



数字时代图书馆情报学青年论丛

信息适应性 原理、机制与方法

Information Flexibility:

Principles, Mechanisms and Methods

■ 陆 泉 编著



WUHAN UNIVERSITY PRESS

武汉大学出版社



数字时代图书馆情报学青年论丛

信息适应性 原理、机制与方法

Information Flexibility:
Principles, Mechanisms and Methods

■ 陆 泉 编著



WUHAN UNIVERSITY PRESS

武汉大学出版社

图书在版编目(CIP)数据

信息适应性原理、机制与方法/陆泉编著. —武汉: 武汉大学出版社,
2008. 6

数字时代图书馆学情报学青年论丛

ISBN 978-7-307-06212-2

I. 信… II. 陆… III. 信息技术—研究 IV. G202

中国版本图书馆 CIP 数据核字(2008)第 055945 号

责任编辑:陈 红 责任校对:刘 欣 版式设计:马 佳

出版发行: 武汉大学出版社 (430072 武昌 珞珈山)

(电子邮件: wdp4@whu.edu.cn 网址: www.wdp.com.cn)

印刷: 武汉中远印务有限公司

开本: 720 × 980 1/16 印张: 17.375 字数: 248 千字 插页: 2

版次: 2008 年 6 月第 1 版 2008 年 6 月第 1 次印刷

ISBN 978-7-307-06212-2/G · 1089 定价: 29.00 元

版权所有, 不得翻印; 凡购我社的图书, 如有缺页、倒页、脱页等质量问题, 请与当地图书销售部门联系调换。

内 容 简 介

本书从信息适应性的原理、机制与方法三个方面对信息适应性进行了全面的阐述。首先分析了信息洪水问题和软件危机问题的内在原因,指出现有信息理论的不足,从定义、理论模型和实现机制的理论框架三个方面阐述了信息适应性的基本原理。然后从基于软件重用、机器学习、进化和复杂适应系统的信息适应性机制等四个层面介绍了信息适应性的实现机制和方法。最后综合阐述了信息适应性的实现技术手段和方法问题。本书是首部全面、深刻地介绍信息适应性的力作,既具有理论启发性和创新性,又具有全面而深入的实践性,对信息管理研究以及信息洪水和软件危机的解决有广泛和深入的影响。本书可作为信息管理、情报学、计算机等专业的研究生和高年级本科生的教学用书,也可供相应专业的科技人员学习和参考。



前言

随着计算机和信息技术的普遍应用，信息洪水和软件危机问题给人类社会带来了负面效应和现实危机。软件危机和信息洪水问题是世界各国必须面对的日益严重的共同现实威胁，虽然目前相关研究很多，但是还没有很好的解决方法。信息洪水一般被理解为这样的一种现象：信息是海量的，而可用的、可信的信息却很少，或者很难获得。软件危机可以理解为：落后的软件生产方式无法满足迅速增长的计算机软件需求从而导致软件开发与维护过程中出现一系列严重问题的现象。

信息洪水起始于 Internet 和个人信息管理软件的普及，目前已经成为一个世界性的亟待医治的网络顽症。其主要解决方法是信息组织、检索和服务技术，如数据质量管理、信息生命周期管理、语义 Web 和搜索引擎等，然而信息洪水问题并没有被真正遏制住。

20 世纪 60 年代中期，计算机应用范围的迅速扩大导致了软件危机的产生，40 多年来其表现形式在不断改变，而且有愈演愈烈的趋势。其主要解决方法是以软件重用为代表的软件适应性研究，具体方法从早期的模块化、算法选择、面向对象到现在的构件、基于代理、软件再工程、基于本体的领域工程和敏捷软件等在一定程度上缓解了软件危机。

本书专注于以信息科学的观点研究解决软件危机和信息洪水问题的软件理论和技术，提出了信息适应性的理论、机制和方法。信息科学认为：软件的意义是信息，软件的变化反映的是信息的变



化。鉴于此,本书把信息洪水问题和软件危机问题合并起来,从信息的本质特征上研究了信息洪水和软件危机的共同根源,提出信息的适应性问题。本书从信息适应性的原理、机制与方法三个方面对信息适应性进行了全面的阐述。首先分析了信息洪水问题和软件危机问题的内在原因,指出现有信息理论的不足,认为现有的解决方法存在理论认识上的误区,不可能真正地解决问题;从而提出了“信息的适应性”研究的解决方法,从定义、理论模型和实现机制的理论框架三个方面阐述了信息适应性的基本原理。然后从基于软件重用、机器学习、进化和复杂适应系统的信息适应性机制等四个层面介绍了信息适应性的实现机制和方法。最后详细阐述了信息适应性的实现技术手段和方法问题。

