致远 交大致远教材系列

Computer Science Theory for the Information Age

信息时代的计算机科学理论（英文版）

John Hopcroft, Ravindran Kannan

（美）约翰·霍普克罗夫特 拉文德兰·坎南 著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

致远 交大致远教材系列

Computer Science Theory for the Information Age

信息时代的计算机科学理论（英文版）

John Hopcroft, Ravindran Kannan

（美）约翰·霍普克罗夫特 拉文德兰·坎南 著

上海交通大学出版社

SHANGHAI JIAO TONG UNIVERSITY PRESS

内 容 提 要

本书是上海交通大学致远教材系列之一，由国际著名计算机科学家约翰·霍普克罗夫特教授和拉文德兰·坎南教授编写。本书包含了高维空间、随机图、奇异值分解、随机行走和马尔可夫链、学习算法和 VC 维、大规模数据问题的算法、聚类、图形模型和置信传播等主要内容，书后有附录及索引。从第 2 章开始，每章后面均附有适量的练习题。

本书可作为计算机及相关专业高年级本科生或研究生的教材，也可供相关专业技术人员参考。

图书在版编目(CIP)数据

信息时代的计算机科学理论：英文/(美) 霍普克罗夫特，(美) 坎南著. —上海：上海交通大学出版社，2013

ISBN 978-7-313-09609-8

Ⅰ. ①信… Ⅱ. ①霍… ②坎… Ⅲ. ①计算机科学—理论—英文 Ⅳ. ①TP3-0

中国版本图书馆 CIP 数据核字(2013)第 081062 号

信息时代的计算机科学理论(英文版)

(美) 霍普克罗夫特
(美) 坎南 著

上海交通大学出版社出版发行

(上海市番禺路 951 号 邮政编码 200030)

电话：64071208 出版人：韩建民

上海交大印务有限公司印刷 全国新华书店经销

开本：710 mm×1000 mm 1/16 印张：25.25 字数：620 千字

2013 年 5 月第 1 版 2013 年 5 月第 1 次印刷

ISBN 978-7-313-09609-8/TP 定价：35.00 元

Journey to Success

This is the first book of the "Zhiyuan Textbook Series" to be published by Shanghai Jiao Tong University (SJTU) Press. In my capacity as SJTU President and the Dean of Zhiyuan College, I am delighted to write this preface.

The name of the college, "Zhiyuan", was derived from the inscription "思源致远(Si Yuan Zhi Yuan)" by Zemin Jiang, the former President of the People's Republic of China, who is also an alumnus of SJTU. "思源(Si Yuan)" is part of SJTU's motto which means respect to the heritage and "致远(Zhi Yuan)" means aim high and achieve big. These four words in Chinese embody the ultimate goal of the educational mission of the College.

Zhiyuan College and SJTU have always encouraged and inspired outstanding students to embrace innovation and adopt pioneering spirit in research. In addition to the many gifted undergraduate students, there are also a number of world-class scholars and faculty pursuing their career goals at Zhiyuan College. They work with immense zeal for science, and through concerted and unremitting efforts, they explore ways to nurture gifted students who have the promise to become top-notch scientists in the years ahead.

I can recall vividly the situation when Zhiyuan College was first launched. Back in July 2008, I met with Professors Weinan E and David Cai, both prominent mathematicians. Realizing how important undergraduate education is, all of us agreed that SJTU should create a special teaching and learning environment that would allow us to compete with the best universities in the world and to foster the most talented undergraduates in China. Three months later, SJTU Science Class was created with a mission to bring up scientific leaders of tomorrow with creativity, critical thinking, solid knowledge of mathematics and physics and willing to explore nature and serve society. In December 2010, the Pilot Program for Training Outstanding Students in Basic Sciences supported by Zhiyuan College was included in National Pilot Project of Educational System Reform. With tremendous efforts made by

professors and students alike, we have achieved much success since then. More importantly, our passion, commitment and pursuit of excellence have become the hallmark of Zhiyuan College.

It is truly an honor for Zhiyuan College to publish *Computer Science Theory for the Information Age* as the first of the Zhiyuan Textbook Series. The book was co-authored by Professor John Hopcroft of Cornell University and Dr. Ravindran Kannan, a principal researcher with Microsoft Research Labs located in India. Many professors and students at Zhiyuan College have become familiar with the subject matter since May 2012, when John taught his second course in Zhiyuan College. Indeed, when John and Ravindran learned of the Zhiyuan Textbook Series, they graciously offered SJTU Press the privilege to publish their textbook.

