

大数据抽样算法



童咏昕

计算机学院 软件开发环境国家重点实验室

yxtong@buaa.edu.cn

课程提纲

- 浅谈前沿热点
- 文献搜索技巧
- 随机抽样算法
- 水库抽样算法 (Reservoir Sampling)

课程提纲

- 浅谈前沿热点
- 文献搜索技巧
- 随机抽样算法
- 水库抽样算法 (Reservoir Sampling)

前沿与热点

- 前沿 vs. 热点
 - 时间维度
 - 影响力度

前沿与热点

- 前沿 vs. 热点
 - 时间维度
 - 影响力度
- 计算机领域前沿热点何处寻?
 - 科技媒体
 - MIT Technology Review、科学网、等等
 - 科技文献
 - 综合性期刊(Science、Nature、PNAS等等)



前沿与热点

- 前沿 vs. 热点
 - 时间维度
 - 影响力度
 - 计算机领域前沿热点何处寻?
 - 科技媒体
 - MIT Technology Review、科学网、等等
 - 科技文献
 - 综合性期刊 (Science、Nature、PNAS等等)
 - 计算机领域顶级期刊与会议 (CCF A类)



中国计算机学会推荐国际学术会议和期刊目录

网址: <http://www.ccf.org.cn/sites/ccf/paiming.jsp>

The screenshot shows the homepage of the CCF Recommended International Academic Conference and Journal Catalog. The header features the CCF logo and the text "中国计算机学会推荐 国际学术会议和期刊目录". Below the header are navigation links: 首页 (Home), 关于目录 (About Catalog), 意见反馈 (Feedback), and 联系我们 (Contact Us). The left sidebar lists categories: 计算机体系结构/并行与分布计算/存储系统, 计算机网络, 网络与信息安全, 软件工程/系统软件/程序设计语言, 数据库/数据挖掘/内容检索, 计算机科学理论, and 计算机图形学与多媒体. The main content area displays the title "中国计算机学会推荐国际学术会议和期刊目录" and a historical note about the YOCSEF forum in 2005. It also includes a detailed description of the catalog's evolution and the classification of conferences and journals into A, B, and C categories.

中国计算机学会推荐
国际学术会议和期刊目录

首页 关于目录 意见反馈 联系我们

计算机体系结构/并行与分布计算/存储系统

▶ 计算机网络

▶ 网络与信息安全

▶ 软件工程/系统软件/程序设计语言

▶ 数据库/数据挖掘/内容检索

▶ 计算机科学理论

▶ 计算机图形学与多媒体

中国计算机学会推荐国际学术会议和期刊目录

2005年12月17日，中国计算机学会青年计算机科技论坛（CCF YOCSEF）举办了“从SCI反思中国的学术评价体制”的专题论坛，探讨为何SCI会成为衡量大学、科研机构和科学工作者学术水平的最重要的、甚至是唯一的尺度；提出了如何建立中国公正合理的学术评价体制的问题，这次论坛在国内引起了强烈的反响。李国杰理事长在各种场合多次呼吁要重视在顶级国际学术会议上发表论文，希望YOCSEF拿出顶级学术会议和重要学术期刊的目录，提供给各高校和科研单位作为学术水平评价的参考。

经过几届YOCSEF 学术委员会的努力，经过调研、分析、选择试点方向，初步完成了大部分学科方向的推荐目录。后来，CCF常务理事会委托CCF学术工委组织此项工作，通过进一步收集、整理、研讨，行成初稿后向学术界公开征集意见，2010年8月，发布了《中国计算机学会推荐国际学术会议和期刊目录》（第一版）。2011年7月，CCF学术工委又根据广大学术同行的反馈意见和建议，修订发布了《中国计算机学会推荐国际学术会议和期刊目录》（第二版）。目录中的刊物和会议分为A、B、C三档。A类指国际上极少数的顶级刊物和会议，鼓励我国学者去突破； B类指国际上著名和非常重要的会议、刊物，有重要的学术影响，鼓励国内同行投稿；C类指国际学术界所认可的重要会议和刊物。

A类 (CCF A) 指国际上极少数顶级刊物和会议，代表计算机学科国际最前沿的发展动态与趋势，鼓励我国学者图突破！

中国计算机学会推荐国际学术会议和期刊目录

• 十类研究方向

序号	研究方向
1	计算机体系结构 / 高性能计算 / 存储系统
2	计算机网络
3	网络与信息安全
4	软件工程/系统软件/程序设计语言
5	数据库/数据挖掘/内容检索
6	计算机科学理论
7	计算机图形学与多媒体
8	人工智能
9	人机交互与普适计算
10	交叉/新兴 / 综合等

中国计算机学会推荐国际学术会议和期刊目录

• 十类研究方向

序号	研究方向
1	计算机体系结构 / 高性能计算 / 存储系统
2	计算机网络
3	网络与信息安全
4	软件工程/系统软件/程序设计语言
5	数据库/数据挖掘/内容检索
6	计算机科学理论

一、A类

序号	刊物简称	刊物全称
1	TODS	ACM Transactions on Database Systems
2	TOIS	ACM Transactions on Information and Systems
3	TKDE	IEEE Transactions on Knowledge and Data Engineering
4	VLDBJ	VLDB Journal

10

二、B类

序号	会议简称	会议全称	出版社	网址
1	SIGMOD	ACM Conference on Management of Data	ACM	http://www.sigmod.org
2	SIGKDD	ACM Knowledge Discovery and Data Mining	ACM	http://www.acm.org/sigkdd/
3	SIGIR	International Conference on Research and Development in Information Retrieval	ACM	http://www.acm.org/sigir/
4	VLDB	International Conference on Very Large Data Bases	Morgan Kaufmann/ACM	http://www.vldb.org
5	ICDE	IEEE International Conference on Data Engineering	IEEE	http://www.icde.org/

中国计算机学会推荐国际学术会议和期刊目录

- CCF A类文献的优势
 - 经典图灵奖的贡献



Alan M. Turing

2014年11月13日前奖金为250,000美元。Google反而将奖金提高到1,000,000美元，和诺贝尔奖奖金相近。

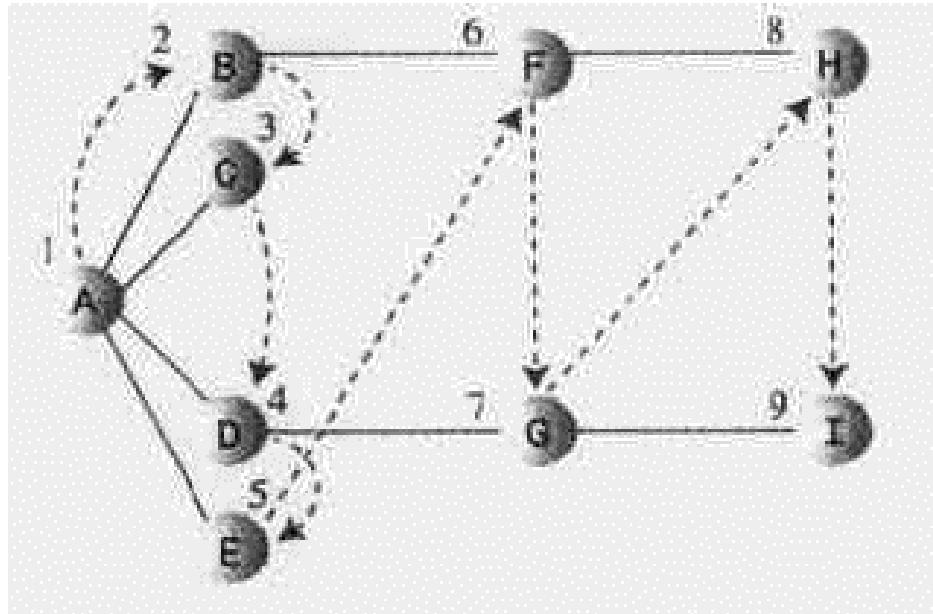


图灵奖
计算机领域的诺贝尔奖

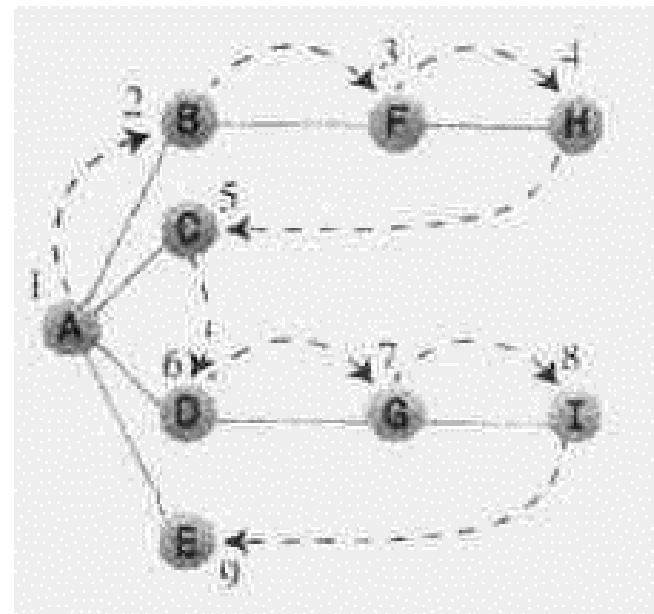
图灵奖自1966设立以来，至今已有64位伟大的计算机科学家获奖！
今年更是图灵奖设立50周年！

深度优先搜索(DFS)

- 深度优先 vs. 广度优先



广度优先搜索

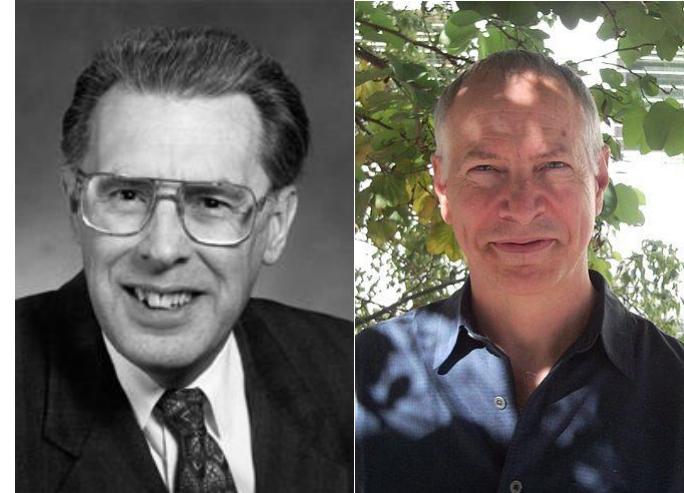


深度优先搜索

深度优先搜索(DFS)

- Robert Tarjan

- 1986 年与 Prof. John Hopcroft 由于在算法与数据结构的设计分析上的卓越贡献获得图灵奖。
- 他发现了解决最近公共祖先 (LCA) 问题、强连通分量问题、双连通分量问题的高效算法，设计斐波那契堆、伸展树等经典数据结构。



John
Hopcroft

Robert
Tarjan

Robert Tarjan. Depth-First Search and Linear Graph Algorithms. SIAM Journal on Computing 1(2), 146-160, 1972. (CCF A类期刊)

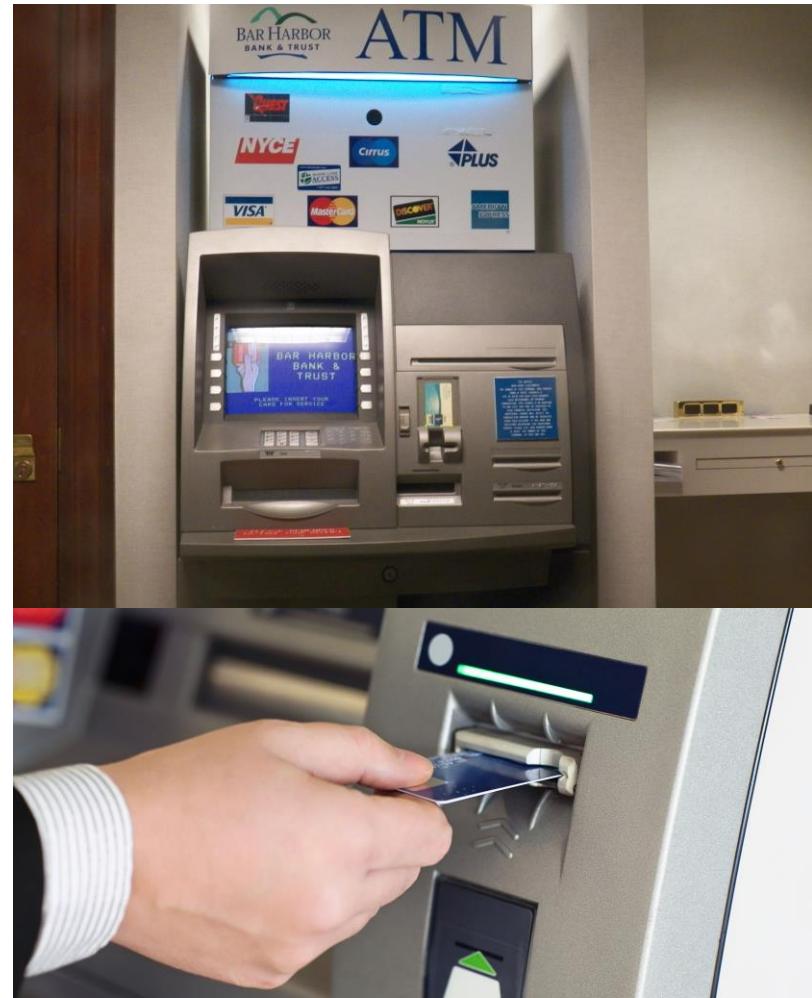
事务处理

- 典型应用

- ATM机
- 银行金融领域交易

- 事务处理四要素
(ACID) :

- 原子性(Atomicity)
- 一致性(Consistency)
- 隔离性(Isolation)
- 持久性(Durability)

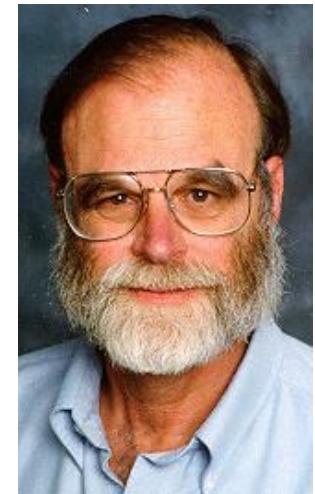


Jim Gray. The Transaction Concept: Virtues and Limitations.
In VLDB, Pages 144-154, 1981 . (CCF A类会议)

事务处理

- James Nicholas "Jim" Gray

- 由于在数据库系统于事务处理领域的卓越贡献，1998年获得图灵奖。
- 微软公司为其建造Redmond研究院，1995年加入微软。
- 2007年，他独自航向法拉伦岛，打算撒散母亲的骨灰，1月28日，他的船失踪。2012年，他在法律意义上被认定已经死亡。



Jim Gray



Jim Gray. The Transaction Concept: Virtues and Limitations.
In VLDB, Pages 144-154, 1981 . (CCF A类会议)

中国计算机学会推荐国际学术会议和期刊目录

- CCF A类文献的优势
 - 经典图灵奖的贡献
 - 新方向的奠基工作

“啤酒-尿布”与数据挖掘

• 超市典型购买记录

小童



苹果



可乐



咖啡

小王



尿布



啤酒

...

