ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Bloom Filters Driven Hybrid Query Propagation Schemes in Unstructured P2P Networks

L.Yuvana[1], Dr. Khalim Amjad Meerja[2]

[1] Student, K.L.University,Vaddeswaram,Guntur(dt),, Andhra Pradesh, India

[2] Professor, K.L.University,Vaddeswaram,Guntur(dt), Andhra Pradesh, India

ABSTRACT: Peer-to-Peer (P2P) file sharing applications, such as Napster and Gnutella supports millions of users to search and download desired data. Replication strategies are extensively utilized to improve search performance in unstructured P2Ps. To address the problems of the query popularity independent replication strategies, previously a novel strategy, called BloomCast, that implements Bloom Filters in WP(With Pointers) scheme to support efficient and effective full-text retrieval over unstructured p2p networks was developed. BloomCast hybridizes a lightweight DHT with an unstructured P2P overlay to support random node sampling and network size estimation. Although efficient these techniques are implemented irrespective of topologies and network size considerations. So we propose to use Flooding, Long random walk, General search scheme(suitable in the case of clustered topologies), Short random walk with local flooding(decreases the response time and is particularly suitable when combined with 1-step replication) schemes in accordance with varying p2p topologies and network sizes. Combined with Bloom filters these Hybrid Query Propagation schemes offers optimal performance over unstructured p2p networks and a practical implementation validates the claim.

Keyword--Peer-to-peer systems, Bloom Filter, replication.

1. INTRODUCTION

P2P networks can be classified as *unstructured*, *loosely structured*, and *highly structured* based on the control over data location and network topology Peer-to-Peer (P2P) file sharing applications, such as Napster and Gnutella supports millions of users to search and download desired data. Flooding is the predominant search technique in unstructured peer-to-peer (P2P) networks. 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.
- Federated Search Engines over unstructured P2Ps
  In federated search engines over unstructured P2Ps, queries are processed based on flooding. Unstructured P2Ps are commonly believed to be the best candidate for supporting full-text retrieval because the query evaluation operations can be handled at the nodes that store the relevant documents. However, search recall is not guaranteed with acceptable communication cost using a flooding-based scheme.

If we measure performance as the number of exchanged messages per distinct response, flooding with small time-to-live performs well in regular networks. Replication strategies are extensively utilized to improve search performance in unstructured P2Ps. The existing replication strategies can be divided into different categories.

- strategies towards Query popularity awareness
- WP(With Pointers) scheme

The items with high query rate are highly replicated for future query searching in query popular aware repletion strategies, thus popular items are improved for search performance. However, the strategy is inefficient for solving insoluble queries, the queries for rare items. Moreover, in practice, the query frequency is difficult or even impossible to obtain in

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

a distributed P2P system. The second type of replication strategy is independent of the popularity of the query, such as the WP scheme. By replicating data and query replicas randomly across a P2P network regardless of the query rate of the data, such kind of schemes improve search recall of queries no matter they are popular or not. In WP scheme, the term query replica is used to differentiate a query message transferred across the network without performing and a query that evaluated in a node. A query replica will be performed by the node holding it. The WP scheme utilizes random walk technique to deploy replicas. The problem of random walk based scheme is that it is not fault-tolerant. Another problem of the existing replication strategies is that simply replicating document reference or selected metadata cannot successfully support full text retrieval. To support full text retrieval, the existing replication strategies need to replicate the full document across the network, raising possibly unacceptable communication and storage costs.So a better system is required that can support full-Text Retrievals in Unstructured P2P Networks without the shortcomings of Popularity Aware and WP schemes. To address the problems of the query popularity independent replication strategies, we propose a novel strategy, called BloomCast, that implements Bloom Filters in WP scheme to support efficient and effective full-text retrieval over unstructured p2p networks. Different from the WP scheme, BloomCast hybridizes a lightweight DHT with an unstructured P2P overlay to support random node sampling and network size estimation. Furthermore, we propose an option of using Bloom Filter encoding instead of replicating the raw data. Using such an option, BloomCast replicates Bloom Filters (BF) of a document. The Bloom Filters is a lossy 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. WP scheme utilizes random walk techniques which is an overkill of resources of the query initiating peer, because these techniques are implemented irrespective of topologies and network size.