本书从信息本质特征及其管理角度研究和解决信息洪水问题和软件危机问题,是首部介绍并全面阐述“信息适应性”理论、机制和技术的著作,首部全面和深入地介绍了实现或提高信息适应性的理论和技术的著作,首部全面和深入地介绍了个性化、本体、Web2.0、Agent等前沿热点技术与信息适应性的结合方法的著作。本书具有较高的学术价值,其“信息适应性”观点具有理论新颖性,突破了现有信息理论对信息本质特征认识的局限性,对解决日趋紧迫的信息洪水和软件危机问题有重要意义,对推动和促进信息管理科学的相关研究有较高价值,对信息类科学的相关研究和发展有启发意义。

本书前沿性和现实性强,集信息管理与计算机应用专业之所长,既具有理论启发性和创新性,又具有全面而深入的实践性。本书可作为信息管理、情报学、计算机等专业的研究生和高年级本科生的教学用书,也可供相应专业的科技人员学习和参考。

本书是在博士学位论文“信息组织的适应性机制与方法研究”的基础上修改、充实而成的,进行了详细的理论阐述、增加了一些新的研究成果,如基于Wiki的信息适应性系统构建,基于人工心理的个性化方法的应用等。

衷心感谢我的博士研究生导师武汉大学计算机学院孟波教授的悉心指导,感谢武汉大学信息管理学院马费成教授、焦玉英教授、



唐晓波副教授、邓仲华教授等多位领导和老师对我的研究给予的大量无私的建议和帮助。

此外，武汉大学信息管理学院与武汉大学出版社对此书的编写和出版给予了大力支持，书中还多次引用国内外同行的研究成果，谨此一并感谢。

由于水平有限，书中错误或不妥之处在所难免，希望读者给予批评指正。

陆 泉

2007 年 12 月于珞珈山下

目 录

前 言	1
第 1 章 概 论	1
1.1 信息洪水问题	1
1.2 软件危机问题	6
1.3 有关工作的研究进展	10
1.3.1 软件适应性研究	10
1.3.2 信息管理学科中的有关研究	13
第 2 章 信息适应性原理	15
2.1 信息适应性	15
2.1.1 信息及其特征	15
2.1.2 信息适应性的提出	18
2.1.3 信息适应性的特征	22
2.1.4 信息适应性的理论意义	23
2.2 信息适应性的理论模型	24
2.2.1 模型描述	24
2.2.2 模型的理论分析	26
2.2.3 模型的运行机制	32
2.3 信息适应性实现机制的理论框架	33
2.3.1 信息适应性的软件重用实现机制	36



2.3.2	基于机器学习的信息适应性实现机制	37
2.3.3	基于进化的信息适应性实现机制	38
2.3.4	基于复杂适应系统的信息适应性实现机制	38
2.4	小 结	39
第3章	软件重用信息适应性机制	40
3.1	软件重用介绍	40
3.1.1	软件危机分析	40
3.1.2	软件重用的定义	43
3.1.3	软件重用的目标	43
3.1.4	软件重用的适应性思想	44
3.2	软件重用信息适应性机制和方法	45
3.2.1	面向对象的软件重用信息适应性机制	45
3.2.2	基于代理的软件重用信息适应性机制	48
3.2.3	基于构件的软件重用信息适应性机制	51
3.2.4	软件再工程信息适应性方法	56
3.2.5	领域工程信息适应性方法	61
3.2.6	敏捷软件信息适应性方法	62
3.3	以软件重用度量为基础的信息适应性度量	66
3.3.1	软件重用度量的借鉴可行性分析	66
3.3.2	重用成熟度模型	67
3.3.3	重用考核指标	69
3.4	软件重用信息适应性机制的主要研究方向	70
3.5	小 结	71
第4章	基于机器学习的信息适应性机制	73
4.1	机器学习介绍	73
4.2	机器学习的信息适应性机制	76
4.2.1	指导式学习信息适应性方法	77
4.2.2	归纳学习信息适应性方法	78
4.2.3	基于解释的学习信息适应性方法	80



