FOUNDATION OF MATCHING IN MASSIVE STRINGS

巨量串匹配基础

高庆狮 高小宇 胡 玥 李 莉 王培凤 著

科学出版社

巨量串匹配基础

Foundation of Matching in Massive Strings

高庆狮 高小宇 胡 玥 李 莉 王培凤 著

科学出版社

北 京

内 容 简 介

本书讨论如何把巨量字符串的串匹配问题自动生成一个优化的完全自动机，以及其简化和有效硬件的实现，进一步讨论模糊化的 U-不确定控制下的巨量字符串和干扰条件下的 V-不确定控制下的巨量字符串的串匹配，以及超长字符串的部分匹配的算法和硬件实现方法。

本书的有关研究工作前后得到国家自然科学基金 GJZRJJ-60873002 的资助，973 课题 2007CB311103 的资助。

本书可作为计算机科学技术相关专业领域研究人员提高理论素质的参考书，也可以作为相关专业研究生学习专业基础的研究资料。

图书在版编目(CIP)数据

巨量串匹配基础/高庆狮等著. —北京：科学出版社，2012

ISBN 978-7-03-033060-4

Ⅰ.①巨… Ⅱ.①高…②高…③胡…④李…⑤王… Ⅲ.①电子计算机-算法理论 Ⅳ.①TP301.6

中国版本图书馆 CIP 数据核字（2011）第 264955 号

责任编辑：王淑兰 / 责任校对：马英菊

责任印制：吕春珉 / 封面设计：耕者设计工作室

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2012 年 1 月第 一 版 开本：850×1168 1/32

