Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Ranking of Data Using BloomCast and Stemming Algorithm in Unstructured P2P Networks

Miss.P.Aruna
Hindustan University, Chennai, India

Mrs.P.Ranjana
Hindustan University,Chennai, India

Abstract---Efficient and effective full-text retrieval and ranking process in unstructured Peer-to-Peer networks remains a challenge in the research community because it is difficult, if not impossible, for unstructured P2P systems to effectively locate items with guaranteed recall and existing schemes to improve search success rate often rely on replicating a large number of item replicas across the wide area network, incurring a large amount of communication and storage costs. Due to the exact match problem of DHTs and federated search problem, such schemes provide poor full-text search capacity. It proposes replication of BloomFilters for efficient and effective data retrieval and ranking of data in unstructured P2P networks. Ranking that provides the needs of the users vary, so that what may be interesting for one may be completely irrelevant for another. To retrieve the best results using ranking process based on the frequency of keywords present in the document instead of number of user clicks. The list of document identifiers of the document with high popularity of keywords in the query, highly ranked list of document displayed in the top of the result followed by the rest of the results. By replicating the encoded term sets using BFs and stemming of words instead of raw documents among peers, the communication and storage costs are greatly reduced, while the full-text multikeyword searching is supported and best ranking will be performed.

Keywords--- BloomCast, Peer-to-Peer systems, Bloom Filter, replication, stemming.

I. INTRODUCTION

With the emergence of (P2P) file sharing applications, such as Napster and Gnutella, millions of users have used P2P systems to search desired data. A P2P network has also shown great potential to become a popular network tool for sharing information on the internet [1], [8]. 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. 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. Replication strategies [1] are extensively utilized to improve search performance in unstructured P2Ps. The existing replication strategies can be divided into two categories. The first type is the query popularity aware strategies [3].Such strategies assume that the access frequencies of the items are known and the number of replicas is determined by the query's popularity. Cohen and Shenker [3] claimed th at the square-root replication strategy, where the number of the replicas is proportional to the square-root of the query popularity/rate, has the optimal search performance. In query popularity aware replication strategies, the items with high query rate are highly replicated for future query searching, thus the search performances for popular items are improved. However, the strategy is inefficient for solving insoluble queries, the queries for rare items [3]. Moreover, in practice, the query frequency is difficult or even impossible to obtain in a distributed P2P system.

A. *Problem Statement:* Existing P2P full-text search schemes can be divided into two types: DHT-based global index and federated search engine over unstructured protocols. Due to the exact match problem of DHT, such schemes provide poor full-text search capacity. In federated search engines over unstructured P2Ps, search recall is not guaranteed with acceptable communication cost using a flooding-based scheme. 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 [9], and any participating DHTs form an infrastructure that can be used to build more complex services, such as any cast, cooperative Web caching, distributed file systems, domain name services, instant messaging, multicast, and also Peer-to-Peer file sharing and content distribution systems. Node can efficiently retrieve the value associated with a given key. Federated search is an information retrieval technology that allows the simultaneous search of multiple searchable resources. A user makes a single query request which is distributed to the search engines participating in the federation. The federated search then aggregates the results that are received from the search engines for presentation to the user.

B. *Objective of the Work:* To improve search success rate often rely on replicating a large number of item replicas across the wide area network, incurring a large amount of communication and storage costs. In this paper, I propose

BloomCast, an efficient and effective full-text retrieval scheme, in unstructured P2P networks. By leveraging a hybrid P2P protocol, BloomCast replicates the items uniformly at random across the P2P networks, achieving a guaranteed recall at a communication cost. Furthermore, by casting Bloom Filters instead of the raw documents across the network, BloomCast significantly reduces the communication and storage costs for replication [6]. Ranking that provides the needs of the users vary, so that what may be interesting for one may be completely irrelevant for another. The role of ranking process is thus crucial: select the pages that are most likely be able to satisfy the user's needs, and bring them in the top positions. Ranking of data is performed based on the term frequency and keywords.