小张



牛奶



饼干

...

小刘

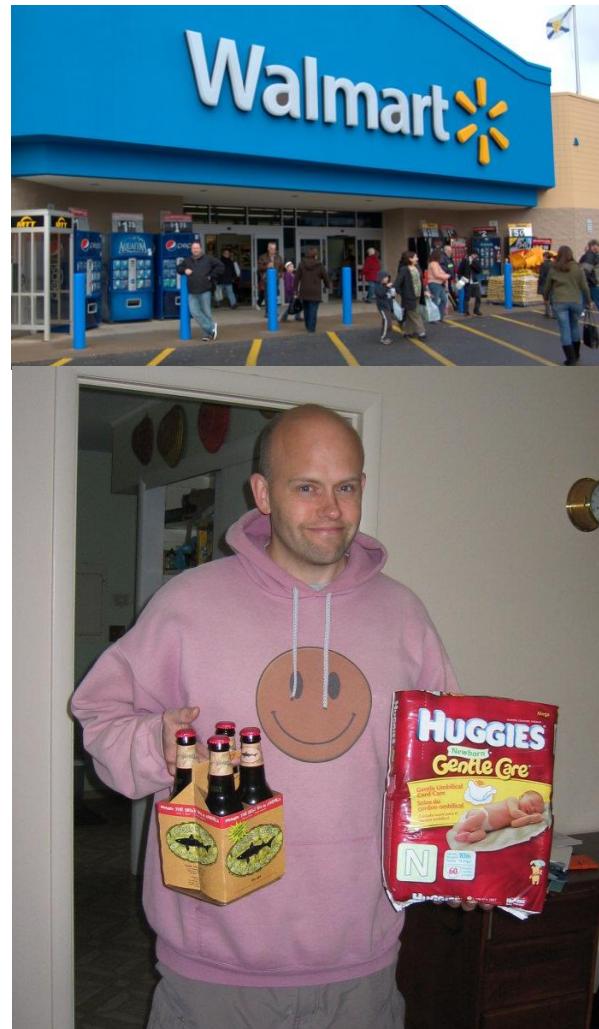


可乐



牛奶

...



“啤酒-尿布”与数据挖掘

- Rakesh Agrawal

- 由于在数据挖掘领域中的先驱性研究，特别是关于关联规则挖掘与隐私保护的研究，其被称为数据挖掘之父。
- 其关于关联规则挖掘的论文，至今已经被全球学者引用超过3.7万次，并分别获得数据库领域国际顶级会议SIGMOD和VLDB的“时间检验奖”与“最具影响力奖”。



Rakesh Agrawal

Rakesh Agrawal, Tomasz Imielinski, Arun N. Swami. Mining Association Rules between Sets of Items in Large Databases.
In SIGMOD, Pages 207-216, 1993 . (CCF A类会议)

“啤酒-尿布”与数据挖掘



中国计算机学会推荐国际学术会议和期刊目录

- CCF A类文献的优势

- 经典图灵奖的贡献
- 新方向的奠基工作
- 新应用的创新源泉

PageRank与Google

- Google诞生简史
 - 他们是斯坦福大学博士生，1994–1998年从事上述“关联规则挖掘”的研究；
 - 他们的研究聚焦于从一篇学术论文在其他论文中的引用量来推断其重要性，这一概念就是PageRank的核心，继而所开发的系统就是Google的原型



Sergey Brin
Lawrence Page

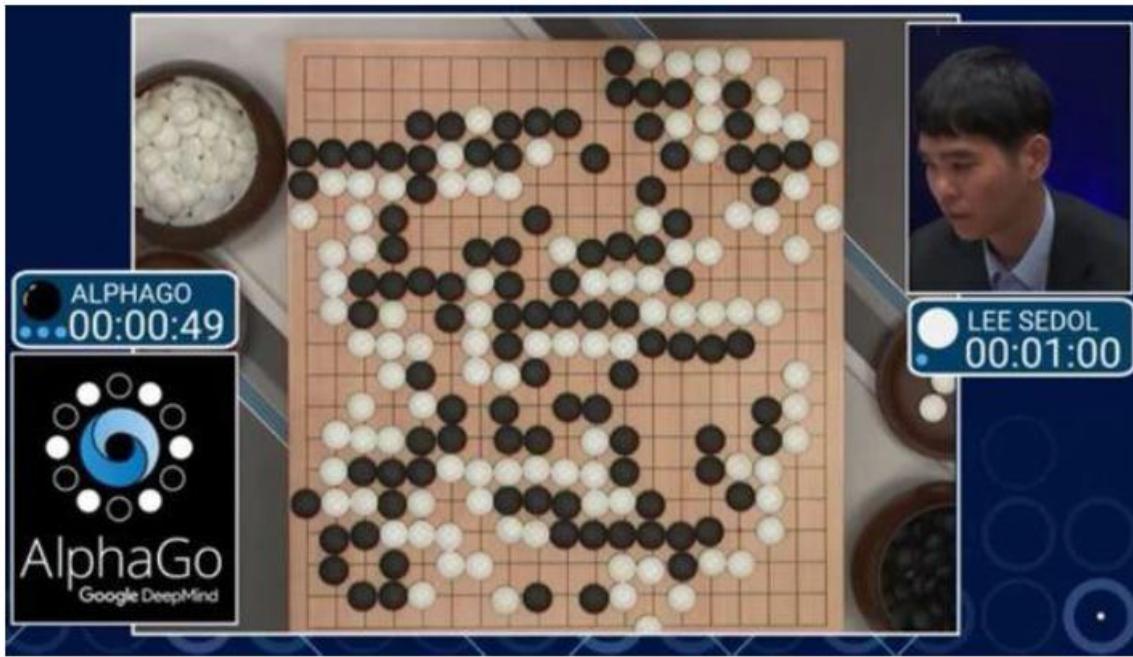
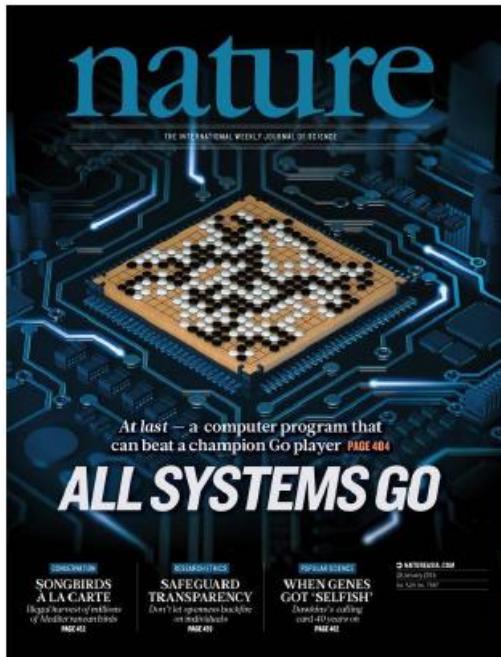


Sergey Brin, Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In WWW, Pages 107-117, 1999. (CCF A类会议)

深度学习与AlphaGo

• AlphaGo（阿尔法围棋，又昵称阿尔法狗）

- 是由英国伦敦Google DeepMind开发的人工智能围棋程序。2015年10月，它成为第一个无需让子，即可在19路棋盘上击败围棋职业棋士的电脑围棋程序
- 2016年3月，在一场五番棋比赛中，AlphaGo于前三局以及最后一局均击败顶尖职业棋手李世乭，成为第一个无需让子而击败围棋职业九段棋士的电脑围棋程序



中国计算机学会推荐国际学术会议和期刊目录

- CCF A类文献的优势

- 经典图灵奖的贡献
- 新方向的奠基工作
- 新应用的创新源泉

Science、Nature、PNAS与CCF A类文献并非是追踪
科技应用前沿技术引用的唯一手段，但有很大概率可
捕获到前沿热点！

课程提纲

- 浅谈前沿热点
- 文献搜索技巧
- 随机抽样算法
- 水库抽样算法 (Reservoir Sampling)

文献搜索技巧

- 搜索方式

- 基于来源
- 基于主题
- 基于作者

- 搜索工具

- 中文工具

<http://www.ccf.org.cn/sites/ccf/>

- 科普网站：科学网、中国计算机学会官网
- 学术刊物：计算机学报、软件学报
- 搜索平台：中国知网、万方数据、等等

<http://lib.buaa.edu.cn/resources?cid=26&pid=19>



文献搜索技巧

• 搜索工具

• 中文工具

。 科普网站：科学网、**中国计算机学会官网**

① www.ccf.org.cn/sites/ccf/

中国计算机学会
China Computer Federation

为计算领域的专业人士服务
Serving the professionals in computing

首 页 | 关于CCF | 学会动态 | 学会活动 | 会员 | 数字图书馆 | 在线培训 | CCF学术空间 | 专业委员会 | 工作委员会 | CCF奖励 | 学术评价

《类脑计算》讲习班开始报名 CCF面向技术前沿讲座TF《网络安全》开始报名

学会新闻

CCF U397: 藏银林走进宁波工程学院
CCF 邀请您一起壮大会员队伍
CCF 重庆成功举办西部“云计算·智能制造”高峰论坛
CCF 山西大学生分会成功举办换届选举会议
CCF 天津大学生分会举办新老生见面会
中国计算机学会（CCF）招聘文字编辑
CCNC2016专题论坛“信息社会—中国准备好了吗？”活动预告

专委活动 · 征文 会员推荐会议

第一届中国计算机学会生物信息学会议（CBC 201...
CCF第六届文字计算学术研讨会会议通知
首届语言与智能高峰论坛会议邀请函
大会征文：2016中国计算机应用大会暨2016年互...
会议通知：2016中国计算机应用大会暨2016年互...
2016年全国高性能计算学术年会（HPC CHINA 201...
征文通知：2016中国互联网+工业4.0助推传统企...

公告栏 会员成就

CCF面向技术前沿讲座TF《网络安全》开始报名
CCF长沙走进高校活动之如何利用互联网深化教学...
2016CCF合肥为新研究生导航活动预告
ACM-IV 研究大会参会学生奖学金
ACM学术中心资源介绍
ACM2016年9月活动预告
山西省高校计算机院长/系主任论坛将在太原举行

CNCC 2016中国计算机大会
China National Computer Congress

计算改变未来
2016年10月20-22日 山西·太原

招聘 求职

【寒武纪科技_智能时代的引领者】诚聘英才
招聘技术总监/合伙人
Java架构开发工程师
中科院计算所《计算机科学技术学报(英)》招聘
杭州电子科技大学自动化学院
同济大学计算机科学与技术系黄德双课题组招聘博

数字图书馆最新资料

企业视角看到的开源——华为开源5年实践经验
开源软件的量化分析
移动数字取证技术
开源软件缺陷管理及自动修复
开源软件和开源社区的反思
移动平台用户隐私保护技术

加入CCF

CCF 会员资格延续 通知

CCF YOCSEF

ADL 中国计算机学会 学科前沿讲习班

欢迎报考：CCF 计算机职业资格认证

CCF 活动计划

《技术动态》2016年第257期

CCF《技术动态》
IT领域重要科学成果与发现，真实反映IT领域的实力和水平...

中国计算机学会通讯

中国计算机学会通讯
本期专题是网络领域新兴课程的建设。网络技术的发展日新月异，现代工程...

