International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:6, 2008

Exploiting Query Feedback for Efficient Query Routing in Unstructured Peer-to-peer Networks

Iskandar Ishak, and Naomie Salim

***Abstract*—**Unstructured peer-to-peer networks are popular due to its robustness and scalability. Query schemes that are being used in unstructured peer-to-peer such as the flooding and interest-based shortcuts suffer various problems such as using large communication overhead long delay response. The use of routing indices has been a popular approach for peer-to-peer query routing. It helps the query routing processes to learn the routing based on the feedbacks collected. In an unstructured network where there is no global information available, efficient and low cost routing approach is needed for routing efficiency.

In this paper, we propose a novel mechanism for query-feedback oriented routing indices to achieve routing efficiency in unstructured network at a minimal cost. The approach also applied information retrieval technique to make sure the content of the query is understandable and will make the routing process not just based to the query hits but also related to the query content. Experiments have shown that the proposed mechanism performs more efficient than flood-based routing.

***Keywords*—**Unstructured peer-to-peer, Searching, Retrieval, Internet.

I. Introduction

PEER-to-peer networking has faced rapid development and becoming one of the most popular Internet applications during these recent years. It has gained a tremendous popularity especially on the use of sharing resources between peers in the internet. Peer to peer application in its earlier years was made popular by file sharing applications such as Napster [1] and Gnutella [2]. Through this application, users can share files with other peers that is connected to the network. Napster allows users to share mp3 music files, while Gnutella enable users to share any digital files (e.g. music files, documents and images).

Peer-to-peer has taken advantages on advancement of current computing and storage capacity of ordinary PCs which are now getting more powerful. The advancements of the high-speed and wireless networking added the ability of these ordinary PCs to become more adept to be used for peer-to-peer application. As the peer-to-peer becomes more popular, efficient routing is needed for the users to have better retrieval when querying data items they want.

There are two types of routing in peer to peer network: structured and unstructured [3]. Unstructured peer-to-peer mostly employ flooding approach towards all its neighbors while random walk forwards a peers' only to randomly selected neighbors. Routing in unstructured networks did not impose any structure in the network.

Structured peer-to-peer is developed to improve the performance of the flooding and random selection approaches. Structured peer-to-peer network uses the distributed hash table (DHT) for routing. Structured peer-to-peer systems as CAN [4] and CHORD [5] use the DHTs to provide data location management in a structured way. Whenever peers join or leave the network, peers will be updated to maintain desirable properties for quick lookup. In DHT, each peer has its own hash table and stores keys that are mapped to them. DHT implements an operation, the lookup (key), which routes the request to the peer responsible for storing the key. However, DHT based structured network suffers in terms of larger overhead than unstructured peer-to-peer and cannot support partial-match queries efficiently [6].

In this paper, we present a query-feedback based peer-to-peer routing for unstructured peer-to-peer network. It learns from the query feedback and its content for improving the query routing in unstructured peer-to-peer network at a very minimal cost.

II. Related Work

The earliest technique for peer-to-peer routing is based on the Naïve Breadth-First Search (BFS) algorithm or Flooding Mechanism. This technique is used in file-sharing peer-to-peer application Gnutella [2]. In this approach, each query from a peer will be broadcasted to all the peers in the network but restricted by the TTL (Time to Live) value. Lookup for this approach may generate O(N) message where N is the number of node. As a result, the query consumes a great deal of processing resources and excessive network. In a low bandwidth network, this technique could make the network become a bottleneck. It is a robust and simple technique for query routing but it involves a great deal of communication overhead, that is, high in number of messages. Hop number or hop count is also increased exponentially. Some of the messages might visit the same node that has been searched previously. Therefore, communication overhead and scalability are the main problems in this approach.

Manuscript received June 30, 2008.

Iskandar Ishak is with the Universiti Putra Malaysia, 43400, UPM, Selangor, Darul Ehsan Malaysia (phone: +603-89466585; fax: +603-8946-6577; e-mail: iskandarishak@ gmail.com).

Naomie Salim is with the Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Darul Takzim, Malaysia (e-mail: naomie@utm.my).

In the random BFS approach [7], each peer forwards a search message to only a fraction of its peers. Each node randomly selects a subset of peers connected to it and then propagates the search message to those peers. The advantage of this technique is that it does not require any global knowledge. Every node is able to make local decision in a quick manner since it needs only small portion of connected peers to route the query.

