The International Journal of Engineering And Science (IJES)
||Volume|| 1 ||Issue|| 2 ||Pages|| 318-323 ||2012||
ISSN: 2319 – 1813 ISBN: 2319 – 1805

Content Based Retrieval in Unstructured P2P Overlay Networks

[1]Yoha Lakshmi.M, [2]Priya Ponnusamy.P

[1]*Department of Computer Science, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India*
[2]*Assistant professor, Department of Computer Science, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India*

----------**ABSTRACT**----------

The emergence of peer to peer (P2P) file sharing has led millions of users to search their desired data's easily. Napster and Gnutella are the two major applications in the P2P network. P2P network stands among one of the best and popular network tool. Sharing of contents by the user through internet has become easier through this. Bloom Cast is an efficient technique used for full-text retrieval scheme in unstructured P2P networks. By using the fullest of a hybrid P2P protocol, Bloom Cast makes copies of the contents in the network uniformly at a random across the P2P networks in order to achieve a guaranteed recall at a communication cost of the network. Bloom Cast model works only when the two constraints are met: 1) the query replicas and document replicas are randomly and uniformly distributed across the P2P network; and 2) every peer knows N, the size of the network. To support random node sampling and network size estimation, Bloom Cast mixes a lightweight DHT into an unstructured P2P network. Further to reduce the replication cost, Bloom Cast utilizes Bloom Filters to encode the entire document. Bloom Cast hybridizes a lightweight DHT with an unstructured P2P overlay to support random node sampling and network size estimation. Since P2P networks are self-configuring networks with minimal or no central control, P2P networks are more vulnerable to malwares, malicious code, viruses, etc., than the traditional client-server networks, due to their lack of structure and unmanaged nature. All peers in a P2P network is identified by its identity certificates (aka identity). The identity here is attached to the repudiation of a given peer. Self-certification helps us to generate the identity certificate, thus here all the peers maintain their own and hence trusted certificate authority which issues the identity certificate to the peer.

Keywords - *aka identity, Bloom Cast, Bloom Filters, Self-Certification, Self-Configuring Networks, Unstructured P2P network*

Date of Submission: 15, December, 2012 ⟺ Date of Publication: 30, December 2012

I. INTRODUCTION

When compared to the structured networks, searching process in unstructured networks are considered to be more challenging because of the lack of global routing guarantees provided by the overlay. More importantly, they are simple to implement and it provide virtually no overhead in topology maintenance. Even though, many real-world large-scale peer-to-peer networks are constructed as unstructured.

There are some typical unstructured P2P networks such as Gnutella, in which the peer searching is done by flooding a (hop) limited neighborhood. But this simple method does not provide any guarantee that an object that is exists in the peers in the network which is required. Flooding does not scale well in terms of message overhead, since each query may generate a significant amount of traffic. To solve this completeness and to test the capability issues Cohen studies how replication techniques can be used to improve search in

unstructured P2P networks. Objects are replicated based on their access frequencies.

Current P2P search mechanisms can be classified into three major types. First, at the server side a centralized index is maintained, and all queries are directed to the server.

Second, the query will be flooded across the network to other peers then the query node will broadcast the query to its neighboring nodes who will then broadcast to their neighbors, and so on. Such an approach will lead to poor network utilization.

Third approach is the Distributed Hash Table (DHT) based scheme where the Peers and data are structurally organized so that the location of a data

can be determined by a hash function. Chord is an example that uses a DHT-based scheme.

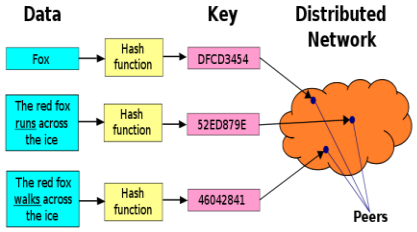

Fig.1 DHT (Distributed hash Table)

A distributed hash table (DHT) is a class of a decentralized distributed system that provides a lookup service similar to a hash table; (key, value) pairs are stored in a DHT, and any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from keys to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. This allows a DHT to scale to extremely large numbers of nodes and to handle continual node arrivals, departures, and failures.

DHT based P2P systems have several advantages such as it is scalable, robust and efficient. As a result, DHT has become a general infrastructure for building many P2P distributed applications. Applications such as content delivery, exchange of physical goods, services or spaces, networking, science and searches adds more advantage to DHT.

