实证软件工程 - Digital Archeology: 程序员成熟度模型

周明辉

zhmh@pku.edu.cn

http://sei.pku.edu.cn/~zhmh

北京大学

提纲

- 软件工程的历史
  - 研究对象
    - 从技术性(technical)
      到社会性(social)
  - 研究方法
    - 从软件度量 (Software Measurement)
      到数字考古(Digital Archeology)
- 数字考古的挑战

软件工程

- 目标：提高生产效率和质量
- 手段：提供各种技术/工具/过程帮助提高生产效率和改进软件质量
  - 流水线的过程：需求->设计->实现->
  - 构件化的组装
  - 类比：建造桥梁，建造房屋
- 原则：
  - 细节抽象
  - 关注点分离

软件开发工程化！

软件工程的尴尬

- **Brook's law:** 人月神话
  - 当向延迟的项目增加人手，效率降低，项目延迟

- **IBM: Rational, 5%** 项目使用

软件开发真的可以工程化吗?

A Plea

Please stop comparing creating software to building bridges and buildings (or even hardware)!

软件工程的非工程性

- 软件开发是知识密集型活动(Software development is a knowledge intensive activity)
- 个体差异是导致效率差异的最大因素(personality has a marked effect on the performance of employees)
- 软件开发是一项本质复杂的活动，技术无法解决软件开发复杂性

软件开发非技术性因素——人的因素

“人是最重要的因素”

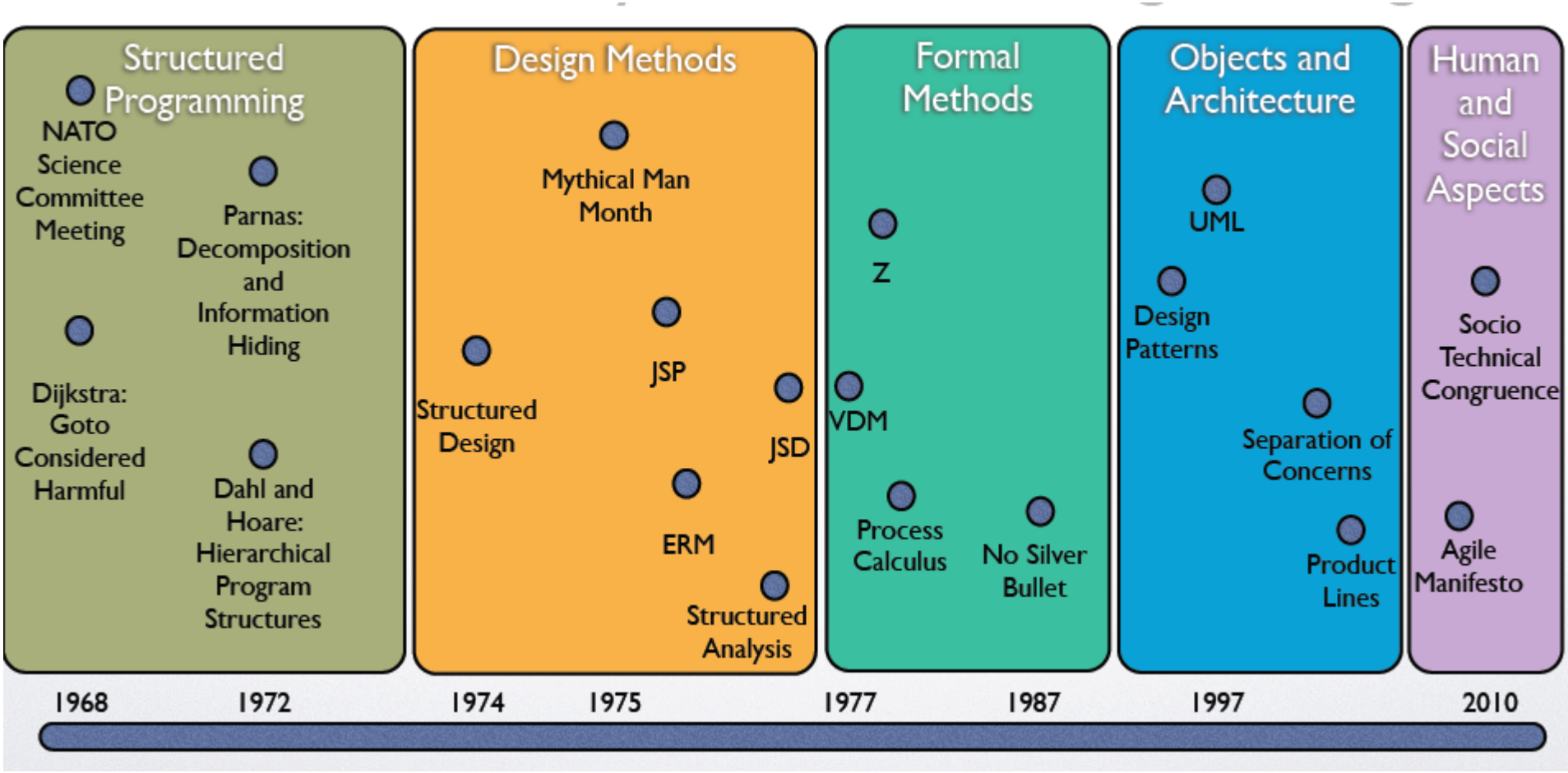
- 人的因素有哪些?
- 人的因素是怎么影响软件开发的?
- 怎么控制软件开发中人的因素?