201	4.2.4	基于神经网络的学习信息适应性方法	82
201	4.2.5	基于遗传算法的学习信息适应性方法	84
081	4.2.6	强化学习信息适应性方法	85
081	4.2.7	多 Agent 学习信息适应性方法	88
181	4.2.8	小 结	89
	4.3	个性化信息适应性方法	90
881	4.3.1	个性化与信息适应性	90
	4.3.2	感性科学与信息适应性	94
081	4.3.3	基于人工心理模型的网站信息个性化方法	100
801	4.3.4	基于群体心理的网站信息个性化方法	112
	4.3.5	个性化信息适应性方法的应用	115
401	4.4	小 结	121
401			
	第 5 章	基于进化的信息适应性机制	123
001	5.1	进化方法介绍	123
100	5.2	进化计算	125
800	5.2.1	进化计算的产生与发展	125
000	5.2.2	几种进化计算思想的信息适应性	127
000	5.3	基于进化的信息适应性机制分析	130
810	5.3.1	遗传算法方法信息适应性	130
210	5.3.2	DNA 计算方法信息适应性	135
710	5.3.3	协同进化方法信息适应性	138
810	5.3.4	进化博弈方法信息适应性	145
000	5.4	小 结	151
000			
	第 6 章	基于复杂适应系统的信息适应性机制	152
000	6.1	复杂性介绍	152
800	6.1.1	复杂性科学的产生与发展	152
800	6.1.2	复杂系统建模工具	157
800	6.2	复杂适应系统	161
800	6.2.1	复杂适应系统概述	161



58	6.2.2 回声模型	162
48	6.2.3 复杂适应系统的主要理论	165
28	6.3 基于复杂适应系统的信息适应性机制的理论选取	180
88	6.3.1 信息适应性与复杂适应系统理论	180
98	6.3.2 自组织临界性与混沌边缘的等价性	181
00	6.3.3 自组织临界性理论宏观规律的普适性	188
00	与代表性	188
40	6.3.4 自组织临界性理论微观机制的普适性	189
001	与代表性	189
11	6.4 小 结	193
112	第7章 信息适应性的方法实现	194
	7.1 概 述	194
63	7.2 信息适应性实现的技术手段	195
151	7.2.1 基于本体的信息适应性方法	196
251	7.2.2 基于代理的信息适应性方法	201
251	7.2.3 基于语义网的信息适应性方法	203
27	7.3 基于本体的信息适应性系统构建	206
081	7.3.1 本体概述	206
081	7.3.2 基于本体的信息适应性系统分析	213
281	7.3.3 动态本体管理的时间模型设计	215
881	7.3.4 基于SOC的动态本体管理算法	217
241	7.3.5 模型与算法分析	219
121	7.4 基于代理的信息适应性系统构建	220
	7.4.1 Agent 概述	220
151	7.4.2 MAS 体系结构与适应性	223
121	7.4.3 复杂系统的多 Agent 建模	226
121	7.4.4 多 Agent 系统的共进化方法	228
121	7.5 基于语义网的信息适应性系统构建	233
101	7.5.1 语义网概述	233
101	7.5.2 Web 2.0 技术	235

7.5.3 基于 Wiki 的信息适应性系统构建	241
7.6 小 结	249
参考文献	250

第1章 概 论

本章与第二章一起构成全书的第一部分：理论部分。第一章提出要研究的现实问题，分析了信息洪水问题和软件危机问题的内在原因，引出信息的适应性问题，并综述了相关研究现状；第二章从定义与特征、理论模型和实现机制的理论框架三个方面阐述了信息适应性的基本原理。

1.1 信息洪水问题

1. 信息洪水的产生与现状

信息洪水^[187] (Information Flood) 一般被理解为这样的一种现象：信息是海量的，而可用的、可信的信息却很少，或者很难获得。

随着信息时代科学技术革命的飞速发展，新兴学科大量涌现，知识量急剧膨胀，知识更新过程空前加快，出现了知识急剧膨胀的现象，据联合国教科文组织的统计，人类近 30 年来所积累的科学知识占有史以来积累的科学知识总量的 90%，而在此之前的几千年中所积累的科学知识只占 10%^[49]。知识急剧膨胀是信息技术广泛应用的现实环境下信息洪水产生的内因，是社会创新力的源泉，是人类社会进步的一个表现，是值得肯定的因素。

