Effective Data Retrieval System with Bloom in a Unstructured p2p Network

Thumil Vannan .P.S, S. Uvaraj

Abstract- ***Bloomcast, an efficient and effective full-text retrieval scheme, in unstructured P2P networks. Bloomcast is effective because it guarantees perfect recall rate with high probability. It is efficient because the overall communication cost of full-text search is reduced below a formal bound. Furthermore casting bloom filters instead of raw data across the network. Bloomcast reduces the communication cost and storage cost for replication. Bloomcast replicates the items uniformly at random across the P2P networks, achieving a guaranteed recall at a communication cost of O(√N), where N is the size of the network. The main contribution of this proposal of Bloomcast design through both mathematical proof and comprehensive simulations.***

Keywords- P2P O(√N),

I. INTRODUCTION

A P2Pnetwork has also shown great potential to become a popular network tool for sharing information on the Internet. Existing P2P full-text search schemes can be divided into two types: DHT based global index and federated search engine over unstructured protocols. DHT-based searching engines are based on distributed indexes that partition a logically global inverted index in a physically distributed manner. Due to the exact match problem of DHTs, such schemes provide poor full-text search capacity. In federated search engines over unstructured P2Ps, queries are processed based on flooding. The best candidate for supporting full-text retrieval because the query evaluation operations can be handled at the nodes that store the relevant documents. recall is not guaranteed with acceptable communication cost using a flooding-based scheme.

II. OUR CONTRIBUTIONS

In this paper, to propose BloomCast, an efficient and effective full-text retrieval scheme, in unstructured P2P networks. BloomCast is effective because it guarantees the recall with high probability. It is efficient because the overall communication cost of full-text search is reduced below a formal bound. By replicating Bloom Filters instead of the raw documents across the network,
BloomCast significantly reduces the communication cost for replication.

III. SYSTEM APPROACH

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

Manuscript received April 2013.

Thumilvannan .P.S, M.E., A.P/CSE Dept., Arulmigu Meenakshi Amman College of Engineering, Thiruvannamalai Dt, Near Kanchipuram, India.

S.Uvaraj, M.E, Arulmigu Meenakshi Amman College of Engineering, Thiruvannamalai Dt, Near Kanchipuram, India.

1. Normal peers
2. Structured peers and
3. Bootstrap peers

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.

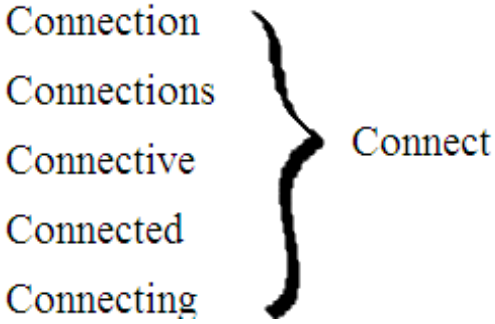
Stemming Algorithm

Fundamental concepts

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.



IV. BLOOMFILTER

The bloom filter utilizes the hashing technique for the search of best document. The bloom filter gets the Query from the node, it performs multiple hashing in the query and as a result it converts the query into URLs. A BF is a loss but succinct and efficient data structure to represent a set S, which can efficiently process the membership query such as "is element x in set S." By replicating the encoded term sets using BFs instead of raw documents among peers, the communication/storage costs are greatly reduced, while the full-text multi keyword searching are supported. Bloom Filter encoding can greatly reduce the communication cost for data replication.

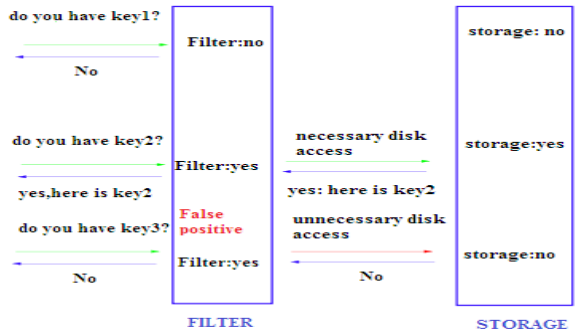


Fig. 1. Bloom Filter

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.