研究"人"是一件很难的事情

- 人是影响项目成功，以及变化性最大的因素
  - Sackman et al, 1968, 28:1
  - Curtis, 1981, 23:1
  - Boehm, 1981
- 人的因素?
  - Knowledge structure
  - Personality
  - Motivation, emotion
  - Technical needs, social needs
- 缺乏数据，难以研究
  - "Until the many sources of variation among individuals have been compared in the same set of data, it will not be possible to determine precisely ... the most important predictor of success"

    - Curtis, 1984

A brief history of software engineering



IBM, Clay Williams: The social side of software engineering

提纲

- 软件工程的历史
  - 研究对象
    - 从技术性(technical) 到社会性(social)
  - 研究方法
    - 从软件度量 (Software Measurement) 到数字考古(Digital Archeology)
- 数字考古的挑战

为什么度量？

- **"... the art of *measurement would do away with the effect of* appearances, and, showing the *truth, would fain teach the soul at last* to find rest in the truth, and would thus save our life."**

  - *Protagoras, Plato*

- **The absence of romance in my history will, I fear, detract somewhat from its interest; but if it be judged useful by those inquirers who desire an *exact knowledge of the past as an aid to the interpretation of* the future, which in the course of *human things must resemble if it* does not *reflect it, I shall be content.***

  - The History of the Peloponnesian War, Thucydides (伯罗奔尼撒战争, 修昔底德)

研究问题

Through measurement obtain the exact knowledge of the past as an aid to the interpretation

of the future, which in the course of human things must resemble it.

“and thus save our life” or at least “be content”

软件开发中的度量

- A small-scale version of real life
- Supported by tools leaving traces to measure
- A maturing discipline ripe for more reproducibility

软件度量

❑研究问题

- ➢软件开发中的最佳实践?
- ➢如何度量生产效率?
- ➢如何度量软件质量?