技术动态

学会通讯

文献搜索技巧

- 搜索工具

- 中文工具

- 学术刊物：计算机学报、**软件学报**

① www.jos.org.cn/ch/index.aspx ☆



访问次数: 2093894
在线出版
各期目录
纸质出版
分辑系列
论文检索
论文排行
综述文章
专刊文章
美文分享
各期封面
E-mail订阅

期刊介绍 | 编委会 | 编辑部 | 服务介绍 | 相关网站 | 在线审稿 | 编委办公 | 编辑办公

微信服务介绍 最新一期: 2016年

软件学报 *Journal of Software*

信息发布 投稿指南 问题解答 下载区 收费标准 在线投稿

[《软件学报》专刊/题一览表\[2015/4/8\]](#)
[《软件学报》2016-2017年专刊出版计划\[2015/5/8\]](#)
["中国计算机期刊网"正式上线, 欢迎访问! \[2013/3/29\]](#)
[《软件学报》专刊征文: 复杂环境下的机器学习研究\(第一轮截稿时间: 2017年1月10日\)\[2016/7/8\]](#)
[招聘信息: 《计算机系统应用》编辑\[2016/9/1\]](#)
[2016年第10期专刊预出版: 面向高精度的快速三维建模——快速三维建模技术专刊\[2016/8/11\]](#)
[2016年第8期专题已出版: 数据开放与隐私管理\[2016/8/9\]](#)
[2016年第7期专刊已出版: 大数据可用性理论、方法和技术\[2016/7/8\]](#)
[Call for Papers: SecureComm2016\[2016/6/8\]](#)
[2016年第6期专刊已出版: 云计算安全研究\[2016/6/6\]](#)
[征文通知: RE'16国际需求工程会议系列研讨会\[2016/6/6\]](#)
[5月19日在线出版综述论文: 大数据可用性的研究综述\(李建中,王宏志\)\[2016/5/19\]](#)

文献搜索技巧

• 搜索工具

• 中文工具

。 学术刊物：计算机学报、软件学报

The screenshot shows the homepage of the Journal of Software. On the left, there's a sidebar with links for '访问次数: 20938941', '在线出版', '各期目录', '纸质出版', '分辑系列', '论文检索', '论文排行', '综述文章' (which is highlighted in blue), '专刊文章', '美文分享', '各期封面', and 'E-mail订阅'. The main content area features the journal's logo 'Journal of Software' and several navigation buttons: '投稿指南', '问题解答', '下载区', '收费标准', and '在线投稿'. Below these buttons, a banner states '这里所列的文章是所有综述文章, 供读者方便阅读。' followed by a list of review articles from 2016, each with its title and page range. The titles include: '邱天宇, 申富饶, 赵金熙. 自组织增量学习神经网络综述. 2016, 27(9): 2230-2247', '罗东, 刘铁, 钱德沛. 内存计算技术研究综述. 2016, 27(8): 2147-2167', '王鹤澎, 王宏志, 李佳宁, 孙欣欣, 李建中, 高宏. 面向新型处理器的数据密集型计算. 2016, 27(8): 2048-2067', '吴垚, 曾菊儒, 彭辉, 陈红, 李翠平. 群智感知激励机制研究综述. 2016, 27(8): 2025-2047', '廖湘科, 李姗姗, 董威, 贾周阳, 刘晓东, 周书林. 大规模软件系统日志研究综述. 2016, 27(8): 1934-1947', '于洋, 王之梁, 毕军, 施新刚, 尹雷. 软件定义网络中北向接口语言综述. 2016, 27(4): 993-1008', '沈国华, 黄志球, 谢冰, 朱羿全, 廖莉莉, 王飞, 刘银陵. 软件可信评估研究综述: 标准、模型与工具. 2016, 27(4): 955-968', '林伟伟, 吴文泰. 面向云计算环境的能耗测量和管理方法. 2016, 27(4): 1026-1041', '周维, 周可人, 余钟治, 姚绍文, 钱德沛. 基于共享内存的多核时代数据结构研究. 2016, 27(4): 1009-1025', and '王蒙蒙, 刘建伟, 陈杰, 毛剑, 毛可飞. 软件定义网络: 安全模型、机制及研究进展. 2016, 27(4): 969-992'. At the top right, there's a link to '微信服务介'.

文献搜索技巧

- 搜索工具
- 中文工具

- 学术刊物：计算机学报、软件学报

① www.jos.org.cn/ch/reader/view_abstract.aspx?file_no=5068 ☆

访问次数: 2093894

- 在线出版
- 各期目录
- 纸质出版
- 分辑系列
- 论文检索
- 论文排行
- 综述文章
- 专刊文章
- 美文分享
- 各期封面
- E-mail订阅
- RSS

[旧版入口](#)


软件学报 Jou

[投稿指南](#) [问题解答](#) [下载区](#) [收费标准](#) [在线投稿](#)

邵天宇, 申富饶, 赵金熙. 自组织增量学习神经网络综述. 软件学报, 2016, 27(9):2230-2247

自组织增量学习神经网络综述

Review of Self-Organizing Incremental Neural Network

投稿时间: 2015-11-18 最后修改时间: 2016-01-25

DOI: [10.13328/j.cnki.jos.005068](https://doi.org/10.13328/j.cnki.jos.005068)

中文关键词: 神经网络 自组织 竞争学习 增量学习

英文关键词: [neural network](#) [self-organizing](#) [competitive learning](#) [incremental learning](#)

基金项目: 国家自然科学基金(61375064, 61373001); 江苏省自然科学基金(BK20131279)

作者	单位	E-mail
邵天宇	计算机软件新技术国家重点实验室(南京大学), 江苏南京210023;南京大学计算机科学与技术系, 江苏南京210023	
申富饶	计算机软件新技术国家重点实验室(南京大学), 江苏南京210023;南京大学计算机科学与技术系, 江苏南京210023	frshen@nju.edu.cn
赵金熙	计算机软件新技术国家重点实验室(南京大学), 江苏南京210023;南京大学计算机科学与技术系, 江苏南京210023	

摘要点击次数: 656

全文下载次数: 503

文献搜索技巧

- 搜索工具
 - 中文工具
 - 搜索平台：中国知网、万方数据、等等

The screenshot shows the homepage of the Beihang University Library website. At the top, there is a blue header bar with links for '咨询我们' (Consultation), '用户名' (Username), '登录' (Login), and '统一认证' (Unified Authentication). Below the header, the Beihang University logo and name are displayed. The main navigation menu includes '首页' (Home), '关于本馆' (About the Library), '文献资源' (Literature Resources), '服务指南' (Service Guide), and '读者留言' (Reader Comments). A red rectangular box highlights the '文献资源' link. On the left side, there is a sidebar with links for '馆藏检索' (Collection Search), '中文发现' (Chinese Discovery), '外文发现' (Foreign Discovery), '百链云服务' (Baidu NetCloud Service), and '机构库' (Institutional Repository). The central part of the page features a search bar with the placeholder '在此输入关键词' (Enter keyword here) and a search button. Below the search bar, there is a note: '使用说明：图书馆书目检索系统，可查找图书馆馆藏的印刷版文献资源' (Usage instructions: Library catalog search system, can find printed version of library collection resources). To the right, there is a '快速链接' (Quick Link) section with icons and links for various services: '查收查引' (Check收Check引), '常用工具' (Common Tools), '读者培训' (Reader Training), '馆际互借' (Interlibrary Loan), '开馆时间' (Opening Hours), '科技查新' (Technology Information Search), '论文提交' (Paper Submission), '论文检测' (Paper Detection), '图书荐购' (Book Recommendation), '图书借阅' (Book Borrowing), '校外访问' (校外访问), and '移动服务' (Mobile Services). At the bottom, there are sections for '新闻动态' (News Dynamic) and '资源动态' (Resource Dynamic), along with a list of recent news items.

新闻动态 资源动态 更多>>

北航图书馆 欢迎你

- 魏志敏副校长带队赴图书馆进行安全防... 2016-09-21
- 图书馆国庆放假通知 2016-09-20
- 人文社科类外文文献优惠活动 2016-09-18
- 汤森路透WOS在线大讲堂2016年秋季课... 2016-09-13
- 图书馆中秋放假通知 2016-09-13
- 图书馆迎新季开设“新生图书专题书架... 2016-09-02

数字资源

- 数据库列表 试用数据库
- 期刊导航 北航SCI
- 北航机构库 开放获取
- 其它资源 学术站占

文献搜索技巧

- 搜索工具
 - 中文工具
 - 搜索平台：中国知网、万方数据、等等

The screenshot shows the homepage of the Beihang University Library website. At the top, there is a blue header bar with the library's logo, a search bar, and navigation links for '咨询我们' (Consultation), '用户名' (Username), '登录' (Login), and '统一认证' (Unified Authentication). Below the header, the main content area features the university's name '北京航空航天大学' (Beihang University) and its library logo. A prominent yellow button labeled '馆藏检索' (Collection Search) is on the left. In the center, there is a search form with fields for '题名' (Title) and '在此输入关键词' (Enter Keyword), and a large yellow search button. To the right of the search form, a dropdown menu for '文献资源' (Literature Resources) is open, showing options like '数据库列表' (Database List), '馆藏目录' (Collection Catalog), '分馆资源' (Branch Resources), and '馆际互借' (Interlibrary Loan). On the far right, there is a '快速链接' (Quick Links) section with icons and links for various services such as '查收查引' (Citation Monitoring), '常用工具' (常用 Tools), '读者培训' (Reader Training), '馆际互借' (Interlibrary Loan), '开馆时间' (Opening Hours), '科技查新' (Technology Information Search), '论文提交' (Paper Submission), '论文检测' (Paper Detection), '图书荐购' (Book Recommendation), '图书借阅' (Book Borrowing), '校外访问' (校外 Access), and '移动服务' (Mobile Services). At the bottom, there are sections for '新闻动态' (News Dynamic) and '资源动态' (Resource Dynamic), along with a list of recent news items.

新闻动态 资源动态 更多>>

更多>>

北航图书馆 欢迎你

- 魏志敏副校长带队赴图书馆进行安全防... 2016-09-21
- 图书馆国庆放假通知 2016-09-20
- 人文社科类外文文献优惠活动 2016-09-18
- 汤森路透WOS在线大讲堂2016年秋季课... 2016-09-13
- 图书馆中秋放假通知 2016-09-13
- 图书馆迎新季开设“新生图书专题书架... 2016-09-02

文献搜索技巧

● 搜索工具

● 中文工具

。 搜索平台: **中国知网**、万方数据、等等

The screenshot shows a library database search interface. At the top, there's a navigation bar with links for '首页' (Home), '数据库列表' (Database List), '快速定位数据库' (Quick Database Location), and '快速检索' (Quick Search). Below the navigation is a search bar with fields for '用户名' (Username) and '密码' (Password), and buttons for '登录' (Login) and '统一认证' (Unified Authentication). A large blue header banner features a photograph of a library interior. The main content area is titled '数据库列表' (Database List) and contains a table with two columns: '外文数据库' (Foreign Language Databases) and '中文数据库' (Chinese Databases). The table lists 30 databases, each with a small thumbnail image, the database name, and its type in brackets. Some names are partially redacted with a red box.

序号	外文数据库	序号	中文数据库
1	ACM [全文]	1	51CTO学院 [多媒体][试用中]
2	ACS [全文]	2	CSSC中文社会学引文索引 [文摘]
3	AIAA Electronic Library [全文]	3	KUKE数字音乐图书馆 [多媒体]
4	APS [全文]	4	阿帕比电子图书 [电子图书]
5	ASME [全文]	5	百链云图书馆 [全文]
6	ASP [全文]	6	北航博硕士论文 [文摘]
7	ASTM [全文]	7	超星电子图书 [电子图书]
8	Begell [全文,电子图书]	8	读秀学术搜索 [全文]
9	Cambridge [全文]	9	国家军用标准全文数据库 [全文]
10	CRC-Mechanical ENGINEERINGnetBASE [全文]	10	国研报告 [全文]
11	EBSCOhost(ASP/BSR/ERIC...) [全文]	11	龙源人文电子期刊 [全文]
12	Elsevier ScienceDirect [全文]	12	律商网 [全文]
13	Emerald [全文]	13	人大复印报刊资料 [全文]
14	EV2(Ei Compendex) [文摘]	14	锐思金融数据库 [全文]
15	Fortune [全文]	15	商劳印书馆精品工具书 [全文]
16	Frontiers in China [全文]	16	书生之家电子图书 [电子图书]
17	HeinOnline [全文]	17	万方视频公开课服务系统 [多媒体][试用中]
18	IEEE/EE Electronic Library [全文]	18	万方数据-期刊/会议/学位论文 [全文]
19	IMechE [全文]	19	网上报告厅 [多媒体]
20	Intel Technology Journal [全文]	20	维普考试资源库 [全文]
21	IOP [全文]	21	维普中文科技期刊数据库 [全文]
22	IOP ebook [电子图书,全文]	22	新东方多媒体学习库 [多媒体]
23	JournalStorage [全文]	23	雅乐经典高清影院 [多媒体]
24	LexisNexis[全文]	24	英语学练在线服务平台 [多媒体,全文][试用中]
25	Maney Publishing Online Journals [全文]	25	正保远程教育多媒体资源库 [多媒体][试用中]
26	MathSciNet [全文]	26	知迅视界 [多媒体]
27	MIT [全文]	27	中国航空行业标准全文数据库 [全文]
28	NASA/AD/PB/DE [全文]	28	中国航天标准数据库 [全文]
29	Nature [全文]	29	中国知网 [全文]
30	Nature Materials [全文]	30	[部分条目被红框遮挡]

文献搜索技巧

- 搜索工具
 - 中文工具
 - 搜索平台：中国知网、万方数据、等等

① lib.buaa.edu.cn/resourceinfo?id=71

The screenshot shows the Beijing Jiaotong University Library website. At the top, there is a blue header bar with a search icon, a star icon, and a link to 'lib.buaa.edu.cn/resourceinfo?id=71'. Below the header are navigation links for '咨询我们' (Consultation), '用户名' (Username), '密码' (Password), '登录' (Login), and '统一认证' (Unified Authentication). The main content area features the university's logo and several photographs of library interiors. A blue banner at the bottom left reads '数据库详细信息' (Database Detailed Information) and includes a breadcrumb trail: '首页 > 数据库列表 > 中国知网'. The right side of the banner has a red background. Below this, the text '中国知网' (CNKI) is displayed in red, followed by the text '更新时间：2016-04-07 11:28:23'. A table below contains various details about the database, with some entries highlighted by red boxes and red lines.

其他名称：	CNKI,知网,大百科,工具书
您的当前IP：	您的IP地址是222.129.37.249，属于[校外地址]
校内访问入口：	中国知网 [本地镜像] CNKI工具书[本地镜像] CNKI大百科全书[本地镜像]
学科类型：	人文 经济 法政 理工
资源类型：	全文
咨询方式：	
备注：	

文献搜索技巧

- 搜索工具
 - 外文工具
 - 科普网站： MIT Technology Review 、 ACM/IEEE官网
 - 搜索平台： Google Scholar、 DBLP、 等等

文献搜索技巧

- 搜索工具

- 外文工具

- 搜索平台: **Google Scholar**、DBLP、等等

<http://dir.scmor.com/google/>

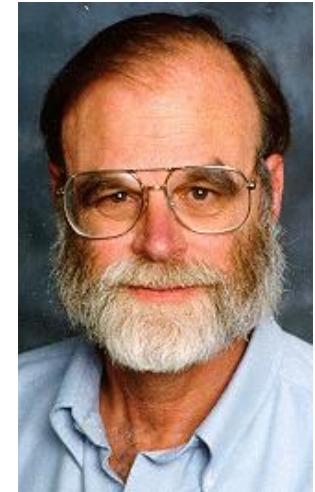
The screenshot shows a web browser window with the URL dir.scmor.com/google/ in the address bar. The page features the classic Google logo with the Chinese characters "思谋导航" underneath. Below the logo, there are two main sections: "学术搜索" (Academic Search) and "网页搜索" (Web Search). Both sections list eight items, each with a speed value and a "现在访问" (Visit Now) link. At the bottom of each section, there is a specific URL for the US and Hong Kong versions of the service.

