A Class based Piece Selection for Multi-Dimensional Aggregated Data Distribution in Peer to Peer Network

N.T.Renukadevi
Assistant Professor, Dept. of CT-UG
Kongu Engineering College, Perundurai
Erode, Tamilnadu, India,

P.Thangaraj, PhD.
Professor and Head, Dept. of CSE
Bannariamman Institute of Technology
Sathy, Tamilnadu,India,.

ABSTRACT
A peer-to-peer (P2P) is a distributed application architecture, where each computer (node) in the net acts as a client/ server for the other nodes in the network. It allows sharing right to different sources and partitions tasks or workloads among peers. All the peers are equally privileged participants in the application. The searching performance, query expressivity, multi-dimensional distributed indexing are the emerging aspects of P2P. Its popularity is due to its scalability and distribution of large files efficiently without undue pressure on the origin server. In this work, it is proposed to investigate the performance of P2P networking when used to store multidimensional compressed aggregated data which can be used to download required data based on the query rather than downloading the entire database. For real-time query from users the specific pieces should be downloaded based on the query. To achieve this goal, it is critical to efficiently schedule the order in which pieces of the desired data are downloaded. Simply downloading pieces in sequential (earliest-first) order is prone to bottlenecks. Consequently it is proposed to implement aggregated class based scheduling ensuring high piece diversity while at the same time prioritizing pieces needed to maintain uninterrupted download based on query

General Terms
Peer-to-peer, Multi-dimensional Range Queries, Minimum bounding regions, Decision Support System, Cryptographic hash functions

Keywords
Distributed Hash Table (DHT), multidimensional routing indices (MRI), Distributed aggregation scheduling algorithm (DAS)

1. INTRODUCTION
A peer-to-peer (P2P) is a distributed application architecture, where each computer (node) in the net acts as a client/ server for the other nodes in the network. It allows sharing right to different sources Viz., peripherals, files and sensors devoid of a central server [1, 2]. It can be set within the home, a business, or over the Internet. All the nodes in the net of a particular type use a similar program to access among themselves. It partitions tasks or workloads among peers. All the peers are equally privileged participants in the application. They are anonymous in managing very large databases made up by thousands. The searching performance, query expressivity, multi-dimensional distributed indexing are the emerging aspects of P2P.

The ensuing solutions can be effectively employed in the forthcoming new distributed database systems to be used in large grid computing networks and in clustering database management systems. The success is related to the use of lossy data compression techniques (such as MPEG formats) that significantly reduces data transmission costs.

As the huge amount of resources such as storage capacity, computing power, and data transmission capability, supports data management which benefits the analysis of multidimensional data [3]. The multidimensional data model is represented as points in a multidimensional space whose dimensions correspond to different perspectives over data (hname, file i pairs).

Here, the users explore data and retrieve aggregates by issuing range queries and evaluate the data domain from which the aggregate information should be retrieved. But they can rely on lossy data compression in analytical data. In Decision Support Systems (DSSs) or statistical databases, users pave new way of expressing data rather than for accurate information extraction. So here high accuracy in little bit of queries is unnecessary. Perhaps, quick and effective answers to these basic queries makes users explorations better and also saves huge system resources.

A decentralized and self-organizing P2P systems challenge is efficiency, scalability & fault-tolerance. Yet, the issue of well-structured data storage and position is a challenge and so, the researchers have anticipated the structured P2P approach.

Here, the tightly controlled overlay topology and the data items mapping is defined properly. Their major focal point is on lookup efficiency i.e., the lookup paths reduction and routing state minimized for every node. A scalable distributed hash function to properly map data items to nodes is implemented by distributed hash tables (DHTs) [4]. They are termed good because of, their routing efficiency, scalability & entirety exploring. Each node assumes responsibility for a portion of the key space that is proportional to its power, and this property is maintained as nodes join and leave the system.

The data distribution and routing loads among nodes is another issue, Load balancing [5]. The uniform distribution of hash values among peers can be achieved by adopting any one strategy,

(i) The DHT address randomization (with better hash function) and

(ii) Make each DHT node responsibility for a balanced portion (DHT address space).