Another unstructured peer-to-peer routing approach is the Directed BFS combined with the most result in past by Yang & Molina [8]. In this approach, a query is defined to be satisfied if X for some constant X or more results is returned. A peer forwards a search message to a number of peers which returned the most results for the last 10 queries. The nature of this approach is it allows peers explore larger network segments and find most stable neighbors.

Interest based routing [9] tries to avoid the blindness of flood-based routing by favoring nodes sharing similar interest in the source. In this approach, nodes which have similar interest is grouped together and the queries are routed to these nodes in hoping that it will shorten the time for the queries to get the answer.

Koloniari et al. [10] proposed a content-based routing for peer-to-peer based system. In this approach, each peer will have a special index called filters to facilitate query routing only to those that may contain relevant information. Each peer maintains one filter that summarizes all documents that exist locally in the peer, called local filters. A merged filters is the filter that summarizing the document of a set of its neighbors. When a query reaches a peer, the peer will check its local filter and uses the merged filter to route the query to the peers whose filters match the query.

Zeinalipour-Yazti et. al [11] proposed a routing technique based on the similarity of the query. In this approach, each peer has its own profile table that stores the information they get from peers that answered their queries. The information stored in this table is the query ID, peer ID, and the query keywords that have been answered and also the query hit. Only the latest peer that answered the query will be kept into the table of a size t. Routing is based on the similarity values of the query word with the keyword from the past queries stored in the profile. Peers that have high similarity with the query will be selected for routing.

III. Query Feedback based Routing

The proposed approach will be based on the unstructured peer-to-peer. There will be no global knowledge shared between all the peers but each peer will also have a list of data collected from the answered query and store it in Neighbor Profile Table (Table I).

TABLE I
Neighbor Profile Table

Query	ID	Connection and hits	Timestamp
Amazon rain forest	E2343	(P1,25), (P3,1),(P5,20)	10123
Arabian gulf oil	D2334	NULL	10224
Waste disposal	G2343	(P11,15), (P13,11),(P15,20)	10979

The ranking of peers will be based on two parameters, query hits and the similarity value between the keywords to be routed and the stored keywords. Query hits determine peer connection stability with the processing peers. The more query hits, the more stable the peer is connected and thus giving the impression of the particular peers connection reliability. Similarity value will determine the content that the particular peer has in its storage. If the peers have answered the similar query, meaning that it contains data object that is related to the peers. Therefore, both parameters are needed to determine the relevance of a peer to be routed.

Peers that have higher query hits but less similarity will also be considered to be rank higher. The peer ranking will be based on the relevance value in which the smaller the relevance value the higher possibility the peer will be rank higher and selected for query routing for that particular query.

A. Neighbor Profile Table

The Neighbor Profile or the query feedback table is based on the work done by Zeinalipour-Yazti et. al [11].The list will contain the ID of the answering peer, connection ID, the query keywords that have been answered by other peers and a timestamp of the returned query. These keywords are actually the words that match the query sent by this peer, and this shows that these words are contained in the peer that answered this query. The list will keep the last M queries and a Least Recently Used (LRU) policy will keep the most recent queries in the table.

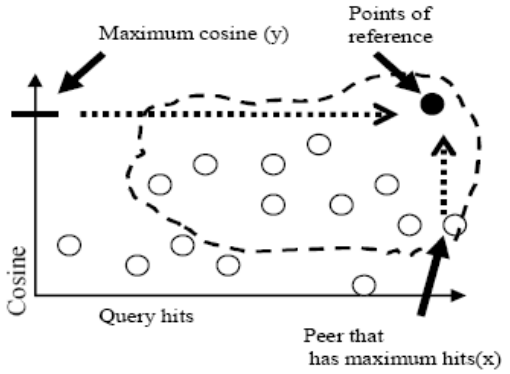

Fig. 1 Routing selection based on query hits and cosine value

Fig. 1 shows the plotted graph of similarity value of incoming query with past queries and past query hits. Each point represents a number of nodes that have answered past queries. A point of reference to determine a peer's relevance is selected based on the optimal point of both parameters. Maximum point on the y-axis is the highest cosine value, which is 1. Therefore, a point that is near to 1 has more similarity with the incoming query. While maximum point on the x-axis is the highest recorded query hits. The higher the query hits, the more reliable this peers are in the network.