类别	内容	速度	操作
学术搜索	学术镜像1:	speed:0.04s.	现在访问
	学术镜像2:	speed:0.49s.	现在访问
	学术镜像3:	speed:0.67s.	现在访问
	学术镜像4:	speed:1.57s.	现在访问
	学术镜像5:	speed:1.62s.	现在访问
	学术镜像6:	speed:0.23s.	现在访问
	学术镜像7:	speed:0.04s.	现在访问
	学术镜像8:	speed:0.39s.	现在访问
美国官网(US): scholar.google.com			
香港官网(HK): scholar.google.com.hk			
网页搜索	网页镜像1:	speed:3.99s.	现在访问
	网页镜像2:	speed:0.04s.	现在访问
	网页镜像3:	speed:0.43s.	现在访问
	网页镜像4:	speed:1.48s.	现在访问
	网页镜像5:	speed:0.19s.	现在访问
	网页镜像6:	speed:1.23s.	现在访问
	网页镜像7:	speed:1.47s.	现在访问
	网页镜像8:	speed:0.04s.	现在访问
美国官网(US): www.google.com			
香港官网(HK): www.google.com.hk			

Jim Gray回顾

- James Nicholas "Jim" Gray

- 由于在数据库系统于事务处理领域的卓越贡献，1998年获得图灵奖。
- 微软公司为其建造Redmond研究院，1995年加入微软。
- 2007年，他独自航向法拉伦岛，打算撒散母亲的骨灰，1月28日，他的船失踪。2012年，他在法律意义上被认定已经死亡。



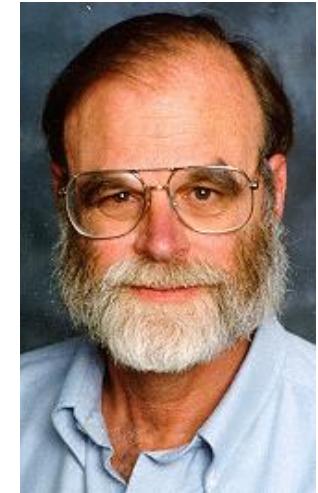
Jim Gray



Jim Gray. The Transaction Concept: Virtues and Limitations.
In VLDB, Pages 144-154, 1981 . (CCF A类会议)

文献搜索技巧

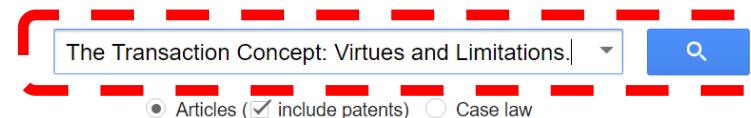
- 搜索工具
 - 外文工具
 - 搜索平台: Google Scholar、DBLP、等等



Jim Gray. The Transaction Concept: Virtues and Limitations. In VLDB, Pages 144-154, 1981.

Jim Gray

A screenshot of a web browser showing the Google Scholar search results for the paper "The Transaction Concept: Virtues and Limitations". The URL in the address bar is <https://scholar.google.com>. The search bar contains the query "The Transaction Concept: Virtues and Limitations.". Below the search bar, there are two radio buttons: "Articles" (selected) and "Case law". At the bottom of the page, the tagline "Stand on the shoulders of giants" is visible.

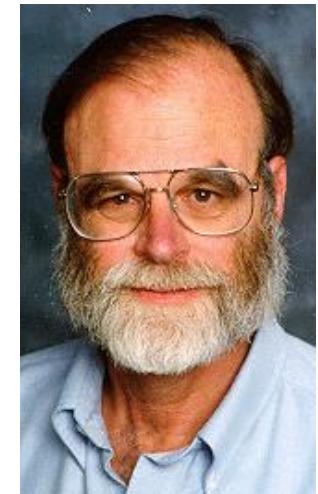


文献搜索技巧

- 搜索工具
 - 外文工具

- 搜索平台: **Google Scholar**、DBLP、等等

Jim Gray. The Transaction Concept: Virtues and Limitations. In VLDB, Pages 144-154, 1981.



Jim Gray

The screenshot shows a Google Scholar search results page. The search query is "The Transaction Concept: Virtues and Limitations". The top result is a paper by Jim Gray from 1981, titled "The Transaction Concept: Virtues and Limitations", published in VLDB. The abstract describes a transaction as a transformation of state with properties of atomicity, durability, and consistency. The page includes filters for "Articles", "Case law", and "My library", and a sidebar for "Any time" search parameters. At the bottom, there are links for "About Google Scholar", "Privacy", "Terms", and "Provide feedback".

文献搜索技巧

- 搜索工具
 - 外文工具
 - 搜索平台: Google Scholar、DBLP、等等

The screenshot shows a Google Scholar search results page. The URL in the address bar is https://scholar.google.com/scholar?cites=15553426371650547409&as_sdt=2005&sciodt=0,5&hl=en. The search term in the search bar is "The transaction concept: Virtues and limitations". The results page displays several academic papers, with the first result being a book by PA Bernstein, V Hadzilacos, and N Goodman from 1987. The interface includes filters for citation counts, publication years, and document types, along with options to sort results and create alerts.

https://scholar.google.com/scholar?cites=15553426371650547409&as_sdt=2005&sciodt=0,5&hl=en

Web Images More... yongxintong@gmail.com

Google

Scholar About 1,080 results (0.06 sec) My Citations

All citations Articles Case law My library

Any time Since 2016 Since 2015 Since 2012 Custom range...

Sort by relevance Sort by date

include citations Create alert

The transaction concept: Virtues and limitations

Search within citing articles

PA Bernstein, V Hadzilacos, N Goodman - 1987 - [osti.gov](#)

This book is an introduction to the design and implementation of concurrency control and recovery mechanisms for transaction management in centralized and distributed database systems. Concurrency control and recovery have become increasingly important as ...

Cited by 5564 Related articles All 10 versions Cite Save More

Principles of distributed database systems

MT Özsu, P Valduriez - 2011 - [books.google.com](#)

This third edition of a classic textbook can be used to teach at the senior undergraduate and graduate levels. The material concentrates on fundamental theories as well as techniques and algorithms. The advent of the Internet and the World Wide Web, and, more recently, ...

Cited by 3460 Related articles All 27 versions Cite Save More

Database management systems

R Ramakrishnan, J Gehrke - 2000 - [202.74.245.22](#)

Page 1. Page 2. CONTENTS PREFACE xxii Part I BASICS 1.1 INTRODUCTION TO DATABASE SYSTEMS 3 1.1 Overview 4 1.2 A Historical Perspective 5 1.3 File Systems versus a DBMS 7 1.4 Advantages of a DBMS 8 1.5 Describing and Storing Data in a DBMS 9 ...

Cited by 2684 Related articles All 12 versions Cite Save More

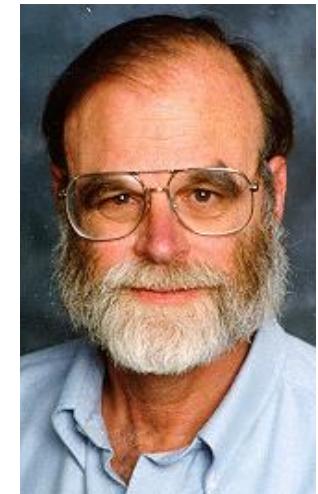
[PDF] centennialcollege.ca [PDF] 202.74.245.22

文献搜索技巧

- 搜索工具
 - 外文工具

- 搜索平台: **Google Scholar**、DBLP、等等

Jim Gray. The Transaction Concept: Virtues and Limitations. In VLDB, Pages 144-154, 1981.



Jim Gray

Scholar

My Citations

Articles Case law My library Any time Since 2016 Since 2015 Since 2012 Custom range...

The Transaction Concept: Virtues and Limitations.

[PDF] usc.edu

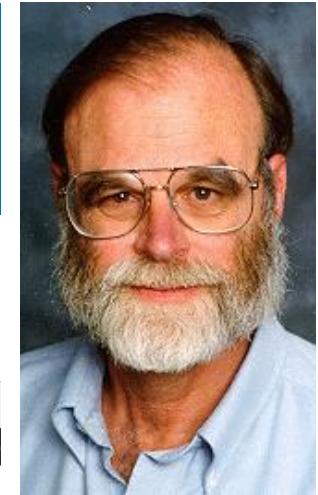
The transaction concept: Virtues and limitations
J Gray - VLDB, 1981 - infolab.usc.edu
ABSTRACT A transaction is a transformation of state which has the properties of atomicity (all or nothing), durability (effects survive failures) and consistency (a correct transformation). The transaction concept is key to the structuring of data management applications. The ...
Cited by 1080 Related articles All 41 versions Cite Save More

About Google Scholar Privacy Terms Provide feedback

文献搜索技巧

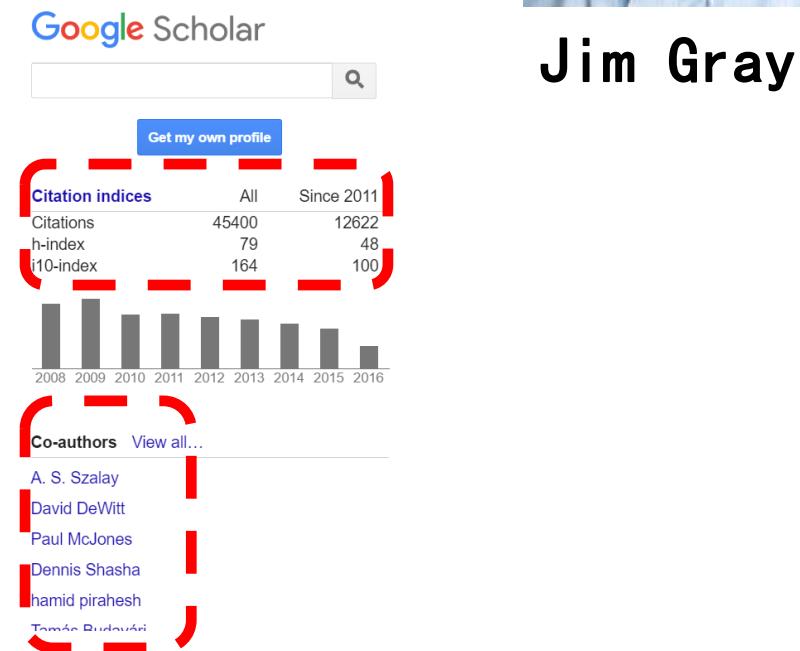
- 搜索工具
- 外文工具
 - 搜索平台: Google Scholar、DBLP、等等

基于“主题”与“作者”的搜索采用
Google Scholar更为适宜，但基于“来
源”的搜索有更好的选择！



Screenshot of a web browser showing Jim Gray's Google Scholar profile page. The URL in the address bar is <https://scholar.google.com/citations?user=Dn4kkYUAAAAJ&hl=en&oi=sra>. The page displays his profile picture, name, affiliation (IBM, Tandem, DEC, Microsoft databases), and a note about no verified email. It also shows his citation statistics: 45400 total citations, an h-index of 79, and an i10-index of 164. A chart shows his citation indices from 2008 to 2016. Below this, a section lists his publications with titles, authors, citation counts, and years. The publications include "Transaction processing" (1993), "Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals" (1997), "The fifth data release of the Sloan Digital Sky Survey" (2007), and "Notes on data base operating systems" (1978).

Title	Cited by	Year
Transaction processing	4822 *	1993
Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals	2810	1997
The fifth data release of the Sloan Digital Sky Survey	2758 *	2007
Notes on data base operating systems	2576	1978



文献搜索技巧

- 搜索工具
 - 外文工具
 - 搜索平台: Google Scholar、**DBLP**、等等

<http://dblp.uni-trier.de/>



Michael Ley

The screenshot shows the homepage of the dblp computer science bibliography. At the top, there is a navigation bar with links for 'home', 'browse', 'search', and 'about'. Below the navigation bar is a search bar with the placeholder 'search dblp'. To the left of the search bar is the dblp logo, which consists of three overlapping geometric shapes (blue, yellow, and green) followed by the text 'dblp' and 'computer science bibliography'. A large red dashed rectangle highlights the search bar area. On the left side of the page, there is a sidebar with several browse categories: 'browse authors | editors' (with letters A-Z), 'browse journals' (with letters A-Z), 'browse conferences | workshops' (with letters A-Z), 'browse series' (listing CoRR, LNCS, CEUR-WS, LNEE, IFIP, LNI, EPTCS, LIPICS, other), and 'browse monographs' (listing books & theses, reference works, edited collections). On the right side, there is a section titled 'About dblp' which provides information about the service, mentioning it is a joint service of the University of Trier and Schloss Dagstuhl, and includes a link to the F.A.Q. There is also a section titled 'dblp statistics' with a list of metrics: # of publications: 3,495,783, # of authors: 1,780,555, # of conferences: 4,901, and # of journals: 1,491.