❑研究方法

- ➢了解过去，预测未来
- ➢观察项目开发

软件度量的障碍

- 缺乏对软件度量的关注
  - Low priority except in emergencies
  - Need for immediate results (short time horizon)
  - Lack of resources for measurement/improvement
  - Multiple stakeholders (developer/support/product management)
  - Difficulty of comparison among projects, even earlier versions within the same project

实证软件工程

- 分析海量的开源/闭源软件数据，发现和总结软件制品、人员、工具、活动的特点及其所反映的软件工程实践效果，寻找软件开发规律

缺陷数据

- Gnome (517793), mozilla (620479), jboss (93149), apache (228870)

邮件数据

- Gnome (1343372), jboss(1052291), apache (6323416)

代码变迁数据

- gnome, mozilla, jboss, apache, netbeans，openjdk，opensolaris，symbian，andriod，debian，github，gitkernel，gitorious，savannash，repo.or.cz

Forge	Type	Files	File/Ver.	Unique File/Ver.	From
Large cmpny.	Var.	3,272K	12,585K	4,293K	1988
SourceForge	CVS	**26,095K**	81,239K	**39,550K**	1998
code.google	SVN	5,675K	14,368K	8,584K	1996
repo.or.cz	Git	2,519K	11,068K	5,115K	1986
Savannah	CVS	852K	3,623K	2,345K	**1985**
git.kernel.org	Git	12,974K	**97,585K**	856K	1988
OpenSolaris	Hg	77K	1,108K	91K	2003
FreeBSD	CVS	196K	360K	75K	1993
Kde	SVN	2,645K	10,162K	527K	1997
gnome.org	SVN	1,284K	3,981K	1,412K	1997
Gcc	SVN	3,758K	4,803K	395K	1989
Eclipse	CVS	729K	2,127K	575K	2001
OpenJDK	Hg	32K	747K	60K	2008

Digital Archeology

- **The study of developer cultures and behaviors through the *recovery, documentation and analysis of digital remains***
  - ***Tomography is image reconstruction from multiple projections***
  - **What is the reconstruction of developer behavior from the digital traces they leave in the code and elsewhere?**

Audris Mockus,
http://digitalarchaeology.info/

软件开发的投影(projection)

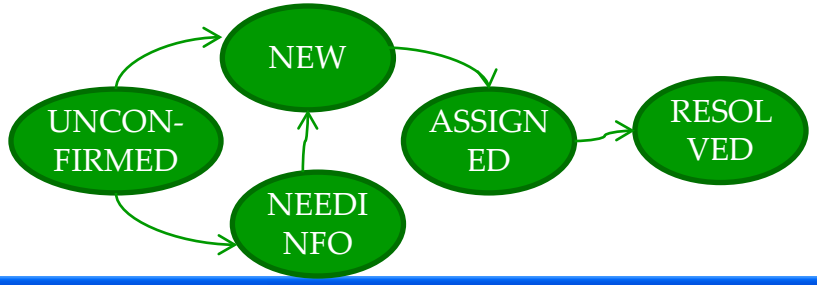
- Task tools: issue (MR) tracking systems
  - (e.g., Bugzilla, Jira, ClearQuest)
- Code tools: Version Control Systems
  - (e.g., CVS/SVN/Git/Mercurial/Bazaar/ClearCase)
- Other tools: communication, testing, organization
  - (e.g., wiki, mailing lists, coverage, pass rate, ...)

问题追踪系统(PROBLEM TRACKING SYSTEMS)中的投影

软件开发是解决问题 (MRs)

- **Stages**
  - Opened / Created
  - A user (or tester) complains
  - A new feature is started
  - A developer decides to change code
  - Assigned: a person is assigned / self-assigns to solve the task
  - Submitted (code changed) / NoChanged / ReAssigned
  - Verified
- **Example MR systems: Bugzilla**