同时，人们对以往经历过的信息饥荒的恐惧和在信息经济时



代、知识经济时代对信息富裕的渴望表现得越来越强烈，这一趋势和信息科学技术相互推动、相互促进，逐渐形成了所谓“信息洪水”、“信息过剩”或“信息超载”的现象。这一现象主要体现在以下三个方面：①明明知道相关信息确实存在而且是大量存在，但就是找不到的挫折感；②生怕有些重要的东西被忽视了的恐惧感；③由于相关信息太多而无法阅读、处理和采取有意义的行动的无力感。所有这些感觉似乎都在证明：信息正在干扰我们的正常工作，信息已经成为我们的负担。在 1942 年，英国剑桥皇后学院的世界上最好的图书馆也只藏有 199 本图书；而根据最近的一份国际数据公司（IDC）的报道^[191]，2006 年全球“创建、收集和复制”的数字信息已达到 1 610 亿 GB，这是常人难以想象的数量级，这个容量大约相当于把书从地球到太阳来回堆 12 遍的全部容量。IDC 估计，到 2010 年信息流量会高达 9 880 亿 GB。

知识急剧膨胀是对人类有利的客观存在，而信息洪水是人们要避免和解决的现实问题。但是，信息洪水问题仍然缺乏足够的研究和重视。信息洪水不是个新问题，但今日信息的组成成分和来源已与过往不可同日而语，如来自网络日志（Blog、Wiki）、即时通信（IM）、电子邮件、电子表单、无线射频识别（RFID）标签、视频、电子商务交易、技术支援、供应链等各方面的数据都大得惊人。IDC 预计，2007 年人们生产的数字信息将首次超过存储容量。信息洪水的危害已经到了无法忽视的程度，如 IDC 分析师苏珊·费尔德曼（Susan Feldman）就明确指出：“当数据容量的威胁已达到骇人听闻的程度时，我们就该有所行动了。”

虽然信息洪水已不是个新问题，但因此就掉以轻心是极其危险的。很多企业主管和 IT 精英认为信息洪水只是他们经历并处理过的一个日常问题，忽视了信息洪水问题的新发展。如麻省理工学院（MIT）的 IT 教授斯图尔特·马德尼克（Stuart Madnick）认为，信息洪水最大的危险在于，海量信息排山倒海而来时，人们却还熟视无睹、麻痹大意。他表示：“很可能有些高层就简单粗暴地把这个问题拒之门外，这会错失良机。”

信息洪水的现实威胁在于过量且快速增长的信息会超出组织的



信息管理能力，更别提怎样去利用它了。

2. 信息洪水的产生原因

分析信息洪水的产生原因，除了知识急剧膨胀这个内因，一般还有：

①信息缺乏管理或管理不善，信息的发布、传播失去控制，产生了大量虚假信息、无用信息，造成信息环境的污染和“信息垃圾”的产生。

②计算机病毒复制、变异和传播而造成的错误信息。

③信息生产者站在不同角度，用不同视角对同一事物做出不同的描述与反映，有时甚至大相径庭，使信息使用者难以抓住事物发展变化的本质与主流。

④网络上的垃圾站点故意散布的垃圾信息。具体就网络信息获取中遇到的信息超载而言，其产生原因一般有^[208]：

·难以决定信息之间的相关性。语言的差异，糟糕的标签和不精确的阐述，造成信息需求和信息供给之间不能形成紧密联系。再加上供给的过度，使得决定信息相关性变得十分艰难。一则信息对你来说是微不足道的，对别人却可能是至关重要的。

·信息来源过多和信息条块分割。在互联网上或企业内联网上，相同、相似或相关的信息储存在多个资料库里，这种分散使得要获取有关某一主题的最完善的信息的全部甚至一小部分都要花很多时间，有时甚至花了很多时间也一无所获。

·信息文盲。信息文盲并不是指不认识字或不会使用电脑的人，而是指不知道从哪里寻找与某一课题相关的高质量信息的人，这其中甚至包括许多信息从业人员。许多人从来没有接受过使用搜索引擎的训练，不知道搜索的根本因素。大多数企业和个人缺乏寻找信息和理解信息的技能。

