Exploiting WordNet for Semantic Routing in Unstructured Peer-to-peer Information Retrieval

Iskandar Ishak Naomie Salim
Faculty of Computer Science and Information System
Universiti Teknologi Malaysia
Johor Bahru, Skudai, 81310, Malaysia

iskandarishak@gmail.com naomie@utm.my

ABSTRACT
Unstructured peer-to-peer networks offer attractive benefits in large-scale information retrieval and search systems because of its fault-tolerance, scalability and decentralization. However, its architecture makes it hard to share useful semantic knowledge among peers. In this paper we propose a semantic based method for unstructured peer-to-peer query routing. Here we exploit the semantic correlation among peers based on the query terms. The term expansion mechanism is the key part of this approach where the term is expanded based on the use of WordNet. Simulation results show that, our approach improves the retrieval and use small number of messages in query routing and retrieval.

Categories and Subject Descriptors
H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval – Retrieval Models

General Terms
Algorithms, Experimentation, Performance

Keywords
Peer-to-peer, semantic routing, information retrieval, WordNet

1. 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 strictly structured way. Whenever peers join or leave the network, peers will be updated to preserve desirable properties for fast lookup. In DHT, each peer has its own hash table and stores keys that are mapped to them. DHT implements one 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 efficiently support partial-match queries [6].

This research will propose an efficient routing scheme for better retrieval in an unstructured peer-to-peer network. The routing will be based on the semantic similarity of the current query with some recorded past queries with the help of an ontology. In this approach, the content of each peer will also be taken into account.

2. RELATED WORK

Earlier peer-to-peer routing is based on random and statistical routing data. However, the use of semantic information is a trend in recent papers. In this section we will review some related works.

The earliest technique for peer-to-peer routing is based on the Naïve Breadth-First Search (BFS) algorithm. 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 is also increased exponentially. Some of the messages might visit the same node that has been searched. Therefore, communication overhead and scalability are the main problems in this approach.

Another routing approach for unstructured peer-to-peer is the is called the random BFS [7]. In this approach, each peer forwards a search message to only a fraction of its peers. Each node randomly selects a subset of peers connected to it to propagate search message. The advantage of this technique is that it does not require any global knowledge. Every node is able to make local decision in a very 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 Z for some constant Z 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 exploring larger network segments and the most stable neighbors.

The other common technique in unstructured peer to peer retrieval is the interest based routing [9]. This technique 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 group 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 indexes called filters to facilitate query routing only to those that may contain relevant information. Each peer maintains one filter that summarize all documents exist locally in the peer, called local filters. A merged filters is another 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 direct the query only to those nodes 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 q. 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.

In [12], a semantic-based peer similarity measurement for efficient query routing is used. A global dictionary using WordNet is built and is updated using flooding approach. They also used some heuristic rules to detect semantic similarity of the peers in the network and answering of queries. The semantic similarity between peers is determined through local schemas.

In a work done by Rostami et al. [13], they used ontology based local indexing to limit the growing size of local indexes without losing indexing information. In this approach, they have the ontology based local index (OLI) which contains a distributed data structure and a routing algorithm. Each peer is able to know the semantic contents of other peers through the links stored in the index. Each semantic is represented by the concept of the data. The routing will be based on the concept that the message have and peers that have similar concept will be chosen to be routed. They have tested this approach only on structured peer-to-peer network.

3. WORDNET

WordNet [14] is an online lexical reference system whose design is inspired by psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into sets of synonyms (synsets), each representing one underlying lexical concept.

The goal of WordNet is to create a dictionary and thesaurus which could be used intuitively. it is also used to determine semantic connections between sets of synonyms, for tracing morphological connections between words. The ontology is organized not only by the “is-the-synonym-of” relation; the verbs and nouns are hierarchically organized through hypernym/hyponym relation too.