MR format



OW2 Forge - Windows Internet Explorer

http://forge.ow2.org/tracker/?atid=100005&group_id=5&func=browse

Tracker: Bugs

Submit New | Browse | Admin | ExportToXml

Assignee: (?) Any | State: (?) Open | Category: (?) Any | Group: (?) Any

Order by: (?) ID | Ascending | Browse

Items 1 to 50

ID	Summary	Open Date	Assigned To	Submitted By	Status	Resolution	Priority
304411	Incorrect statistics on transactions	* 2005-12-21 09:51	David A. Egolf	Benoit PELLETIER	Open	None	1
304751	JMX services boot error in multiprotocol environnement jrmp/irmi with JDK 1.4	* 2006-03-08 18:09	Archit Shah	Benoit PELLETIER	Open	None	3
305269	JOnAS fires timers too often when running in clustered environment	* 2006-05-15 12:31	Philippe Durieux	Markus KARG	Open	Accepted	3
305291	All the beans of the application must be available at the same time	* 2006-05-17 07:31	Philippe Durieux	Xavier Bugaud	Open	Accepted	3
305418	XSLT Transformation crashing	* 2006-06-01 11:36	Guillaume Sauthier	Markus KARG	Open	None	3
305505	'run-as' not used when persistent timer are triggered after deployment	* 2006-06-15 12:00	Philippe Durieux	Xavier Bugaud	Open	Accepted	3
305605	[genic] GenIC Fails to update JAR due to URLClassLoader lock	* 2006-06-28 01:32	Guillaume Sauthier	Gary Moselen	Open	None	5
306023	[resource] jonas stop crashes if user tried to load RA twice	* 2006-08-23 15:24	Philippe Durieux	Markus KARG	Open	None	5
306141	JOnAS 4.7.6's GenIC enforces use of <key-jdbc-name> with single field primary key	* 2006-09-08 16:52	Philippe Durieux	Markus KARG	Open	Postponed	1
306267	Recurrence: Read-Only access produces Deadlock	* 2006-10-09 09:33	Philippe Durieux	Markus KARG	Open	Accepted	3
306289	ResourceException are not logging in the ConnectionManager.registerXAResource()	* 2006-10-17 07:47	Philippe Durieux	Benoit PELLETIER	Open	None	3
306294	joram server can not start in NON persistent mode if another was (in the same working directory)	* 2006-10-17 17:35	Benoit PELLETIER	SOUILLARD	Open	None	3
306303	DMQ objects are not visible in jonasAdmin	* 2006-10-20 08:59	Adriana Danes	Benoit PELLETIER	Open	Postponed	3
		* 2006-11-06				Works For	

Back to bug 412233

Who	When	What	Removed	Added
eitan@monotonous.org	2007-02-27 03:01:08 UTC	CC		eitan@ascender.com
		Assignee	lsr-maint@gnome.bugs	eitan@ascender.com
parente@cs.unc.edu	2007-03-21 00:22:49 UTC	Status	NEW	ASSIGNED
parente@cs.unc.edu	2007-03-30 23:55:35 UTC	Component	accerciser	general
		Product	lsr	Accerciser
eitan@monotonous.org	2007-04-06 20:37:31 UTC	Target Milestone	---	0.1.1
eitan@monotonous.org	2007-04-11 16:59:21 UTC	Blocks		422164
eitan@monotonous.org	2007-04-16 19:10:33 UTC	Target Milestone	0.1.1	0.1.2
eitan@monotonous.org	2007-05-08 22:36:21 UTC	Target Milestone	0.1.2	0.1.3
eitan@monotonous.org	2007-06-03 06:10:13 UTC	Target Milestone	0.1.3	0.1.4
eitan@monotonous.org	2007-07-05 19:40:47 UTC	Status	ASSIGNED	RESOLVED
		Resolution		FIXED

Back to bug 412233

版本控制系统(VERSION CONTROL SYSTEMS)中的投影

程序员通过changes开发软件

- All changes are recorded
- The product/code is simply a dynamic superposition of

Before:

```
int i = n;
while(i++)
    prinf(" %d", i--);
```

After:

```
//print n integers
int i = n;
while(i++ && i > 0)
    prinf(" %d", i--);
```

- one line deleted
- two lines added
- two lines unchanged
- Other attributes: date, developer, defect number, . . .
- Version Control System (VCS) track them, e.g., CVS/SVN/git

