International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 3, Issue 4 (November 2013) PP: 51-55

Optimizing of Bloom Filters by Automatic Bloom Filter Updating and Instantly Updating Service Availability on Peers in P2P Network

Amulia P.M[1], Jisha G[2]
[1]Department of Information Technology, RSET, Kakkanad,

Abstract: *P2P networks have become popular in the distributed computing paradigm. Different search mechanisms are used in P2P networks. Although flooding search technique, often provide required search results, this technique incurs lot of unnecessary traffic in the network. Bloom filter technique was developed to reduce the network traffic and thereby reduce communication cost. The main drawback of bloom filter is false positive results. False positives increase when data becomes stale. To reduce the stale data problem the paper proposes a mechanism to automatically updating the bloom filter information. The optimized bloom filter technology we propose provides automatic registering and deregistering of services. So information in the encoded bloom filters will be current and only contain the membership from the currently service-available registered peers. This reduces overloading the network with search request and also guarantees required search result with minimum overhead and wastage of resources. The expiry-date based encoded bloom filters we propose provide automatic refreshing of the bloom filter set. This also minimizes stale information and there by reduces the chance of false positive results.*

Keywords: *P2P, bloom filter, service discovery, optimizing, query*

I. INTRODUCTION

Peer-to-peer (P2P) network is a popular information sharing tool where data is located in a distributed manner mostly in geographically separate locations. It also includes distribution of resources.
Currently, the common search process in p2p network includes the following steps:

1. Locating the service providing peers.
2. Forwarding the query to peers providing the service.
3. Local processing of the query.
4. Retrieving the results for the query.

The search efficiency depends greatly on the time taken to provide the result. The common blind search technique is flooding. Flooding uses the basic Breadth First Search (BFS) and floods the network with query. This requires contacting unnecessary nodes that in no way can contribute to the search result. Query search requires the result of the query to be received with minimum delay and lesser bandwidth wastage. With user queries becoming broad and complex multi-keyword searches are becoming popular. Contacting only the peer nodes that can provide the required search result becomes crucial in reducing communication cost. In a query that includes “distributed computing” the multi keywords are separated into individual keywords “distributed” and “computing”. The traditional flooding method includes that each keyword be separately searched at all the peers in the network and the results to be merged at the selected peer node. This causes flooding technique to contact large number of unnecessary peers and also waste bandwidth and other scarce resources. The search will be heavily time consuming leading to user frustration. The bloom filter search mechanism reduces the wastage of resources by using encoded filters. This reduces unnecessary traffic in the network as it requires contacting only the required peers for the keyword result. The main drawback of bloom filter is false positive results. False positives increase when data becomes stale. Optimizing bloom filters is very important to reduce communication cost in multi keyword search that require intersection or union (AND or OR) [1] query operations. The mechanism we propose uses checking service availability by polling peer nodes in the network. By getting the current service-providing peers, only these peers need to be contacted with query request. Using this information about the service providing peers encoded bloom filter [3] data can be updated automatically with timestamp based current information. The current information available with bloom filter minimizes stale data in the bloom filter set. This provides current search result and reduces search delay and wastage of resources. The number of unnecessary nodes contacted is reduced and there by reduces the communication cost. Optimizing bloom filters using expiry date method improves multi keyword search.

The rest of this paper is organized as follows. Section 2 elaborates on flooding and bloom filter search techniques and the requirement for optimizing bloom filter search technique. Section 3 elaborates on the design of optimizing bloom filter search technique using automatic polling of peers and expiry date based bloom filter data. Implementation procedures for generating optimized bloom filter based search mechanism and the graphical performance results are given in section 4. Conclusion is given in section 5.

II. FLOODING AND BLOOM FILTER SEARCH MECHANISMS AND THE REQUIREMENT FOR OPTIMIZING BLOOM FILTER SEARCH

The common search technique [6] used in a P2P network is flooding. Flooding comes under blind search [2]. The flooding technique uses simple broadcasting method where each node contacts all its neighbors. The query is propagated throughout the network or until the search result is obtained. This provides unnecessary load in the network and drastic increase in search time. The method to reduce the load in the network and to reduce the search time is to include Time to Live (TTL) [5] or hop count with the query search. TTL or hop count can be used to control the number of hops the query needs to be propagated in case the required search result is not obtained. Choosing the appropriate TTL or hop count is difficult. If the TTL is too high it creates unnecessary burden on the network. If it is too low the required result may not be obtained even though the result might be somewhere in the network. Bloom Filter reduces the unnecessary overload generated in the network when using flooding search technique.