2. RELATED WORK

Basically, every new technique is inspired from either flooding or random walk. Mainly they are distinguishing as techniques based on selection and techniques based on underlying topology changes Most searching schemes in unstructured P2P networks are forwarding-based and are different variations of flooding. Flooding is the predominant search technique in unstructured peer-to-peer (P2P) networks. If we measure performance as the number of exchanged messages per distinct response, flooding with small time-to-live performs well in regular networks. Forwarding schemes in unstructured P2Ps can also be classified as blind search or informed search. Blind searches, such as *k*-walker random walk and modified random BFS, nodes do not keep any information about the data location. Informed searches, for example, directed BFS and SQR, nodes store some hints that facilitate the search. The directed BFS utilizes simple hints while SQR takes advantage of complicated hints. Hence we use the random walk technique for the search. Simulating a random walk has been proposed as an alternative search technique. The performance of the random walk simulation method appears to be better than the performance of flooding in the Regular topologies. In addition, the random walk simulation method scales well and has excellent granularity. We consider hybrid schemes which can be viewed as a random walk of substantially shorter length combined with very shallow floodings on every step of the random walk (fig . 1d)

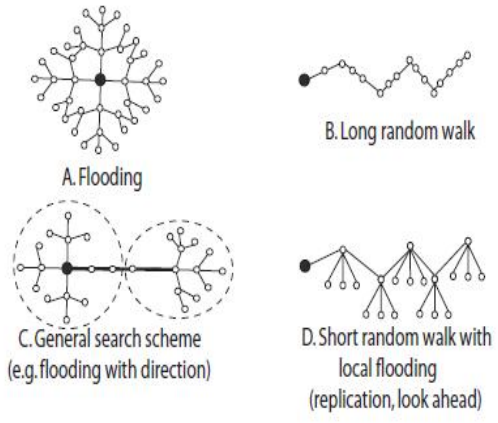

Fig. 1. Figure 1a represents search by flooding. Flooding has good performance for small values of time-to-live. Figure 1b represents search by random

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

walk. The response time is proportional to the length of the walk. Figure 1c represents a general search scheme, which is flooding amplified towards a critical direction. This is suitable in the case of clustered topologies, where the critical direction leads flooding outside a cluster. Figure 1d represents a shorter random walk with local floodings. This decreases the response time and is particularly suitable when combined with 1-step replication.

The simulation of a short random walk with shallow floodings on every step performs particularly well. The idea is the following. Naturally, we expect that the time to discover a certain number of nodes using a random walk with shallow floodings will be somewhat smaller than in the simulation of a random walk without local floodings. We study normalized flooding, where a vertex of small degree forwards a query to all his neighbors, while a vertex of large degree forwards a query to a small subset of its neighbors chosen uniformly at random. On the other hand, we noted that flooding has poor performance when there is discrepancy in the degrees of the underlying network. The performance of flooding in the case of a sparse network with a few vertices of large degrees. We study normalized flooding, where a vertex of small degree forwards a query to all his neighbors. Normalized flooding with 1-step replication achieves performance comparable to random walk with 1-step replication, further indicating that the gaining in 1-step replication comes from the bias of large degrees, and further strengthening the suggestion to use a small number of super nodes. We use the graph models some crucial structural properties that will be later used in the proofs.

2.1 RANDOM GRAPH MODELS

We introduce random regular graphs, which aim to capture the behavior of a typical regular topology, and random graphs with supernodes. We review the graph theoretic notion of expansion and relate to the performance of flooding. This is a standard model in the theory of random graphs as well as networking. Particularly for $d_1 \geq d_2 \geq \ldots \geq d_n$ denoting the degrees of a graph on n nodes. First consider $D = \sum_{i=1}^{n} d_i$ mini-vertices corresponding to nodes in the natural way: the first d1 minivertices correspond to node 1, the next d2 mini-vertices correspond to node 2, and so on. Let d be a constant. By *random regular graph*, denoted Gn,d. A random graph in the configurational model, with $d_i = d,\ 1 \leq i \leq n.$ Futhur we introduce the Random graph model for graphs with the super nodes. Consider the α and β and $\alpha n^{\frac{1}{2}}$ and $\beta n^{\frac{1}{2}}$, called *large vertices*, and all the remaining nodes of degree d, called *small vertices*. Random regular graphs and random graphs with supernodes have sum of degrees D=dn and $D \simeq (\alpha\beta + d)n$ respectively, hence they are sparse, in the sense that the sum of the degrees of their vertices is $\Theta(n)$. Throughout this paper, ' means 1±o(1). We need the following definitions. Let G(V,E) be an undirected graph, with |V |=n.

Let S be a subset of vertices, S $\subset$ V , and let $\bar{S}$ be its complement, $\bar{S}$ = V \ S. The structural facts for graphs with supernodes are Lemmas are used and a detailed below.