文献搜索技巧

- 搜索工具
 - 外文工具
 - 搜索平台: Google Scholar、DBLP、等等

① dblp.dagstuhl.de/pers/hd/g/Gray:Jim

[+] Jim Gray ▲ ↴ ↵

> Home > Persons

[+] Person information

- affiliation: Microsoft Research
- award: Turing Award, 1998

[+] Other persons with the same name ⓘ

[+] 2010 – today ⓘ

2010

■ [j49] Andreas Terzis, Razvan Musaloiu-Elefteri, Joshua Cogan, Katalin Szlavecz, Alexander S. Szalay, Jim Gray, Stuart Ozer, Chieh-Jan Mike Liang, Jayant Gupchup, Randal C. Burns: **Wireless sensor networks for soil science.** IJSNet 7(1/2): 53-70 (2010)

[+] 2000 – 2009 ⓘ

2008

■ [j48] Ani R. Thakar, Alexander S. Szalay, George Fekete, Jim Gray: **The Catalog Archive Server Database Management System.** Computing in Science and Engineering 10(1): 30-37 (2008)

■ [j47] Alexander S. Szalay, Ani R. Thakar, Jim Gray: **The sqlLoader Data-Loading Pipeline.** Computing in Science and Engineering 10(1): 38-48 (2008)

■ [j46] Jim Gray: **Distributed Computing Economics.** ACM Queue 6(3): 63-68 (2008)

■ [j45] Jim Gray, Bob Fitzgerald: **Flash Disk Opportunity for Server Applications.** ACM Queue 6(4): 18-23 (2008)

2007

■ [j44] Magdalena Balazinska, Armol Deshpande, Michael J. Franklin, Phillip B. Gibbons, Jim Gray, Mark H. Hansen, Michael Liebhold, Suman Nath, Alexander S. Szalay, Vincent Tao: **Data Management in the Worldwide Sensor Web.** IEEE Pervasive Computing 6(2): 30-40 (2007)

2006

zoomed in on 110 of 170 records

refine by search term

refine by type

- Books and Theses (only)
- Journal Articles (only)
- Conference and Workshop Papers (only)
- Parts in Books or Collections (only)
- Editorship (only)
- Informal Publications (only)

select all | deselect all

refine by coauthor

- Alexander S. Szalay (14)
- Irving L. Traiger (11)
- Bruce G. Lindsay 0001 (10)
- Raymond A. Lorie (9)
- Gianfranco R. Putzolu (9)
- Michael Stonebraker (8)
- Mike W. Blasgen (7)
- Ani Thakar (7)
- Donald R. Slutz (7)
- David J. DeWitt (7)

... more authors

refine by venue

- SIGMOD Conference (18)
- SIGMOD Record (7)
- Commun. ACM (7)
- VLDB (7)

Jim Gray

文献搜索技巧

- 搜索工具
 - 外文工具

。 搜索平台: Google Scholar、DBLP、等等

① dblp.dagstuhl.de/pers/hd/g/Gray:Jim

[+] Jim Gray ↓ ↕ 🔍

[Home](#) > Persons

🕒 by year ☰ Dagstuhl

[−] Person information

- affiliation: Microsoft Research
- award: Turing Award, 1998

[+] Other persons with the same name ?

[−] 2010 – today ?

2010

■ [j49] grid down right left Andreas Terzis, Razvan Musaloiu-Elefteri, Joshua Cogan, Katalin Szlavecz, Alexander S. Szalay, Jim Gray, Stuart Ozer, Chieh-Jan Mike Liang, Jayant Guchupur, Randal C. Burns:
Wireless sensor networks for soil science. IJSNet 7(1/2): 53-70 (2010)

[−] 2000 – 2009 ?

2008

■ [j48] grid down right left Ani R. Thakar, Alexander S. Szalay, George Fekete, Jim Gray:
The Catalog Archive Server Database Management System. Computing in Science and Engineering 10(1): 30-37 (2008)

■ [j47] grid down right left Alexander S. Szalay, Ani R. Thakar, Jim Gray:
The sqlLoader Data-Loading Pipeline. Computing in Science and Engineering 10(1): 38-48 (2008)

2007

■ [j44] grid down right left Magdalena Balazinska, Armol Deshpande, Michael J. Franklin, Phillip B. Gibbons, Jim Gray, Mark H. Hansen, Michael Liebhold, Suman Nath, Alexander S. Szalay, Vincent Tao:
Data Management in the Worldwide Sensor Web. IEEE Pervasive Computing 6(2): 30-40 (2007)

2006

■ [j43] grid down right left Gordon Bell, Jim Gray, Alexander S. Szalay:
Petascale Computational Systems. IEEE Computer 39(1): 110-112 (2006)

■ [c56] grid down right left Stuart Ozer, Jim Gray, Alexander S. Szalay, Andreas Terzis, Razvan Musaloiu-Elefteri, Katalin Szlavecz, Randal C. Burns, Joshua Cogan:

[−] Refine list ?

zoomed in on 14 of 170 records

refine by search term

refine by type

- Books and Theses (only)
- Journal Articles (only)
- Conference and Workshop Papers (only)
- Parts in Books or Collections (only)
- Editorship (only)
- Informal Publications (only)

[select all](#) | [deselect all](#)

refine by coauthor

Alexander S. Szalay (14) ✓

Ani R. Thakar (1)

Peter Z. Kunszt (4)

Donald R. Slutz (2)

Katalin Szlavecz (2)

Joshua Cogan (2)

Maria A. Nieto-Santisteban (2)

Razvan Musaloiu-Elefteri (2)

Christopher Stoughton (2)

Randal C. Burns (2)

22 more options

refine by journal

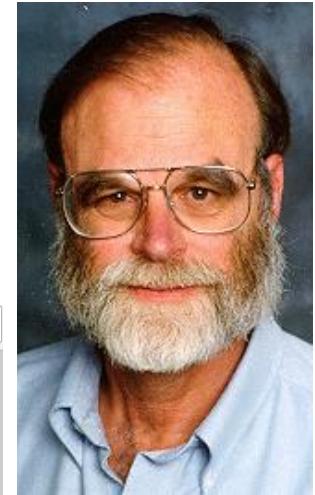
Computing in Science and Engineering (3)

SIGMOD Conference (2)

SIGMOD Record (1)

IEEE Data Eng. Bull. (1)

ACM SIGMOD Record (1)



Jim Gray

文献搜索技巧

- 搜索工具
 - 外文工具
 - 搜索平台: Google Scholar、DBLP、等等

① dblp.dagstuhl.de/pers/hd/g/Gray:Jim

[+] Jim Gray ▲ ↴ ↵

> Home > Persons

[+] Person information

- affiliation: Microsoft Research
- award: Turing Award, 1998

[+] Other persons with the same name ⓘ

[+] 2010 – today ⓘ

2010

■ [j49] Andreas Terzis, Razvan Musaloiu-Elefteri, Joshua Cogan, Katalin Szlavecz, Alexander S. Szalay, Jim Gray, Stuart Ozer, Chieh-Jan Mike Liang, Jayant Gupchup, Randal C. Burns: **Wireless sensor networks for soil science.** IJSNet 7(1/2): 53-70 (2010)

[+] 2000 – 2009 ⓘ

2008

■ [j48] Ani R. Thakar, Alexander S. Szalay, George Fekete, Jim Gray: **The Catalog Archive Server Database Management System.** Computing in Science and Engineering 10(1): 30-37 (2008)

■ [j47] Alexander S. Szalay, Ani R. Thakar, Jim Gray: **The sqlLoader Data-Loading Pipeline.** Computing in Science and Engineering 10(1): 38-48 (2008)

■ [j46] Jim Gray: **Distributed Computing Economics.** ACM Queue 6(3): 63-68 (2008)

■ [j45] Jim Gray, Bob Fitzgerald: **Flash Disk Opportunity for Server Applications.** ACM Queue 6(4): 18-23 (2008)

2007

■ [j44] Magdalena Balazinska, Armol Deshpande, Michael J. Franklin, Phillip B. Gibbons, Jim Gray, Mark H. Hansen, Michael Liebhold, Suman Nath, Alexander S. Szalay, Vincent Tao: **Data Management in the Worldwide Sensor Web.** IEEE Pervasive Computing 6(2): 30-40 (2007)

2006

zoomed in on 110 of 170 records

refine by search term

refine by type

- Books and Theses (only)
- Journal Articles (only)
- Conference and Workshop Papers (only)
- Parts in Books or Collections (only)
- Editorship (only)
- Informal Publications (only)

select all | deselect all

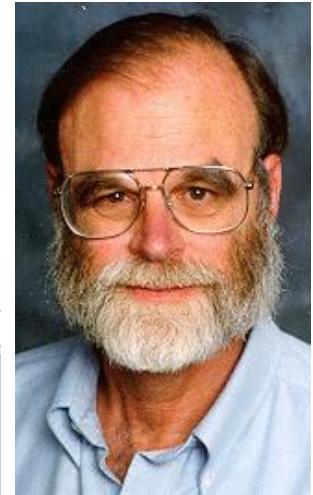
refine by coauthor

- Alexander S. Szalay (14)
- Irving L. Traiger (11)
- Bruce G. Lindsay 0001 (10)
- Raymond A. Lorie (9)
- Gianfranco R. Putzolu (9)
- Michael Stonebraker (8)
- Mike W. Blasgen (7)
- Ani Thakar (7)
- Donald R. Slutz (7)
- David J. DeWitt (7)

160 more options

refine by venue

- SIGMOD Conference (18)
- GMCS (7)
- Commun. ACM (7)
- VLDB (7)



Jim Gray

文献搜索技巧

- 搜索工具
 - 外文工具
 - 搜索平台: Google Scholar、DBLP、等等

dblp.dagstuhl.de/pers/hd/g/Gray:Jim

[+] Jim Gray 🔍 ↴ ↵ ↶ ↷

> Home > Persons

[+] Person information

- affiliation: Microsoft Research
- award: Turing Award, 1998

[+] Other persons with the same name ⓘ

[+] 2010 – today ⓘ

no results

[+] 2000 – 2009 ⓘ

2006

■ [c55] 🔍 ↴ ↵ ⓘ Naga K. Govindaraju, Jim Gray, Ritesh Kumar, Dinesh Manocha:
GPUTeraSort: high performance graphics co-processor sorting for large database management. SIGMOD Conference 2006: 325-336

2004

■ [c51] 🔍 ↴ ↵ ⓘ Jim Gray:
The Next Database Revolution. SIGMOD Conference 2004: 1-4

2003

■ [c48] 🔍 ↴ ↵ ⓘ Jim Gray, Hans-Jörg Schek, Michael Stonebraker, Jeffrey D. Ullman:
The Lowell Report. SIGMOD Conference 2003: 680

2002

■ [c47] 🔍 ↴ ↵ ⓘ Alexander S. Szalay, Jim Gray, Ani Thakar, Peter Z. Kunszt, Tanu Malik, Jordan Raddick, Christopher Stoughton, Jan vandenBerg:
The SDSS skyserver: public access to the Sloan Digital Sky Server data. SIGMOD Conference 2002: 570-581

2000

■ [c43] 🔍 ↴ ↵ ⓘ Tom Barclay, Donald R. Slutz, Jim Gray:
TerraServer: A Spatial Data Warehouse. SIGMOD Conference 2000: 307-318

by year

Dagstuhl

[+] Refine list ⓘ

zoomed in on 18 of 170 records

refine by search term

refine by type

- Books and Theses (only)
- Journal Articles (only)
- Conference and Workshop Papers (only)
- Parts in Books or Collections (only)
- Editorship (only)
- Informal Publications (only)

select all | deselect all

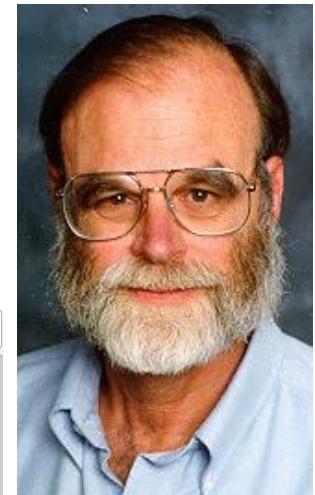
refine by coauthor

- Donald R. Slutz (2)
- Alexander S. Szalay (2)
- Peter Z. Kunszt (2)
- Michael Stonebraker (2)
- Tom Barclay (2)
- Patrick E. O'Neil (2)
- Ani Thakar (2)
- Ritesh Kumar (1)
- Jeffrey D. Ullman (1)
- Tanu Malik (1)

30 more options

refine by venue

SIGMOD Conference (18) ✓



Jim Gray

课程提纲

- 浅谈前沿热点
- 文献搜索技巧
- 随机抽样算法
- 水库抽样算法 (Reservoir Sampling)

随机抽样算法

抽样场景	单一对象抽样 ($ S = 1$)
离线场景 (已知对象总数N)	离线随机抽样
在线场景 (未知对象总数N)	在线随机抽样

随机抽样算法

抽样场景	单一对象抽样 ($ S = 1$)
离线场景 (已知对象总数N)	离线随机抽样
在线场景 (未知对象总数N)	在线随机抽样

离线随机抽样

问题定义：从一个包含**N(N已知)**个元素的集合中**随机选取其中任意一个元素**

- **抽样算法**

- 从1到N中随机选取一个整数；
- 返回这个整数索引的元素；

- **代码实现 (Python)**

```
import random

def random_element( a, N ):
    return a[ int( random.random() * N ) ]
```

其中**random.random()**返回[0, 1]区间随机实数

随机抽样算法

抽样场景	单一对象抽样 ($ S = 1$)
离线场景 (已知对象总数N)	离线随机抽样
在线场景 (未知对象总数N)	在线随机抽样

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 抽样算法(伪代码)

```
Initialize N = 0;  
for each new item  
    N += 1;  
    generate a random integer p from 1 to N;  
    if p<=1  
        element = item;  
return element;
```

在线随机抽样

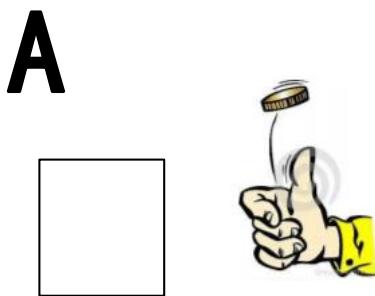
问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 算法实例