In multi keyword search the difficulty is in locating the peer nodes which is responsible for the keywords. The multi-word query is separated into single keywords and each keyword is separately searched. If the result requires an intersection (AND) operation the common documents of the entire keyword search is taken. Only those document identifiers that have all the keywords in the query are sent to the searching node as the final result. For a union operation the keywords could be searched in separate nodes and all the document identifiers that have at least one keyword in the query could be transmitted to the searching node. The merged result of all the documents is the final result. The regular multi keyword search requires large replication of documents at peers across the network resulting in storage and consistency problems. The large volume of raw data transmission could produce unnecessary traffic in the network leading to bandwidth wastage. Thus P. Reynolds and A.Vahdat [7] suggested bloom filters to recursively encode the transmitted lists. Bloom filter [3,4] is an efficient data structure that can represent a set S and also process a membership query 'is x in set S'. Bloom filters are space efficient data structures. This minimizes storage and transmission cost that is a huge problem in flooding based search technique. The mapping of which location to look for the result, greatly reduces unnecessary load in the network and wastage of resources. The problem with bloom filter is, these encoded sets could include stale or false information. The documents that are old or removed from the network could be included in the bloom filter encoded set. The peers located for performing the search could be removed from the network. There should be a mechanism to automatically update the current service providing peers. There should also be a mechanism to reduce the possibility of obtaining stale result for the search.

Bloom Filter reduces the unnecessary overload generated in the network when flooding search technique is used. There should be some mechanism to improve bloom filter search mechanism. With multi keyword search becoming popular in peer-to-peer network, there should be techniques to guarantee required search results with minimum overhead and delay. This will reduce the communication cost and wastage of resources.

III. DESIGN FOR THE OPTIMIZING OF BLOOM FILTER BASED SEARCH TECHNIQUE

In this section we give a brief overview of the design of our proposed system for multi keyword search in a peer-to-peer network. In search the important aspect is to locate the appropriate peer node which is responsible for the search result.

The design uses a P2P network. The technique to locate the peer nodes for the search service is the random probing. The service discovery and update mechanism sends automatic messages to the peer nodes in the network. The peer nodes that respond to the message are added to the service registry. These peers are considered registered for the multi keyword search service. The nodes in the network that no longer respond to the requests are deregistered from the service registry. When service discovery is performed an automatic refreshing of the service registry is also done, that is registering new services and deregistering of services if they are no longer available. This removes the chance of any deleted or unauthorised service to be added to the service registry. It also adds new peers that have started providing the service. Only the registered peers need to be contacted for the required document service search. This reduces overloading the network with search request and also guarantees required search result with minimum overhead.

The encoded bloom filters are generating so as to reduce stale data in the bloom filter set. This uses an expiry date based mapping technique. If the current date is greater than expiry date the encoded bloom filter table is populated again with new mapping information. This reduces the chance of stale membership in the bloom filter set.

IV. IMPLEMENTATION PROCEDURES AND GRAPHICAL RESULT COMPARISON FOR THE TIME TAKEN FOR AND/OR (INTERSECTION/UNION)QUERY USING BLOOM FILTER Vs OPTIMIZED BLOOM FILTER BASED SEARCH

Table 1
Expiry Date based Bloom Filter Optimization Table with Keyword-Document-Peer Mapping

Keyword	Document	Peer Nodes	Keyword Occurrence in Each Document	Expiry Date
Keyword1	Doc1.txt	Peer1	2	Date1
	Doc2.txt	Peer2	3	Date1
	Doc3.txt	Peer3	1	Date1

The implementation of a p2p network was done by creating an overlay network with automatic Service Discovery those polls for any new service. The new service is registered for responding peers and any non-responding peer nodes are deregistered from the registry table. This is done using a polling mechanism. Bloom filters are generated using query-document-peer mapping. The Optimized Bloom Filter checks with the registry service, for available peers that provide the search. Then from the expiry date based encoded bloom filter optimization table the peers with documents that contain the “search keyword” is selected. For the search the query containing multi keywords is split into a set of individual keyword and searched separately. For AND query the intersection operation for the document results for each individual keyword is obtained by using a document-name matching mechanism. For OR query all the documents for each individual keyword search result are merged. The merge result does not include duplicates. Our results show about 185 times improvement in search time when compared to regular bloom filter search.

A test suite was created to check the performance of Flooding Vs. Regular Bloom Filter search technique and also check the performance of Bloom Filters Vs. Optimised Bloom Filters. We compared the time taken to obtain query search results for both intersection and union operation. The testing was done for various number of test runs. The ratio of run-time against the number of tests ran stayed almost the same. Even during stress testings the improvement in time for optimised bloom filters were same.

Steps Involved in the Implementation of Optimised Bloom Filter Search Mechanism in a P2P Network.