John is a world-renowned scientist and an expert on education in computer science. He was awarded the A. M. Turing Award in 1986 for his contributions in theoretical computing and data structure design. I should mention here the interesting story of how I came to know John. In the early summer of 2011, I learned from one of my friends, Professor Shanghua Teng of University of Southern California, who is also an alumnus of SJTU, that John was in Chongqing on an academic trip. My first thought was that if I could invite such a distinguished scholar to Zhiyuan College, it would benefit the entire computer science program at SJTU enormously. Therefore, I contacted him immediately and flew to Chongqing right away for a face-to-face conversation with him. To this day I can still recall lively how surprised John was at the first sight of me in his hotel lobby. When I briefed him about Zhiyuan College, SJTU and my vision for the development of the Computer Science Department, he showed great interest and readily accepted my invitation to visit Zhiyuan College and SJTU. In December 2011, he began to teach courses to undergraduates at SJTU. Since then, he has spent at least two months a year teaching at Zhiyuan College. For our undergraduates, it is definitely a rare and precious opportunity to learn from such a famous professor. One of the students said after taking his course, "It is a treasured opportunity for us to talk face-to-face with such a great scientist. I think John's purpose for our class was to lead us to discover our research interests, to explore simple, workable ideas that we find engaging. I think he really did that. Now I am passionate about scientific research!"

In addition to teaching courses, John identified another ten world-class scientists in computer science and with them formed a "Chaired Professors' Group". This group worked very hard to improve the undergraduate education program in computer science at Zhiyuan College. In May 2012, John accepted my invitation to become a special counselor to the President of SJTU and offered valuable support in recruiting high quality faculty members for the Computer Science Department in SJTU. He has worked tirelessly in performing his job. I am more than happy to avail myself of this opportunity to express my gratitude to John and Ravindran on behalf of both SJTU and Zhiyuan College.

Zhiyuan College aspires to become not only the home to world-class faculty, but also the springboard for talented undergraduates to success. I would like to express my gratitude to all those who have worked so hard to help us make this dream come true. I firmly believe that the publication of this textbook will prove to be an important milestone along the way. As one student in Zhiyuan College said, "The greatest thing that I have found here is a place full of dreams, a place where great scientists share their passion, a place for new thoughts and opportunities. My four years studying at Zhiyuan College have allowed me to find my dreams, my passion, most importantly new opportunities for research."

I sincerely hope Zhiyuan College and SJTU will do a still better job to help talented students succeed in their academic endeavor.

为天下英才成就梦想

致远系列教材将由交大出版社出版。本书是其中的第一部。作为交大校长兼致远学院院长，我愿做序以推广成才阶梯，尤其是见证过去5年来全体致远人的办学经历。这是一段充满激情与梦想的岁月。

“致远”得名于江泽民学长出席交大校庆时的题词：思源致远。“思源”是交大的校训，“致远”是交大人的目标。

致远学院是一个培养创新型拔尖人才的地方。在这里，汇聚了一群充满激情的追求者和坚持不懈的梦想家。在他们当中，既有国内外知名的学者教授，也有一批批出类拔萃的同学。这些人的共同特征，是有献身科学事业的梦想和激情。他们集聚在这里，用百折不挠的毅力和上下求索的精神，探索着一条中国特色创新型拔尖人才的成长道路。

回首来路，难忘创业同仁。2008年7月，国际知名数学家鄂维南教授、蔡申瓯教授和我聚首交大，一起深入讨论了创新型拔尖人才的培养问题。我们达成的共识是：拔尖人才需要优质教育。上海交通大学有责任为热爱科学事业并有志于科学研究的莘莘学子，率先营造世界一流的成才环境。我们的愿景，是为中国的未来培养科学大师。三个月后，“上海交大理科班”项目正式启动。2010年1月，在理科班基础上，学校成立了基础学科拔尖创新人才培养特区——“致远学院”，致力于培养热爱科学研究，具有创新意识、质疑精神和社会责任感的创新型拔尖人才。同年12月，以致远学院为依托，上海交通大学“基础学科拔尖学生培养试验计划”正式纳入教育部国家教育体制改革试点项目。从理科班到致远学院，交大的老师和同学们付出了巨大的努力，同时赢得了无数的荣誉和赞赏，办学初见成效。尤其令人欣慰的是，我们看到激情、梦想、求实、卓越，已经成为致远人的内在追求。

这本《信息时代的计算机科学理论》得以成为“致远系列”的首部教材，同样令致远人倍感骄傲。教材由图灵奖获得者、康奈尔大学教授John Hopcroft和微软研究院首席研究员、卡耐基梅隆大学兼职教授Ravindran Kannan共同编撰。其实早在2012年5月，John教授第一