Randomized data distribution yields an elevated probability load disparity among peers (logarithmic factor [6]). By peer removals and/ or joins randomization [7] can be made uniform. As DHT protocols prop up exact-match lookups the randomization fails to protect the data locality and their

schemes are devoid to hold up composite queries (range queries & nearest-neighbor searches). But efficient load balance and data locality achievements are highly required by some applications and there are methods regarding it.

DHTs one core functionality i.e., the key k is routed to an appropriate node n in the overlay network that holds the given key. It can decide lookups for a 1-dimensional entity with algorithmic overlay routing hops [4]. The issues in the DHT designs are

(i) Creation of keys (node-ids & object-ids) using cryptographic hash functions (SHA-1),

(ii) Structuring routing information (routing tables) at nodes in the network, and

(iii) A good look-up query resolution scheme.

A node in the overlay must resolve a look-up request within log n time. By routing and storing the information in the routing table is vital. The core DHTs may look similar but there are differences in the algorithm implementation in building the overlay network (network graph structure), maintaining routing table and handling node join/depart. The fault-tolerance, efficiency of lookups, load-balancing, and inserts and closeness are the performance metrics in DHT evaluation. It is a right choice for single-dimensional queries but extending DHTs to hold (MRQ) is tiresome.

The way the query processing load is spread over the peers shows the distribution mechanisms proficiency as it divide the attribute space (multidimensional) among the peer set. MRQs are based on attribute ranges of values than on certain values. As there is no total ordering in the attribute space points resolving MRQs is a cumbersome task. As a window query the query interval has varying size, aspect ratio and position. The key challenges to adapt MRQs in a DHT network involve

(i) Data distribution mechanisms and

(ii) Data indexing or query routing techniques.

A distribution mechanism has these characteristics

(i) locality-Tuples/data points nearby in the attribute space that's mapped to the same node (limits the lookup complexity),

(ii) Load balance-number of data points indexed by each peer be the same (ensures uniform distribution of query processing) and

(iii) Minimum metadata-Prior information maps the attribute space to the peer space..

At present, P2P multidimensional data distribution mechanisms based (i) space filling curves, (ii) tree-based structures (iii) variant of SHA-1/2 hashing have been proposed.

In this work, it is proposed to investigate the performance of P2P networking when used to store multidimensional compressed aggregated data which can be used to download required data based on the query rather than downloading the entire database.

2. RELATED WORKS

Because of the occurrence of contention the inherent concurrent nature of peer-to-peer search techniques are mainly disregarded and the performance evaluation is performed on the basis of basic performance metrics, like total network traffic, and message hop counts.

Spyros Blanas et al., [8] focused on the methods for multidimensional range search by considering the contention effect in complex P2P network search. By presenting two new metrics namely responsiveness and throughput that are concerned to concurrency and contention, the peer-to-peer networks are evaluated.

The effect of contention on these networks is seen from the results obtained and the presence of contention that leads to no scale in networks is also illustrated. Some of the network properties are considered to be desired such as peer accesses or uniform data distribution will not be crucial like before are the indications obtained from the results.

In P2P networks the essential problem to be addressed is the estimation of the global data distribution. Several P2P applications like query processing, load balancing analysis, data mining, achieves advantage because of it. Minqi Zhou et al., [9] proposed a new algorithm based on compact multidimensional histogram information for the purpose of attaining extreme accuracy in estimation with less cost for estimation based on the compact multidimensional histogram information.

To obtain the estimation of the global data density with extreme accuracy and efficiency in a multi-dimensional histogram, maintaining data distribution that is distributed between peers without overlapping and all the part is additionally condensed using a set of discrete cosine transform coefficients that all peer are able to accumulate the compact information to the entire histogram hierarchically by means of information exchange.

Along with detailed theoretical investigation and validation the algorithms on discrete cosine transform coefficients hierarchically accumulating and density estimation error are presented. The efficacy and competence of the proposed methods on density estimation in dynamic P2P networks are validated from the wide-spread performance study.

In P2P networks, for the purpose of document retrieval applications and to maintain the aggregated one-dimensional values that represents the obtainable number of documents in a particular direction in the network is designed into the classical routing indices. Doulkeridis et al., [10] proposed the multidimensional routing indices (MRIs) concept that is appropriate to manage the multidimensional data represented by minimum bounding regions (MBRs). The aggregation of the MBRs might cause MRIs to reveal very deprived performance depending on the data distribution in peers that renders them to become inefficient. Therefore, the focus is mainly on a hybrid unstructured P2P network and to construct MRIs of extreme selectivity the parameters are evaluated.