- Create an overlay P2P network with automatic service discovery.
- Allocate evenly the available documents (files) to be searched for the keyword on each peer node.
- Create a hash-map with keyword-document-peer mapping.
- Using the hash-map, generate bloom filters showing membership of each document on the available peers.
- Also give an expiry date to the mapping so as to populate the bloom filter table if the expiry date has reached.
- Perform the multi keyword search by splitting the query into individual keywords.
- Get the multi keyword search results using AND (intersection of individual keyword search results) or OR (union of individual keyword search results) operation.
- Retrieve the result document from the peer node.

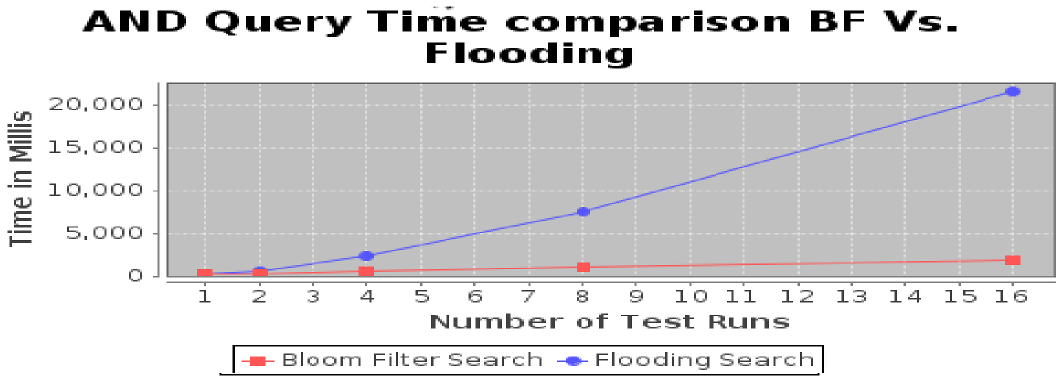

Fig 1. Graphical representation of the time taken for a intersection (AND query) operation in Flooding and Bloom Filter based search against the number of test runs.

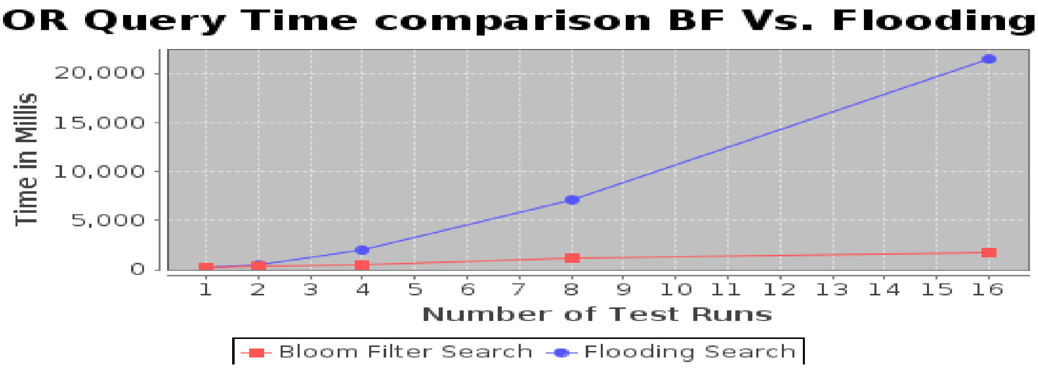

Fig 2. Graphical representation of the time taken for a union (OR query) operation in Flooding and Bloom Filter based search against the number of test runs.

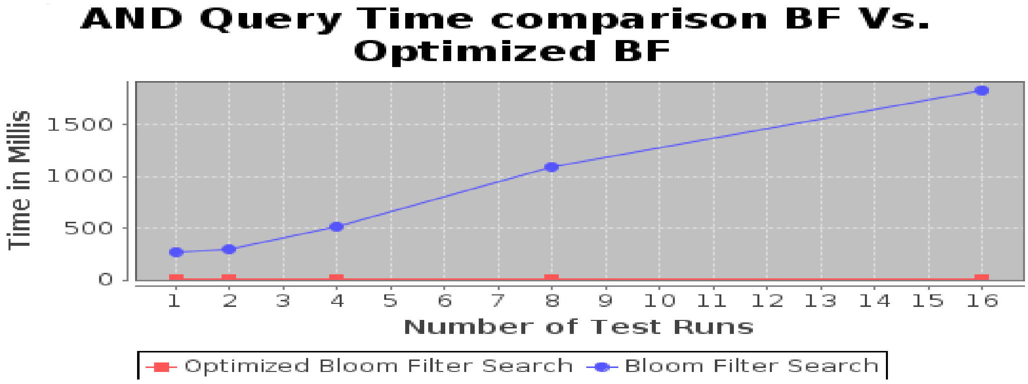

Fig 3. Graphical representation of the time taken for intersection operation (AND query) Bloom Filters and Optimized Bloom Filter based search against the number of test runs.