·信息的不合格。到目前为止帮助使用者理解文件的诸多方面（比如说涵盖的广度和深度，来源的可靠性，资料的准确性和及时性等）的可靠而又全面的反馈机制尚不存在。

信息洪水发生的原因是多样的，本书以信息本质特征的观点来



看：信息管理不具有适应性是其中一个根本原因。具体来说，现有的信息管理理论不符合信息的变化规律，信息管理技术不能适应现实与环境变化，维护不及时、不准确，同一系统中的信息变化不一致等设计与维护上的技术问题；这会使得管理量和复杂性增加，同时降低用户对信息管理的信任，加重用户对信息匮乏的恐惧，从而导致更多人为的信息洪水。

3. 信息洪水的解决办法

下面分别从企业和互联网角度分析信息洪水的解决办法。对于企业内部普遍存在的信息洪水问题，目前企业主要的解决方法是采用信息生命周期管理和数据质量管理等方法。

目前，信息生命周期管理（ILM）正在成为企业管理的一个重要课题，企业需要在重新规划业务流程的同时，使信息与业务流程实现完美的融合。信息生命周期管理是以成本与信息不断变化的价值相对应的方法，从创建到最终处置对信息进行全程管理。信息生命周期管理使企业可以根据商务、监管和法律方面的要求对信息进行分类、搜索、存档、恢复和删除。简言之，信息生命周期管理就是：在信息的整个生命周期（从产生到最终归档或销毁），如何把合适的信息，在合适的时间，存储在合适的介质上。

也有人认为，信息生命周期管理的概念不仅仅局限于如何在存储介质上更好地管理数据，更重要的观点是，数据的保存应该基于业务目标以及数据如何被利用和改变等方面。因此，良好的信息生命周期管理应该包括贯穿各种存储技术的内容分类、备份与恢复、复制、归档、数据移植、数据分发等，同时具有较好的索引和搜索功能。

数据质量管理是通过管理、控制和改进数据质量，促进信息资源的开发、利用和共享。对于什么是数据质量可从两个方面来理解：

① 数据本身的数据质量：包括数据的真实性，即数据必须真实准确地反映实际发生的业务；数据的完备性，是指数据是充分的，任何有关操作的数据都没有被遗漏；数据的自洽性，是指数据并不是孤立存在的，数据之间往往存在着各种各样的约束，这种约



束描述了数据的关联关系，数据必须能够满足这种数据之间的关联关系，而不能相互矛盾。

数据的真实性、完备性、自洽性是数据本身应具有的属性，一般把这三点称为数据的绝对质量。数据的绝对质量是保证数据质量的基础。

②数据的过程质量：除了数据的绝对质量外，还有我们在利用和存储数据的过程中所产生的数据质量问题，包括使用质量、存储质量和传输质量，一般称之为过程质量。其中，数据的使用质量是指数据被正确地使用。再正确的数据，如果被错误地使用，就不可能得出正确的结论。数据的存储质量是指数据被安全地存储在适当的介质上。所谓安全是指采用了适当的方案和技术来抵制外来的因素，使数据免受破坏，备份是我们常使用的技术，是保护信息资源的重要手段，通常包括异地备份和双机备份等，目前越来越多的企业领导意识到备份尤其是异地备份的重要性；所谓存储在适当的介质上是指当需要数据的时候能及时方便地取出。数据的传输质量是指数据在传输过程中的效率和正确性。在现代信息社会中，数据在异地之间的传输越来越多，保证传输过程中的高效率 and 正确性非常重要。为了提高数据在网络上的传输质量，除了研究更快更稳定的网络传输协议（如 FAST TCP 协议）、建设更好的网络基础设施以外，还可以使用一些常用软件方法如校验、数字签名等。

从本质上分析，信息生命周期管理和数据质量管理等方法都属于数据信息的一致性、完整性管理技术，在这些管理中都没有涉及信息的适应性问题。

信息洪水的另一个典型代表是 Web。Web 从 1991 年出现以来，经过 17 年发展已经成为一个巨大的全球化信息资源库。Web 上的信息量以几何级数的速度增长，使得用户发现其所需要的信息变得非常困难。在这种情况下，如何有效地组织和检索 Web 信息也就成为一项重要的研究课题。

一般认为，提高 Web 信息检索的质量包括两方面内容：一方面是如何在现有的资源上设计更好的检索技术；如何开发和利用 Web 信息资源已经成为人们所关注的焦点。目前最成功的是国际