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 算法实例



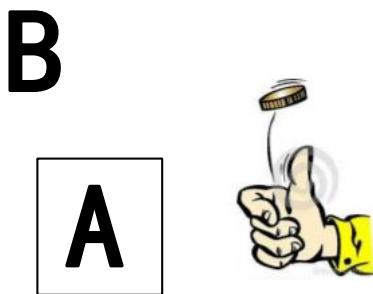
- 数据流:A
- 产生[1, 1]间随机整数p=1

仅当p为1时替换，即 $1/N$ 的概率的事件发生

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 算法实例



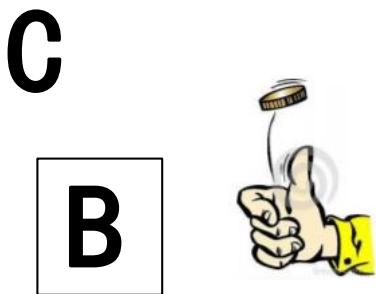
- 数据流:A, B
- 产生[1, 2]间随机整数p=1

仅当p为1时替换，即 $1/N$ 的概率的事件发生

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 算法实例



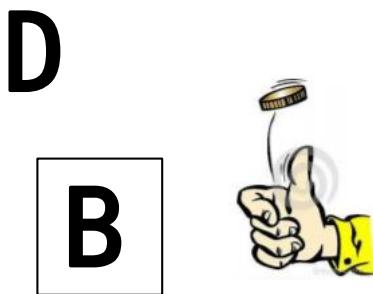
- 数据流:A, B, C
- 产生[1, 3]间随机整数p=3

仅当p为1时替换，即 $1/N$ 的概率的事件发生

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 算法实例

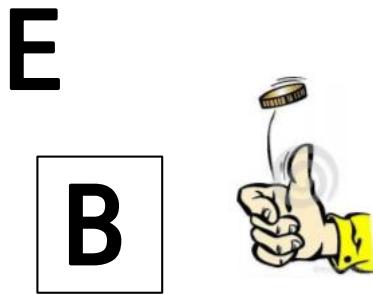


- 数据流:A, B, C, D
- 产生[1, 4]间随机整数p=2

仅当p为1时替换，即 $1/N$ 的概率的事件发生

在线随机抽样

- 算法实例



- 数据流:A, B, C, D, E
- 产生[1, 5]间随机整数p=1

仅当p为1时替换，即 $1/N$ 的概率的事件发生

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 算法正确性证明

- 证明：当前数据流中每项元素均等概率以 $1/N$ 抽到
- 归纳法：
 - 对 $N=1$ ：以1的概率抽到
 - 假设 $N=k$ 成立
 - 当 $N=k+1$ 时：
 - 第 $k+1$ 项以 $1/(k+1)$ 概率被抽到
 - 下讨论第1至 k 项中任一项为最后样本的概率。要为最后样本，首先根据归纳假设，在第 $k+1$ 轮之前以 $1/k$ 的概率为样本，其次第 $k+1$ 轮时以 $1-1/(k+1)$ 的概率没被第 $k+1$ 项替换，综合第 $k+1$ 轮完以 $1/k*k/(k+1)=1/(k+1)$ 概率为最后样本。

在线随机抽样

问题定义：从一个数据流中随机选取当前已知元素中任意一个元素

- 总结

- 上述算法是在线算法 (online algorithm)
- 无需事先已知数据流中的元素总个数
- 无需将数据流中所有元素保存在内存中，只需储存一个元素的内存空间

(单一对象) 随机抽样的扩展

扩展问题：如何从包含 N 个元素的集合中随机选取其中**任意 k 个元素**？

- 上述扩展问题未能描述清楚
- 若选取 k 个元素有放回：
 - 只需要重复上述选随机1个元素的算法 k 次即可。
- 若选取 k 个元素无放回呢？



课程提纲

- 浅谈前沿热点
- 文献搜索技巧
- 随机抽样算法
- 水库抽样算法 (Reservoir Sampling)

随机抽样算法

抽样场景	单一对象抽样 ($ S = 1$)	k对象抽样 ($ S = k$)
离线场景 (已知对象总数N)	离线随机抽样	枚举可能组合， 随后抽样
在线场景 (未知对象总数N)	在线随机抽样	水库抽样

随机抽样算法

抽样场景	单一对象抽样 ($ S = 1$)	k对象抽样 ($ S = k$)
离线场景 (已知对象总数N)	离线随机抽样	枚举可能组合， 随后抽样
在线场景 (未知对象总数N)	在线随机抽样	水库抽样

离线水库抽样

离线水库抽样问题：从包含 N 个元素的集合中随机选取其中不同的 k 个元素，也就是所有 k 大小的子集中随机的一个

- 问题先保留，先学习随机洗牌

洗牌问题：给定 N 元数组，如何随机产生所有可能排列（共 $N!$ 种）的一种！

洗牌算法

洗牌问题：给定N元数组，如何随机产生所有可能排列（共 $N!$ 种）的一种！

- Fisher–Yates algorithm

for $i=1$ to N :

generate a random integer p from 1 to i
swap($a[i]$, $a[p]$)

- $O(N)$ 时间 $O(N)$ 空间

洗牌算法

洗牌问题：给定N元数组，如何随机产生所有可能排列（共 $N!$ 种）的一种！

- 算法实例

当前循环轮数	随机整数范围	产生的随机整数	数组下标
1	1-1	1	123456789
2	1-2	1	213456789
3	1-3	3	213456789
4	1-4	2	243156789
5	1-5	1	543126789
...
9	1-9	7	

离线水库抽样 vs. 洗牌算法

离线水库抽样问题：从包含N个元素的集合中随机选取其中不同的k个元素，也就是所有k大小的子集中随机的一个

洗牌问题：给定N元数组，如何随机产生所有可能排列（共 $N!$ 种）的一种！

- Fisher-Yates algorithm的结果取前k项即可
- 显然任意随机的k大小的排列也对应着随机的k大小的组合

离线水库抽样

离线水库抽样问题：从包含N个元素的集合中随机选取其中不同的k个元素，也就是所有k大小的子集中随机的一个

- 算法描述(伪代码)
- 下述算法前k轮循环没有起作用，因为我们只需要k大小的组合

```
for i=1 to N:  
    generate a random integer p from 1 to i  
    if p<=k:  
        swap(a[i], a[p])
```

离线水库抽样

离线水库抽样问题：从包含N个元素的集合中随机选取其中不同的k个元素，也就是所有k大小的子集中随机的一个

- 算法实例 ($k=4$)

当前循环轮数	随机整数范围	产生的随机整数	数组下标	结果
1	1-1	1	123456789	1234
2	1-2	1	213456789	1234
3	1-3	3	213456789	1234
4	1-4	2	243156789	1234
5	1-5	1	543126789	1345
...	
9	1-9	7		

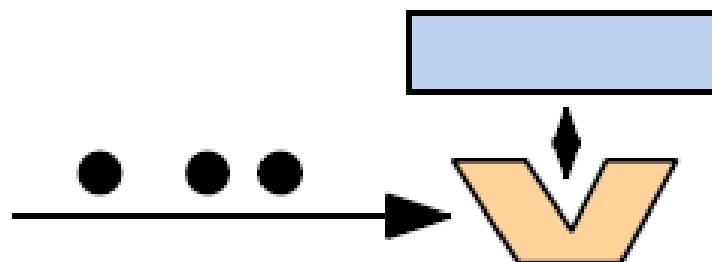
在线水库抽样

- 问题：如何从包含 N 个元素的集合中随机选取其中不同的 k 个元素，也就是所有 k 大小的子集中随机的一个
- 上述讨论是已知集合大小为 N 的情况，若是考虑在线(online)的情况，即 N 未知呢？



在线水库抽样

- 重新叙述问题
- 有一个数据流大小为 N , N 非常大
- 集合元素高速出现，一个一个来，每个元素只能处理一次
- 只有有限的大小为 k 的内存
- 需要始终在内存保留一个样本，样本为当前已看到的数据集的随机样本



在线水库抽样

- 问题：面对在线的数据流，在任意时刻截止时，在有限内存（大小为 k ）下，保持着当前已有 N 个元素的均匀随机样本
- 上述问题又称为水库抽样问题 (Reservoir Sampling)
- 包含一类问题
- 带权水库抽样
- 分布式水库抽样
- ...

在线水库抽样

- 问题：面对在线的数据流，在任意时刻截止时，在有限内存（大小为k）下，保持着当前已有N个元素的均匀随机样本
- 算法：舍弃前k轮，并没有起作用。N为当前轮数，element为当前处理元素

```
if N<=k:  
    mem[N]=element  
else:  
    generate a random integer p from 1 to N  
    if p<=k:  
        mem[p]=element  
    N+=1
```

在线水库抽样

- 问题：面对在线的数据流，在任意时刻截止时，在有限内存（大小为 k ）下，保持着当前已有 N 个元素的均匀随机样本
- 例子：
 - $k=3$, 数据流为 {A, B, C, D, E...}
 - 最初的 $k=3$ 项被装进水库，水库为 {A, B, C}
 - 给第四项D, 产生一个1至4的随机整数，如4，因为 $4>k=3$ ，所以舍弃D，水库不变
 - 第五项E来时，产生1至5的随机整数，如2，因为 $2< k=3$ ，所以把水库第2项替换为E，水库为 {A, E, C}

在线水库抽样

- 问题：面对在线的数据流，在任意时刻截止时，在有限内存（大小为k）下，保持着当前已有N个元素的均匀随机样本
- 上述算法在Knuth的TAOCP中又称为算法R，由Vitter在1985年提出。
- Vitter, Jeffrey S. (1 March 1985). "Random sampling with a reservoir "
- ACM Transactions on Mathematical Software. 11 (1): 37 - 57.

在线水库抽样

- 证明：当前的N个元素每一项均等概率以 k/N 概率抽到（归纳法）
- 对 $N=k$ 时，以1的概率在水库
- 假设 $N=t$ 成立
- 当 $N=t+1$ 时，
 - 第 $t+1$ 项以 $k/(t+1)$ 概率存在于水库
 - 下讨论第1至 t 项中任一项存在于水库的概率。要存在于最后的水库，首先根据归纳假设，在第 $t+1$ 轮之前以 k/t 的概率在水库，其次第 $t+1$ 轮时以 $1-1/(t+1)$ 的概率没被第 $t+1$ 项替换，综合第 $t+1$ 轮完以 $k/t*t/(t+1)=k/(t+1)$ 概率存在于水库。

在线水库抽样

- 证明：当前的N个元素每一项均等概率以 k/N 概率抽到（归纳法）
 - 对 $N=k$ 时，以1的概率在水库
 - 假设 $N=t$ 成立
 - 当 $N=t+1$ 时，
 - 第 $t+1$ 项以 $k/(t+1)$ 概率存在于水库
 - 下讨论第1至 t 项中任一项存在于水库的概率。
要存在于最后的水库，首先根据归纳假设，在第 $t+1$ 轮之前以 k/t 的概率在水库，其次第 $t+1$ 轮时以 $1-1/(t+1)$ 的概率没被第 $t+1$ 项替换，综合第 $t+1$ 轮完以 $k/t*t/(t+1)=k/(t+1)$ 概率存在于水库。

在线水库抽样

- 正确性证明的误区：
- 认为只要证明了在任一轮停止时，当前已知每一个元素存在于内存的概率为 k/N
- 这并不等价于算法可以选取随机 k 大小子集
- 例子： 设当前 N 为4， $k=2$ 。

要求算法能从六个集合中均匀随机选出一个，即 $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}$ 。

然而若算法只是从 $\{1, 2\}, \{3, 4\}$ 中随机选出了一个，每个元素以 $k/N=2/4=1/2$ 的概率存在于水库中的证明也是可以推出的，所以这一证明并不能充分说明我们取得是六个集合中的均匀随机一个。

在线水库抽样

• 水库抽样错误证明

Google reservoir sampling

All Images Videos News Maps More Search tools

About 38,000,000 results (0.42 seconds)

Reservoir sampling - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Reservoir_sampling ▾
Reservoir sampling is a family of randomized algorithms for randomly choosing a sample of k items from a list S containing n items, where n is either a very large ...
Relation to Fisher-Yates shuffle - Example implementation - Statistical properties



Gregable: Reservoir Sampling - Sampling from a stream of ...
gregable.com/2007/10/reservoir-sampling.html ▾
Oct 8, 2007 - **Reservoir Sampling** is an algorithm for sampling elements from a stream of data. Imagine you are given a really large stream of data elements ...
You visited this page.



3. Algorithms Every Data Scientist Should Know: Reservoir ...
https://blog.cloudera.com/.../hadoop-stratified-random-sampling-algorithm/ ▾
Apr 23, 2013 - You can find an excellent overview of a set of algorithms for performing **reservoir sampling** in this blog post by Greg Grothaus. I'd like to focus ...



Reservoir Sampling - GeeksforGeeks
www.geeksforgeeks.org/reservoir-sampling/ ▾
Reservoir sampling is a family of randomized algorithms for randomly choosing k samples from a list of n items, where n is either a very large or unknown ...



[PDF] Random Sampling with a Reservoir - Berkeley Database ...
db.cs.berkeley.edu/cs286/papers/reservoirsampling-toms1985.pdf ▾
by JS VITTER - 1985 - Cited by 846 - Related articles
Random Sampling with a Reservoir. JEFFREY SCOTT VITTER. Brown University. We introduce fast algorithms for selecting a random **sample** of n records ...



algorithm - Reservoir sampling - Stack Overflow
stackoverflow.com/questions/2612648/reservoir-sampling ▾
Apr 10, 2010 - I actually did not realize there was a name for this, so I proved and implemented this from scratch: import random def random_subset(iterator, ...



reservoir sampling
https://xlinux.nist.gov/dads/HTML/reservoirSampling.html ▾
Definition of **reservoir sampling**, possibly with links to more information and implementations.



What is an intuitive explanation of reservoir sampling? - Quora
https://www.quora.com/What-is-an-intuitive-explanation-of-reservoir-sampling...
Imagine the following "dating" game show. The contestant, a bachelorette, is seated at a table with an empty chair. The host introduces the first suitor; the ...



Reservoir Sampling | Math ∩ Programming
jeremykun.com/2013/07/05/reservoir-sampling/ ▾
Jul 5, 2013 - Discussion: This is one of many techniques used to solve a problem called **reservoir sampling**. We often encounter data sets that we'd like to ...