There are a number of semantic similarity measures for WordNet such as the Hirst & Onge [15], Resnik [16], Leacock and Chodorow[17], and Lin [18]. Peng-Yuan et al. [19] have proved that Resnik approach is the best similarity measure to be used for an English thesaurus.

4. PROPOSED APPROACH

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 of the peers will have a list of keywords extracted from the documents it stores or the indexed words. Each peer will also have a list of data collected from the answered query and store it in Neighbor Profile Table (Table 1).

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.

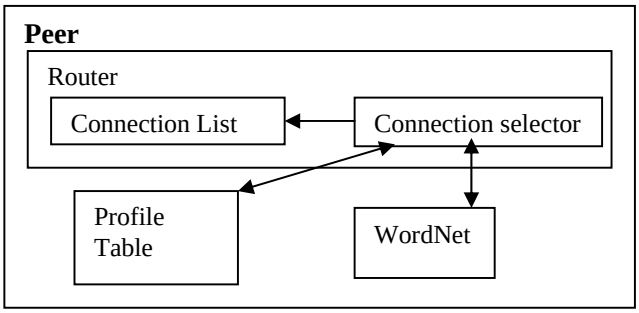

Figure 1: Router framework

The search mechanism will be similar to the mechanism used in the Gnutella Network. A node utilizes its pre-established socket connections to its peers to forward the query messages. This approach will utilize the neighbor profiles to determine which peer should be chosen to route based on the similarity function. When the relevant peers are selected, the query message will be forwarded to these peers. Then it will search within the local repository in each selected peers. Similarly like Gnutella, peers that match the query will generate a Query Hit message and transmits the results along the same path that the query messages had gone through.

Query routing will be based on the peer that has similarity with the words in the index list which means that the peer selected has similarity in terms of its content. The keywords from the list will be further expanded using WordNet to be compared with the keywords in the thesaurus.

Based on the work by Zeinalipour-Yazti et al. [11], each peer will keep a list of query that has been sent and answered by a number of other peers. After a query has been processed locally, process of connection selection will take place. Connection that contains similar keywords with the query will be selected by Connection Selector (Figure 1).

Table 1: An example of a profile table in a peer

Query keywords	Query ID	Connection and hits	Timestamp
Amazon rain forest	E2343	(P1,25), (P3,1),(P5,20)	1000123
Arabian gulf oil	D2334	NULL	1000224
Waste disposal	G2343	(P11,15), (P13,11),(P15,20)	1000979

4.1 Similarity

In order to find the most likely peers to answer a query, a similarity metric can be use, such as the cosine similarity which is also used in information retrieval. Assumption is made on peers that have answered a given query also likely to have other documents that are relevant to that query.

The cosine similarity metric [20] (formula 1) between two vectors has been used extensively in information retrieval for nearest neighbor searches and this measure is also used in this research in this setting as queries consists of keywords. For example, let say we have query word L which contains the words (V, W, X, Y, Z) and we have query W,Z, then the vector that corresponds to this query is (0,1,0,0,1). Similarly to the vector that corresponds to query V, Y is (1,0,0,1,0). The cosine similarity model of the two queries is the cosine angle between two vectors.

$$sim(q, q_i) = \cos(q, q_i) = \frac{\sum (\vec{q} * \vec{q}_i)}{\sqrt{\sum (\vec{q})^2} * \sqrt{\sum (\vec{q}_i)^2}} \qquad (1)$$

4.2 Term expansion

In each peer, before being routed, each receiving query will be expanded using the WordNet. However, only noun synsets will be considered to be expanded. The WordNet will generate a number synsets that contains a number of terms. Identical term to the query will be left out. However, the similar terms produced could be in large number and some might not be relevant. Therefore, there should be a mechanism to filter out the terms, and only the most related or most similar terms will be selected.

