Effective Use of Multiple Random Walks in P2P Networks

Zita Maria Almeida do Vale1), Carlos Ramos2), Rosslin John Robles3)

Abstract

Nowadays, millions of users search and download desired data such as Napster and Gnutella as they are Peer-to-Peer (P2P) files sharing applications. In order to improve performance in unstructured P2Ps replication strategies are used. Efficient and effective full-text retrieval over unstructured p2p networks was developed in order to address the problems of the query popularity independent replication strategies, previously a novel strategy. In order to support random node sampling and network size estimation a lightweight DHT with an unstructured P2P overlay. However these well-organized techniques are executed irrespective of topologies and network size concerns. To overcome this problem, we propose a query algorithm based on multiple random walks that resolve queries almost as quickly as unstructured P2P overlay method while reducing the network traffic by two orders of magnitude in many cases. We also present simulation results on a distributed replication strategy.

Keywords : Peer-to-Peer, multiple random walks, lightweight DHT, General search scheme, random node sampling.

1. Introduction

A peer-to-peer (P2P) network is a type of decentralized and distributed network architecture in which individual nodes in the network (called "peers") act as both suppliers and consumers of resources, in contrast to the centralized client–server model where client nodes request access to resources provided by central servers.

In a peer-to-peer network, tasks (such as searching for files or streaming audio/video) are shared amongst multiple interconnected peers who each make a portion of their resources (such as processing power, disk storage or network bandwidth) directly available to other network participants, without the need for centralized coordination by servers [1].

1GECAD, Institute of Engineering- Polytechnic of Porto, Portugal.
email: zav@isep.ipp.pt

2GECAD, Institute of Engineering- Polytechnic of Porto, Portugal.
email: car@isep.ipp.pt

3(Corresponding Author) Chair, Computer Science Department, University of San Agustin, Philippines
email: rosslin.john@yahoo.com
A peer-to-peer network is designed around the notion of equal peer nodes simultaneously functioning as both "clients" and "servers" to the other nodes on the network. This model of network arrangement differs from the client–server model where communication is usually to and from a central server. A typical example of a file transfer that uses the client-server model is the File Transfer Protocol (FTP) service in which the client and server programs are distinct: the clients initiate the transfer, and the servers satisfy these requests.

2. Routing and resource discovery

Peer-to-peer networks generally implement some form of virtual overlay network on top of the physical network topology, where the nodes in the overlay form a subset of the nodes in the physical network. Data is still exchanged directly over the underlying TCP/IP network, but at the application layer peers are able to communicate with each other directly, via the logical overlay links (each of which corresponds to a path through the underlying physical network). Overlays are used for indexing and peer discovery, and make the P2P system independent from the physical network topology. Based on how the nodes are linked to each other within the overlay network, and how resources are indexed and located, we can classify networks as unstructured or structured (or as a hybrid between the two) [2-4].

3. Unstructured networks

Unstructured peer-to-peer networks do not impose a particular structure on the overlay network by design, but rather are formed by nodes that randomly form connections to each other [5] (Gnutella, Gossip, and Kazaa are examples of unstructured P2P protocols [6]).

Because there is no structure globally imposed upon them, unstructured networks are easy to build and allow for localized optimizations to different regions of the overlay [7]. Also, because the role of all peers in the
network is the same, unstructured networks are highly robust in the face of high rates of "churn"—that is, when large numbers of peers are frequently joining and leaving the network [8][9].

![Overlay network diagram for an unstructured P2P network, illustrating the ad hoc nature of the connections between nodes](image)

However the primary limitations of unstructured networks also arise from this lack of structure. In particular, when a peer wants to find a desired piece of data in the network, the search query must be flooded through the network to find as many peers as possible that share the data. Flooding causes a very high amount of signaling traffic in the network, uses more CPU/memory (by requiring every peer to process all search queries), and does not ensure that search queries will always be resolved. Furthermore, since there is no correlation between a peer and the content managed by it, there is no guarantee that flooding will find a peer that has the desired data. Popular content is likely to be available at several peers and any peer searching for it is likely to find the same thing. But if a peer is looking for rare data shared by only a few other peers, then it is highly unlikely that search will be successful [10].