V. CAST BLOOM FILTER

Algorithm:

Require: $EstimatedNetworkSize = N$ is achieved

1: **for all** documents in local collection **do**
2: create an empty bit vector with m bits for document x, BF_x;
3: **for all** terms in a document **do**
4: insert term t into BF_x by setting the $h_j(t)$th bits of BF_x to 1, where $\{h_j(\cdot), 1 \le j \le k\}$ is the set of hash functions used by BF_x;
5: **end for**
6: **end for**
7: sample an optimal number of $r = c\sqrt{N \cdot \frac{S_q}{S_d}}$ random peers in the network by the lightweight DHT;
8: replicate BF_x together with url_x, the URL of document x, to the set of randomly sampled nodes;
9: **return**

VI. BLOOMCAST

In Unstructured P2P networks, BloomCast is an effective and efficient full text retrieval scheme. By leveraging a hybrid P2P protocol, Bloom Cast replicates the items uniformly at random across the P2P networks. BloomCast hybridizes a lightweight DHT with an unstructured P2P overlay to support random node sampling and network size estimation.

VII. QUERY EVALUATION OF BLOOMCAST

Algorithm:

1: $R \leftarrow \emptyset$;
2: **for all** BFs replicated in this peer **do**
3: $BooleanContainFlag \leftarrow True$;
4: **for all** terms in Q **do**
5: **if** $\exists(j)(1 \le j \le k) s.t. BF_x[h_j(t)] = 0$ **then**
6: $ContainFlag \leftarrow False$;
7: **end if**
8: **end for**
9: **if** $ContainFlag = True$ **then**
10: $R \leftarrow R \cup \{url_x\}$;
11: **end if**
12: **end for**
13: **return** R.

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, the nodes will check the for the data or Urls related to the requested data. Once search has been finished, the best results will display to the user.

VIII. DATAFLOW DIAGRAM

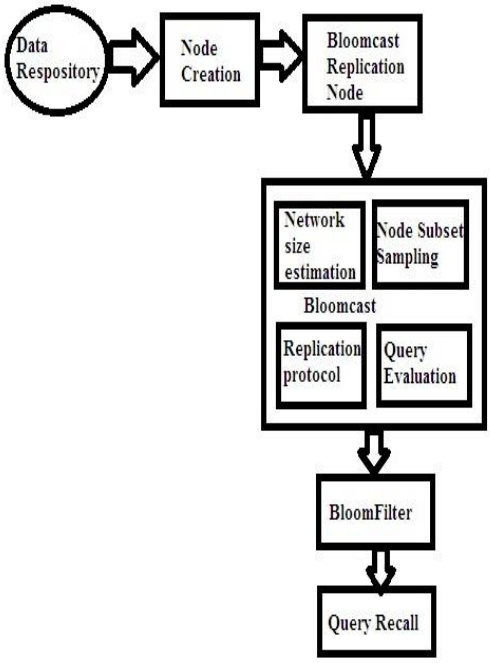


Fig. 2. Dataflow Diagram

IX. 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 and it is given to the user efficiently.

Retrieval of Data

After ranking the documents, the user can choose the required data that they wanted. By using the Bloom Filter Concept, an effective and efficient data retrieval process is achieved in the Unstructured P2P Networks.

X. CONCLUSION

A Bloomcast, an efficient and effective full-text retrieval scheme, in unstructured P2P networks. Bloomcast is effective because it guarantees the recall with high probability. It is efficient because the overall communication cost of full text search is reduced below a formal bound. Bloomcast reduces the communication cost for replication. All such limitations can be overcome by using this application. This provides enhanced security.

Aim:

The main sources of this project efficient and effective full text retrieval over unstructured p2p networks. Our goal is reduces the network size, communication cost and storage cost.

XI. FUTURE ENHANCEMENT

The future work should be focused towards increasing the efficiency to downloading the documentation with secret bloom filter value; each peer can store bloom filter values. This bloom values are unique identifiers for peers and any other device on a network. Therefore, if the bloom filter values are detected occurring at wrongly, user peer will be automatically barred from downloading the text file. When this method works together with the current algorithm it can provide an acceptable amount of security and privacy during data transmission.

