American Journal of Networks and Communications
2013; 2(3): 62-66
Published online June 20, 2013 (http://www.sciencepublishinggroup.com/j/ajnc)
doi: 10.11648/j.ajnc.20130203.12

Textual based retrieval system with bloom in unstructured Peer-to-Peer networks

E. Mohan[1], S. Uvaraj[2], S. Suresh[3], U. Helen Monisha[4], N. Kannaiya Raja[5]

[1]Pallavan College of Engineering, Kanchipuram
[2]Arulmigu Meenakshi Amman College of Engineering, Kanchipuram
[3]Sri Venkateswara College of Engineering, Chennai
[4]Jei Mathaajee College of Engineering, Kanchipuram
[5]Defence Engineering College, Ethiopia

Email address:

emohan1971@gmail.com(E. Mohan), ujrj@rediffmail.com(S. Uvaraj), ss12oct92@gmail.com(S. Suresh), monishalazarav@gmail.com(U. H. Monisha), kanniya13@hotmail.co.in(N. K. Raja)

To cite this article:

E. Mohan, S. Uvaraj, S. Suresh, U. Helen Monisha, N. Kannaiya Raja. Textual Based Retrieval System with Bloom in Unstructured Peer-to-Peer Networks. *American Journal of Networks and Communications.* Vol. 2, No. 3, 2013, pp. 62-66.
doi: 10.11648/j.ajnc.20130203.12

Abstract: P2P network is the best and popular network sharing of contents by the user through internet. 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: Bloom Cast, Bloom Filters, Self-Certification, Self-Configuring Networks, Unstructured P2P Network

1. Introduction

Recently, P2P (Peer to Peer) systems, direct file sharing systems among the peers, are one of the most attractive file Sharing system. P2P architectures have high scalability and high performance due to the fact that its architectures which have characteristics of distributed file processing.

However, P2P Architecture is infamous for distribution channel of illegal contents. So we must apply the DRM (Digital Rights Management) system to the P2P architecture, and we should keep advantages of P2P even if after DRM system applied. Here, we propose new type of DRM applied P2P system architecture that keeps existing P2P system's advantages. Often referred to simply as peer-to-peer, or abbreviated P2P, peer-to-peer architecture is a type of network in which each workstation has equivalent capabilities and responsibilities. This differs from client/server architectures where some computers are dedicated to serving the others. Peer-to-peer networks are generally simpler but they usually do not offer the same performance under heavy loads. The P2P network itself relies on computing power at the ends of a connection rather than from within the network itself. In P2P networks, all clients provide resources, which may include bandwidth, storage space, and computing power. As nodes arrive and demand on the system increases, the total capacity of the system also increases. This is not true of client-server architecture with a fixed set of servers, in which adding more clients could mean slower data transfer for all users.

Once you have downloaded and installed a P2P client, if you are connected to the Internet you can launch the utility and you are then logged into a central indexing server. This central server indexes all users who are currently online connected to the server. This server does not host any files for downloading. The P2P client will contain an area where you can search for a specific file. The utility queries the index server to find other connected users with the file you are looking for. When a match is found the central server will tell you where to find the requested file. You can then choose a result from the search query and your utility when then attempt to establish a connection with the computer hosting the file you have requested. If a successful connection is made, you will begin downloading the file. Once the file download is complete the connection will be broken.

2. 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 fulfil 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 signalling traffic in the network. These networks typically have very poor Content Based Retrieval in Unstructured P2P Overlay Networks www.theijes.com The IJES Page 320 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.

3. System Approach

Types of Nodes are an interactive system which provides detect text file from copyright infringement in P2P file sharing by using bloomcast scheme.

Bootstrap node maintains a local repository and maintains the partial list of bloom cast nodes.

Normal peers to provide services of random node sampling and network size estimation.

Good connectivity and long uptimes are promoted to structured peers by bootstrap peers to forms a global DHT.

3.1. Node 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.

3.2. 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.

Algorithm 1 Query Evaluation of Bloom Cast

```
1: R ←Ø;
2: for all BFs replicated in this peer do
3: BooleanContainFlag ← True
4: for all terms in Q do
5: if □(j)(1 ≤ j ≤ k)s.t.BFx[hj(t)] = 0 then
6: ContainFlag ← False;
7: end if
8: end for
9; if ContainFlag True then
10: R ← R U {urlx};
11: end if
12: end for
13: return R
```

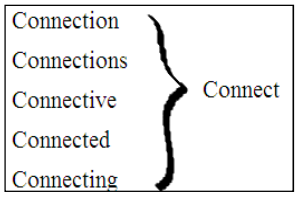
3.3. Stemming Algorithm

Stemming is the process for reducing inflected words to their stem, base or root form generally a written words form. Many search engines treat words with the same stem as synonyms as a kind of query broadening a process called conflation.

Function: Stemming is a process of reducing a word by removing some pattern. For example : when user searches with keyword 'Searching' then the stemming process will remove the 'ing' from 'searching' and you will get the 'search'. Then you can use this keyword 'search' to use for searching in the index server. It's done using porter algorithm.

Input: A query with collection of keywords.

Output: keywords are stemmed to their roots and used for the search system.



3.4. Bloom Filter

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.