By considering the detection of similar peers and to group and to reassign these peers to other parts of the hybrid network by means of distributed and scalable manner, the proposed methods introduced is able to boost the query routing performance.

Using large scale simulations, the benefits of the proposed method is illustrated. Finally, the results obtained from the experiments show that P2P similarity search is advanced due to implementation of the proposed method.

For multi-dimensional data applications the existing indexing and search methods are not appropriate as the goal of existing distributed indices for P2P networks is based on one-dimensional data and for centralized environments.

Kant ere et al., [11] introduced a completely decentralized indexing and searching framework called SPATIALP2P which is more appropriate for spatial data information sharing in structured P2P systems. The P2P applications are supported by SPATIALP2P where different sizes spatial information are inserted or deleted dynamically, and connect or disconnect the peers. Suitable locality and directionality of space are preserved by the proposed method. Thus, for point and range query operations SPATIALP2P executes remarkably.

3. METHODOLOGY

Different scheduling strategies are adopted by the P2P. The most commonly used scheduling methods are:

- Earliest-first scheduling: In Earliest-first the download of pieces is prioritized by its position in a media file. The disadvantage of this method is that the rare piece may not be propagated and thus the peers requiring the rare piece will encounter bottleneck.
- Rarest-first scheduling: In Rarest-first the rare pieces are prioritized for the download. This scheduling is used by Bit torrent networks.
- Random-selection scheduling: Random-selection is used when there is no available information to use for alternative strategies.
- Distributed aggregation scheduling algorithm (DAS) consists of two phases. One is to construct a distributed aggregation tree. Another one is to perform the distributed aggregation scheduling. The schedule generated by DAS is distributed and collision-free thus data aggregation can be done even without MAC protocol. A schedule of a node N in a sensor network is a sending time slot for N to send data during the slot. The second phase of DAS is to determine the schedules for all the nodes in a sensor network in a distributed manner to solve the distributed aggregation scheduling problem.

4. EXPERIMENTAL SETUP AND RESULTS

The performance of P2P networking when multidimensional compressed aggregated data is stored which can be used to download as required on data based on the query rather than downloading the entire database is investigated.

For real-time query from users the specific pieces should be downloaded based on the query. To achieve this goal, it is critical to efficiently schedule the order in which pieces of the desired data are downloaded.

Simply downloading pieces in sequential (earliest-first) order is prone to bottlenecks. Consequently it is proposed to implement aggregated class based scheduling ensuring high piece diversity while at the same time prioritizing pieces needed to maintain uninterrupted download based on query.

Simulations to evaluate the proposed method were conducted with the 1000 peers. Size of each block of data is 256Kb. The size of synthetic file used for simulation is 100. The bandwidth capability of node is given in Table 1. For Join/leave process, a flash crowd where all queries come within a 10-second interval was used.

Querying nodes depart as soon as they finish downloading. Number of neighbors of each node (degree d) is assigned 7. Maximum number of concurrent upload transfers is 5. All the simulation run were conducted for 5000 second.

Table 1: Bandwidth Capability of Node

Downlink (kbps)	Uplink (kbps)	Fraction of Nodes
784	128	0.1
1500	384	0.6
3000	1000	0.15
10000	5000	0.15

The following Figures 1-3 give the simulation results for mean upload and download bandwidth utilization over time, percentage error for queries and number of replicas created.