XII. RESULTS

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

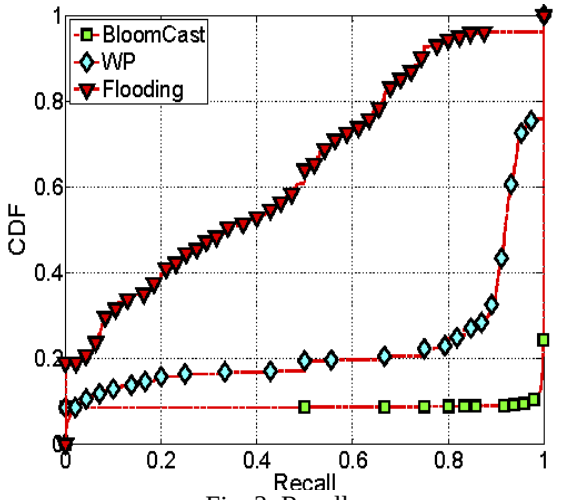


Fig. 3. Recall.

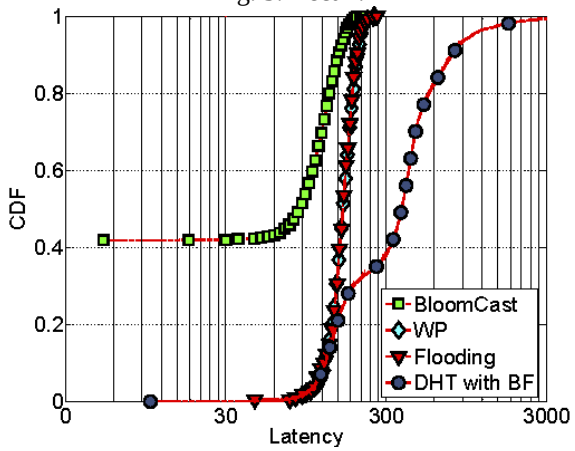


Fig. 4. Latency.

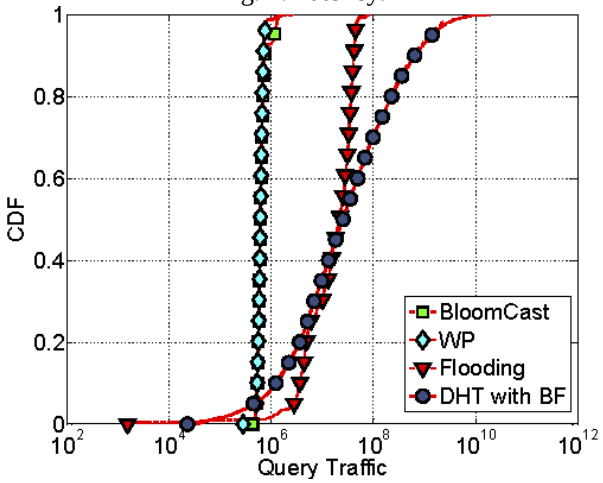


Fig. 5. Query traffic..

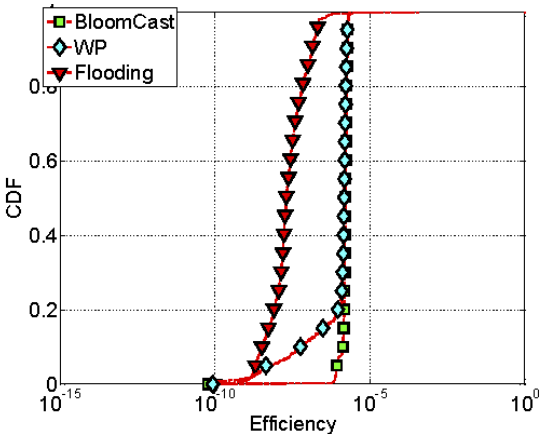


Fig. 6. Efficiency

The result in Fig. 5 shows that the average query traffic of BloomCast 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.

[8] 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.

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

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