of Database Management Systems (IJDMS), Vol. 2, No. 2, pp. 116-139, May 2010.

[5] Gopal Pandurangan, Prabhakar Raghavan, Eli Upfal, "Building Low-Datameter Peer-to-Peer Networks", IEEE Journal on Selected Areas in Communications, Vol. 21, No. 6, pp. 995-1002, August 2003.

[6] Krishna Kant, Ravi Iyer, Vijay Tewari, "A Framework for Classifying Peer-to-Peer Technologies", Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, May 2002.

[7] Swarup Acharya, Phillip Gibbons, Viswanath Poosala, "Aqua: A Fast Decision Support System Using Approximate Query Answers", Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999.

[8] R. Saravanan, "Processing of Query in Peer to Peer Networks", International Journal of Computer Applications, Vol. 9, No. 6, pp. 12-16, November 2010.

[9] Surajit Chaudhuri, Gautam Das, Vivek Narasayya, "A Robust Optimization-Based Approach for Approximate Answering of Aggregate Queries", ACM SIGMOD 2001, California, USA, 2001.

Suraj N.Arya is a Post Graduation student pursuing M.E in Computer Science and Engineering from Walchand Institute of Technology, Solapur. He received his Bachelor of Technology degree from Shri.Guru Gobind Singhji Institute of Engineering and Technology, Nanded affiliated to Swami Ramanand Teerth Marathwada University in 2010. His research interests lie in the area of Query Processing in peer to peer Networks.

Mr. Rajesh V. Argiddi is an Associate Professor in Computer Science and Engineering Department at Walchand Institute of Technology, Solapur. He received his B.E degree from Shivaji University, Kolhapur and M.E degree from Shivaji University, Kolhapur. He is currently doing his Ph. D from Solapur University, Solapur. His research area lies in Data Mining. Currently he is working for Indian stock market behavior analysis using Data Mining techniques. For this, he has published papers in various renowned journals.