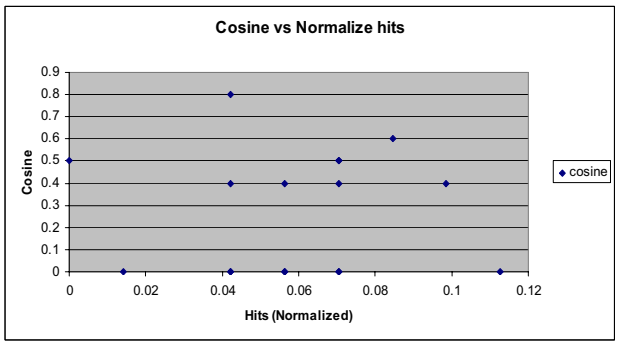

Fig. 2 Snapshot showing query hits and the cosine similarity with recorded past query of a peer processing incoming query "crude oil"

Fig. 2 depicts the similarity and query hits data in a peer during query processing retrieved from the profile table. In this paper, we exploit the similarity and query hits data to rank the peers to be routed. Each point in the figure represents list of connection to other peers. Connections that are represented by these points which then have small Peer Relevance value will be selected to be routed.

B. Peer Relevance

A reference coordinate must be selected to be calculated with all the query hits and cosine similarity vector. The distance between these points, will determined the relevance of a peer to be routed. Maximum query hit, *H* will be selected from the list of query hits for all recorded past query. Similarity between the incoming query and the stored past query is based on the cosine similarity. The max function selects the highest query hits of a query from the profile table.

$$sim(q,q_i) = \frac{\sum(q * q_i)}{\sqrt{\sum(q)^2 * \sum(q_i)^2}} \quad (1)$$

$$H_p = \max(h_i) \quad (2)$$

The peer relevance is determined as follows:

$$R(q,q_i) = \sqrt{\left(\frac{H_p - h_i}{N_p}\right)^2 + \left(M - sim(q,q_i)\right)^2} \quad (3)$$

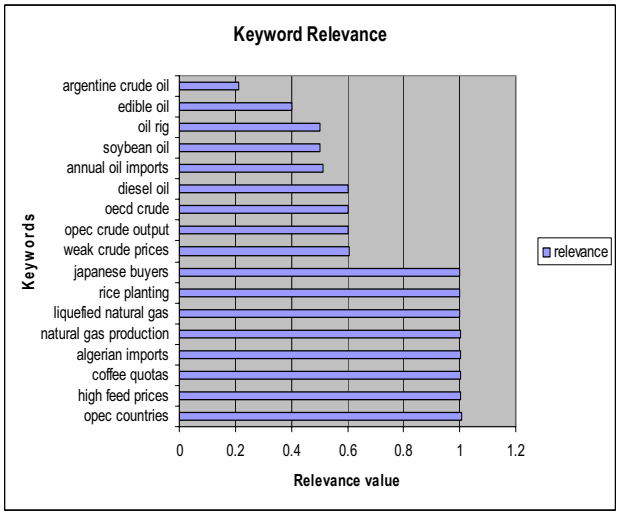

Fig. 3 Snapshot showing query hits and the cosine similarity with recorded past query of a peer processing incoming query "crude oil"

M is the maximum cosine value, but since the maximum value is 1, therefore we decide $M = 1$. h_i is the returned hits values for a particular query, while H_p is the maximum hits retrieved from all *h* that have been recorded. N_p is the total number of query hits of all peers stored in the Neighbor Profile Table. Fig. 3 shows the relevance value of recorded past query with the query "crude oil". We can see that query that has high similarity queries and high query hits will be ranked higher and we can also see that query that has similarity value will have more weight as it guarantee a related content to the query rather than only based on query hits.

IV. Performance Evaluation

We evaluate the performance of the relevance based query routing by extending a peer-to-peer simulator Peerware[12]. The number of nodes generated in this simulation is 230 nodes and the number of documents used is 23336 in total which is generated using a random graph. The documents used in the simulation are part of the Reuters-21578 document collection which appeared on the Reuters newswire in 1987. The documents for each node is categorized by the country attribute and more than one node can have document for one country. A total of 30 queries are used in the experiment. In the simulation, we use Gnutella-based search manner and we compare our approach with the Flooding Mechanism or Breadth-First Search (BFS) routing approach.