Lemma 2.1: Let G = G n,d,α,β be a random graph with supernodes, and let € be any constant € < max{ α,β }. Then, with all but exponentially vanishing probability, every large vertex of G has $\frac{(\alpha-\epsilon)\beta}{d+\alpha\beta}\frac{\epsilon n^{\frac{1}{2}}}{2}$ distinct large neighbors.

Lemma 2.2: Let G = G n,d,α,β be a random graph with supernodes. Then, with all but exponentially vanishing probability, every large vertex of G has $\frac{d}{d+\alpha\beta}\frac{\beta n^{\frac{1}{2}}}{2}$ edges incident to (not necessarily distinct) small neighbors.

Lemma 2.3: Let G = G n,d,α,β be a random graph with supernodes. Then, with all but exponentially vanishing probability, every large vertex v has $\Gamma(\{v\}) \cup \Gamma(\Gamma(\{v\})) \geq \frac{(\alpha-\epsilon)\epsilon\beta^2}{4(d+\alpha\beta)^2}n.$

2.2FLOODING AND NORMALIZATION

Flooding is the predominant search technique in unstructured peer-to-peer networks. In particular, a node initiates a search by propagating a request, together with a time-to-live ┐ , to all his neighbors Flooding proceeds as follows. The first time that a node receives a request with time-to-live t, the node responds to the request and, if t > 0, the node propagates the same request to all his neighbors. If a node receives the same request multiple times,

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

then it will neither respond nor propagate it. We quantify the performance of flooding by the *number of responses*, the *response time* (we assume that the delay of a particular response is proportional to the number of hops between the initiator of the query and the responding node), and by the *number of propagated messages*. When a graph is not regular, then the performance of flooding deteriorates. Tthese cause a sudden sharp increase in the number of neighbors they introduce (hence poor granularity), which, in turn, causes a lot of shared edges. Hence therefore consider *normalized flooding* which is the following algorithm. In normalized flooding, when a node of degree dmin receives a query, the node propagates the query to all his neighbors (except the one which forwarded the query). In the case of higher degree has receive the query, Then the node propagates the query only to the dmin to all of the neighbour nodes.

Hence as we introduce the good behavior of flooding in regular graphs. The proof of this theorem is directly based on known structural properties of random regular graphs. Our intension is to translate these properties in the context of the flooding. As we want the maintain the upper bounds in the theorem to differentiate the time-to-live, and suggest to guarantee the flooding. To generate the distinct responses we maintain the lower bounds for flooding in the graphs with supernodes to indicate that, without normalization. A large vertex is discovered for a very small value of time-to-live, hence even poorer granularity. Flooding with effective time-to-live for distinctive responses the no.of minigroups can be seen and expectable small vertices for to attain the flooding with propability. It Indicates the normalized flooding in graphs with supernodes can rectify the performance of flooding. It brings the performance of normalized flooding, up to order of magnitude, to the performance of flooding in regular graphs.

2.3 RANDOM WALKS AND REPLICATION

Everything else being equal, the best way to search a graph would be by uniform sampling. Assuming that a random node of the network could be generated efficiently, we could take k such samples simultaneously at cost one message per sample. By the well known coupon collection theorem. For any 1< K < n, the expected number of the samples to visit all the nodes of nlogn, for any constant ϵ < 1, then expected no.of samples visit to ϵn. The amount of network overhead per distinct response can comearbitrarily close to 1. In additional we can retrieve drawn samples simultaneously. The *random walk* method has been proposed as a practical alternative to implement uniform sampling. In several random graph models mixing time of the random walk, which is the number of simulation steps in order for the random walk to reach a distribution close to uniform, is O(log n). Which means that we may simulate the *k* uniform sampleswith O(logn) random walk steps for uniform sampling. the random walks can be simulated in parallel, and assuming that the response delay of a random walk is proportional to the number of simulation steps of the walk. Hence we may get maximum response time O(logn), Overhead atmost of O(klogn). The drawback of this approach is the network overhead which scales as O(log n). On the positive side, the theory of cover times , complexity theory, and extensive experimentation suggest that this overhead can be reduced to a constant by taking O(log n) steps to randomize and then using k *successive steps* of the random walk

3.EXISTING SYSTEM

In the Existing system, To address the problems of the query popularity independent replication strategies, we propose a novel strategy, called BloomCast, that implements Bloom Filters in WP scheme to support efficient and effective full-text retrieval over unstructured p2p networks. Different from the WP scheme, BloomCast hybridizes a lightweight DHT with an unstructured P2P overlay to support random node sampling and network size estimation.

