Scalability of Findability: Decentralized Search and Retrieval in Large Information Networks

Weimao Ke
College of Information Science and Technology
Drexel University, Philadelphia, PA 19104, USA
wk@drexel.edu

Abstract

Amid the rapid growth of information today is the increasing challenge for people to survive and navigate its magnitude. Dynamics and heterogeneity of large information spaces such as the Web challenge information retrieval in these environments. Collection of information in advance and centralization of IR operations are hardly possible because systems are dynamic and information is distributed.

While monolithic search systems continue to struggle with scalability problems of today, the future of search likely requires a decentralized architecture where many information systems can participate. As individual systems interconnect to form a global structure, finding relevant information in distributed environments transforms into a problem concerning not only information retrieval but also complex networks. Understanding network connectivity will provide guidance on how decentralized search and retrieval methods can function in these information spaces.

The dissertation studies one aspect of scalability challenges facing classic information retrieval models and presents a decentralized, organic view of information systems pertaining to search in large scale networks. It focuses on the impact of *network structure* on search performance and investigates a phenomenon we refer to as the *Clustering Paradox*, in which the topology of interconnected systems imposes a scalability limit.

Experiments involving large scale benchmark collections provide evidence on the *Clustering Paradox* in the IR context. In an increasingly large, distributed environment, decentralized searches for relevant information can continue to function well only when systems interconnect in certain ways. Relying on partial indexes of distributed systems, some level of network clustering enables very efficient and effective discovery of relevant information in large scale networks. Increasing or reducing network clustering degrades search performances. Given this specific level of network clustering, search time is well explained by a poly-logarithmic relation to network size, indicating a high scalability potential for searching in a continuously growing information space.