Small World Peer Networks in Distributed Web Search

Ruj Akavipat †
rakavipa@cs.indiana.edu

Le-Shin Wu†
lewu@cs.indiana.edu

Filippo Menczer †‡
fil@indiana.edu

†Department of Computer Science and ‡School of Informatics
Indiana University
Bloomington, IN 47405

ABSTRACT

In ongoing research, a collaborative peer network application is being proposed to address the scalability limitations of centralized search engines. Here we introduce a local adaptive routing algorithm used to dynamically change the topology of the peer network based on a simple learning scheme driven by query response interactions among neighbors. We test the algorithm via simulations with 70 model users based on actual Web crawls. We find that the network topology rapidly converges from a random network to a small world network, with emerging clusters that match the user communities with shared interests.

Categories and Subject Descriptors: C.2.4 [Computer-Communication Networks]: Distributed Systems; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Measurement

Keywords: Peer collaborative search, topical crawlers, small world

1. INTRODUCTION

Peer network are increasingly seen as a candidate framework for distributed Web search applications. YouSearch [1] maintains a centralized search registry for query routing, making it difficult to adapt the search process to the heterogeneous and dynamic contexts of the peer users. A more distributed approach is to completely decentralize search, as in Gnutella. Queries are sent and forwarded blindly by each peer. But since the peer network topology is uncorrelated with the interests of the peer users, peers are flooded by requests and cannot effectively manage the ensuing traffic. Adaptive, content based routing has been proposed to overcome this difficulty in the file sharing setting. NeuroGrid [2] employs a learning mechanism to adjust metadata describing the contents of nodes. A similar idea has been proposed to distribute and personalize Web search using a query-based model and collaborative filtering [4]. Search however is disjoint from crawling, making it necessary to rely on centralized search engines for content.

To address the scalability limitations of centralized search engines, both Web crawling and searching must be distributed. This allows for symbiotic interactions whereby a peer can vertically adapt to its users' search interests [3], while horizontally peers can achieve better coverage by learning to collaboratively route and respond to queries. In this context, here we present preliminary results to test the hypothesis that even the simplest collaboration method should lead to a topology in which clusters emerge to match communities of peers with shared interests. We predict that the ideal topology for such a network is a small world [6], allowing for any two peers to reach each other via a short path (small diameter) while maximizing the efficiency of communication within clustered peer communities.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

2. PEER COLLABORATION MODEL

Each peer has a unique local search database where it can retrieve local hits. A peer can send queries to other peers and respond to queries from other peers with messages containing search results, scores, and a peer location. A profile can be requested from other known peers. A peer may respond to such a request with a list of most frequent keywords in its search database.

For query forwarding, we give each peer a fixed number of slots for neighbors, N_n, that a peer can forward a query to, but a peer can know more than N_n other peers. The actual set of N_n neighbors, i.e. those to whom queries are sent, is selected dynamically for each query at time t among the $N_k(t)$ known peers. The queried neighbors are chosen among the known peers as those with highest similarity between query and peer profile. The standard TTL mechanism is used to limit forwarding and prevent loops.

When a response is received, the sender is entered into the known peer list. If the peer was not known, a profile is requested. Finally, the scores of hits received are associated with the query keywords and entered into the peer profile.

Many other details of the proposed peer collaborative framework for distributed Web search are omitted for brevity.

3. EXPERIMENTAL SETUP

We created a simulator that allows us to model synthetic users and run their queries over real indexes obtained from actual distributed Web crawls. Our simulator takes a snapshot of the network for every time step. In a time step of the simulator, all of the peers process all of their buffered incoming messages and send all of their buffered outgoing messages. This may include the generation of a local query as well as forwarding and responding to the queries received by other peers.

There are $N = 70$ peers in our simulation. To study whether the adaptive routing algorithm can generate network topologies that capture the interests shared by user communities, we model synthetic users belonging to 7 different groups of 10 users each. Each group is associated with a general topic. Each user has its own peer search engine, but for the peers in a given group the search engines are built by topical crawlers focusing on the same topic.

The group topics are chosen from the Open Directory (ODP, `dmoz.org`) to simulate the group structure, according to a simple methodology developed to evaluate topical crawlers [5]. For each group, we extract a set of 100–150 URLs from the ODP subtree rooted at the category node corresponding to the group's topic. These URLs are used to seed the crawlers of the peers in the group.

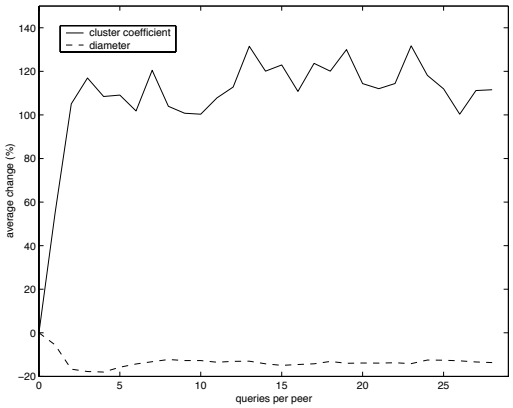

Figure 1: Small world statistics of the peer network.

Each peer has 10 local 2-word queries related to its group topic, generated from the descriptions of the group's seed URLs. The peer uses one of its queries and its group's seed URLs to crawl 10,000 pages (for a total of 700,000 pages). The topical crawlers employ a *best-N-first* crawling algorithm [3, 5]. The `nutch.org` indexer is used to build each peer's search engine from its crawled pages.