Furthermore, we propose an option of using Bloom Filter encoding instead of replicating the raw data. Using such an option, BloomCast replicates Bloom Filters (BF) of a document. A BF is a lossy 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. WP scheme utilizes random walk techniques which is an overkill

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

of resources of the query initiating peer, because these techniques are implemented irrespective of topologies and network size

4. Proposed System

So we propose to use Flooding, Long random walk, General search scheme(suitable in the case of clustered topologies), Short random walk with local flooding(decreases the response time and is particularly suitable when combined with 1-step replication) schemes in accordance with varying p2p topologies and network sizes. we are interested in characterizing the performance of searching. We choose some distinct random nodes and perform searching starting from these nodes with the algorithms. We measure the number of distinct peers visited per searching per node, which we call *hits*. The hits directly related to the standard definition discovered the no.of specific object copies, assuming that the copies of the requested information are placed at random in the network. In addition, we measure the response time of the searching, i.e. the maximum time it takes for the query to complete.We use the following Specific metrics

- *Median and Mean number of distinct peers discovered (hits).*

  It should maximize the median and mode no.of distinct peers and median is the most more robust metrics degrees, it is possible to measure relatively large mean values because few searches may reach a very large number of users and increase the mean value.
- *Minimum, Maximum, and Standard Deviation of the number of hits.*

  A large minimum value is important in order to guarantee that the algorithm will have a good worst case performance. The variation of the maximum value and minimum value is measured using the standard deviation.
- *Number of messages*

  In order to perform a fair comparison of the different searching algorithms we require that they use the same number of messages. we require that the expected number of messages used in each experiment is approximately the same for all algorithms.
- *Granularity of number of messages.*

  This is a qualitatively and not quantitative metric, it
  is difficult to control the parameters of the algorithm, usually the time-to-live, to use a pre-specified number of messages. Algorithms with finer granularity are preferable for searching.
- *Response time.*

  We also measure the maximum running time of each algorithm. Each node receives queries from its neighbors and at the same time processes them and forwards copies of the queries, if necessary, to its neighbors.

We are interested in studying the performance of the searching algorithms in networks with irregularities in the node degrees and in networks. Both cases are typical in complex and unstructured communication networks. Typically, these users have a much larger number of neighbors. A common pattern that appears in every unstructured communication network is the clusterness of the topology. We will use the following synthetic topologies to compare the searching algorithms.

Random d-regular Graphs. Extensive analytical work has shown that random d-regular graphs have good properties, including low diameter, good connectivity. We will use d-regular random graphs as a canonical model of a well connected network. The topologies of third generation peer-to-peer networks, like BitTorrent, how the topology is formed.

Power Law Graphs. Many seemingly complex networks, including the Internet, the Web Graph, and many others, have been shown to be characterized by powerlaws. The power-law is usually characterized by a parameter λ called the powerlaw exponent. For λ = 3.0 the largest degree in 1M nodes graph is less than 100, which brings the parameters close to real peer-to-peer networks.

Clustered topologies. We assume that there are clusters of users with very good connectivity inside each cluster. In particular, we assume that the network is composed of a small number of clusters, and each cluster is a 3-regular random graph.

Bimodal topologies. We assume that there are two types of nodes in the network. Few nodes are connected to a large number of other nodes and Such nodes are typically called ultra-peers. The rest of the users have few neighbors.

The performance of normalized flooding to standard flooding, Both schemes perform similarly in regular topologies. When the topology contains nodes with high degrees, which is common in large unstructured communication networks. With normalization it is easier to control how many nodes will be reached by the flooding. We give the mean number of unique peers discovered as a function of the initial time-to-live for topologies. Where as in the other topologies however, which contain nodes of high degree, the increase in the number of peers is very fast after the search reaches a high degree node. Increase in the TTL by 1 or 2 will result in discovering a large part of the network. Hence it shows the tremendous increase comes at the cost of

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

reduced efficiency in the search process. A large number of messages reach already discovered nodes. Normalized flooding behaves better than standard flooding with respect to other metrics. Consider the example the standard deviation is much smaller with normalized Flooding. Moreover, the number of peers discovered as a function of the initial time-to-live follows a more predictable behavior. In Figure 2 we plot the number of nodes visited by the flooding as a function of the initial time-to-live.