Here we use bloom filters, which is a hash based data structure used to reduce the amount of communication required. It has more benefit that it compares the keyword with the entire match list and found the exact match of the keyword. Here we can easily find the locality of the document where it is actually present. We can search the content with less amount of time.

This allows to achieve higher lookup performance for a given memory bandwidth, without requiring large amounts of buffering in front of the lookup engine.

II. RELATED WORKS

The content retrieval scheme is an important issue in the distributed P2P information sharing systems. There are two content searching schemes in the existing P2P systems. For a structured P2P networks it is DHT-based distributed global inverted index, for unstructured P2P networks we use federated search engines.

2.1. Pure P2p Systems

A pure P2P network does not have the notion of clients or servers but only equal peer nodes that simultaneously function as both "clients" and "servers" to the other nodes on the network. The network arrangement of this model differs from the client–server model. Here the communication is from and to the central server. File Transfer Protocol (FTP) is a typical example of a file transfer that does not use the P2P model. Here the client and server programs are distinct: the clients initiate the transfer, and the servers fulfill these requests.

The P2P overlay network consists of all the participating peers as network nodes. If there exists a link between any two nodes that know each other: i.e. if a peer knows the location of another peer in a P2P network, then there forms a directed edge between the former node and the latter in the overlay network. Based on the linking between the various nodes in the overlay network, we can classify the P2P networks as structured or unstructured.

2.2. Searching In Structured Networks

Structured P2P networks employ a globally consistent protocol to ensure that any node can efficiently route a search to some peer that has the desired resource or data, even if it a rare one. But this process needs more structured pattern overlay links. The most commonly seen structured P2P network is to implement a distributed hash table (DHT), in which deviating of hashing is used to assign ownership of files to that particular peer. It is not similar to the traditional hash table assignment in which in a for a particular array slots a separate key is assigned. The term DHT is generally used to refer the structured overlay, but DHT is a data structure that is implemented on top of a structured overlay.

2.3. Searching In Unstructured Networks

Unstructured P2P networks are formed when the overlay links are established randomly. The networks here can be easily constructed by copying existing links of another node and then form its own links over a time. In an unstructured P2P network, if a peer wants to find out a desired data in the network, the query is flooded through the network which finds many peers that share their data. The major disadvantage here is that the queries may not be resolved frequently. If there exists popular content then the available peers and any peer searching for it is likely to find the same thing. In cases where a peer is looking for rare data shared by only a few peers, then it is highly improbable that search will be successful. Since the peer and the content management are independent of each other, there is no assurance that flooding will find a peer that has the desired data. Flooding causes a high amount of signaling traffic in the network. These networks typically have very poor

search efficiency. Most of the popular P2P networks are unstructured.

2.4. Indexing and Source Discovery

The older P2P networks replicate the resources across each node in the network that is configured to carry out the type of information. This will provides the local searching, but it requires much more traffic.

Nowadays, the modern networks using the central coordinating servers and it provide the search requests directly. Central servers are mainly used for list out the potential peers that are present in the network, organizing their activities, and searching purposes. In decentralized type of networks, searching was first done by flooding the search requests along the peers. But now the new and more efficient search strategies called super nodes and distributed hash tables are used.

2.5. Overlay Networks

An overlay network is a type of the computer network which is built on the top of another existing network. Nodes that are present in the overlay network can be thought of as being connected as virtual links or logical links, each one of which is corresponds to a appropriate path, connected through many physical links, in the existing network. The major applications of overlay networks are, distributed systems such as cloud computing, peer-to-peer systems, and client-server systems, because they are run on top of the Internet. Initially the internet was built as an overlay network upon the telephone network whereas nowadays with the invention of VoIP, the telephone network is turning into an overlay network that is built on top of the Internet. The area in which the overlay networks used is telecommunication and internet applications.

III. PEER CREATION

Peer-to-peer (P2P) computing or networking is a distributed application architecture that divides the tasks among the peers. Peers are active and more privileged participants in the application. They are said to form a P2P network of nodes.

These P2P applications become popular due to some files sharing systems such as Napster. This concept paved a way to new structures and philosophies in many areas of human interaction. Peer-to-peer networking has no restriction towards technology. It covers only social processes where peer-to-peer is dynamic. In such context peer-to-peer processes are currently emerging throughout society.

