Efficient Search for Peer-to-Peer Information Retrieval Using Semantic Small World

Hai Jin, Xiaomin Ning, Hanhua Chen
Cluster and Grid Computing Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
hjin@hust.edu.cn

ABSTRACT
This paper proposes a semantic overlay based on the small world phenomenon that facilitates efficient search for information retrieval in unstructured P2P systems. In the semantic overlay, each node maintains a number of short-range links which are semantically similar to each other, together with a small collection of long-range links that help increasing recall rate of information retrieval and reduce network traffic as well. Experimental results show that our model can improve performance by 150% compared to Gnutella and by up to 60% compared to the Interest-based model - a similar shortcut-based search technique.

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, Small World, Information Retrieval

1. INTRODUCTION
Efficient search and locating appropriate peers that can answer specific queries in P2P *information retrieval* (IR) systems still remain a challenging problem. Social networks [1] exhibit the small-world phenomenon [2, 3] in which people are willing to have friends with similar interests as well as friends with many social-connections. In our system, we assume that each peer maintains a collection of documents and each document is categorized into one or more known topics. A peer searches by issuing a query that contains a set of keywords and a topical entry. As in social networks, each node maintains a number of short-range links which are semantically similar to the node, together with a small collection of long-range links that help increasing recall of IR and reduce network search traffic as well.

2. SEMANTIC SMALL WORLD MODEL
We consider the model in an unstructured P2P document-sharing system with n nodes and the average degree γ. In this system, there are totally m topics $T = \{T_1, T_2, \ldots, T_m\}$ and each peer P maintains a collection of documents $D = \{D_1, D_2, \ldots, D_d\}$. For the following example, peer 1 has similar topics with peers 6, 8, 13 and 14 in a 2-hops distance. Thus peer 1 has short-range links 6, 8, 13 and 14. Peer 18 has a very strong interest in a particular topic and links as a long-range link to peer 1 with probability.

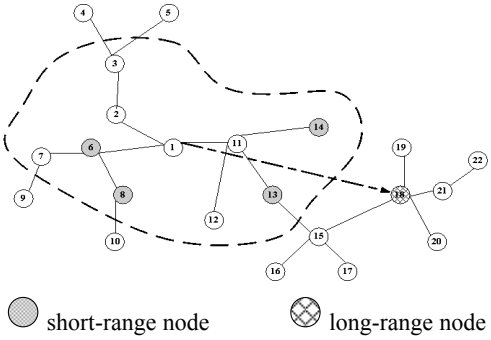

Figure 1: Network topology of semantic small world overlay.

Semantic Summarization
The semantic representation of a peer is based on the stored documents collection. We define the semantic summarization of peer P as $S = \{N, Pr\}$, where N is the total number of documents on P and $Pr = \{Pr_1, Pr_2, \ldots, Pr_m\}$ denotes the fraction of documents belonging to each topic.

Semantic Similarity
We expand cosine similarity in IR to measure the semantic similarity between two peers. For peer P_1 and peer P_2, if the semantic summarizations of peer P_1 and peer P_2 are $S_1 = \{N_1, Pr_1\}$ and $S_2 = \{N_2, Pr_2\}$, respectively, the semantic similarity between P_1 and P_2 is measured as follow:

$$Sim(P_1, P_2) = \frac{1 + \log \min(N_1, N_2)}{1 + \log \max(N_1, N_2)} \frac{Pr_1 \cdot Pr_2}{\| Pr_1 \| \times \| Pr_2 \|} \quad (1)$$

Ultra-Semantic Peer
Long-range links have more "richer" semantic summarizations, thus having much more information useful to answer a specific query. According to the definition of semantic summarization, these peers should have the following properties: (1) large total number of documents N and (2) a distinctly large proportion in a specific topic. For peer P with $S = \{N, Pr\}$, we define the ultra-semantic measurement metric as follow:

$$U(P) = \frac{1 + \log N}{1 + \log \max_{i \in [1..n]} N_i} \max_{Pr_i \in Pr} Pr_i \quad (2)$$

If $U(P)$ exceeds a predefined ultra-semantic threshold $Ultra_{threshold}$, peer P will be called an ultra-semantic peer. We can replace $\max N_i$ with a large value (e.g., 1000) because it is impossible and

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

unnecessary to retrieve the maximal total number of documents in a large-scale unstructured P2P system.

3. ALGORITHMS

Constructing Semantic Overlay

The construction of a semantic small world overlay involves two major tasks: (1) setting up short-range links and (2) establishing long-range links.

Short-range links. When a peer joins the network, it first establishes its semantic summarization and then pulls semantic summarizations from 2-hops neighbors and chooses those peers which are semantically similar to the peer according to formula 1 as short-range links, i.e., for node P, $Sim(P, P_i) > Sim_{threshold}$ where P_i is a 2-hops neighbor of peer P.