Resnik define the similarity between two concepts lexicalized in WordNet to be the information content of their lowest super-ordinate (most specific common subsumer) lso (c_1,c_2) as:

$$sim\ (c_1,c_2) = -log\ p(lso(c_1,c_2)) \qquad (2)$$

where *p(c)* is the probability or encountering an instance of a synset c in some specific corpus.

Since the maximum similarity value for each term is dynamic, there will be no fixed value to put any single threshold value that can be applied to all terms. To select the most semantically close terms, terms that have the nearest value to the maximum value of similarity will be selected. Each time a query arrived to a peer, steps below will be taken to expand the query.

1. Find the similar terms of term c in WordNet
2. Calculate the distance between term c and all the similar terms in 1
3. Find max distance from all the distance value calculated in 2
4. Calculate nearest similarity values with some threshold and terms that have nearest values are to be selected to support the current query

When A is the array of similarity distance between the synset terms produced from term c_i and term c_k. the terms that will be selected will have the values greater than the max distance. To get the max distance:

$$Max\_distance(c_k) = max(sim\ (c, c_k)\ ) \qquad (3)$$

Distance to max with max_distance > 0

$$n = X\ (Max\_distance(c_k))\ -\ sim\ (c, c_k) \qquad (4)$$

X is the threshold value that determines the closeness of the similarity value with the maximum similarity value of a term. The terms that have $n > 0$, will be the terms that have nearest similarity with term c. Figure 2 shows an example how a nearest semantic term is selected for term expansion and which is not selected after the semantic calculation.

As an example, the term 'petroleum', WordNet will retrieve similar words of: '*crude oil', 'crude', 'oil', 'fossil oil'*, and '*rock oil'*. By using the above formula and X=0.8 and Resnik similarity measure, *n* for '*crude oil', 'crude', 'fossil oil'*, and '*rock* oil' yields 7.146, while *n* for term *'oil'* yields 1.865. Therefore term 'oil' is the nearest to the term 'petroleum' and it will be selected for the term expansion.

4.2.1 Term Expansion Strategy

Even though the term expansion is used only for calculating similarity between query terms to be route and local profile table we use Boolean joint OR, AND as in other query expansion research such as in [21].

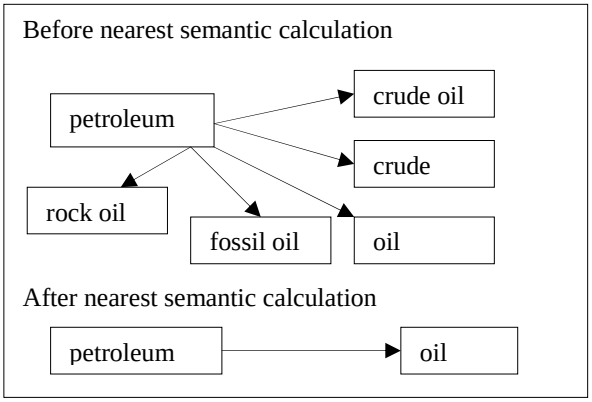

Figure 2: Selection nearest semantic term

For example let say we have a query 'algerian petroleum output'. In WordNet, Algerian do not have any synsets, 'petroleum' produces 'crude oil', 'crude', 'oil', 'fossil oil' and 'rock oil'. Therefore, the query will be joined as in 'Algerian (petroleum OR oil)'.

4.3 Routing Scheme

Semantic routing in the network will be based on several aspects: semantic similarity and availability. Routing of query will be based on the selected terms as explained in Section 4.2.