次来致远学院授课时，该书的内容已经作为讲义在师生间广泛流传。听说交大正在策划致远系列教材方案，John 和 Ravi 欣然同意将这本讲义在中国的版权交给了上海交通大学出版社。

既为序，不能不提到我与 John Hopcroft 教授近乎传奇般的相识经历。2011 年初夏，一个偶然的机会，我从一位好友那里得知 John 正在重庆讲学。John 是享誉世界的计算机科学家和教育专家。1986 年，基于他在算法及数据结构设计和分析方面的成就，他被授予计算机科学领域的世界最高荣誉——图灵奖。当时我的第一个念头是，如果能邀请到这样一位大师级人物加盟致远学院，并为交大整个计算机学科进行规划，那将带来多么珍贵的发展契机。于是，我第一时间与 John 取得了联系。得知他所下榻的宾馆之后，我从上海专程飞赴重庆登门拜访。当 John 打开房门看到我时的惊讶神情，至今留在我的脑海里。我介绍了致远、交大，以及我对发展交大计算机学科的设想。John 不仅表现出浓厚的兴趣，而且立刻接受了我的访问邀请。2011 年 6 月，John 正式加盟致远学院，亲自为本科生授课。自 2011 年 12 月以来，他不远万里，每年坚持两个月来校上课。为了致远的学生，他放弃了与子女圣诞团聚的机会，放弃了夏天计划中的旅行。说心里话，能坐在课堂上聆听图灵奖获得者当面授课，对于本科生而言，确实机会难得。一位致远学院的同学课后这样说，“能听到一位如此杰出的科学家亲自授课，是我之前从未想到的事情。John Hopcroft 教授真正关注的，是激发我们对科学研究的兴趣。他鼓励我们从简单问题着手，找到我们真正感兴趣的科学问题。我想他做到了。现在，我对科学研究充满了激情。”

除此之外，John 还以他个人的感召力，汇聚了另外 10 位世界级计算机科学家，成立了致远首批讲座教授组，极大地提升了致远学院计算机科学本科教学的水平。2012 年 5 月，他接受我的邀请，开始担任校长特别顾问，承担起交大计算机学科高层次人才引进的全球招募任务。为了做好这项工作，他倾注了大量的时间和心血，在我们共同追求梦想的过程中，我们成为了无话不谈的好朋友。在此，我要代表全体交大人，特别是致远人，向 John Hopcroft 教授表示衷心的感谢。同时，也要向 Ravindran Kannan 先生的理解和支持，表示诚挚的谢意。

梦想是生活的动力。致远学院的梦想，就是汇聚世界名师，成就天下英才。感谢所有与我们一起为梦想奋斗的人！我相信，这本教材的出版，将成为记录在致远人追梦道路上的一块丰碑。正如一名致远学子毕业时所言：“求学致远的最大收获，是找到了一个充满梦想和激情的地方。这里有的是大师，有的是题目，有的是空间，有的是机会。四年来做致远人的耳濡目染，使我更加坚定地走上了献身科学、献身事业的路。我确信，这条路通向梦想！”

愿致远学院和交通大学，为天下英才成就梦想。

Preface

One of Zhiyuan College's main objectives is to develop curriculum for emerging scientific disciplines. One such discipline is data science. As "big data" is becoming a household word, data science is poised to be the next scientific frontier. It will not only affect mathematics, computer science, biology, medicine, social science, economics and a host of other disciplines, it will, by turns, bring fundamental changes to our daily lives.

The foundation of data science is rooted in mathematics and computer science. Together they provide the fundamental principles for collecting, storing and analyzing data. To a large extent, the success of data science hinges upon the collaborative effort between mathematicians and computer scientists.

There are four fundamental challenges in dealing with the data that is coming to us: its large size, the fact that it is often in high dimensional spaces, its complex types, and the ever-presence of noise. In order to deal with these issues, we need to develop a basic intuition about high dimensional spaces, be familiar with probabilistic thinking, and be able to think about modeling and algorithms, all at the same time. This is a significant departure from the traditional curriculum of either computer science or mathematics.

This book, written by two leading theoreticians, comes at a time when the international scientific community is in urgent need of such a text. While presented in computer science theory, its focus is on the mathematical foundations of data science. More than anything else, it lays down a framework upon which computer scientists and mathematicians can discuss primary issues encountered in trying to extract information out of data. It centers upon the probabilistic approach, rather than the traditional discrete mathematics set of ideas. It gives a concise and concrete introduction to basic features of high dimensional spaces, random graphs, singular value decomposition, Markov chains, and, as well, basic

models and algorithms for classification, clustering, on-the-fly sampling and belief propagation. Other current research topics such as compressive sensing, sparsity, and low rank properties are also discussed.