2012 年 1 月第一次印刷 印张：2 3/4

字数：80 000

定价 15.00 元

（如有印装质量问题，我社负责调换<新科>）

销售部电话 010-62134988 编辑部电话 010-62130750（VA03）

献给我的启蒙老师们：

母校北京大学数学力学系——

传授美丽简洁的仿射空间的江泽涵老师

在万变中寻找不动点和不变性质的吴光磊老师

用抽象难题严训学生的冷生明老师

严格、精确、绝妙的教学方法的丁石荪老师

感人的热诚和严谨的陈杰老师

讲授神奇宇宙的戴文赛老师

才华横溢的段学复和程民德代课老师

讲授让万物循规法则的钱敏老师

揭示宏观和微观世界奥秘的钱尚武老师

燃烧自己、照亮别人的张世龙老师

百年母校漳州一中——

覃景芬代数老师、陈“妈妈”常玉几何老师、钱学正化学老师、魏德亨音乐老师等老师们

漳州钟芬小学我的数学老师杨校长

永远感谢你们，永远怀念你们，我的启蒙老师们！

献给我有幸能够得到较长时间直接教诲的共和国泰斗们：

中国巨型机指路人，把物理概念融化在数学方程中的“两弹一星”功勋科学家钱学森老师

平时和声细语，战时慷慨激昂怒斥台上“四人帮”的“两弹一星”功勋科学家钱三强老师

充满父辈关爱的“两弹一星”功勋科学家王淦昌老师和“两弹一星”功勋科学家杨嘉墀老师

生活十分简朴，平易近人，深受日本物理界敬重的周培源老师

思维敏捷，精力充沛，永远在思考的华罗庚老师

平易近人，风雨无阻骑车从中关村到端王府上班的“两弹一星”功勋科学家陈芳永老师

八十多岁高龄时仍然热情豪放，热爱专业，身材魁梧，声音宏亮的汪德昭老师

年迈时仍然不知疲倦，白天开全国“人大”会，晚上工作到深夜，日日夜夜抢时间的黄炳维老师

永远感谢你们，永远怀念你们，我的共和国泰斗的老师们！

学生　高庆狮

2011年1月1日

前　言

有一些研究生问我如何找研究课题，下面的轨迹或许是一个答复。

发现 Łukasiewicz 多值命题逻辑理论的错误和缺点是因为 Zadeh先生的同僚喜欢用错误的 Łukasiewicz 多值命题逻辑理论的所谓“$A \vee \neg A \neq$真，和 $A \wedge \neg A \neq$假”来解释“$A \cup \neg A \neq \Omega$（全集）和 $A \cap \neg A \neq \varnothing$（空集）”。而 Zadeh 模糊集合论缺点和错误是在给学外语专业跨大学科研究生开一门指导学习课程——“语言学进展”是阅读模糊语言学时偶然发现的。从 20 世纪 80 年代在中科院计算技术研究所，90 年代在北京科技大学，到 21 世纪在大连理工大学招外语专业跨大学科研究生是因为发现机器翻译巨大经济效益和社会效益的前提是没有语无伦次、没有正错混杂，合乎语言表示规律。其关键是必须从自然科学的角度来研究语言学。这就需要培养一批跨学科的语言学人才。研究机器翻译是因为研究人类智能及其模拟和应用。智能是指能够自动学习知识和自动并且有效地利用学习到的知识去解决问题的能力。人类在数学、智力游戏、写论文等等活动表现出的能力都能够反映出智能。人类的语言能力同样能反映出人类的智能。一个小孩放在东京、伦敦、北京，他就会流利地讲日语、英语、汉语。不需要大人们干预，更不要修改他们的“程序”。由于 1980 年日本东京一本小册子，谈到自然语言之间的翻译未来市场很大，才意识到机器翻译未来是具有巨大的经济效益和社会效益。之后，才把它作为一个独立研究课题。研究人类智能是因为 1980 年为国防科

工委的研制巨型机任务转到国防科技大学之后，科学院转向研究“未来面向智能领域应用的巨型机”（注：该项研究经过科学院、国家科委和国防科工委推荐，民口论证、军民联合论证和中南海论证和钱学森支持之后，被列入“863”项目）。之后发现全世界人工智能50多年来所研究的系统，没有一项有智能，因而重新研究人类的智能。而研究巨型机任务是从1973年3月中科院计算所刚刚“解放”的老所长阎沛霖带我到国防科委钱学森主任接收亿次机设计任务开始。虽然之前的几年，钱老曾经布置过国防科委下属的两个研究所进行研究，但是，反馈回来只有一批调研资料和国内技术条件不成熟的意见。当时国内实际进行计划仍然是200万～500万次，不能满足需要，国际上巨型计算机ILLIAC-IV因为不可靠及难于使用正处于一片批评声中，而SRAR-100的条件国内难于满足。因为我们在阅读国防科技情报所和计算所情报室所提供国外有关巨型机材料后，发现这两种截然不同的两种巨型机是等价的。关键在于向量必须进行分段、流水线处理，使用必须依靠向量语言，所以两个月后的1973年5月提出了可行的解决方案，正式承担这项巨型机设计任务及其模型机——中国第一台向量计算机757的研制任务。并且得到两星期向钱老汇报一次和聆听钱老讲述他如何把物理概念和数学方程融合在一起解决问题的经验的机会。而有机会承担国防部门巨型机任务是因为1957年我从北京大学数学力学系毕业之后被分配到中国第一个计算机系统结构研究和设计小组，承担中国第一台自行设计大型电子管计算机、第一台自行设计大型晶体管计算机和专为“两弹一星”服务、被誉为“功勋计算机”的109丙机的系统结构设计。而1957年我被分配到中国第一个计算机系统结构研究和设计小组是因为我在1955年被动员改学计算数学，1956年参加北京大学、清华大学与计算所筹备处合并第一届训练班的计算数学组，而且根据苏联计算机领域的领导人列贝捷夫院士的意见，要安排

数学专业的人员从事计算机系统结构研究和设计工作。（这点很重要，是关键。比具体承担者是谁更重要。）一切似乎是偶然，偶然背后又有必然。这是一名北京大学数学系学生毕业后众多的工作轨迹中的一个。20 世纪 50～70 年代，课题是国家给的，努力去完成就是。80 年代开始，国家只给意向，甚至不给，要靠自己独立思考。关键是：独立思考，不人云亦云。独立判断是非曲直，独立判断经济效益、社会效益和理论价值。冷对众说纷纭。或许以下三句话有参考价值：