II. RELATED WORK

The emergence of Peer-to-Peer (P2P) file sharing applications, such as Napster and Gnutella, millions of users has used P2P systems to search desired data. Existing P2P full-text search schemes can be divided into two types: DHT-based global index and federated search engine over unstructured protocols. Due to the exact match problem of DHTs, such schemes provide poor full-text search capacity [9]. In federated search engines over unstructured P2Ps, search recall is not guaranteed with acceptable communication cost using a flooding-based scheme. Replication strategies are extensively utilized to improve search performance in unstructured P2Ps. The existing replication strategies can be divided into two categories. The first type is the query popularity aware strategies. In query popularity aware replication strategies, 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 a distributed P2P system. The second type of replication strategy is independent of the popularity of the query, such as the WP scheme. The WP scheme utilizes random walk technique to deploy replicas. The problem of random walk based scheme is that it is not fault-tolerant.

A. *Demerits:* Due to the exact match problem of DHTs, such schemes provide poor full-text search capacity [9]. In federated search engines over unstructured P2Ps, search recall is not guaranteed with acceptable communication cost using a flooding-based scheme. 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 a distributed P2P system. Not effective retrieval of information and more communication and storage cost.

III. MODEL

Efficient and effective full-text retrieval in unstructured Peer-to-Peer networks remains a challenge in the research community. First, it is difficult, if not impossible, for unstructured P2P systems to effectively locate items with guaranteed recall. Second, existing schemes to improve search success rate often rely on replicating a large number of item replicas across the wide area network, incurring a large amount of communication and storage costs [1]. To overcome these issues we propose a novel strategy, called BloomCast, to support efficient and effective full-text retrieval in this paper. BloomCast hybridizes a lightweight DHT with an unstructured P2P overlay to support random node sampling and network size estimation. Furthermore, I propose an option of using Bloom Filter encoding instead of replicating the raw data. Using such an option, BloomCast replicates BF of a document. A BF is a loss but succinct and efficient data structure to represent the data. By replicating the encoded term sets using BFs instead of raw documents among peers, the communication and storage costs are greatly reduced, while the full-text multikeyword searching are supported [6]. Ranking of data is performed based on the term frequency and keywords.

A. *Merits:* BloomCast, an efficient and effective full-text retrieval scheme, in unstructured P2P networks, by leveraging a hybrid P2P protocol, BloomCast replicates the items uniformly at random across the P2P networks, achieving a guaranteed recall at a communication cost. Furthermore, by casting Bloom Filters instead of the raw documents across the network, BloomCast significantly reduces the communication and storage costs for replication. Ranking that provides the needs of the users vary, so that what may be interesting for one may be completely irrelevant for another. The role of ranking process is thus crucial: select the pages that are most likely be able to satisfy the user's needs, and bring them in the top positions.

IV. CONSTRUCTION

A. *File Uploading to Peer in Network and Stemming:* Network has many numbers of node and their details. It maintains the connection details also. Nodes are interconnected and exchange data directly with each other nodes. Nodes are connecting with other nodes in the network. Network server maintains the node IP address, port details and status. The popularity of P2P multimedia file sharing applications such as Gnutella and Napster has created a flurry of recent research activity into P2P architectures. A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form. It is important to appreciate that I use stemming with the intention of improving the performance of IR systems. It is not an exercise in etymology or grammar. In fact from an etymological or grammatical viewpoint, a stemming algorithm is liable to make many mistakes.

B. *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 [1].

C. BloomFilter, Ranking and Retrieval of Data

The bloom filter utilizes the hashing technique for the search of best document [9]. 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 [9]. 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. 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.