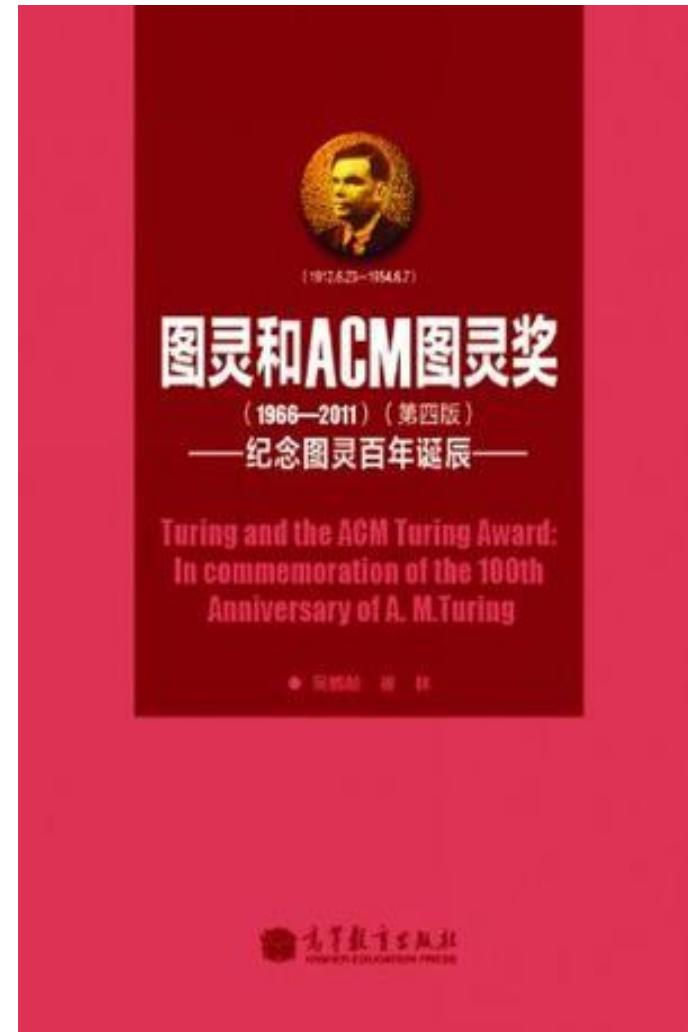
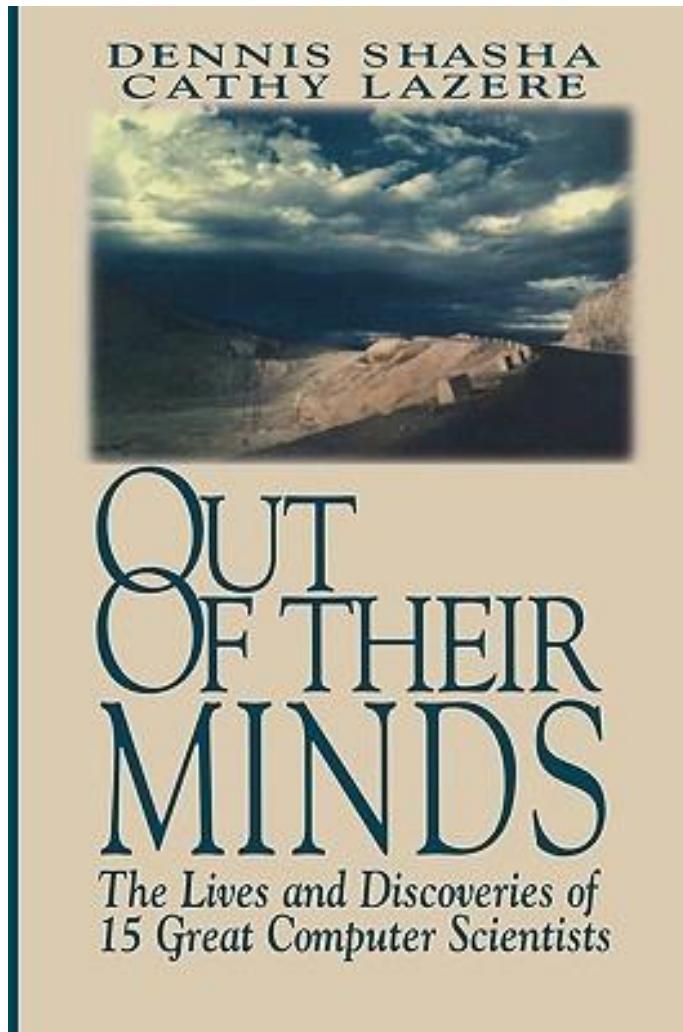
Reservoir Sampling in MapReduce | Had00b
had00b.blogspot.com/2013/07/random-subset-in-mapreduce.html ▾
Aug 5, 2013 - One of the most common sequential approaches to this problem is the so-called **reservoir sampling**. The algorithm works as follows: the data is ...

在线水库抽样

- 水库抽样的应用：
 - 对于存储在硬盘上的数据抽样
 - 路由器对每日流量的ip地址抽样
 - ...

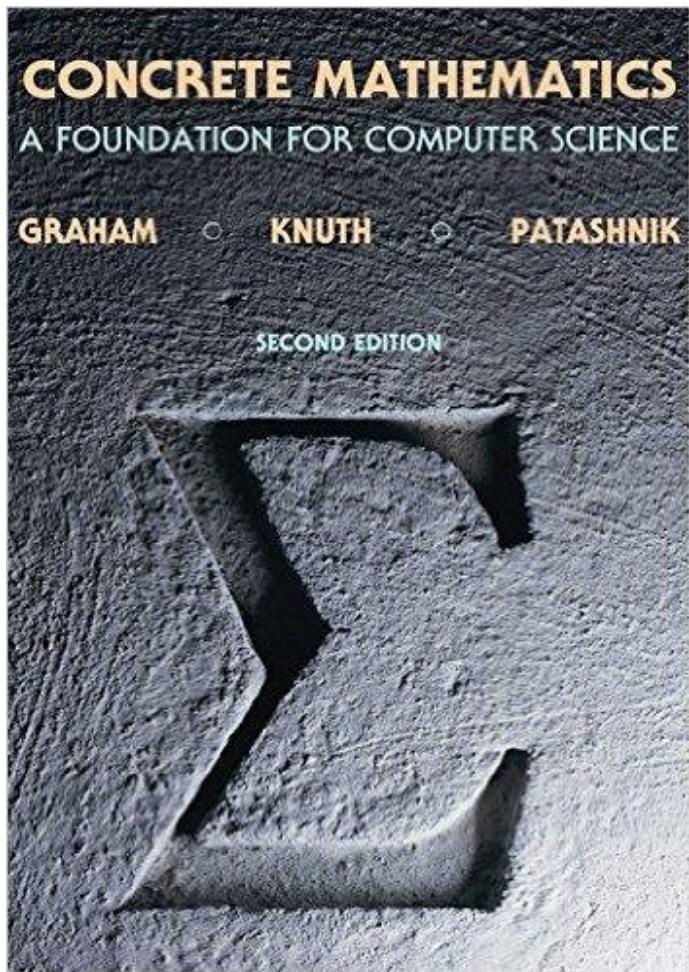
计算机基础书籍推荐

- 计算机发展史（图灵奖传记类）



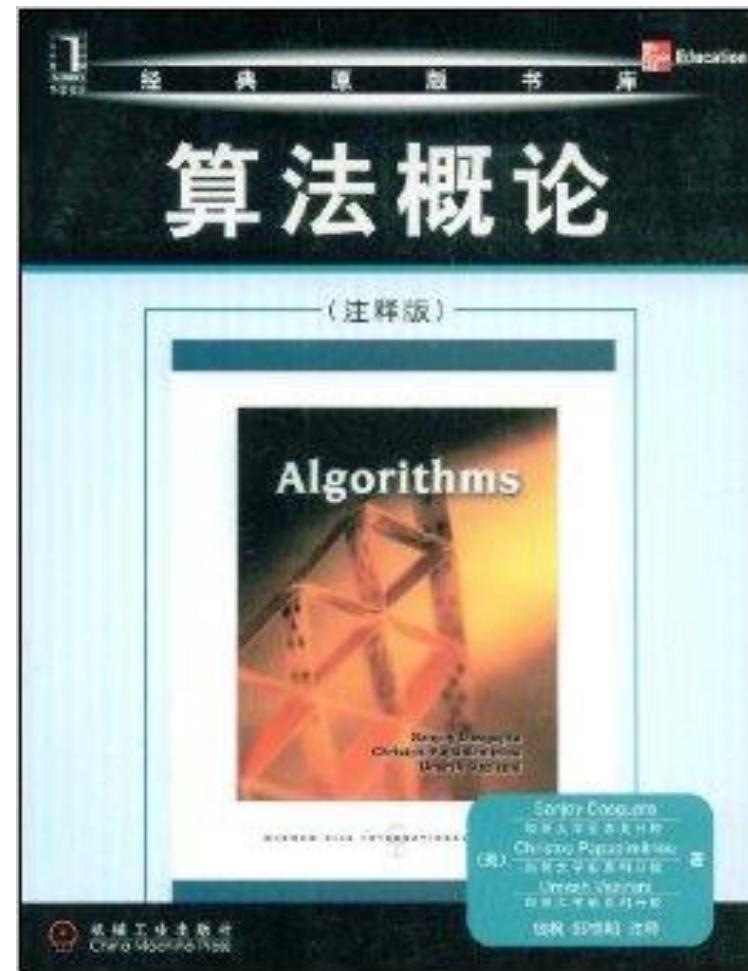
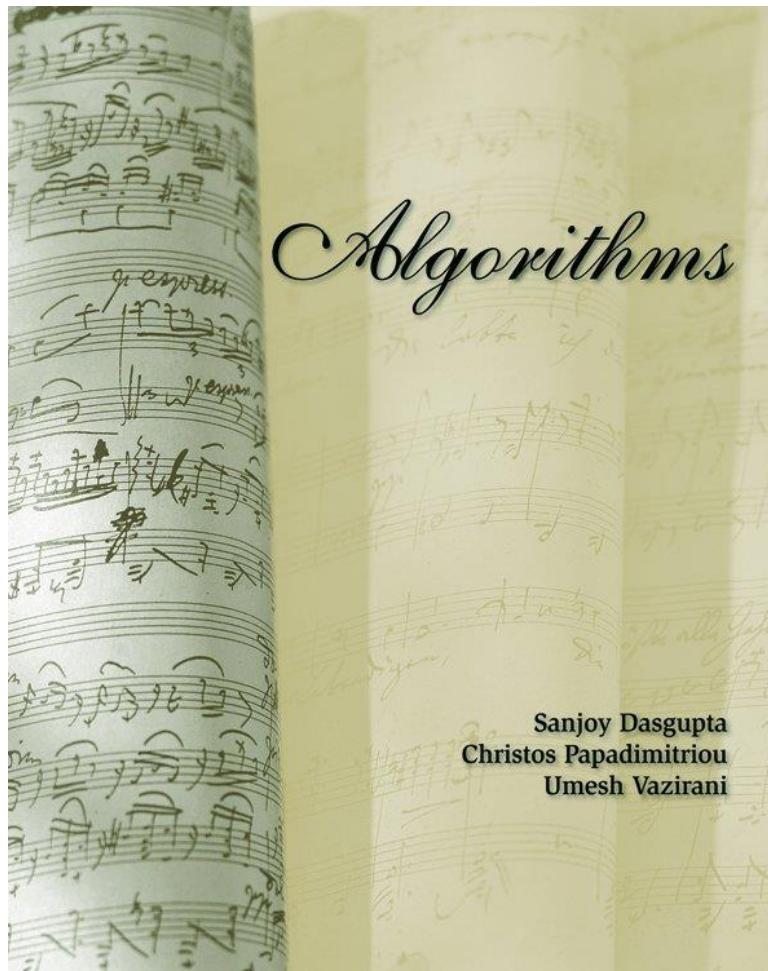
计算机基础书籍推荐

- 数学基础



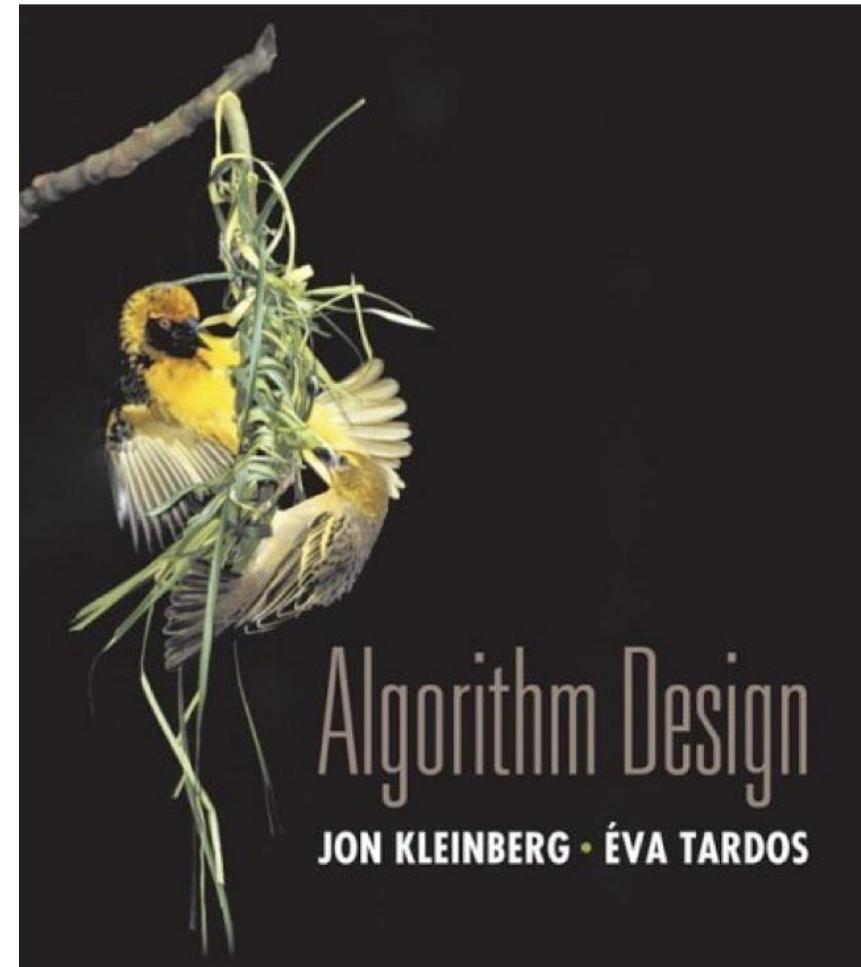
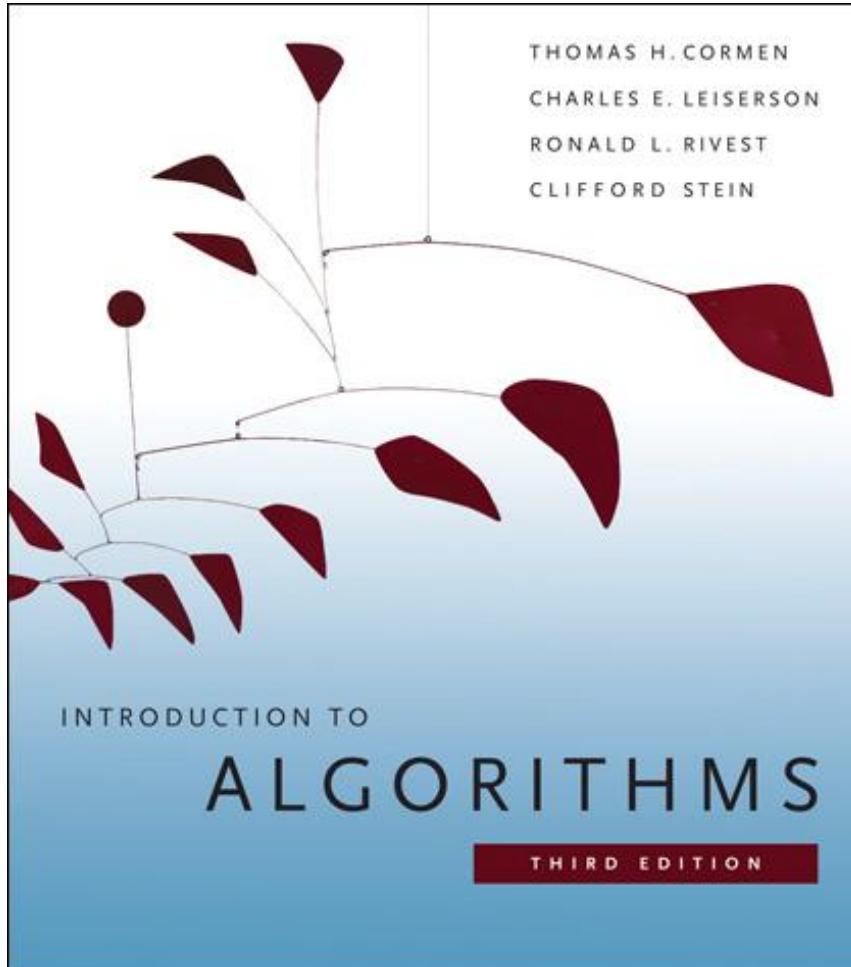
计算机基础书籍推荐