### 4. Structured Networks

In structured peer-to-peer networks the overlay is organized into a specific topology, and the protocol ensures that any node can efficiently search the network for a file/resource, even if the resource is extremely rare.

The most common type of structured P2P networks implement a distributed hash table (DHT), in which a variant of consistent hashing is used to assign ownership of each file to a particular peer [11]. This enables peers to search for resources on the network using a hash table: that is, (key, value) pairs are stored in the DHT, and any participating node can efficiently retrieve the value associated with a given key.
5. Related Work

Dongsheng Li states that with the increasing popularity of the peer-to-peer (P2P) computing paradigm, many general range query schemes for distributed hash table (DHT)-based P2P systems have been proposed in recent years. Although those schemes can support range query without modifying the underlying DHTs, they cannot guarantee to return the query results with bounded delay. The query delay in these schemes depends on both the scale of the system and the size of the query space or the specific query. In this paper, we propose Armada, an efficient range query processing scheme to support delay-bounded single-attribute and multiple-attribute range queries. We first describe the order-preserving naming algorithms for assigning adjoining Object IDs to objects with close attribute values. Then, we present the design of the forwarding tree to efficiently match the search paths of range queries to the underlying DHT topology. Based on the tree, two query processing algorithms are proposed to, respectively, process single-attribute and multiple attribute range queries within a bounded delay. Analytical and simulation results show that Armada is an effective general range query scheme on constant-degree DHTs, and can return the query results within 2 logN hops in a P2P system with N peers, regardless of the queried range or the size of query space.

Ion Stoica states that a fundamental problem that confronts peer-to-peer applications is to efficiently locate the node that stores a particular data item. This paper presents Chord, a distributed lookup protocol that addresses this problem. Chord provides support for just one operation: given a key, it maps the key onto a node. Data location can be easily implemented on top of Chord by associating a key with each data item, and storing the key/data item pair at the node to which the key maps. Chord adapts efficiently as nodes join and leave the system, and can answer queries even if the system is continuously changing. Results from theoretical
analysis, simulations, and experiments show that Chord is scalable, with communication cost and the state maintained by each node scaling logarithmically with the number of Chord nodes.

Edith Cohen States that the Peer-to-Peer (P2P) architectures that are most prevalent in today's Internet are decentralized and unstructured. Search is blind in that it is independent of the query and is thus not more effective than probing randomly chosen peers. One technique to improve the effectiveness of blind search is to proactively replicate data.

We evaluate and compare different replication strategies and reveal interesting structure: Two very common but very different replication strategies uniform and proportional yield the same average performance on successful queries, and are in fact worse than any replication strategy which lies between them. The optimal strategy lies between the two and can be achieved by simple distributed algorithms. These fundamental results order a new understanding of replication and show that currently deployed replication strategies are far from optimal and that optimal replication is attainable by protocols that resemble existing ones in simplicity and operation.

Christos Gkantsidis states that we quantify the effectiveness of random walks for searching and construction of unstructured peer-to-peer (P2P) networks. For searching, we argue that random walks achieve improvement over flooding in the case of clustered overlay topologies and in the case of re-issuing the same request several times. For construction, we argue that an expander can be maintained dynamically with constant operations per addition. The key technical ingredient of our approach is a deep result of stochastic processes indicating that samples taken from consecutive steps of a random walk can achieve statistical properties similar to independent sampling (if the second eigenvalue of the transition matrix is bounded away from 1, which translates to good expansion of the network; such connectivity is desired, and believed to hold, in every reasonable network and network model). This property has been previously used in complexity theory for construction of pseudorandom number generators. We reveal another facet of this theory and translate savings in random bits to savings in processing overhead.