“任何正常人都有优点和特长，检查检查你的优点和特长在哪里。人类的需要是阳光，你的优点和特长是水和土壤。有了阳光、水和土壤，你的兴趣就会带你在事业上飞翔。没有阳光，或者没有水和土壤，兴趣只能给你幻想。”

“要特别注意那种经济效益和社会效益很大，人们认为做不到，难度很大但可能做到的事，因为这往往是重要的生长点和突破口。”

“不要幻想经过成千上万个聪明人没有搜索到的重要的科技宝藏，会突然从天上掉到你的口袋里。首先想一想解决它需要什么先决条件？例如跨学科知识，你是否具备？你是否有决心和有条件去具备？”

在北京大学数学力学系的同年级同学中，水平与我一样或者比我高者，至少有 50 名。除了被安排任务存在“运气”之外，就只有独立思考。

由于网络安全研究工作的需要，近几年，在学习中科院计算技术研究所网络信息安全研究组译 Navarro G. 和 Raffinot M. 著的《柔性字符串匹配》，在此书基础上，进行巨量字符串的串匹配研究工作。本专著就是根据其研究成果写出的，属于抛砖引玉性质。

第 1 章绪论介绍前人与本专著有关的工作；第 2 章讨论巨量

字符串匹配完全自动机的具体自动生成算法及算法的计算复杂性，特别是详细讨论了其中的第 4 步的状态连接补全；第 3 章讨论了面向巨量字符串匹配完全自动机新的、简单可实现的专用系统结构，并讨论其中的并行处理机制；第 4 章进一步讨论模糊化的 U-不确定控制下的巨量字符串和干扰条件下的 V-不确定控制下的巨量字符串的串匹配；最后一章讨论超长字符串的部分匹配的算法和硬件实现方法。

中国科学院院士

高庆狮

2011 年 1 月 1 日

目　录

第1章 绪 论

1.1 需 求

拼写检查、语言翻译、数据压缩、搜索引擎、网络入侵检测、计算机病毒特征码匹配，以及DNA序列匹配等的应用，都需要字符串匹配。

字符串的长度可以是等长，也可以不等长；组成字符串的基本字符可以是位、字节、双字节、汉字、日语字母、日语字、日语词、英文字母、英文字、英文词，及某种混合，等等。这里的英文字可以是一个或者多个英文字母所组成，英文词可以是一个或者多个英文字所组成。

字符串匹配可以是它匹配，也可以是自匹配。它匹配可以是一个随机的有穷字符串或者无穷字符串，与确定的一个或者多个给定的等长或者不等长，包括超长的字符串相匹配，目的是从随机的有穷长字符串或者无穷字符串中找出与一个或多个给定的字符串，所有的相同的段或者部分连续相同的段。

1.2 半个世纪研究工作（1951～2001年）的总结

2002年，Navarro G. 和 Raffinot M. 总结了1951～2001年有关字符串匹配的论文，出版了《柔性字符串匹配》一书。中国科学院计算技术研究所网络信息安全研究组刘萍等人翻译了该书，中译本由电子工业出版社于2009年出版。

《柔性字符串匹配》的中译本共分 7 章。其中导言介绍了该书的目的、侧重点、概况和基本概念；前 5 章讨论了字符串匹配，多字符串匹配，扩展字符串匹配，正则表达式匹配和近似匹配；最后一章是总结。

1.3 Shift-Or 算法

下面介绍的 Shift-Or 算法［A. Aho，M. Corasick，1975］的例子取自《Flexible pattern matching in strings：practical on-line search algorithms for texts and biological sequence. Gonzalo Navarro，Mathieu Raffinot》［Cambridge University Press，2002］中的 2.2 节，其中，$ 表示匹配成功。