Figure 3 shows the Procedure Connection shown in peers that have answered past queries and higher similarity with current query to be routed will be selected. Therefore, the routing is selective in which it only selects semantically related peers to route.

```
Procedure ConnectionList
for i=0 to profile.size
{
 If similarity (query, profile[n].query_answered)>k
 {
        resulted_connection = profile[n].connection
 }
}
return resulted_connection
```

Figure 3: Procedure to select links that semantically related

Figure 4 shows the Procedure Routing. If the peers in selected to route the message are unavailable, peers will be selected to route the message. If the peers selected are not enough, the list will be filled with random peers.

```
Procedure Routing
if (connection.size == 0)
        connection = generateRandom
else if (connection.size < half)
        connection = generateRandom
else
    for i=0 to connection.size
     {
        If ((connection ==not null) && connection!=
        originator))
        Send(connection, message)
     }
```

Figure 4: Procedure to route messages

5. SIMULATION AND ANALYSIS

The semantic query routing which based on term expansion is studied by a simple simulation. The number of nodes generated in this simulation is 20 nodes and the number of documents used is 2597 in total. 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. A total of 30 queries are used in the experiment.

For comparison of performance, the proposed technique will be compared with the performance of the popular flooding routing technique, Most Query Hits technique, and directed BFS, and the other one is the Intelligent Semantic Mechanism by Zeinalipour-Yazti et al [11]. Performances of all the methods are compared based on three aspects: retrieval, time consumed to answer queries, and number of messages used to answer all the queries.

Figure 5 shows the number of messages used to answer all queries. Gnutella or Random BFS approach recorded highest number of messages used to answer all the queries in the experiment. Most Query Hits have a slight advantage from other approach as it recorded the lowest number of messages. The semantic based approach recorded slightly better than the ISM and random BFS.

Figure 6 shows the time taken for each routing method to answer all queries. Semantic based approach has shown tremendous increase in time consumption to answer all queries. The expanded terms have made the similarity calculation takes longer times than other approaches.

In terms of query hits, other than the proposed semantic based routing approach, all of the methods recorded the same number of query hits. The network is fixed so that no nodes will be disconnected to make sure that the peers found the same documents and so that we can identify which method can get any differences in retrieval.

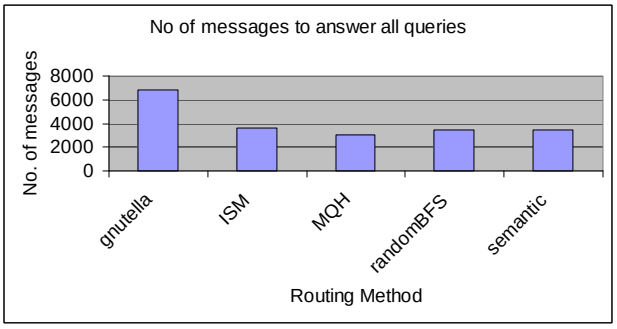

Figure 5: Number of messages used

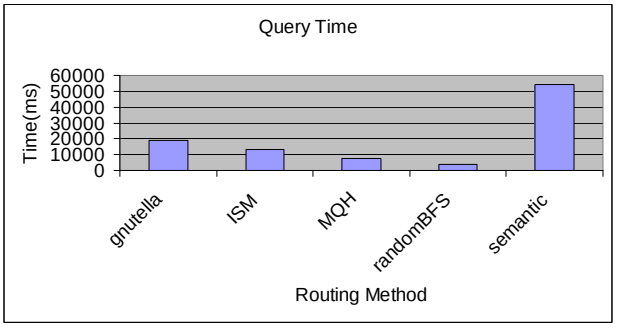

Figure 6: Query Time

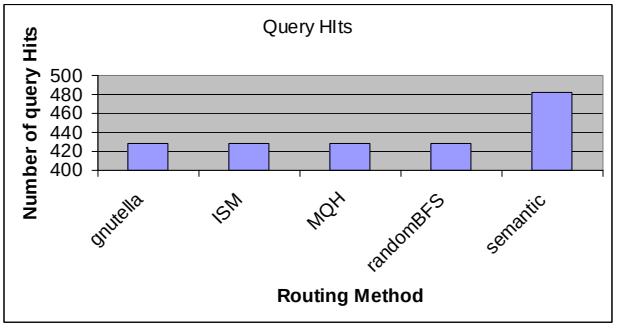

Figure 7: Query Hits

Figure 7 shows that only semantic based approach retrieve more documents. This is because the keywords have been expanded with similar words and thus the retrieval has been increased slightly.