Hopcroft and Kannan have done a great service to the community by providing such a valuable resource.

Weinan E

Contents

1 Introduction

Computer science as an academic discipline began in the 1960's. Emphasis was on programming languages, compilers, operating systems and the mathematical theory that supported these areas. Courses in theoretical computer science covered finite automata, regular expressions, context free languages, and computability. In the 1970's, algorithms was added as an important component of theory. The emphasis was on making computers useful. Today, a fundamental change is taking place and the focus is more on applications. There are many reasons for this change. The merging of computing and communications has played an important role. The enhanced ability to observe, collect and store data in the natural sciences, in commerce and other fields calls for a change in our understanding of data and how to handle it in the modern setting. The emergence of the web and social networks, which are by far the largest such structures, presents both opportunities and challenges for theory.

While traditional areas of Computer Science are still important and highly skilled individuals are needed in these area, the majority of researchers will be involved with using computers to understand and make usable massive data arising in applications, not just how to make computers useful on specific well-defined problems. With this in mind we have written this book to cover the theory likely to be useful in the next 40 years just as automata theory, algorithms and related topics gave students an advantage in the last 40 years. One of the major changes is the switch from discrete mathematics to more of an emphasis on probability and statistics and numerical methods.

The book is intended for either an undergraduate or a graduate theory course. Significant background material that will be needed for an undergraduate course has been put in the appendix. For this reason the appendix has homework problems.

The book starts with the treatment of high dimensional geometry since much modern data in diverse fields such as Information Processing, Search, Machine Learning etc. is represented to advantage as vectors with a large number of components. This is so even in cases when the vector representation is not the natural first choice. Our intuition from two or three dimensional space can be surprisingly off the mark when it comes to high dimensional space. Chapter 2 works out the fundamentals needed to understand the differences. The emphasis of the chapter (as well as the book in general) is to get across the mathematical foundations

rather than dwell on particular applications which are only briefly described.

The mathematical area most relevant to dealing with high-dimensional data is Matrix Algebra and Algorithms. We focus on Singular Value Decomposition a central tool in this area. Chapter 4 gives a from-first-principles description of this. Applications of singular value decomposition include Principal Component Analysis (a widely used technique) which we touch upon as well as modern applications to statistical mixtures of probability densities, discrete optimization etc. which are described in more detail.

Central to our understanding of large structures like the web and social networks is building models to capture essential properties of these structures. The simplest model is that of a random graph formulated by Erdös and Rényi, which we study in detail proving that certain global phenomena like a giant connected component arise in such structures with only local choices. We also describe other models of random graphs.

One of the surprises of Computer Science over the last two decades is that some domain-independent methods have been immensely successful in tackling problems from diverse areas. Machine Learning is a striking example. We describe the foundations of machine learning — both learning from given training examples as well as the theory of Vapnik-Chervonenkis dimension which tells us how many training examples suffice for learning. Another important domain-independent technique is based on Markov Chains. The underlying mathematical theory as well as the connections to electrical networks forms the core of our chapter on Markov Chains.

The field of Algorithms has traditionally assumed that the input data to a problem is presented in Random Access Memory, which the algorithm can repeatedly access. This is infeasible for modern problems and the streaming model and other models have been formulated to better reflect this. In this setting, sampling plays a crucial role and indeed we have to sample on the fly. We study how to draw good samples efficiently and how to estimate statistical as well as linear algebra quantities with such samples in Chapter 7.

One of the most important tools in the modern toolkit is Clustering — dividing data into groups of similar objects. After describing some of the basic methods for clustering such as the k-means algorithm, we focus on modern developments in understanding these as well as newer algorithms. The chapter ends with a study of clustering criteria.

The book also covers graphical models and belief propagation, ranking and voting, sparse vectors, and compressed sensing. The appendix includes a wealth of background material.

A paragraph about notation in the book. To help the student we have adopted certain notations and with a few exceptions adhered to them. We use low case letters for scaler variables and functions, bold face lower case for vectors, and bold face upper case letters for matrices. Lower case near the beginning of the alphabet tend to be constants, in the middle of the alphabet such as i, j, and k are indices in summations, n and m for integer sizes, and x, y and z for variables. Where the literature traditionally used a symbol for a quantity, we also used the symbol even if it meant abandoning our convention. If we have a set of points in some vector space and work with a subspace we use n for the number of points, d for the dimension of the space and k for the dimension of the subspace.