Each peer can forward queries to $N_n = 5$ neighbors. At the beginning of each experiment, the peer network is initialized as a random *Erdos-Renyi* graph, i.e., each peer is assigned 5 random neighbors drawn from a uniform distribution, irrespective of groups. A query can be forwarded at most `TTL`=3 times from one peer. We ran the simulator for about 300 time steps, corresponding to 30 queries issued per peer. Since there are only 10 distinct queries per peer, each query is submitted 3 times in the course of a simulation.

4. ANALYSIS OF RESULTS

Let us define two network statistics. The *cluster coefficient* for a node is the fraction of a node's neighbors that are also neighbors of each other. This is computed in the directed graph based on each peer's N_n neighbors, with a total of $N_n(N_n - 1) = 20$ possible directed links between neighbors. The overall cluster coefficient C is computed by averaging across all peer nodes. The *diameter* is defined as the average shortest path length ℓ_p across all pairs p of nodes. We compute the average shortest path as $D = N(N-1)/\sum_p \ell_p^{-1}$, a measure that is robust with respect to the fact that the network is not necessarily strongly connected, and therefore some pairs do not have a directed path ($\ell_p = \infty$). We ran three experiments, corresponding to simulation runs with different seeds of the random number generator. C and D were measured at each time step and averaged across simulation runs.

Figure 1 shows that the diameter remains roughly equal to the initial random graph diameter (actually there is a slight decrease), while the cluster coefficient increases rapidly and significantly, stabilizing around a value 100–130% larger than that of the initial random graph. These conditions define the emergence of a small world topology in the peer network [6]. This is a very interesting finding, indicating that the peer interactions cause the peers to route queries in such a way that communities of users with similar interests cluster together to find quality results quickly, while it is still possible to reach any peer in a small number of steps.

To illustrate the small world phenomenon and the fact that the emerging clusters correspond to the modeled communities, Figure 2 shows the neighborhood topology for all peers and for one of the groups (others display analogous topology) in one of the runs.

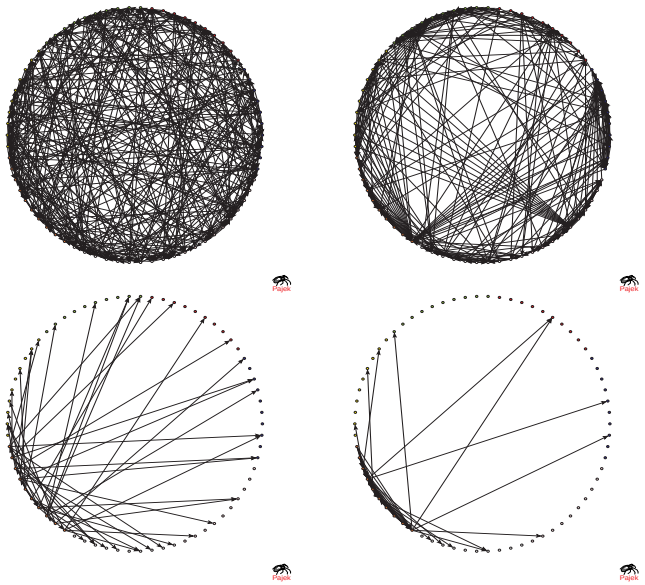

Figure 2: Peer network connectivity for all groups (top) and for one of the groups (bottom). Left: initial neighbor links. Right: Final neighbor links.

Peers are placed along a circle, adjacent within each group; the 10 nodes in the selected group are placed around the 7–8 o'clock position. The total number of links is the same in left and right graphs. The decreasing density of long (cross-group) links indicates that the peers start by routing queries randomly but eventually learn to route queries preferentially to neighbors within groups, so that the network evolves to match the communities with shared interests.

5. CONCLUSION

The experiment results support our hypothesis about the emergence of small world peer networks in collaborative, distributed Web search. The results also support the idea that adaptive routing can work with real Web data. We are currently experimenting with our distributed Web search framework to verify that the critical network structure emerging from local peer interactions leads to improved result quality compared to centralized search.

6. ACKNOWLEDGMENTS

We thank Shannon Bradshaw, Gautam Pant and Padmini Srinivasan for helpful discussions. We are grateful to the Nutch Organization for its open source search engine code, to Gautam Pant for the topical crawler libraries, and to the ODP for the data used to model our simulated users. This work was supported in part by NSF Career Grant IIS-0348940. The AVIDD cluster used in the experiments was funded in part by NSF Grant CDA-9601632.

7. REFERENCES

[1] M. Bawa *et al*. Make it fresh, make it quick — searching a network of personal webservers. *Proc. 12th WWW*, 2003.

[2] S. Joseph. Neurogrid: Semantically routing queries in Peer-to-Peer networks. *Proc. Intl. Work. P2P Computing*, 2002.

[3] G. Pant, S. Bradshaw, and F. Menczer. Search engine - crawler symbiosis. *Proc. 7th ECDL*, 2003.

[4] J. Pujol, R. Sangüesa, and J. Bermúdez. Porqpine: A distributed and collaborative search engine. *Proc. 12th WWW*, 2003.

[5] P. Srinivasan, G. Pant, and F. Menczer. A general evaluation framework for topical crawlers. *Information Retrieval*, Forthcoming.

[6] D. Watts and S. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.