In this paper, we will evaluate on time efficiency through the number of query hits over query time. The bigger the value, the more efficient the routing approach in terms of finding hits in a very small time. Network efficiency is evaluated through the total of query hits over total number of messages. The bigger the value means the more efficient the approach is since few number of messages are needed for getting high query hits.

$$\text{Time Efficiency} = \frac{QueryHits}{QueryTime(ms)} \quad (4)$$

$$\text{Network efficiency} = \frac{QueryHits}{Messages} \quad (5)$$

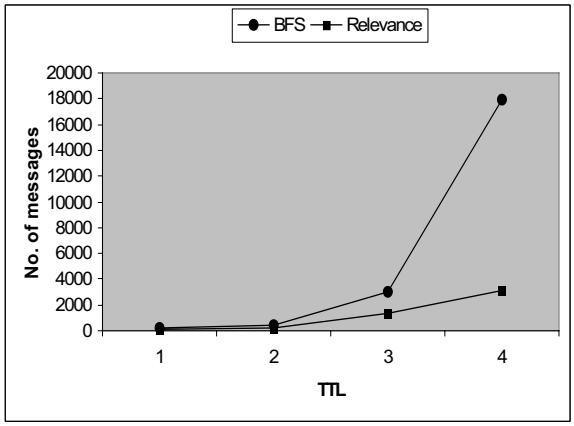

Fig. 4 Number of messages used on different TTL settings

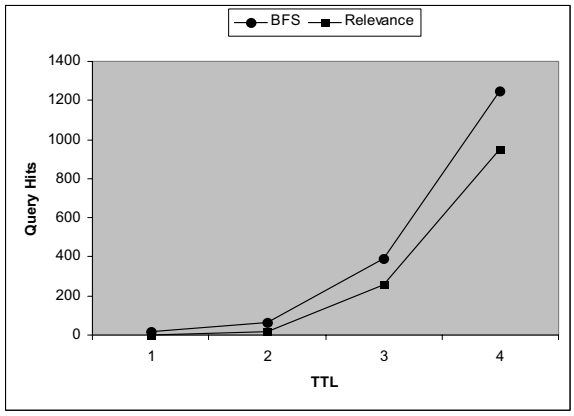

Fig. 5 Number of query hits used on different TTL settings

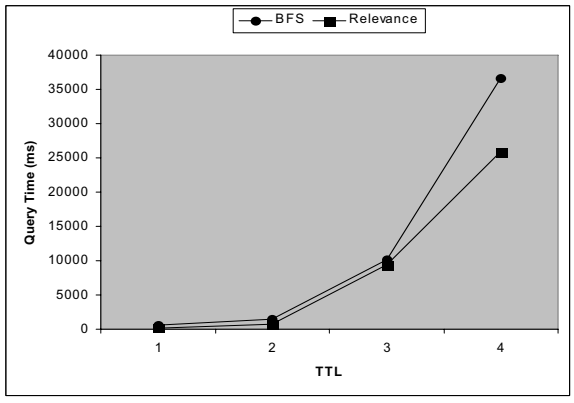

Fig. 6 Query time taken based on different TTL settings

As the number of TTL increases, BFS approach floods to more peers and retrieved the most query hits. Fig. 5 depicts that our approach retrieved a lesser query hits for all TTL settings. Although BFS retrieves more answers than our approach, it has to pay higher cost. At every TTL setting, our approach recorded average 40% less query time than BFS approach as shown in Fig. 6. We also recorded average 13% less use of messages for every TTL settings when compared to flooding approach (Fig. 4).