Peer-to-peer systems implement an abstract overlay network which is built at Application Layer on the top of the physical network topology. These overlays are independent from the physical network topology and are used for indexing and peer discovery. The contents are shared through the Internet Protocol (IP) network. Anonymous peer-to-peer systems are interruption in the network, and implement extra routing layers to obscure the identity of the source or destination of queries.

IV. CONTENT SEARCHING AND REPUTATION MAINTAINANCE

In this section we are going to discuss detail about the concepts of content searching and reputation maintenance.

4.1. Content Searching

In a content search function, the input is a set of keywords representing a user's interests and the output is a set of resources containing these keywords. In the content search context, resources represent text documents or metadata of general resources. Some of these resources are software applications, computer platforms, or data volumes. Content search is useful when a user does not know the exact resource names of interests; this case is common in P2P-based searches as well as in web searches.

Flooding is the basic method of searching in unstructured P2P networks; however, large volume of unnecessary traffic is seen in blind flooding based search mechanism. This greatly limits the performance of P2P systems. The further study shows that a large amount of this unwanted traffic is divinable and can be avoided while searching in P2P networks.

The bloom hash table is used to store the resources which help in effective searching of resources with desired capabilities. It also provides more information about the path-name. This design enables resource discovery without knowledge of where the corresponding data items are stored.

4.2. Reputation Maintenance

The reputation systems based on the client-server model, the server here provides pseudonyms (identities) to users and inducts them into the system. Once logged into the system, a requester (client) selects a service provider (server) (from other users) for a given service, based on the reputation of the service provider. The requester then receives a service from the provider. Once the transaction is complete, the requester gives recommendation to the server based on its satisfaction level from the transaction. In P2P networks, there is no way to ascertain the distinctness of a peer in the absence of a central agency or without using external means.

V. BLOOM CAST

Bloom Cast is a novel replication strategy to support efficient and effective full-text retrieval. Different from the WP scheme, random node sampling of a lightweight DHT is utilized by the Bloom Cast. Here we generate the optimal number of replicas of the content in the required workspace. The size of the networks is not depending on any factor since it is an unstructured P2P network. The size of the network is represent here as N. By further replicating the optimal number of Bloom Filters instead of the raw documents, Bloom Cast achieves guaranteed recall rate which results in reduction of the communication cost for replicating. We can design a query evaluation language to support full-text multi keyword search, based on the Bloom Filter membership verification.

Bloom Cast hybrid P2P network has three types of nodes: they are structured peers, normal peers, and bootstrap peers. A Bloom Cast peer stores a collection of documents and maintains a local storage also known as repository. A bootstrap node maintains a partial list of Bloom Cast nodes it believes are currently in the system. In previous P2P designs, there are different ways to implement the bootstrap mechanism.

Bloom Cast is an inter-domain protocol, operating between border routers in the AS hosting the source (source AS) and the border routers of the ASes hosting the receivers (receiver ASes). However, for the sake of simplicity, we will treat each AS a single node.

5.1. Bloom Filters

Bloom Filters to encode the transferred lists while recursively intersecting the matching document set. A Bloom Filter is an efficient data structure method that is used to test whether the element belongs to that set or not. False positive retrieval results are also possible, but false negatives are not possible; i.e. a query returns either it is 'inside the set' or 'not inside the set'. Elements can only be added to the set and cannot be removed. When more elements are added to the set then the probability of false positives increases. Bloom Casting is a secure source specific multicast technique, which transfers the membership control and per group forwarding state from the multicast routers to the source. It uses in-packet Bloom filter (iBF) to encode the forwarding tree. Bloom Casting separates multicast group management and multicast forwarding.

It sends a Bloom Cast Join (BC JOIN) message towards the source AS. The message contains an initially empty collector Bloom filter. While the message travels upstream towards the source, each AS records forwarding information in the control packet by inserting the corresponding link mask into a collector. After this, it performs a bit permutation on the collector.

The figure for Bloom Filter and their memory storage is designed here to show the interconnections between source and specific multicast protocols. Unlike traditional IP multicast approaches, where the forwarding information is installed in routers on the delivery tree, in Bloom Cast, transit routers do not keep any group-specific state.