• 算法基础(1)



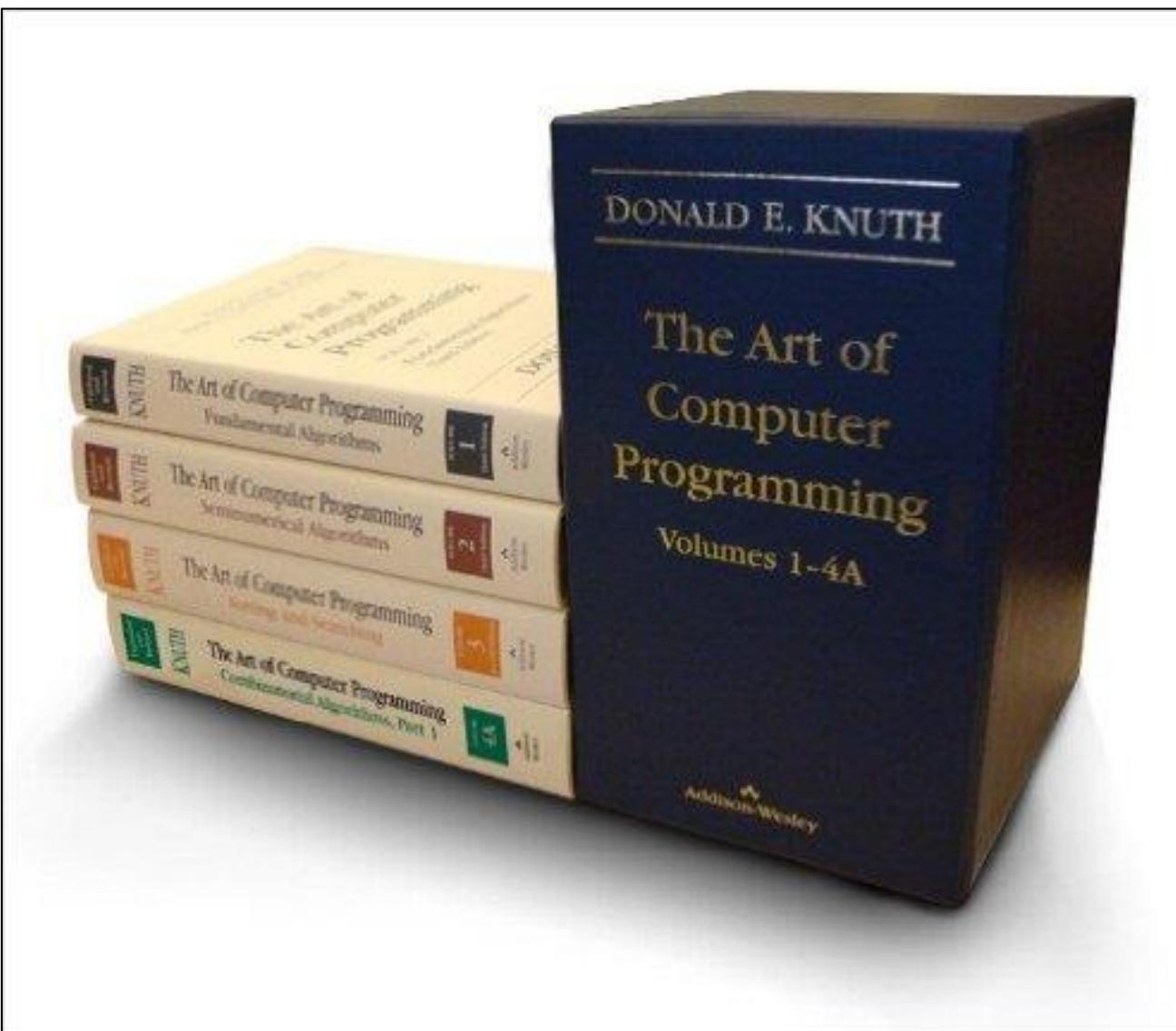
计算机基础书籍推荐

• 算法基础(2)



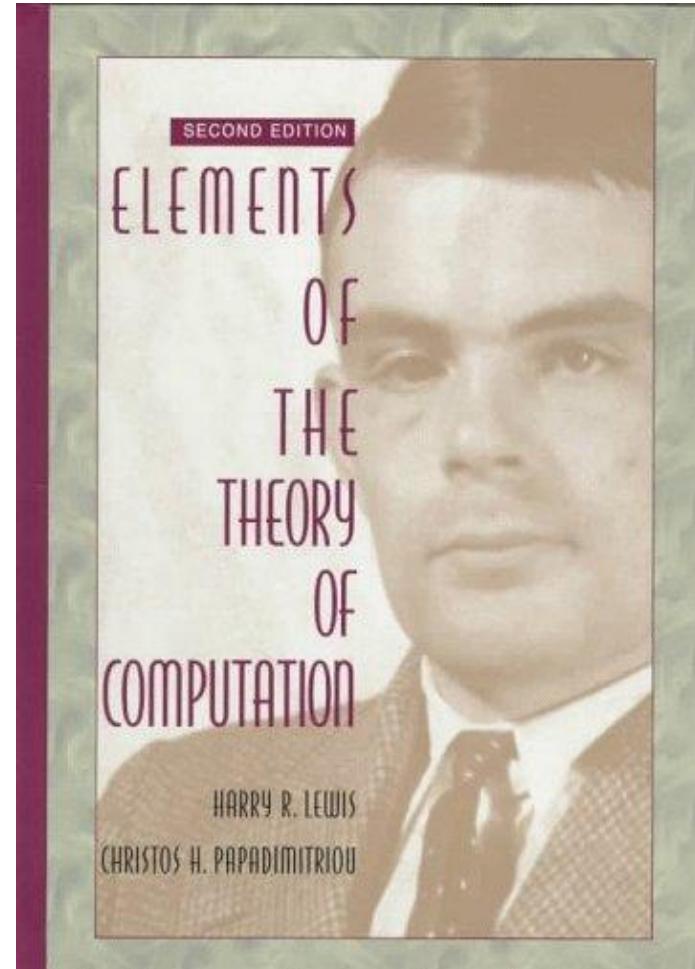
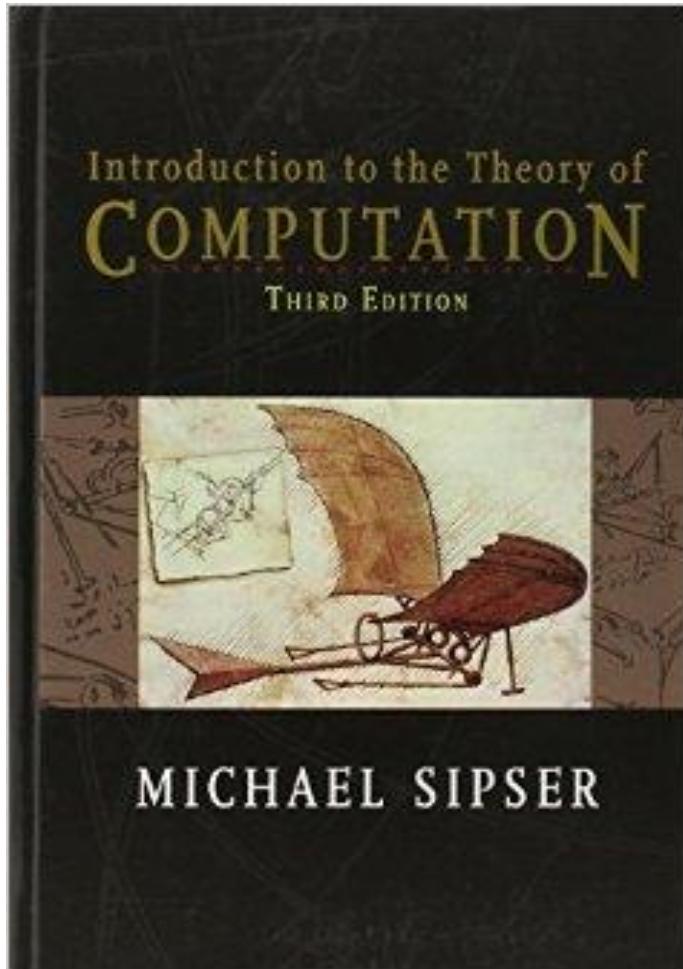
计算机基础书籍推荐

• 算法基础(3)



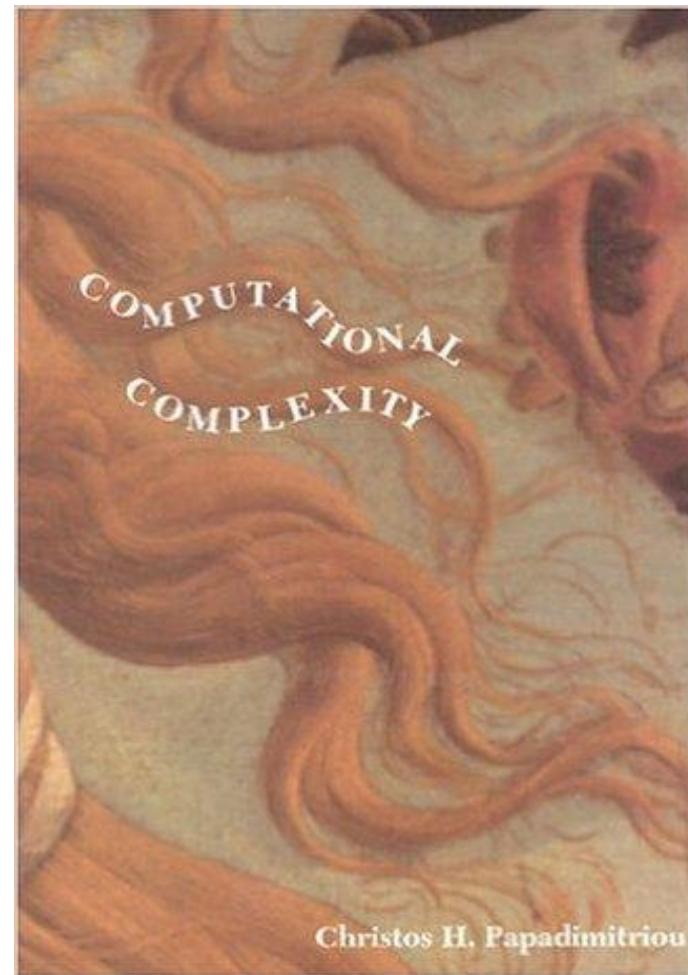
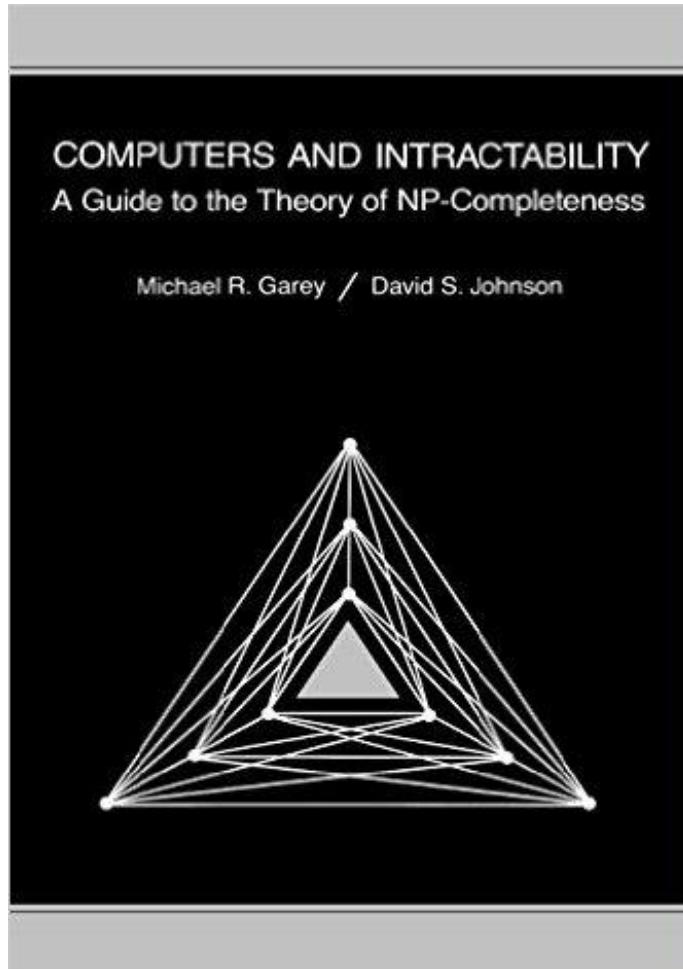
计算机基础书籍推荐

- 计算理论



计算机基础书籍推荐

- 计算复杂性



謝謝