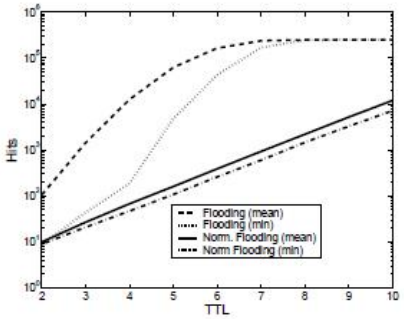

Fig. 2. Number of unique peers discovered as a function of the initial time-tolive. In the case of normalized flooding the number of unique peers increases exponentially with the TTL, and, moreover, the increase is predictable and consistent for all nodes. In the case of regular flooding, the increase is much faster and depends on the node that initiates the flooding.

The straight line in the case of normalized flooding indicates that the horizon of the search increases exponentially. Observe that in order to discover the same number of nodes with normalized flooding as with standard flooding. Since the increase of the horizon is more predictable and can be roughly computed by either knowing the properties of the topology.

Replication of one step is a practical way to improve the performance of searching by allowing each node to answer queries on behalf of its neighbors. The advantage comes at the cost of replicating information about the content of the neighbors. This cost is paid once when a new neighbor arrives and is amortized over a large number of messages that go through the node. The advantage of 1-step replication becomes clear in graphs with large degrees, like the power-law graphs and the bimodal graph. The main reason is that both methods, flooding and random walks, quickly discover the nodes of high degree and through them discover a large portion of the nodes in the network. In the cases of graphs with large degrees the performance of normalized flooding is better than random walk (in the worst case by approximately 20%)

An extension of the previous scheme is to perform a short random walk with local floodings, We call this random walk with lookahead. we do not think that it is realistic to maintain replicas of your neighbors' neighbors, therefore we charge the algorithm for all the messages generated by both the random walk and the local flooding. The main observation is that the performance of the random walk with lookahead is similar in terms of unique peers discovered to performing a long random walk without lookahead.

5.Referrence

[1] Christos Gkantsidis, Milena Mihail, and Amin Saberi, "Random walks in peer-to-peer networks," in *IEEE Infocom*, Hong Kong, 2004.

[2] Jordan Ritter, "Why gnutella can't scale. no, really.," http://www.darkridge.com/_jpr5/doc/gnutella.html, 2001.

[3] Qin Lv, Pei Cao, Edith Cohen, Kai Li, and Scott Shenker, "Searchand replication in unstructured peer-to-peer networks," in *International Conference on Supercomputing*, New York, New York, USA, 2002,pp. 84–95, ACM Press, Extended version in http://www.cs.princeton.edu/_qlv/download/searchp2p full.pdf .

[4] VicentCholvi, AntonioFern´ andez,LuisL´ opez,LuisRodero-Merino," Using Random Walks to Find Resources in Unstructured Self-Organized P2P Networks", Dasson 2007.

[5] William Acosta Surendar Chandra, "Unstructured PeertoPeer Networks Next Generation of Performance and Reliability", IEEE Infocom 2005.

[6]Milena Mihail, Amin Saberi, and Prasad Tetali, "Random walks with lookahead in power law random graphs," Available at http://www.cc.gatech.edu/fac/Milena.Mihail/ lookahead.pdf, 2004.

[7] Colin Cooper and Alan Frieze, "The cover time of random regular graphs," Available at http://www.math.cmu.edu/_af1p/ Cover.ps, 2004.

[8] Colin Cooper and Alan Frieze, "The cover time of sparse random graphs," in *Symposium on Discrete Algorithms (SODA)*, Baltimore, Maryland, 2003, pp. 140 – 147, SIAM/ACM.

[9] R. Impagliazzo and D. Zuckerman, "How to recycle random bits," in *30th IEEE Symposium on the*

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

Foundations of Computer Science, Research Triangle Park, NC, 1989, pp. 248–253, IEEE.

[10] D. Gillman, "A chernoff bound for random walks on expander graphs," *Journal on Computing*, vol. 27, no. 4, pp. 1203–1220, 1998.

[11] Christos Gkantsidis, Milena Mihail, and Ellen Zegura, "Spectral analysis of internet topologies," in *IEEE Infocom*, San Francisco, CA, US, 2003.

[12]Gurmeet Singh Manku, Moni Naor, and Udi Wieder, "Know thy neighbor's neighbor: the power of lookahead in randomized p2p networks," in *ACM Symposium on Theory of Computing (STOC)*, Chicago, IL, USA, 2004, pp. 54 – 63, ACM.