V. STEMMING ALGORITHM

A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form. It is important to appreciate that we use stemming with the intention of improving the performance of IR systems. It is not an exercise in etymology or grammar. In fact from an etymological or grammatical viewpoint, a stemming algorithm is liable to make many mistakes. In addition, stemming algorithms - at least the ones presented here - are applicable to the written, not the spoken, form of the language. For some of the world's languages, Chinese for example, the concept of stemming is not applicable, but it is certainly meaningful for the many languages of the Indo-European group. In these languages words tend to be constant at the front, and to vary at the end. The variable part is the `ending', or `suffix'. Taking these endings off is called `suffix stripping' or `stemming', and the residual part is called the stem. It is important to realise that the stemming process cannot be made perfect. For example, in French, the simple verb endings -ons and -ent of the present tense occur repeatedly in other contexts. -ons is the plural form of all nouns ending -on, and -ent is also a common noun ending. On balance it is best not to remove these endings. In practice this affects -entverb endings more than -ons verb endings, since -ent endings are commoner. The result is that verbs stem not to a single form, but to a much smaller number of forms (three), among which the form given by the true stem of the verb is by far the commonest. It has been traditional in setting up IR systems to discard the very commonest words of a language - the stop words - during indexing. A more modern approach is to index everything, which greatly assists searching for phrases for example, stop words can then still be eliminated from the query as an optional style of retrieval. In either case, a list of stop words for a language is useful.

VI. ARCHITECTURE AND IMPLEMENTATION

The novel strategy called BloomCast is the replication of BloomFilter, to support efficient and effective full-text retrieval in this project. Furthermore, there is an option of using Bloom Filter encoding instead of replicating the raw data. Using such an option, BloomCast replicates BF of a document. A BF is a loss but succinct and efficient data structure to represent the data. By replicating the encoded term sets using BFs instead of raw documents among peers, the communication and storage costs are greatly reduced, while the full-text multi keyword searching are supported.

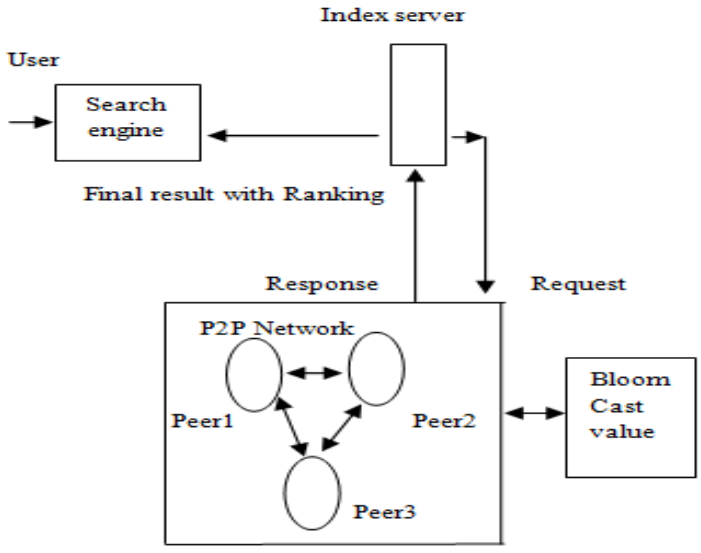

Fig.1. System Architecture diagram

It demonstrates the power of BloomCast design through both mathematical proof and comprehensive simulations and ranking of data is performed. The following architecture diagram describes the concept of generating BloomCast value and the performance of text retrieval. BloomFilter hybridizes a lightweight DHT [9] with an unstructured P2P overlay to support random node sampling and network size estimation. Furthermore, I propose an option of using Bloom Filter encoding instead of replicating the raw data. Using such an option it replicates BF of a document. A BF is a loss but succinct and efficient data structure to represent the data. By replicating the encoded term sets using BFs instead of raw documents among peers, the communication and storage costs are greatly reduced, while the full-text multi keyword

searching are supported [6]. The popularity of P2P multimedia file sharing applications such as Gnutella and Napster has created a flurry of recent research activity into P2P architectures. A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form. It is important to appreciate that i use stemming with the intention of improving the performance of IR systems. It is not an exercise in etymology or grammar. In fact from an etymological or grammatical viewpoint, a stemming algorithm is liable to make many mistakes. 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. 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 [2]. 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. 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.