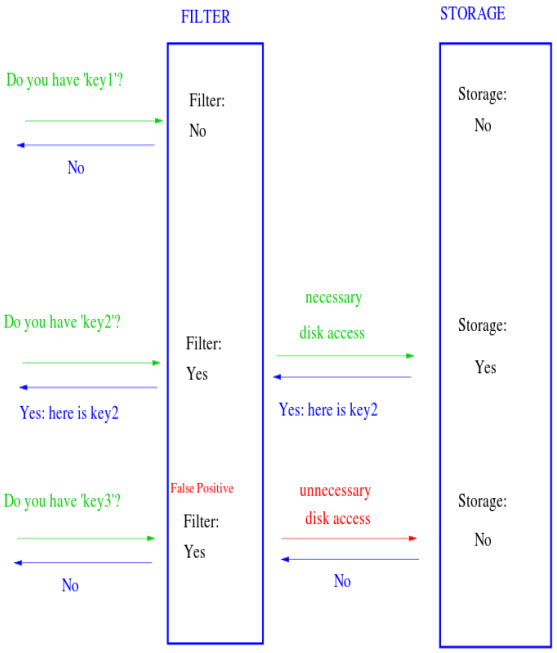

Fig.2 Bloom Filters

The above figure clearly explains the working mechanism of the bloom filters. If the user wants to search content, after giving the query, there available three possibilities of the result. The first of that is, the bloom filter will initially check whether the required content is actually present in the storage area, if it finds the key in storage, then it will gives the positive result to the user. The second is, if the original content is not found in the storage, then it will give the negative result. The third is, the original content may be deleted from the storage area and due to un-updating of the content, the bloom filter may have the chance to show the false positive results.

VI. SELF CERTIFICATION

The reputation of a peer is associated with its handle. This handle is commonly termed as the "identity" of the peer even though it may not A peer receives a recommendation for each transaction performed by it, and all of its recommendations are accumulated together for calculation of the reputation of a given peer.

Self-certification obviates the centralized trusted entity needed for issuing identities in a centralized system. Peers using self-certified identities remain pseudononymous in the system as there is no way to map the identity of a peer in the system to its real-life identity.

A malicious peer can use self-certification to generate a large number of identities and thereby raising the reputation of one of its identities by performing false transactions with other identities. There is no need for the malicious peer to collude with other distinct peers to raise its reputation. It only needs to generate a set of identities for itself.

VII. Content Auditing

Peer to peer (P2P) network systems are not preferred by content providers due to the occurrence of free-of-charge content sharing in such a model. This is sometimes referred to as the ""free rider"" problem. Therefore companies that sell digital content such as music, video or movies typically rely on ""direct download"" methods. In direct download, users download content either from the vendor's website directly or via a contracted CDN (content delivery network). This approach of content distribution is less efficient and powerful than P2P. Some P2P systems do require payment before a user can enter the network. However, once a user has access to the content on the network, she may directly share the purchased content with other users on other networks, without authorization from the content provider. Once these users learn of one another and know that they have similar interests, they can easily form a private community for free future sharing of similar content. Therefore, regular monitoring and auditing of P2P network is required in maintaining a paid and secure content distribution.

By analysing the results of the bloom hash table, we found that this is significantly faster than a normal hash table using the same amount of memory, hence it can support better throughput for router applications that use hash tables.

VIII. Conclusion And Future nhancement

We here propose an efficient and effective full-text retrieval scheme in an unstructured P2P networks using Bloom Cast method. Bloom Cast is effective here because it guarantees the recall with high probability. The overall communication cost of a full-text search is reduced below a formal bound. Thus it is efficient and effective among other schemes. Furthermore the communication cost for replication is also reduced since we replicate Bloom Filters instead of the raw documents across the network. We demonstrate the power of Bloom Cast design through both mathematical proof and comprehensive simulations based on the TREC WT10G data collection and query logs from a real world search engine.

Peer-to-peer (P2P) networks are self-configuring networks with minimal control. P2P networks are more vulnerable to dissemination of malicious code, viruses, worms, and Trojans than the traditional client-server networks, due to their unregulated and unmanaged nature.

All peers in the P2P network are identified by their identity certificates i.e. aka identity. The reputation of a given peer is attached to its identity. The identity certificates are generated using self-certification, and all peers maintain their own (and hence trusted) certificate authority which issues the identity certificate(s) to the peer.

