# （软件数据分析组）研究方向介绍

## 关于领域名称

我们的研究领域是软件工程（software engineering）。鉴于我们研究的很多问题不仅是工程问题，我不确定称这个名是否合适。但是，我们的工作单位，我们的发表venue，都在软件工程范畴。

又因为我们所关注的问题和方法的特性，我们的研究领域有一些更为特定的名称，例如：

数据驱动的软件开发（data-driven software development），

软件仓库挖掘（mining software repositories），

软件度量（software measurement），

实证/经验软件工程（empirical software engineering），

软件数字考古学（software digital archeology），

软件数字社会学（software digital sociology）。

但使用哪个名字应该无妨，不影响我们对本质问题的探索（但也许影响交流和传播）。

## 目标

* 观察和度量大规模软件项目/生态中人们的开发行为，量化开发者与其环境和软件制品之间的关系，以期理解和掌握控制大型复杂软件系统的方法。
* 研究和量化影响个体/项目/社区成功的因素，建立推荐工具，对软件开发各种实践活动进行预测和推荐。

注：鉴于开源项目的广泛存在及其数据的可获取性，我们的研究对象大多是开源项目及社区。我们面向产业项目大多做的是咨询工作，企业一般不会允许公开其数据，但产业咨询可以丰富我们的洞察力并应用我们的研究结果。

## 研究方法

* 目前主要是利用软件开发活动数据，例如由下述软件开发支持工具产生的数据：
* Version control system (cvs, svn, git, hg)
* Issue tracking system (jira, bugzilla)
* Email archives
* Forums ……
* 使用的挖掘和分析技术包括：

Raw data process: shell scripts, perl, python

Data analysis: R language, Machine learning…

Qualitative analysis: Conducting interviews, surveys, grounded theory …

## 研究问题

任何有趣的问题，但需要具有下述三个特性：

* new
* relevant（目前我们还在软件工程领域，因此任何研究问题都需要与提高开发效率和软件质量相关）
* non-trivial（我们不能只是解决一个人或几个人或几十个的问题，即，我们的问题及solution需要有一般性）

## 资源

###### 1，中文的领域科普参见下文，

* 周明辉,郭长国. 基于大数据的软件工程新思维.中国计算机学会通讯.第10卷,第3期. 2014年3月
* 周明辉,张伟,尹刚. 开源软件的量化分析.中国计算机学会通讯.第12卷,第2期. 2016年2月

###### 2，英文有相当多一组文章要读，

先从下述列表开始：使用数据explore开源开发和全球分布式开发是本领域的pioneer work（也是读起来有美感的论文），需要尤其关注其研究问题和研究方法及其跟软件开发效率和质量的关系：

A. Mockus, R. T. Fielding, and J. Herbsleb. Two case studies of open source software development: Apache and Mozilla. ACM Transactions on Software Engineering and Methodology, 11(3):1–38, July 2002.

J. D. Herbsleb and A. Mockus. An empirical study of speed and communication in globally-distributed software development. IEEE Transactions on Software Engineering, 29(6):481–494, June 2003.

A. Mockus and D. M. Weiss. Globalization by chunking: a quantitative approach. IEEE Software, 18(2):30–37, March 2001.

其次，我们组（目前）的代表工作在于研究程序员、制品及环境的关系（跟我个人意趣有关）。研究对象从个体(成熟度、学习途径、环境影响)到群体(项目最佳实践)，到生态(可持续性、可演化性)，列举几篇（觉着读起来有美感的）：

M. Zhou and A. Mockus. Developer fluency: Achieving true mastery in software projects. In ACMSIGSOFT/FSE, Nov 7-11, 2010, pp137-146.

Minghui Zhou, Audris Mockus. What make long term contributors: willingness and opportunity in OSS community, 2012 International Conference on Software Engineering (ICSE 2012), 2012/6/2-2012/6/9, pp518-528.

Jialiang Xie, Minghui Zhou, Audris Mockus. Impact of Triage: a Study of Mozilla and Gnome. ESEM 2013 (Empirical Software Engineering and Measurement). Oct 7-11 2013. pp247-250.

Minghui Zhou, Qingying Chen, Audris Mockus, and Fengguang Wu. On the scalability of Linux kernel maintainers' work. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017). pp27-37.

