Chapter 3.7
Adaptive Peer-to-Peer Social Networks for Distributed Content-Based Web Search

Le-Shin Wu
Indiana University, USA

Ruj Akavipat
Indiana University, USA

Ana Gabriela Maguitman
Universidad Nacional del Sur, Argentina

Filippo Menczer
Indiana University, USA

ABSTRACT

In this chapter we propose a collaborative peer network application called 6Search (6S) to address the scalability limitations of centralized search engines. Each peer crawls the Web in a focused way, guided by its user's information context. Through this approach, better (distributed) coverage can be achieved. Each peer also acts as a search "servent" (server + client) by submitting and responding to queries to/from its neighbors. This search process has no centralized bottleneck. Peers depend on a local adaptive routing algorithm to dynamically change the topology of the peer network and search for the best neighbors to answer their queries. We present and evaluate learning techniques to improve local query routing. We validate prototypes of the 6S network via simulations with 70–500 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 clusters emerging from user communities with shared interests. We finally compare the quality of the results with those obtained by centralized search engines such as Google.

DOI: 10.4018/978-1-59904-543-6.ch008

BACKGROUND AND MOTIVATION

Centralized search engines have difficulties in achieving good coverage of the Web (Lawrence & Giles, 1999) because the Web is large, fast growing, and fast changing (Brewington & Cybenko, 2000; Fetterly, Manasse, Najork, &

Wiener, 2003; Ntoulas, Cho, & Olston,, 2004). Further, various biases introduced to address the needs of the "average" user imply diminished effectiveness in satisfying many atypical search needs. Examples of bias include interfaces (advanced search features are often buried and poorly documented), ranking (in favor of precision and popularity, to please the majority of users who do not look beyond the first few hits), and coverage (well-connected pages are easy for a crawler to find and thus more likely to be indexed (Najork & Wiener, 2001)).

We identify the above limitations as problems of scale in spite of enormous progress in crawling (Cho & Garcia-Molina, 2002), indexing (Dean & Ghemawat, 2004), and retrieval and ranking (Brin & Page, 1998); the "one-engine-fits-all" model does not—cannot—scale well with the size, dynamics, and heterogeneity of the Web and its users.

Topical or vertical search engines are one approach to address this problem. Effective topical crawling algorithms have been designed to support specialized portals (Chakrabarti, Berg, & Do, 1999; Menczer & Belew, 2000; Menczer, Pant, & Srinivasan, 2004). However, these efforts are generally aimed to very limited information spaces—digital libraries, Web sites, databases—and are not designed to scale with searching the Web at large.

It is evident that distributed systems are part of the answer to the scale problem. Peer social networks are increasingly seen as a candidate framework for distributed Web search applications. A social network is a social structure between participants who are connected through various social relationships. In the real world, we discover that people can successfully find relevant information for questions by just asking the "right" people through their social network, although the network is extremely dynamic (for example, people may not be available all time, people may change their interests anytime, or people can decide not to respond to requests, etc.). Thus, a peer-to-peer (P2P) social network searching system is a network that uses the social network as the basis to route queries for information retrieval. Each peer in the network acts just as a person in the social network:

- Peers are independent
- A peer can enter and leave the network at any time
- Peers learn and store profiles of other peers with a view to their potential for answering prospective queries
- Peers discover new peers through their current neighbors

By simulating the information finding mechanism in a social network of people, the peer network collectively tries to route the queries to the "right" peers according to some peer selection algorithms which predict the degree of match between queries and peers.

A P2P computer social network relies on the computing power and bandwidth of the participants in the network rather than concentrating it in a relatively few servers (Wikipedia, 2005). The most popular use of a P2P network is for file sharing. Applications such as Gnutella (http://www.gnutella.com), BitTorrent (http://www.bittorrent.com), and KaZaa (http://www.kazaa.com) allow peers to share content files, mostly media related, among peers without having to set up dedicated servers and acquiring large bandwidth to support the whole community. The P2P file sharing application is by no means replacing the dedicated servers in content distribution. It simply provides an alternative for content distribution by trading the speed and reliability of dedicated servers for the ease of sharing, lower cost, fault tolerance, and lower bandwidth requirement for a file sharer. In a similar way as P2P file sharing applications are used to facilitate content distribution, P2P applications can be developed to facilitate Web search.

There is extensive work on peer network searching applications in the AI and IR literature.