References

Journal Papers:

[1] D. Li, J. Cao, X. Lu, and K. Chen, "Efficient Range Query Processing in Peer-to-Peer Systems," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 1, pp. 78-91, Jan. 2008.

[2] H. Shen, Y. Shu, and B. Yu, "Efficient Semantic-Based Content Search in P2P Network," IEEE Trans. Knowledge and Data Eng.,vol. 16, no. 7, pp. 813-826, July 2004.

[3] A. Broder and M. Mitzenmacher, "Network Applications of Bloom Filters: A Survey," Internet Math., vol. 1, no. 4, pp. 485-509, 2004.

[4] M. Li, W.-C. Lee, A. Sivasubramaniam, and J. Zhao, "SSW: A Small-World-Based Overlay for Peer-to-Peer Search," IEEE Trans. Parallel and Distributed Systems, vol. 19, no. 6, pp. 735-749, June 2008.

[5] B.H. Bloom, "Space/Time Trade-Offs in Hash Coding with Allowable Errors," Comm. the ACM, vol. 13, no. 7, pp. 422-426, 1971.

[6] X. Tang, J. Xu, and W. Lee, "Analysis of TTL-Based Consistency in Unstructured Peer-to-Peer Networks," IEEE Trans. Parallel and Distributed Systems, vol. 19, no. 12, pp. 1683-1694, Dec. 2008.

Proceedings Papers:

[1] R.A. Ferreira, M.K. Ramanathan, A. Awan, A. Grama, and S. Jagannathan, "Search with Probabilistic Guarantees in Unstructured Peer-to-Peer Networks," Proc. IEEE Fifth Int'l Conf. Peer to Peer Computing (P2P '05), pp. 165-172, 2005.

[2] E. Cohen and S. Shenker, "Replication Strategies in Unstructured Peer-to-Peer Networks," Proc. ACM SIGCOMM '02. pp. 177-190, 2002.

[3] P. Reynolds and A. Vahdat, "Efficient Peer-to-Peer Keyword Searching," Proc. ACM/IFIP/USENIX 2003 Int'l Conf. Middleware (Middleware '03), pp. 21-40, 2003.

[4] H. Song, S. Dharmapurikar, J. Turner, and J. Lockwood, "Fast Hash Table Lookup Using Extended Bloom Filter: An Aid to Network Processing," Proc. ACM SIGCOMM, 2005.

[5] C. Tang and S. Dwarkadas, "Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval," Proc. First Conf. Symp. Networked Systems Design and Implementation (NSDI '04),p. 16, 2004.

[6] G.S. Manku, "Routing Networks for Distributed Hash Tables," Proc. ACM 22nd Ann. Symp. Principles of Distributed Computing (PODC '03), pp. 133-142, 2003.

[7] V. King and J. Saia, "Choosing a Random Peer," Proc. ACM 23rd Ann. Symp. Principles of Distributed Computing (PODC '04), pp. 125-130, 2004.

[8] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H.Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," Proc. ACM SIGCOMM '01, pp. 149-160, 2001.

[9] C. Tang and S. Dwarkadas, "Hybrid Global-Local Indexing for Effcient Peer-to-Peer Information Retrieval," Proc. First Conf. Symp. Networked Systems Design and Implementation (NSDI '04), p. 16, 2004.

[10] F.M. Cuenca-Acuna, C. Peery, R.P. Martin, and T.D. Nguyen, "Planetp: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities," Proc. 12th IEEE Int'l Symp. High Performance Distributed Computing (HPDC '03), pp. 236-246, 2003.

[11] K. Sripanidkulchai, B. Maggs, and H. Zhang, "Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems,"Proc. IEEE INFOCOM '03, 2003.

[12] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker, "Making Gnutella-Like P2P Systems Scalable," Proc. ACM SIGCOMM '03, pp. 407-418, 2003.

[13] S. Saroiu, P.K. Gummadi, and S.D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," Proc. Multimedia Computing and Networking (MMCN '02), pp. 156-170, 2002.

[14] J.P.C. Jie Lu, "Content-Based Retrieval in Hybrid Peer-to-Peer Networks," Proc. 12th Int'l Conf. Information and Knowledge Management (CIKM), pp. 199-206, 2003.