Qin Lv states that Decentralized and unstructured peer-to-peer networks such as Gnutella are attractive for certain applications because they require no centralized directories and no precise control over network topology or data placement. However, the flooding-based query algorithm used in Gnutella does not scale; each query generates a large amount of traffic and large systems quickly become overwhelmed by the query induced load. This paper explores, through simulation, various alternatives to Gnutella's query algorithm, data replication strategy, and network topology. We propose a query algorithm based on multiple random walks that resolve queries almost as quickly as Gnutella's flooding method while reducing the network traffic by two orders of magnitude in many cases. We also present simulation results on a distributed replication strategy proposed in [8]. Finally, we end that among the various network topologies we consider, uniform random graphs yield the best performance.
Lada A. Adamic states that many communication and social networks have power-law link distributions, containing a few nodes that have a very high degree and many with low degree. The high connectivity nodes play the important role of hubs in communication and networking, a fact that can be exploited when designing efficient search algorithms. We introduce a number of local search strategies that utilize high degree nodes in power-law graphs and that have costs scaling sub linearly with the size of the graph. We also demonstrate the utility of these strategies on the GNUTELLA peer-to-peer network.

6. Existing System

We evaluate the performance of unstructured P2P overlay simulations. In this section, we describe the simulation setup. First, we introduce the unstructured P2P overlay traces we collected. We then describe the data used for evaluation including the WT10G data collection from NIST and the query logs. Finally, we present the metrics used for performance evaluation.

The topology of a small-world network has the properties of sparseness, short global separation, and high-local clustering of nodes while power law denotes the property of the node degree distribution.

BRITE is a topology generation tool that provides the option of generating topologies based on the AS model. Using BRITE, we generate a physical topology with 100,000 nodes. Using the physical topology generated by BRITE, we can simulate the underlying Internet with rich configuration information, including bandwidth configuration, latency, and so forth.

Using BRITE, we configure the upload bandwidth of a peer according to the measurement study on MSN from Microsoft in 2007. The study has shown that 97.2 percent MSN video users have upstream bandwidth higher than 128 Kbps (16 KBps). On one hand, this conservative configuration about peer bandwidth capacity indeed pushes the system performance examination close to the system limits. On the other hand, in practice a real-world peer-assisted text retrieval system may not want to fully exploit the available bandwidth of a high capacity peer, as doing so might deter their participation. There has been no standard data set established for evaluating the performance of P2P content search. We built one based on TREC WT10G collection, a large test set widely used for performance evaluation in information retrieval research area.

We have used several standard metrics for evaluating the performance of network. The evaluation considers both search quality and system efficiency. Quality focuses on user-perceived qualities such as recall, precision, FMeasure, and latency; while efficiency focuses on resource utilization such as traffic and efficiency.

7. Proposed System

We look at three aspects of a P2P system: P2P network topology, query distribution and replication. By
network topology, we mean the graph formed by the P2P overlay network; each P2P member has a certain number of neighbors" and the set of neighbor connections forms the P2P overlay network. In this paper when we refer to the network" we are referring to the P2P network, not the underlying Internet. For simplicity, we assume that the P2P network graph does not change during the simulation of our algorithms. By query distribution, we mean the distribution of query frequencies for individual files. Again, we assume that this distribution is fixed during our simulations. By replication, we mean the number of nodes that have a particular file.

**Normal Random Graph (Random):** a 9836-node random graph generated by a modified version of GT-ITM topology generator.

**Two-Dimensional Grid (Grid):** a two-dimension \((100 \times 100)\) grid. We choose this simple graph for comparison purposes.

**Power-Law Random Graph (PLRG):** this is a 9230 node random graph. The node degrees follow a power-law distribution: when ranked from the most connected to the least connected, the \(i\)'th most connected node has \(w^i\) neighbors, where \(w\) is a constant. Once the node degrees are chosen, the nodes are connected randomly.

### 8. Experimental Results

In the results we try to show that our proposed system will be having many advantages when compared to the previous methods.

![Graph showing the retrieval of information](image)

[Fig. 4] Graph showing the retrieval of information

In the above graph we will be observing the number of system related retrieval information’s between existing and proposed systems. Systems using the same information at a time in a network are shown in proposed and existing systems.
9. Conclusion

Here in this paper we conclude that by using the naïve method in P2P networks retrieval of information will be difficult as there will be much traffic as many users are using the same network at a time. So as a result the search process will be delayed. To overcome this problem we use our proposed technique multiple random walks so by using this we can decrease the traffic of the systems in the network. And the retrieval of information will be done very much effectively when compared to the traditional method.

References


