Chapter 2
Decentralized Search and the Clustering Paradox in Large Scale Information Networks

Weimao Ke
College of Information Science and Technology, Drexel University, USA

ABSTRACT

Amid the rapid growth of information today is the increasing challenge for people to navigate its magnitude. Dynamics and heterogeneity of large information spaces such as the Web raise important questions about information retrieval in these environments. Collection of all information in advance and centralization of IR operations are extremely difficult, if not impossible, because systems are dynamic and information is distributed. The chapter discusses some of the key issues 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 discusses a phenomenon we refer to as the Clustering Paradox, in which the topology of interconnected systems imposes a scalability limit.

INTRODUCTION

Information distributes in many large networked environments, in which it is rarely possible to collect all information in advance for centralized retrieval operations. The Web, as one of such information spaces, poses great challenges for information retrieval because of its size, dynamics, and heterogeneity. Baeza-Yates et al. (2007) reasoned that centralized IR systems will become inefficient in the face of continued Web growth and a fully distributed architecture is desirable. Today, a web search engine has to have more than one million servers to survive. However, how to

DOI: 10.4018/978-1-4666-0330-1.ch002

coordinate information collection, indexing, and query processing operations among the huge number of computers internally remains a challenging question.

In addition, there are realistic situations in which collection of information in advance is hardly possible, sometimes unnecessary. The deep web, for example, possesses at least half million databases behind their diverse, sometimes complex, interfaces that do not provide information without being properly queried (Mostafa, 2005; He et al., 2007). In other systems, information is not allowed to be collected and indexed for issues such as privacy and copyright. Sometimes, it is useless to store information beforehand because it is transient and might become irrelevant after being gathered.

In these cases, a large number of information collections distributed in a networked environment is inevitable. The traditional notion of knowing where information is and indexing a "known" collection for later retrieval no longer holds (Marchionini, 1995). While an information need may arise from anywhere in the space (from a delegate system, an agent, or a connected peer), relevant information may exist in certain segments but there requires a mechanism to help the two meet each other – by either delivering relevant information to the one who needs it or routing a query (representative of the need) where information can be retrieved. Potentially, intelligent algorithms may be designed to help one travel a *short path* to another in the networked space.

As these information spaces continue to evolve and grow, it has become crucial to study retrieval models that can adapt and scale into the future. While centralized, "one for all" IR systems are unlikely to keep up with the evolving challenges, a decentralized architecture is promising and, due to many additional constraints, is sometimes the only choice (Baeza-Yates et al., 2007). Without a centralized information repository and global control, the new architecture has to take advantage of distributed computing power and allow a large number of systems to participate in the decision making for finding relevant information.

What is potentially useful in such an information space is that individual systems (e.g., sites, peers, and/or agents) connect to one another and collectively form some global structure. Examples of these network structures include the Web graph of hyperlinks, peer-to-peer networks, and interconnected services/agents in the Semantic Web. Understanding these structures will provide guidance on how decentralized search and retrieval methods can function in networks. Seen in this light, finding relevant information in these information spaces transforms into a problem concerning not only information retrieval but also complex networks (Albert & Barabási, 2002; Kleinberg, 2006).

BACKGROUND

Related challenges for search in distributed settings have been studied in areas of distributed (federated) IR, peer-to-peer networks, multi-agent systems, and complex networks (Callan, 2000; Crespo & Garcia-Molina, 2005; Yu & Singh, 2003; Kleinberg, 2006). In peer-to-peer information retrieval research, for example, problems regarding the applicability of federated IR models in fully distributed environments and scalability of various P2P search models were scrutinized (Zarko & Silvestri, 2007).

While traditional IR and federated IR research provides basic tools for attacking decentralized search problems, the evolving dynamics and heterogeneity of today's networked environments have challenged the sufficiency of existing methods and call for new innovations (Baeza-Yates et al., 2007). Whereas peer-to-peer offers a new type of architecture for application-level questions and techniques to be tested, research on complex networks studies related questions in their basic forms (Albert & Barabási, 2002; Zarko & Silvestri, 2007; Barabási, 2009).