Chapter One

Introduction

Distributed, self-organizing networks such as the World Wide Web and peer-to-peer networks allow for fast access to vast quantities of diverse information for a large number of users. However, with such large scale and data diversity comes the challenge of finding relevant data from reputable sources in an efficient manner.

This book, addresses the issues of relevance and reputation by exploiting user preference information to perform reputation management and personalized search. The issues of personalization and reputation management are highly intertwined, in terms of both the basic ideas and the underlying technologies. Personalization exploits the preferences of an individual to bias search toward that individual's preferences, while reputation management aggregates the preferences of all individuals to bias search toward the data sources that are deemed reputable by the group.

The ideas of reputation and personalization are powerful in conjunction. For example, a personalized Web search for the term "giants" would return the official site of the New York Giants to a football fan from New York, while the same query would return the official site of the San Francisco Giants to a baseball fan from San Francisco. Personalization takes advantage of the local context to return the right sports team, and reputation takes advantage of the global context to return the official site of the corresponding team, rather than some random fan page.

In large-scale diverse data networks, a query will often have so many results that the challenge lies in finding those that are most relevant and reputable. When traditional IR keyword matching techniques are combined with the dual techniques of personalization and reputation management, the end user is likely to have to spend less time intelligently formulating a query and filtering through irrelevant data.

This book is written in two parts, the first part focusing on the Web, and the second on peer-to-peer networks. These parts, while they share the themes of reputation and personalization, differ in style as well as application area. Part I has more mathematical proofs, while Part II relies more on simulation and experimentation. You may read them independently, but together they will give a broader view of the world. Both Part I and Part II, however, are meant for an audience comfortable with advanced concepts in math and computer science.

1.1 WORLD WIDE WEB

Google's PageRank algorithm [56] revolutionized Web search by providing a reliable, spam-resistant way to find reputable web pages. The algorithm is based on the

idea that a link from page i to page j confers authority on page j. Therefore, pages with many links from reputable pages are themselves reputable. Part I of this book addresses the issue of personalizing the PageRank algorithm for individual users.

The PageRank algorithm involves the computation of the dominant eigenvector of a Markov matrix describing the behavior of a model Web surfer jumping from page to page on the Web hyperlink graph. Chapter 2 reviews the PageRank algorithm and the random surfer model. Chapters 3 and 4 introduce some mathematical properties of PageRank that guide how we proceed in algorithm design.

It has been suggested that, by biasing the behavior of the model surfer to reflect the biases of a given user, PageRank can be personalized for each individual user [56]. However, due to the sheer size of the web matrix, doing an individual eigenvector computation for each user is prohibitively expensive, and a computationally tractable algorithm for Personalized PageRank has remained an open problem since it was suggested in 1998.

Chapters 5 through 7 discuss techniques for accelerating PageRank in order make the idea of Personalized PageRank computationally tractable. Personalized PageRank is presented at the end of Chapter 7.

Much of the content in Part I represents joint work with Taher Haveliwala, Glen Jeh, Chris Manning, and Gene Golub.

1.2 P2P NETWORKS

Part II addresses the idea of reputation and personalization in the context of file-sharing peer-to-peer networks. Due to the highly distributed nature of P2P networks, the technical challenges here are different from those described for Personalized PageRank. The first challenge is to devise an algorithm that computes and stores reputation in a distributed manner with minimal overhead and that is resistant to malicious users. Chapter 9 describes the EigenTrust algorithm for reputation management in P2P systems. Since queries in a large-scale P2P network have a limited time horizon, personalizing P2P search can be achieved by designing the topology of a P2P network such that each peer is surrounded by peers that are likely to store data of interest to that peer. In Chapter 10, a peer-level protocol is presented for the self-organization of such a P2P topology. These protocols are tested using a P2P simulator called the Query-Cycle simulator, described in Chapter 8.

Much of the content in Part II represents joint work with Mario Schlosser, Tyson Condie, and Hector Garcia-Molina.

1.3 CONTRIBUTIONS

The work presented in this book offers three main contributions to research in information retrieval.

The first is a mathematical analysis of PageRank, including convergence and stability guarantees. While convergence and stability have been observed empirically for PageRank, domain-independent guarantees are useful when proposing

PageRank-like algorithms in other problem domains. Furthermore, convergence and stability analysis generally lays a foundation for future work in numerical algorithms. In this case, the convergence analysis of PageRank suggests that future algorithms should be based on the Power Method.

The second main contribution of this work is the presentation of a scalable, personalized PageRank algorithm for Web search. In particular, we use properties of the problem and the domain to speed up the PageRank algorithm. The properties of the matrix (sparsity and large eigengap) lead us to use algorithms based on the Power Method throughout the book, and the extrapolation algorithms specifically exploit the matrix properties. The domain properties of the Web as a hierarchical dynamic system lead us to the Adaptive PageRank and BlockRank algorithms. And finally, the linearity of PageRank, another property of the problem, allows us to use all these algorithms in conjunction with Topic-Sensitive PageRank. The scalability issues have long been a bottleneck for the successful deployment of personalized search on the scale of the web, and this book addresses those issues.

The third main contribution is bringing the ideas of reputation and personalization to search in P2P networks. As the quantity and diversity of data on P2P networks approaches that of the web, the importance of search quality in P2P networks becomes increasingly important. The recent focus of research in P2P search has been efficiency for point queries (exact-match queries). However, while efficiency for point queries is important, point queries represent only a small fraction of possible queries in today's P2P networks. Three main ideas are presented within this contribution. The first is the understanding that the ideas behind PageRank can also be applied to search in P2P networks. The second is a method of computing the dominant eigenvector in a highly distributed and potentially subversive environment. These are the basis of the EigenTrust algorithm. The third is a recognition that the local neighborhood is more important than a differential quality score for personalization in P2P search, where queries are only broadcast across a limited time horizon. This is the basis of Adaptive P2P Topologies.