

Content Summarization in Document –Sharing Peer-to-Peer Networks

Svetlana Strunjaš¹

Current trends indicate that the peer-to-peer architecture (P2P) is becoming a very important model for information sharing. In a P2P system, a large number of peers (e.g. PC's connected to the Internet) can be pooled together to share resources, information and services. Participating nodes can consume as well as provide information and additionally may join and leave the P2P network at any time resulting in a truly dynamic and *ad-hoc* environment. Music and video files were the primary focus of the first generation of P2P networks, e.g. Napster, Gnutella, and Kazaa. Recently, however, P2P networks have shown a real potential for sharing not only video and music files but enterprise documents as well. Although the P2P information sharing systems have many advantages over traditional, centralized information systems in areas of fault tolerance, decentralization, robustness, and flexibility, there are still many challenges related to modeling these systems. One challenge is how to provide fast and reliable data retrieval and efficient search.

The following metrics characterize P2P search efficiency: *bandwidth* consumed for search and retrieval process; *quality* of retrieved results and *computing resources* consumed for search and retrieval process. One very important factor that significantly affects the bandwidth and quality of retrieved results is quality and compactness of *content summarization* at each peer. *Content summarization of a peer is a compact description of the contents stored at the peer.* Due to the distributed nature of P2P systems, there is no centralized authority that contains information on content stored at *every* peer. Therefore, each peer *disseminates* its own content summarization to a certain number of other peers in order to support the distributed search process. There is a tradeoff between the size and accuracy of the content summarization. Larger content summarization describes peer's content more accurately, but consumes more bandwidth for dissemination and vice versa.

My current research focuses on creating a useful model for content summarization in P2P networks that will provide efficient search. I am using the Vector Space Model (VSM – the method popular in information retrieval) to represent the content summarization at each peer and applying different *approximate dimensionality reduction* methods to make the summarization more compact but still highly accurate.

¹ ECECS Department, University of Cincinnati, Cincinnati, OH 45221-0030. E-mail: strunjs@ececs.uc.edu