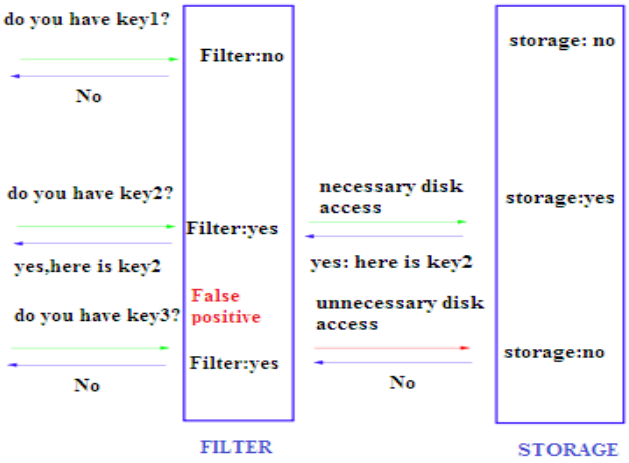


Fig 1. *Bloom Filter*

Efficient probabilistic data structure that is used to test whether an element is a member of a set.

- False positive retrieval results are possible.
- False negative are not.

Algorithm 2 Cast Bloom Filters

Require: Estimated NetworkSize N is achieved

1: for all documents in local collection do

2: create an empty bit vector with m bits for document x, BF,

3: for all terms in a document do

4: insert term t into BF, by setting the hj (t)th bits of BF, to 1, where {hj (.), 1 < j < k} is the set of hash functions used by BF,

5: end for

6: end for

7: sample an optimal number of r, random Peers in the network by the lightweight DHT;

8: replicate BF, together with url, the URL of document x, to the set of randomly sampled nodes;

9: return

3.5. Distribute Bf among the Nodes

Once the data are converted into the URL's, the url's are distributed to all other nodes. Once the node the request for the particular data in the network, once search has been finished, the best results will display to the user.

3.6. Ranking Process

The Ranking of data, so that the new users may able to find the exact data when they search/surfing. Using the chord algorithm, the peer node will do forward and backward search and as a result each document is provided with the rank and hence according to the rank given, the best document is identified by the server.

3.6.1. Retrieval of Data

By analyzing 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.

4. Dataflow Diagram

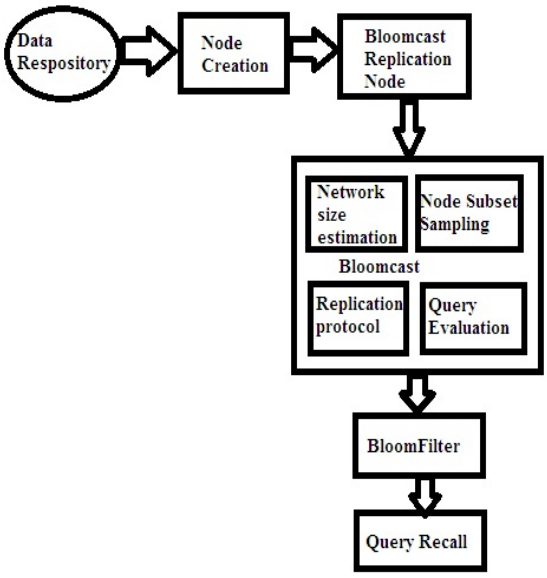


Fig 2. *Dataflow Diagram*

5. Conclusion & Future Enhancement

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.

6. Results

To evaluate the performance of BloomCast, in the simulation implement three baseline schemes.

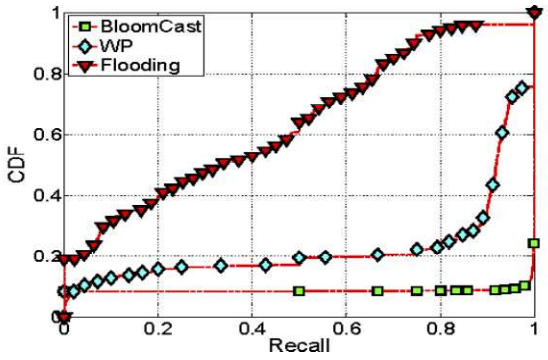


Fig 3. *Recall.*

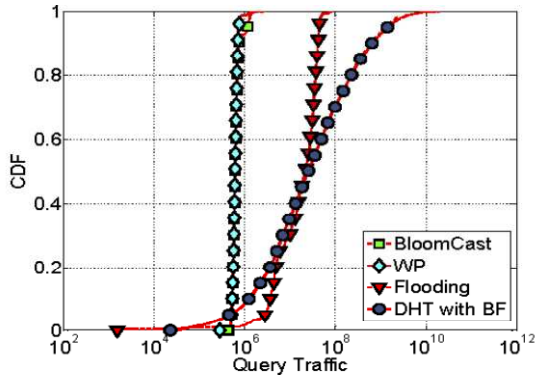


Fig 4. *Query traffic*

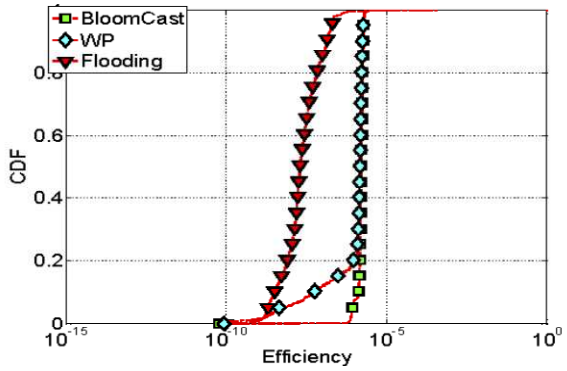


Fig 5. *Efficiency*

The result in Fig. 4 shows that the average query traffic of Bloom Cast is 6:5 _ 105, very similar with that of the WP algorithm. The average traffic of BloomCast is much less than that of flooding.

References

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

[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] 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.

[4] S. Robertson, "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF," J. Documentation, vol. 60, pp. 503- 520, 2004.

[5] 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.

[6] 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. 7. 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.

[7] 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.

[8] E.M. Voorhees, "Overview of Trec-2009," Proc. 16th Text Retrieval Conf. (TREC-11), 2009.

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