6. CONCLUSION

As a conclusion, the proposed approach has tremendous advantage over the flooding technique in terms of all aspects. The proposed routing approach managed to have better query hits than other approaches in the experiment.

However, it consumes a lot of messages to answer all queries as well as even though the number of messages used less than half of the messages used by the flooding technique. The proposed approach also consumes the highest times to answer all queries because of the expanded terms inside each peer that took time to find similarity

terms in WordNet and also to calculate and determine the most nearest terms to that query terms.

As future work we intend to implement it on the larger simulated network and we intend to improve the approach to reduce the time consume for answering queries.

7. REFERENCES

[1] *Napster*, www.napster.com.

[2] *Gnutella*, www.gnutella.com.

[3] J. Mishchke and B. Stiller, *A Methodology for the Design of Distributed Search in P2P middleware*, IEEE Network, 18 (2004), pp. 30-37.

[4] S. Ratnasamy, P. Francis, M. Handley, R. Karp and S. Shenker, *A Scalable Content-Addressable Network*, *SIGCOMM'01*, ACM, 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 (2001), pp. 149-160.

[6] E.Cohen, A. Fiat and H. Kaplan, *Associative search in peer-to-peer networks: Harnessing latent semantics*, *Infocom*, IEEE, San Francisco, 2003.

[7] V. Kalogeraki, D. Gunopulos and D. Zeinalipour-Yazti, *A local search mechanism for peer-to-peer networks*, *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*, *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*, *22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03)*, San Francisco, Californa, USA, 2003, pp. 2166-2176.

[10] G. Koloniari and E. Pitoura, *Content-Based Routing of Path Queries in Peer-to-peer Systems*, Advances in Database Technology, 2992 (2004), pp. 29-47.

[11] D. Zeinalipour-Yazti, V. Kalogeraki and D. Gunopolus, *Exploiting locality for scalable information retrieval in peer-to-peer networks*, Information System, 30 (2004), pp. 277-298.

[12] H. Zhuge, J. Liu, L. Feng, X. SUn and C. He, *Query Routing In A Peer-to-peer Semantic Link Network*, Computational Intelligence, 21 (2005).

[13] H. Rostami, J. Habibi, H. Abolhassani and M. Amirkhani, *An Ontology Based Local Index In P2P Networks*, *Second International Conference on Semantics, Knowledge and Grid (SKG '06)*, IEEE, 2006.

[14] WordNet, http://www.wordnet.princeton.edu

[15] G. Hirst and D. S. Onge, *Lexical chains as representations of context for the detection and correction of malapropism*, MIT Press (1998), pp. 305-32.

[16] P. Resnik, *Using Information Content to evaluate semantic similarity*, *14th International Joint Conference on Artificial Intelligence*, Montreal, Kanada, 1995, pp. 448-453.

[17] C. Leacock and M. Chodorow, *Combining local context with WordNet similarity for word sense identification. In Christiane Fellbaum, editor, WordNet: A Lexical Reference System and its Application*, Fellbaum (1998), pp. 265-283.

[18] D. Lin, *An information-theoretic definition of similarity*, *15th International Conference on Machine Learning*, Morgan Kauffmann, Madison, Wisconsin, 1998, pp. 296-304.

[19] P.-Y. Liu, T.-J. Zhao and X.-F. Yu, *Application-Oriented Comparison and Evaluation of Six Semantic Similarities Measures Based On WordNet*, *International Conference on Machine Learning and Cybernetics*, IEEE, Dalian, China, 2006.

[20] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman, 1999.

[21] D. Parapar, A. Barreiro and D. E. Losada, *International Conference Applied Computing 2005*, Algarve, Portugal, 2005.