【例 1.1】 在字符串 AGATACGATATATAC 中搜索字符串 ATATA。

搜索过程如下：

B=	A	1	0	1	0	1
	T	0	1	0	1	0
	*	0	0	0	0	0

初始：D= 0 0 0 0 0

		0	0	0	0	1
1.	读入 A	1	0	1	0	1
	D=	0	0	0	1	1
		0	0	0	1	1
2.	读入 G	0	0	0	0	0
	D=	0	0	0	0	0
		0	0	0	0	1
3.	读入 A	1	0	1	0	1
	D=	0	0	0	0	1
		0	0	0	1	1
4.	读入 T	0	1	0	1	0
	D=	0	0	0	1	0
		0	0	1	0	1
5.	读入 A	1	0	1	0	1
	D=	0	0	1	0	1
		0	1	0	1	1
6.	读入 C	0	0	0	0	0
	D=	0	0	0	0	0
		0	0	0	0	1
7.	读入 G	0	0	0	0	0
	D=	0	0	0	0	0
		0	0	0	0	1
8.	读入 A	1	0	1	0	1
	D=	0	0	0	0	1

续

		0	0	0	1	1			0	1	0	1	1
9.	读入 T	0	1	0	1	0	13.	读入 T	0	1	0	1	0
	D=	0	0	0	1	0		D=	0	1	0	1	0
		0	0	1	0	1			1	0	1	0	1
10.	读入 A	1	0	1	0	1	14.	读入 A	1	0	1	0	1
	D=	0	0	1	0	1	$	D=	1	0	1	0	1
		0	1	0	1	1			0	1	0	1	1
11.	读入 T	0	1	0	1	0	15.	读入 A	0	0	0	0	0
	D=	0	1	0	1	0		D=	0	0	0	0	1
		1	0	1	0	1							
12.	读入 A	1	0	1	0	1							
$	D=	1	0	1	0	1							

【例 1.2】 在字符串 annual _ announce 中搜索字符串 announce。搜索过程如下：

	a	00000001
	c	01000000
	e	10000000
B=	n	00100110
	o	00001000
	u	00010000

*	00000000

初始： D= 00000000

		00000001			00000001
1.	读入 a	00000001	8.	读入 a	00000001
	D=	00000001		D=	00000001
		00000011			00000011
2.	读入 n	00100110	9.	读入 n	00100110
	D=	00000010		D=	00000010
		00000101			00000101
3.	读入 n	00100110	10.	读入 n	00100110
	D=	00000100		D=	00000100
		00001001			00001001
4.	读入 u	00010000	11.	读入 o	00001000
	D=	00000000		D=	00001000
		00000001			00010001
5.	读入 a	00000001	12.	读入 u	00010000
	D=	00000001		D=	00010000
		00000011			00100001
6.	读入 1	00000000	13.	读入 n	00100110
	D=	00000000		D=	00100000
		Q0000001			01000001
7.	读入—	00000000	14.	读入 c	01000000
	D=	00000000		D=	01000000
					10000001
			15.	读入 e	10000000
			$	D=	10000000

1.4 多字符串匹配

问题：多字符串匹配。

给定：一个有穷字符集 Ω 和一个由 k（数以百万计）字符串组成的字符串集 $\sum$，其中

$$\sum = \{B_i \mid B_i = b_{i1}\ b_{i2}\ b_{i3} \cdots b_{im_i}, i = 1, 2, \cdots, k, b_{ij_i} \in \Omega\}$$

任给：一个无穷字符串 A，$A = a_1\ a_2\ a_3 \cdots a_n \cdots$

求：生成一个完全自动机，该自动机能够自动匹配出所有的 B_i，满足 $B_i = A_t$，其中

$$A_t = a_{t+1}a_{t+2}a_{t+3} \cdots a_{t+mi}, B_i \in \sum, a_t \in \Omega, 0 \leqslant t$$

【例 1.3】 在一个无穷英文字母字符串中，找出与字符串集 $\sum$ 中所有相同的字符串，其中

$$\sum = \{BCEC, AED, BCG, EDFG, OPQ, PQS, MBA, \cdots\}$$

【例 1.4】 在一个无穷汉字串中，找出与汉字串集 $\sum$ 中所有相同的汉字串，其中，$\sum$ = {仓库，仓库标记，仓库储存，仓库船，仓库交货，仓库交货价，仓库交货条件，仓库交货现价，仓库费，仓库基地，仓库价格，仓库理货，仓库内的货物，库存，库房，……交货……货运单货运……}。

【例 1.5】 在一个无穷位串中，找出与位串集 $\sum$ 中所有相同的位串，其中

$\sum$ = {0101，0001110101，001100，11001010，010011000111，111，11110101，001110101，1010101100，11011100，0100110，1001，0110，…}