Long-range links. In our model, only ultra-semantic peers can be taken as long-range links. We know in a k-dimensional lattice, each node has $2k$ neighbors. Faloutsos et al. [5] considers the neighborhood as a H-dimensional sphere with radius equal to the number of hops where H is the hop-plot exponent. We generalize a P2P network with the average degree γ to an abstract multi-dimensional network and determine the dimension of the network as $H = \gamma/2$. Thus we define the distance $d(P, P_i)$ between peer P and ultra-semantic peer P_i as follows:

$$d(P, P_i) = T \times e^{-Sim(P, P_i)} \tag{3}$$

In the above formula, T is the hops from peer P_i to peer P and $Sim(P, P_i)$ is the semantic similarity between P and P_i. Here we use the semantic distance $d(P, P_i)$ instead of Manhattan distance in [3]. As [2] indicated that a very small probability (e.g., < 0.001) is just enough to construct a small world. Peer P has a long-range link to peer P_i with a probability proportional to $d(P, P_i)^{-r}$ where $r = H$. To establish long-range links, these ultra-semantic peers will actively broadcast their semantic summarizations at a large interval time (e.g., 15 minutes) in the network.

Search

Assume the query $Q = \{K, t\}$ where the set of keywords $K = \{k_1, k_2, \ldots, k_l\}$ and the search topic $t \in T$. The main idea of search is through short-range links and long-range links to intelligently guide the search operation to those appropriate peers which are mostly likely to answer the query. The search operation should not be plunged in a local search and should have the ability to rapidly reach other appropriate regions far away in the network, thus increasing recall rate. There are two modes for the search process: the search topic hits with the peer's major interests or not hits. For the first case, the peer forwards the query to its every short-range link and long-range links which topic with highest proportion equals to t. For the latter case, the peer broadcasts the query to its direct neighbors and at the same time forwards it to long-range links which topic with highest proportion equals to t.

4. EXPERIMENTAL RESULTS

The performance metrics for our evaluation are recall rate and efficiency which is the ratio of recall rate and the average number of messages caused by per query. The two models to be compared with are Gnutella (Version 0.4) and the Interest-based shortcut model [4]. We simulate six topological graphs with the numbers of nodes ranging from 1000 to 6000. Each topology accords with a power-law graph with the exponent $\alpha = 3.0$.

Figure 2 shows that when the number of nodes in the network varies from 1000 to 6000, the recall rate of our model excels at least 150% and 60% over Gnutella and the Interest-based shortcut model in every network scale, respectively.

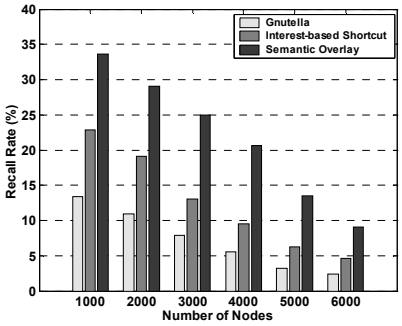

Figure 2: Recall rates comparison for different network scale.

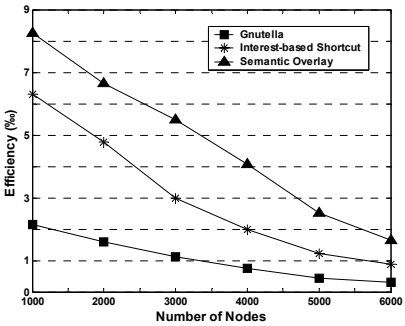

Figure 3: Efficiency comparison for different network scale.

From Figure 3, we can see that the efficiency of our model is much higher than the other two models.

5. CONCLUSIONS

In this paper we present a semantic small world overlay model that facilitates efficient search for P2P information retrieval. Experiments have shown the following results: (1) establishing a semantic small world overlay is feasible and it performs well, and (2) search for IR in the semantic overlay is efficient.

6. ACKNOWLEDGMENTS

This paper is supported by the National 973 Key Basic Research Program under grant No.2003CB317003.

7. REFERENCES

[1] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167-256, 2003.

[2] J. Kleinberg. Navigation in a Small World. *Nature*, 406(845), August 2000.

[3] J. Kleinberg. The Small-World Phenomenon: an Algorithm Perspective. In *Proceedings of ACM Symposium on Theory of Computing*, 2000.

[4] K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient Content Location Using Interest-based Locality in Peer-to-Peer Systems. In *Proceedings of the IEEE INFOCOM'03*, San Francisco, CA USA, 2003.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. *ACM SIGCOMM Computer Communication Review*, 29(4):251-262, 1999.