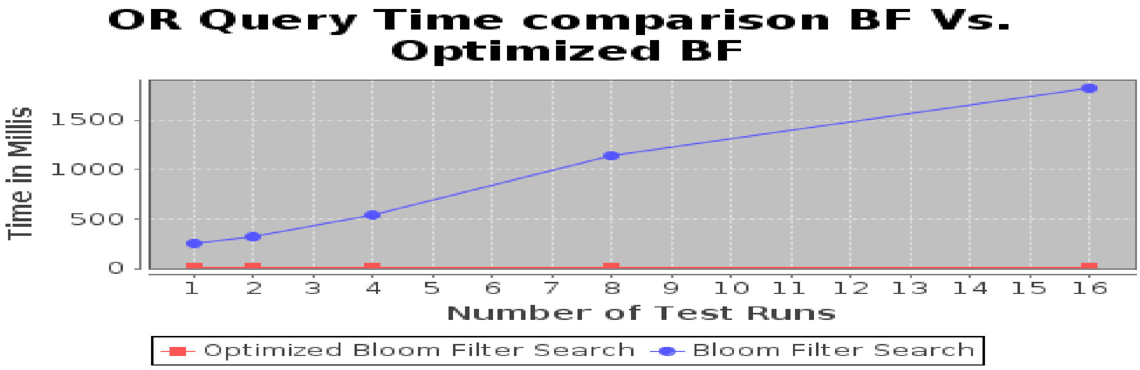

Fig 4. Graphical representation of the time taken for union operation (OR query) Bloom Filters and Optimised Bloom Filter based search against the number of test runs.

The performance analysis shows improvement in search time for multi keyword search using optimized bloom filter based search for both intersection and union operation.

V. CONCLUSION

Significant research has been done in the P2P research field. Efficient search mechanisms have to be developed to improve efficiency and cost effectiveness in multi keyword search. Flooding technique is the basic search mechanism commonly used in a P2P network. This method is reliable but affects the scalability and efficiency of the P2P network. The techniques specified in this paper provide optimum query forwarding with the quick and current search result. In multi keyword search which required AND or OR operation minimizing traffic in the network is crucial. The improvement to the bloom filter search technique proposed in this paper reduces communication cost as it reduces traffic in the network and also avoids wastage of resources.

REFERENCES

[1]. Efficient Multi-keyword Search over P2P Web; Hanhua Chen, Hai Jin , School of Computer Science and Technology Huazhong University of Science and Technology Wuhan, 430073, China {chenhanhua,hjin}@hust.edu.cn Jiliang Wang,Lei Chen,Yunhao

Liu,Lionel Ni Department of Computer Science and Engineering Hong Kong University of Science and Technology Clear Water Bay, Kowloon, Hong Kong {aliang,leichen,liu,ni}@cse.ust.hk. Proceedings of the 17th international conference on World Wide Web, 2008, ISBN: 978-1-60558-0852 DOI:10.1145/1367497.

[2]. Efficient Search Techniques in Peer to Peer Networks, Tarunpreet Bhatia, Dr Deepak Garg, Computer Science Department, Thapar University, Patiala. International Journal of Computer Applications (0975- 8887) Volume 36– No.1, December 2011.

[3]. Optimizing Bloom Filter Settings in Peer-to-Peer Multikeyword Searching Hanhua Chen,Member, IEEE, Hai Jin, Senior Member, IEEE, Lei Chen, Member, IEEE Yunhao Liu,Senior Member, IEEE, and Lionel M. Ni, Fellow, IEEE, IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 4

[4]. Bloomcast Efficient and Effective full-text retrieval in unstructured P2P networks. Hanhua Chen, Member, IEEE, Hai Jin, Senior Member, IEEE, Xucheng Luo, Yunhao Liu, Senior Member, IEEE, Tao Gu, Member, IEEE, Kaiji Chen, and Lionel M. Ni, Fellow, IEEE, Transaction on parallel and distributed systems, Vol.23, No.2, February 2012.

[5]. Search and Replication in Unstructured Peer-to-Peer Networks, Qin Lv, Pei Cao, Edith Cohen, Kai Li, and Scott Sheenier. 2002. Search and replication in unstructured peer-to-peer networks. In *Proceedings of the 16th international conference on Supercomputing* (ICS '02). ACM, New York, NY, USA, 84-95. DOI=10.1145/514191.514206. http://doi.acm.org/10.1145/514191.514206.

[6]. Search Methods in P2P Networks : German Sakaryan, Markus Wulff, Herwig Unger , 4th International Workshop, IICS 2004, Guadalajara,Mexico,June21-23, DOI:10.1007/11553762_6, Online ISBN : 978-3-540-33995-3.

[7]. Efficient Peer-to-Peer Keyword Searching, Patrick Reynolds, Amin Vahdat, Computer Science Department of Duke University. International Conference on Distributed Systems Platforms and Open Distributed Processing, 2003.