Commits format

```
------------------------------------------------------------------------
r14984 | benoitf | 2008-09-04 08:33:37 -0400 (Thu, 04 Sep 2008) | 1 line
Changed paths:
   M /jonas/trunk/jonas/modules/services/ejb/easybeans/src/main/config/easybeans-jonas.xml

Enable Remote JNDI Resolver
------------------------------------------------------------------------
r14983 | pelletib | 2008-09-04 05:44:14 -0400 (Thu, 04 Sep 2008) | 1 line
Changed paths:
   M /jonas/trunk/jonas_doc/src/docbook/doc-en/clustering/principles/management.xml

clustering guide: continue
------------------------------------------------------------------------
r14982 | loris | 2008-09-04 04:25:29 -0400 (Thu, 04 Sep 2008) | 1 line
Changed paths:
   D /jonas/branches/jonas-5.0
   D /jonas/branches/jonas-admin-layer

Unused branches
------------------------------------------------------------------------
r14981 | pelletib | 2008-09-03 12:10:12 -0400 (Wed, 03 Sep 2008) | 1 line
Changed paths:
   M /jonas/trunk/jonas_doc/src/docbook/doc-en/clustering/principles/management.xml
   M /jonas/trunk/jonas_doc/src/resources/images/clustering/clustersolution.svg
   A /jonas/trunk/jonas_doc/src/resources/images/clustering/domainmngt.png
   A /jonas/trunk/jonas_doc/src/resources/images/clustering/domainmngt.svg

clustering guide: continue
------------------------------------------------------------------------
```

其他的一些开发历史纪录

- 商业公司经常使用的系统
  - Sales/Marketing: customer information, customer rating, customer purchase patters, customer needs: features and quality
  - Accounting: Customer/system/software billing information and
  - maintenance support level
  - Maintenance support: Currently installed system, support level
  - Field support: dispatching repair people, replacement parts
  - Call center support: customer call/problem tracking
  - Development field support: software related customer problem tracking, installed patch tracking
  - Development: feature and development, testing, and field defect tracking, software change and software build tracking

数字考古简史

人是影响项目成功，以及变化性最大的因素

- Sackman et al, 1968, 28:1
- Curtis, 1981, 23:1
- Boehm, 1981

缺乏数据，难以研究

“Until the many sources of variation among individuals have been compared in the same set of data, it will not be possible to determine precisely … the most important predictor of success”
- Curtis, 1984

数据积累

实证研究出现

“By using …source code change history and problem reports we quantify aspects of developer participation, …”

–Mockus, 2000

“Some fundamental questions can be answered only by considering the entire universe of publicly available source code and its history”

非技术时代

软件工程

海量数据

Internet时代:
Crowdsourcing

1950’s

1984

2000

2010

Time

提纲

- 软件工程的历史
  - 研究对象
    - 从技术性(technical) 到社会性(social)
  - 研究方法
    - 从软件度量 (Software Measurement) 到数字考古(Digital Archeology)

- 数字考古的挑战

数字考古学

- 理解人们如何生产软件，以提高效率和质量
  - 逐步增加精度和范围

范围	研究问题
个体	成熟度，生产效率
项目	技能，协作，学习
社区/组织	动力，创造力，持久
社会	真理，美，知识

研究方法的主要步骤

- 选择一个研究问题(软件开发中的一个现象)
  - Select a phenomenon for study
- 在一个小规模范围内，观察和验证此现象是如何被记录在数字制品中的
  - Observe and validate on a smaller scale how it projects onto digital artifacts
- 设计和验证投影(实际开发到数字制品)的模型
  - Design and validate models of the projection
- 将投影模型应用到大量项目，重构现象及其影响
  - Apply the suitable tomography method on the entire population to reconstruct the phenomenon and its impact
- 基于验证过的底层重构，构建更高层的软件开发概念、模型和方法
  - Build higher-level concepts of software development based on the validated lower-level reconstructions

目前的一些研究

- **IBM Watson:**
  - A collaborative work environment for developers, architects and project managers
- **CMU:**
  - Socio-technical congruence
- **UBC:**
  - Coordination and requirements
- **Avaya:**
  - Organizational change

目前的挑战

- 数据？big data
  - 数据抓取，过滤，分析
- 观念的改变
  - 人们仍然聚焦于技术因素
  - 人的因素 (human factor)终究不可控?
- 如何度量?
  - 动力，智力，环境，如何度量?