Figure 1: Mean Upload and Download Bandwidth Utilization over time

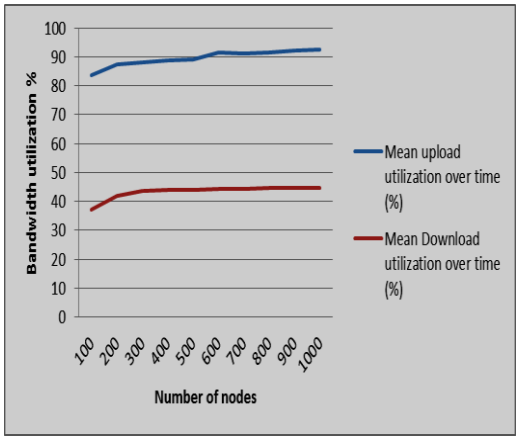


Figure 2: Percentage Error for Query

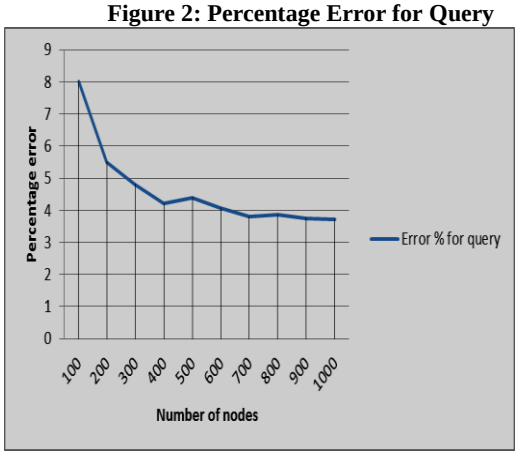
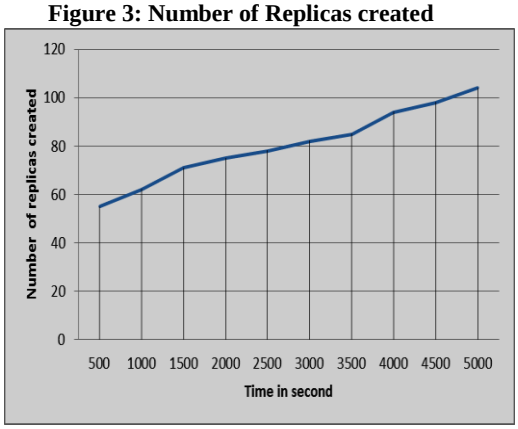


Figure 3: Number of Replicas created



5. CONCLUSION

The performance of P2P networking when used to store multidimensional compressed aggregated data which can be used to download required data based on the query rather than downloading the entire database is investigated. For real-time query from users the specific pieces should be downloaded based on the query. Simple downloading pieces in sequential (earliest-first) order are prone to bottlenecks. Consequently it is proposed to implement aggregated class based scheduling ensuring high piece diversity while at the same time prioritizing pieces needed to maintain uninterrupted download based on query. Simulation results are encouraging. The proposed method establishes efficient download based on query.

6. REFERENCES

[1] Moro, G. & Ouksel, A. M. (2003), G-Grid: A class of scalable and self-organizing data structures for multi-dimensional querying and content routing in p2p networks, in Proceedings of Agents and Peer-to-Peer Computing, Melbourne, Australia', Vol. 2872, pp. 123{137

[2] Milojicic, D. S., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., Rollins, S. & Xu, Z. (2002), Peer-to-peer computing, Technical Report HPL-2002-57, HP Lab.

[3] E. K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim. A survey and comparison of peer-to-peer overlay network schemes. IEEE Communications Surveys and Tutorials, 7(2):72{93, 2005.

[4] D. Boukhelef and H. Kitagawa. Dynamic load balancing in RCAN content addressable network. In ICUIMC, pages 98{106, 2009.

[5] G. Giakkoupis and V. Hadzilacos. A scheme for load balancing in heterogenous distributed hash tables. In PODC, pages 302{311, 2005.

[6] K.Kenthapadi and G. S. Manku. Decentralized algorithms using both local and random probes for p2p load balancing. In SPAA, pages 135{144, 2005.

[7] S. Blanas and V. Samoladas. Contention-based performance evaluation of multidimensional range search in p2p networks. In InfoScale'07, pages 1–8, 2007.

[8] Minqi Zhou, Weining Qian, Xueqing Gang, Aoying Zhou, Multi-dimensional data density estimation in P2P networks, Distributed And Parallel Databases 26 (2–3) (2009) 261–289.

[9] Kantere, V., Skiadopoulos, S., Sellis, T.: Storing and Indexing Spatial Data in P2P Systems. IEEE Transactions on Knowledge and Data Engineering (to appear).

[10] C. Doulkeridis, A. Vlachou, K. Nørvag, Y. Kotidis, and M. Vazirgiannis, "Multidimensional routing indices for efficient distributed query processing," in CIKM, 2009, pp. 1489–1492