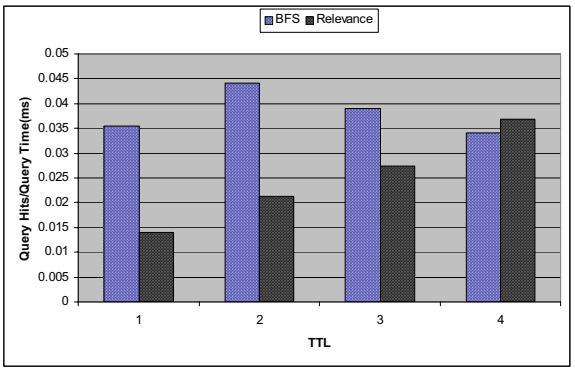

Fig. 7 Efficiency of query over query time

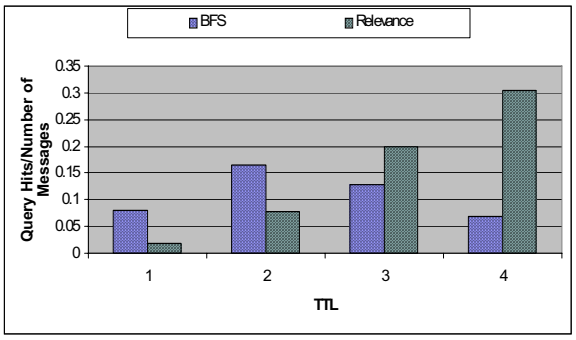

Fig. 8 Efficiency of Query over Use of Messages

As the TTL increases, we can see in Fig. 7 that our approach recorded increasing efficiency. Although BFS retrieve better query hits for all TTL settings, but our approach shown that our approach is efficient as TTL increases. In terms of the use of messages per query hit, our approach recorded gradual increase of efficiency (Fig. 8).

V. Conclusion

This paper researches feedback-based query routing in an unstructured peer-to-peer. Our routing approach uses minimal data to select relevant peers to route queries.

Simulation test shows that our routing approach learn from its previous queries which is stored and select the path to route based on the similarity of the query and the past query as well as the query hits. We show that by using minimal information of query hits and similarity, efficient routing in unstructured peer-to-peer network can be achieved. We have shown that our approach is better in terms of network consumption and time usage than flood-based routing.

Acknowledgment

We wish to thank Demetris Zeinalipour for giving the permission to use the full data set of Reuters-21578 document collection and the Peerware simulator.

REFERENCES

[1] "Napster," http://www.napster.com.
[2] "Gnutella," http://www.gnutella.com.
[3] J. Mishchke and B. Stiller, "A Methodology for the Design of Distributed Search in P2P middleware," IEEE Network, vol. 18, pp. 30-37, 2004.
[4] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network," presented at SIGCOMM'01, San Diego, California, 2001.
[5] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-peer Lokup Service for Internet Applications," ACM, pp. 149-160, 2001.
[6] E.Cohen, A. Fiat, and H. Kaplan, "Associative search in peer-to-peer networks: Harnessing latent semantics," presented at Infocom, San Francisco, 2003.
[7] V. Kalogeraki, D. Gunopulos, and D. Zeinalipour-Yazti, "A local search mechanism for peer-to-peer networks," presented at International Conference on Information and Knowledge Management (CIKM '2002), McLean, Virginia, USA, 2002.
[8] B. Yang and H. Garcia-Molina, "Efficient Search in Peer-to-peer Networks," presented at Proceeding of the International Conference on Distributed Computing System, Vienna, Austria, 2002.
[9] K. Sripanidkulchai, B. Maggs, and H. Zhang, "Efficient content location using interest-based locality in peer-to-peer systems," presented at 22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03), San Francisco, Californa, USA, 2003.
[10] G. Koloniari and E. Pitoura, "Content-Based Routing of Path Queries in Peer-to-peer Systems," Advances in Database Technology, vol. 2992, pp. 29-47, 2004.
[11] D. Zeinalipour-Yazti, V. Kalogeraki, and D. Gunopolus, "Exploiting locality for scalable information retrieval in peer-to-peer networks," Information System, vol. 30, pp. 277-298, 2004.
[12] D. Zeinalipour, "Peerware," http://www.cs.ucr.edu/~csyiazti/peerware.html.

Iskandar Ishak has received a degree in Information Technology from University Tenaga Nasional, Malaysia. He received his Master degree from Royal Melbourne Institute of Technology University, Melbourne. He is currently pursuing Ph.D degree in the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia.

Naomie Salim is an Assoc. Prof presently working as a Deputy Dean of Postgraduate Studies in the Faculty of Computer Science and Information System in Universiti Teknologi Malaysia. She received her degree in Computer Science from Universiti Teknologi Malaysia. She received her Master degree from University of Illinois and Ph.D Degree from University of Sheffield.