公共数据池

- 建立一个大样本的软件项目数据池，回答软件工程，甚至是社会学、组织学中的经典问题
  - 数据丰富： Internet上无数的开源项目
  - 数据开放：可以自由地采用自动脚本获取
- 物理服务器
  - Mem/Proc.: R910(4U), 64GbRAM, 16-cores X7550
  - DELL MD3200, 12*2TB SAS
- 逻辑数据层次
  - Level-0： 原始数据
  - Level-1： 为快速检索和导出，从Level-0中抽取的数据
  - Level-2： 以Level-1为基础，根据研究需要析取的数据
- http://passion-lab.org

最难是观念的改变

- Software science?

总结

- 构造软件的挑战仍然存在
- 软件工程的目标仍然是生产效率和质量
- 软件工程的研究对象呈现社会性/human factor
  - Social thinking forms a new basis for software engineering
  - Technical challenges remain, but must be solved with social context in mind
- 软件工程大数据提供了方法
  - Digital Archeology

Reference

- DIJKSTRA, E. W. The humble programmer. Commun. ACM 15, 10 (1972), 859–866.
- J. D. Couger and R. A. Zawacki. Motivating and Managing Computer Personnel. John Wiley & Sons, Inc., New York, NY, USA, 1980.
- B. Boehm. Software Engineering Economics. Prentice-Hall, 1981.
- B. Curtis. Fifteen years of psychology in software engineering: Individual differences & cognitive science. In ICSE'84, pp 97–106, 1984.
- P. Robillard, "The role of knowledge in software development," Communications of the ACM, vol. 42, no. 1, pp. 87–92, 1999.
- A. Mockus, R. T. Fielding, and J. Herbsleb, "A Case Study of Open Source Development: The Apache Server", 22nd ICSE, pp. 263-272, Limerick, Ireland, June 4-1, 2000.
- D. Cubranic and G. Murphy. Hipikat: A project memory for software development. TSE, 31(6), 2005.
- M. Cataldo, P. Wagstrom, J. Herbsleb, and K. Carley. Identification of coordination requirements: Implications for the design of collaboration and awareness tools. In Conference on Computer Supported Cooperative Work CSCW'06, Banff, Alberta, Canada, 2006.
- A. Mockus. Software support tools and experimental work. In V. Basili and et al, editors, Empirical Software Engineering Issues: Critical Assessments and Future Directions, volume LNCS 4336, pages 91–99. Springer, 2007.

Reference

- C. R. de Souza, S. Quirk, E. Trainer, and D. F. Redmiles. Supporting collaborative software development through the visualization of socio-technical dependencies. In GROUP '07, pages 147–156, New York, NY, USA, 2007. ACM.
- A. Begel and B. Simon. Novice software developers, all over again. In International Computing Education Research Workshop, Sydney, Australia., 2008.
- A. Mockus. Amassing and indexing a large sample of version control systems: towards the census of public source code history. In 6th IEEE Working Conference on Mining Software Repositories, May 16–17 2009.
- M. Zhou and A. Mockus. Developer fluency:Achieving true mastery in software projects. In ACMSIGSOFT/FSE 18, Santa Fe, New Mexico, Nov 7-11,2010, pp 137-146.
- M. Zhou and A. Mockus. Growth of Newcomer Competence: Challenges of Globalization. In FoSER(future of software engineering) on ACMSIGSOFT / FSE, Santa Fe, New Mexico, Nov 7-8, 2010, pp443-448.
- M. Zhou and A. Mockus. Does the initial environment impact the future of developers? ICSE 2011, Honolulu, Havaii, May 21-28, 2011, pp71-80.
- Clay Williams, IBM, : The social side of software engineering
- Audris Mockus. http://digitalarchaeology.info/