The Information Retrieval Series

Series Editor

W. Bruce Croft

Jin Zhang

Visualization for Information Retrieval

Foreword by Edie Rasmussen

Springer

Jin Zhang
University of Wisconsin
School of Information Studies
532 Bolton Hall
53211 Milwaukee, WI, USA
E-mail: zhang@sois.uwm.edu

ISBN: 978-3-540-75147-2 e-ISBN: 978-3-540-75148-9

Library of Congress Control Number: 2007937243

ACM Codes: H.3, H.4, H.5

Cover Design: Künkel Lopka, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

This book is dedicated to my parents, wife Yi, and son Theodore

Foreword

It was my good fortune, as a relatively new professor at the University of Pittsburgh's School of Information Sciences, to meet Jin Zhang when he was first sent to the US for a year of study by his university in Wuhan, China. Jin impressed me with his energy and enthusiasm for research and I welcomed the chance to work with him. Knowing he had only a year, he accomplished in that time what many take two or three years to do, laying the foundation for his PhD in information sciences. Though he had to return to China after that first year, he continued to actively develop his ideas and models on the mathematical foundations of visualization for information retrieval. He was able to return with his family a few years later to complete his degree, when he again worked at a feverish pace to complete his research and thesis on "Visual Information Retrieval Environments". Jin received his PhD from the University of Pittsburgh in 1999 and moved to the University of Wisconsin – Milwaukee to take up a faculty position in the School of Information Studies, where he is now an Associate Professor.

Jin was driven in large part by his passion for his field of study, visualization models for information retrieval. His inspiration at the University of Pittsburgh was Professor Robert Korfhage. A few years earlier Bob had been involved in the design of *VIBE* (Visualization by Example), one of the earliest models for visualization in information retrieval. The problem of projecting an n-dimensional space onto a two-dimensional one is elegantly but simply solved in *VIBE*, but it is only one of many possible solutions. Bob's background in mathematics was a good match for Jin's, and when he introduced Jin to the problem of visualization for information retrieval, they began a collaboration. There is a fascinating challenge in developing models that are mathematically interesting while producing a display that can be unambiguously interpreted to produce effective retrieval, efficiently calculated and capable of handling large databases. Taking *VIBE* as a starting point, Jin developed new models, and—not always the case in those early years—insisted on implementing them and evaluating their performance as well. Their collaboration ended with Bob's death in 1998, six months before Jin completed his PhD. Though his research interests have broadened to include other areas, Jin has continued to work on developing mathematical models and prototypes for information visualization, including some in the Web environment. *Visualization for Information Retrieval* is the result of over ten years of research in this field. In this book Jin presents the models, limitations and challenges of visualization for information retrieval, and he provides a significant resource for new researchers in the field.

Edie Rasmussen
University of British Columbia

Preface

The dynamics, diversity, heterogeneity, and complexity of information on the dramatically growing Internet and other information retrieval systems have posed an unprecedented challenge to traditional information retrieval techniques and theories. These challenges have driven the need for more interactive, intuitive, and effective systems for information retrieval. The situation has necessitated intense interest in looking for new ways to facilitate users in retrieving relevant information. Information visualization techniques, which can demonstrate data relationships in a visual, transparent, and interactive environment, have become our best hope in dealing with this challenge. Information visualization has a very natural relationship to information retrieval. In fact information retrieval is a thread that goes through all information visualization systems. Information visualization offers a unique way to reveal hidden information in a visual presentation and allows users to seek information from the visual presentation. Browsing as a powerful information seeking means is fully utilized and strengthened in such a visualization environment. Visualization techniques hold a lot of promise for information retrieval. Addressing information visualization from an information retrieval perspective would definitely benefit both information retrieval and information visualization.

The book *Visualization for Information Retrieval* provides a systematic explanation of the latest advancements in information retrieval visualization from both theoretical and practical perspectives. It reviews the main approaches and techniques available in the field. It explicates theoretic relationships between information retrieval and information visualization and introduces major information retrieval visualization algorithms and models. The book addresses crucial and common issues of information retrieval visualization such as elusive evaluation, notorious ambiguity, and intriguing metaphorical applications in depth. It takes a detailed look into the theory and applications of information retrieval visualization for Internet traffic analysis, and Internet information searching and browsing as well. At end of this book, it compares the introduced information retrieval visualization models from multiple perspectives. And finally it discusses important issues of information retrieval visualization and research directions for future explorations.

Readers of this book will gain a good understanding of the current status of information retrieval visualization, technical and theoretical findings and advances made by leading researchers, sufficient and practical details for implementation of an information retrieval visualization system, and existing problems for researchers and professionals to be aware of.