1.5 Aho-Corasick 算法与 Aho-Corasick 自动机

Aho-Corasick 算法是针对多字符串的串匹配，对应于算法的自动机是一种特殊的自动机，称作 Aho-Corasick 自动机，其特殊在于引进和使用了特殊函数 S_{AC}。

【例 1.6】 集合 {ATATATA，TATAT，ACGATAT} 的 Aho-Corasick 自动机，如图 1.1 所示。[Navarro G.，Raffinot M.，2002]。

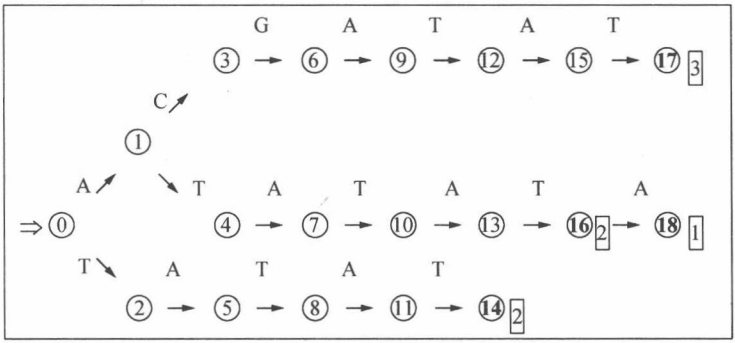

图 1.1 例 1.6 的 Aho-Corasick 自动机

在例 1.6 中，$S_{AC}(q) = q'$。

有：$S_{AC}(9) = 1$，$S_{AC}(12) = 4$，$S_{AC}(15) = 7$，$S_{AC}(17) = 10$，$S_{AC}(4) = 2$，$S_{AC}(7) = 5$，$S_{AC}(10) = 8$，$S_{AC}(13) = 11$，$S_{AC}(16) = 14$，$S_{AC}(5) = 1$，$S_{AC}(8) = 4$，$S_{AC}(11) = 7$，$S_{AC}(14) = 10$。

由于 $S_{AC}(16)=14$ 是终结状态，所以 16 也是终结状态。

【例 1.7】 在文本“AGATACGATATATAC”中搜索模式串集合 P={ATATATA，TATAT，TCGATAT} 中的模式串 [Navarro G.，Raffinot M.，2002]。

搜索过程如下：

Current←0	8. 读入 A Current←9 =δ(6，A)
1. 读入 A Current←1 =δ(0，A)	9. 读入 T Current←12 =δ(9，T)
2. 读入 G δ(1，G)=θ，状态转 0= S_{AC}(1) δ(0，G)=θ，状态转 θ= S_{AC}(0) 然后继续从初始状态 0 开始搜索， Current←0	10. 读入 A Current←15 =δ(12，A)
3. 读入 A Current←1 =δ(0，A)	11. 读入 T Current←**17** =δ(12，T) 状态 17 为终结状态，报告匹配 **F（17）→ACGATAT**
4. 读入 T Current←4 =δ(1，T)	12. 读入 A δ(17，A)=θ，状态转 10= S_{AC}(17) δ(10，A)=13，Current←13。
5. 读入 A Current←7 =δ(4，A)	13. 读入 T Current←**16** =δ(13，T) 状态 16 为终结状态，报告匹配 **F（16）→TATAT**
6. 读入 C δ(7，C)=θ，状态转 5= S_{AC}(7) δ(5，C)=θ，状态转 1= S_{AC}(5) δ(1，C)=3，Current←3	14. 读入 A Current←**18** =δ(16，A) 状态 18 为终结状态，报告匹配 **F(18) →ATATATA**
7. 读入 G Current←6 =δ(3，G)	15. 读入 C δ(18，C)=θ，状态转 13= S_{AC}(18) δ(13，C)=θ，状态转 11= S_{AC}(13) δ(11，C)=θ，状态转 7= S_{AC}(11) δ(7，C)=θ，状态转 5= S_{AC}(7) δ(5，C)=θ，状态转 1= S_{AC}(5) δ(1，C)=3，Current←3