VII. RESULTS

This paper uses several standard metrics for evaluating the performance of BloomCast. The evaluation considers both search quality and system efficiency. Quality focuses on user-perceived qualities such as recall, precision, F Measure, and latency; while efficiency focuses on resource utilization such as traffic and efficiency. Fig. 2 shows that the BloomCast scheme greatly improves the search efficiency by calculating the number of keywords and term frequency. The average search efficiency of BloomCast outperforms that of the flooding strategy by 67 percent, as well as outperforming that of the WP scheme by 24 percent. I vary the size of the network by using different Gnutella traces and evaluate the performance of BloomCast in different network scales and ranking process improves the search capacity.

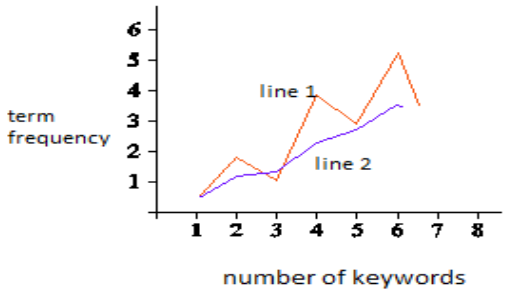

Fig.2.Rranking for efficient searching

Line1 – ranking by number of user clicks
Line2 – ranking by frequency of terms

VIII. CONCLUSIONS

This paper proposes BloomCast, an efficient and effective full-text retrieval scheme and ranking, 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. Furthermore, by replicating Bloom Filters instead of the raw documents across the network, BloomCast significantly reduces the communication cost for replication. It demonstrate the power of BloomCast design through both mathematical proof and comprehensive simulations based on the TREC WT10G data collection and query logs from a real world search engine. Results show that BloomCast outperforms existing schemes in terms of both search results quality and system efficiency and ranking of data is performed based on the term frequency and keywords.

IX. FUTURE WORK

To address the threats to same for more documents ranking, future work should evaluate best ranking using machine learning and apply the method at the start of the network. The mathematical proof and comprehensive simulations based on the query logs the BloomCast design is demonstrated. BloomCast replicates Bloom Filters of a document. A BF is a loss but succinct and efficient data structure to represent the data from a real world search engine. Results show that ranking outperforms existing schemes in terms of both search results quality and system efficiency and quality data can be retrieved. The future work should further increase the performance of ranking process and machine learning can be performed.

REFERENCES

[1] E. Cohen and S. Shenker (Feb 2002) *"Replication Strategies in Unstructured P2P Networks,"* Proc. ACM SIGCOMM '02. pp. 177-190.

[2] D. Li, J. Cao, X. Lu, and K. Chen, (Jan 2008) *"Efficient Range Query processing in P2P Systems,"* IEEE Trans. Knowledge and Data Eng., vol. 21, no. 1, pp. 78-91.

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

[4] N.F. Huang, R. Liu, C.H. Chen, Y.T. Chen, and L.W. Huang, (Dec 2005) *"A Fast Url Lookup Engine for Content-Aware Multi-Gigabit Switches,"* Proc. 19th Int'l Conf. Advanced Information Networking and Applications (AINA).

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

[6] P. Reynolds and A. Vahdat, (Aug 2003) *"Efficient P2P Keyword Searching"* Proc. ACM/IFIP/USENIX 2003 Int'l Conf. Middleware (Middleware '03), pp. 21-40.

[7] H. Shen, Y. Shu, and B. Yu, (July 2004) *"Efficient Semantic-Based Content Search in P2P Network,"* IEEE Trans. Knowledge and Data Eng., vol16 no7 813-826.

[8] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H. Balakrishnan, (Jan 2001) *"Chord: A Scalable P2P Lookup Service for Internet Applications,"* Proc. ACM SIGCOMM '01, pp. 149- 160.

[9] H. Song, S. Dharmapurikar, J. Turner, and J. Lockwood, (May 2005)*"Fast Hash Table Lookup Using Extended Bloom Filter,"* Proc. ACM SIGCOMM.

[10] X. Tang, J. Xu, and W. Lee, (Dec 2008) *"Analysis of TTL-Based Consistency in Unstructured P2P Networks,"* IEEE Trans. Parallel and Distributed Systems, vol. 19, no. 12, pp. 1683-1694.