The book is organized and presented as follows:

Chap. 1 provides answers to the fundamental questions about information retrieval visualization such as why the information visualization technique is vital and necessary for information retrieval, how it enhances information retrieval on two fronts: querying and browsing, what are the basic information retrieval visualization paradigms, what are the potential applications and implications of information visualization in information retrieval, and what are the basic procedures for the development of an information retrieval visualization model.

Chap. 2 covers the basic and necessary concepts and theories of information retrieval. These concepts and theories such as similarity measures, information retrieval models, and term weighting algorithms, are prerequisites for the following chapters about information retrieval visualization models. Putting these concepts and theories together as a chapter would not only avoid unnecessary duplicative introduction of these concepts and theories in the following chapters, but also lay a theoretical foundation and better prepare readers to understand the information retrieval visualization models.

Chaps. 3 through 7 address the multiple reference point based models, Euclidean spatial characteristics based models, self-organizing map models, Pathfinder associative network models, and multidimensional scaling models, respectively. The history, concept definition, categorization, algorithm description, algorithm procedure, and applications and implications of these major information retrieval visualization models on information retrieval are discussed in depth. These chapters are at the heart of the book.

Chap. 8 introduces the application of information retrieval visualization to the Internet. The Internet not only poses unprecedented challenges for information retrieval visualization but also provides an enormous opportunity for its application. Information retrieval visualization techniques can be used to alleviate the notorious lost in cyberspace syndrome or disorientation during navigation, making navigation smoother and more comfortable. In addition information visualization applications in related fields such as hyperlink hierarchies, subject directories, browsing history, visual search engine results presentation, Web user information seeking behavior patterns, networking security, and user online discussions are included.

Chap. 9 addresses the notorious concept of ambiguity in a visual space. Reasons for the ambiguity phenomenon are analyzed in different information retrieval visualization environments, both positive and negative implications on information retrieval are expounded, types of ambiguity are defined, and solutions to the problems are also included.

Chap. 10 discusses the basic elements of a metaphor and cognitive implication of a metaphorical interface on communication among users, system developers, and system designers. Metaphorical applications in information retrieval

visualization in various situations at different levels are analyzed. Procedures and principles of a metaphorical application in the field are presented.

Chap. 11 focuses on the evaluation issue. Evaluation for information retrieval visualization is both important and difficult. Two aspects: visualization environment evaluation and visualization retrieval evaluation are distinguished and analyzed. An evaluation standard system for information retrieval visualization, including information exploration, query search, visual information presentation, and controllability, is proposed.

The last chapter of the book is titled "Afterthoughts". This chapter briefly recapitulates the main ideas of the chapters. It compares the five major information retrieval visualization models from the angles of a visual space, semantic framework, projection algorithm, ambiguity, and information retrieval. And finally, it addresses important issues, challenges, and future research directions of information retrieval visualization.

The selected information retrieval visualization models in this book are based on the following criteria. [1] They are mainstream and mature algorithms or models in information retrieval visualization. These models are widely used and recognized. [2] They are representative for various types of information retrieval visualization. Each of the introduced models is sophisticated enough to derive a cluster of related models. [3] They must reflect information retrieval characteristics. Unique features of information retrieval in the context of information visualization are included. [4] They can reveal deep semantic and comprehensive relationships of displayed objects. Although the five information retrieval visualization models are introduced, many other models are also included in various contexts such as metaphorical application, and information retrieval visualization evaluation in the book. In each of these chapters, a complete example of a visualization model is given and implication of information retrieval is presented. Internet information visualization is an independent chapter because the Internet offers an ideal stage for information visualization techniques and a wide spectrum of information retrieval visualization approaches can be applied to it.

I would like to take this opportunity to thank Dr. Edie Rasmussen for writing a foreword for the book and her inspiration and support; Dr. Robert Korfhage for introducing me to this amazing and intriguing field of information retrieval visualization when I pursued my Ph.D. in University of Pittsburgh; Dr. Dietmar Wolfram for his reviewing this book and providing valuable suggestions; the anonymous proposal reviewers and final manuscript reviewers for their insightful comments; Ralf Gerstner and the staff in Springer who made a contribution to the book for their excellent and professional work; Ms. Lynda Citro for her editing the book; and other people who made a contribution to the book. I am also grateful to these publishers Elsevier, Wiley, and IEEE for permission to use their figures in the book. The work is in part sponsored by the Program of Introducing Talents of Discipline to Universities from the Chinese Ministry of Education and the State Administration of Foreign Experts Affairs of China (Grant No.:B07042). Furthermore, the University of Wisconsin Milwaukee has been very supportive of the work.

Finally, thanks must go to my family for their support.

Contents