我们组另外一部分工作侧重于代码研究，涉及代码克隆、代码搜索（代码与自然语言的转换）、代码生成等。可以阅读的工作：

陈峰宏. 大规模开源代码复用检测及度量工具的设计与实现. 北大硕士论文. 2013

Qingying Chen and Minghui Zhou. A Neural Framework for Retrieval and Summarization of Source Code. ASE’18

###### 3，关于数据集。

我们可以拿到的（软件开发活动）数据是非常多的，例如互联网上成千上万开源项目的数据都是开放可获取的。一般来说，开源项目都会支持在线问题追踪系统，版本控制系统和邮件列表等。

例如，Mozilla的在线issue tracking system可以参看：

<https://bugzilla.mozilla.org/show_bug.cgi?id=1000000>

<https://bugzilla.mozilla.org/show_bug.cgi?id=1450000>1

例如，Linux kernel的在线代码库(mainline code repository)可以参看：

<https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git>

例如，Linux kernel的在线邮件列表(mailing-list)可以参看：

<https://lkml.org/>

我们一般都是从上述在线资源下载数据进行分析和研究，下文举几个例子。

（1）Issue tracking data，即项目中的issue报告和处理流程数据。我们有两份由mozilla社区官方提供的Mozilla Bugzilla dump (分别在2013年和2016年提供)，原始数据如下：

* + pae:/store1/bug/MozDump/Mozilla-Bugzilla-Public-02-January-2013.sql
  + pae:/store/bug//mozilladump20160304/research\_bugzilla\_2016-03-04.sql.gz
  + 网上下载地址参见：<https://zenodo.org/record/971162#.Wcb12BfadCU>
* 有关数据本身的处理流程，可以参见：Jiaxin Zhu, Minghui Zhou, and Hong Mei. Multi-extract and multi-level dataset of mozilla issue tracking history. In Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16). ACM, New York, NY, USA, 472-475.
* 有关如何使用这个数据集做研究，可以参见：Minghui Zhou, Audris Mockus: Who Will Stay in the FLOSS Community? Modeling Participant's Initial Behavior. IEEE Transactions on Software Engineering, vol.41, no.1, pp.82-99, Jan. 1 2015.

（2）commits data，即项目开发过程中的代码提交日志。基本上任何一个开源项目都可以先clone到本地然后获取其commit log。下面以linux kernel为例描述日志获取过程：

$git clone git://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git

$cd linux

linux$git log --numstat

--pretty=format:"STARTOFTHECOMMIT%n%H;%an;%ae;%ad;%cn;%ce;%cd;%s" > log.linux linux$perl extrgit.perl <log.linux >linux.l1

linux$cat linux.l1 | while read

do perl -ane 'use Time::ParseDate qw(parsedate); ($rev,$aname,$cname,$alogin,$clogin,$nadd,$atime,$ctime,$f,$cmt)=split(/\;/,$\_,-1); $at=parsedate("$atime");$ct=parsedate("$ctime"); print "$rev\;$aname\;$cname\;$alogin\;$clogin\;$nadd\;$at\;$ct\;$f\;$cmt"';

done > linux.l2

########extrgit.perl#######################

use strict;

#Extract each revision from FreeBSD-release/10.2.0 log output. -zhouminghui

use Time::Local;

my %paths = ();

my ($rev, $aname, $alogin, $atime, $cname, $clogin, $ctime, $comment) = ("","","","","","","","","","");

my ($getHeader, $getPaths) = (0, 0);

sub output {

foreach my $f (keys %paths){

$comment =~ s/\r/ /g;

$comment =~ s/\;/SEMICOLON/g;

$comment =~ s/\n/\_\_NEWLINE\_\_/g;

print "$rev\;$aname\;$cname\;$alogin\;$clogin\;$paths{$f}\;$atime\;$ctime\;$f\;$comment\n";

}

%paths = ();

($rev, $aname, $alogin, $atime, $cname, $clogin, $ctime, $comment) = ("","","","","","","","","","");}

while(<STDIN>){

chop ();

#catch end of last revision information

if (/^STARTOFTHECOMMIT$/){

&output ();

$getHeader=1;$getPaths=0;

next;

}

#process file header

if ($getHeader){

$\_ =~ s/ \| /\|/g;

($rev, $aname, $alogin, $atime, $cname, $clogin, $ctime, $comment) = split(/\;/, $\_, -1);

$getHeader = 0;

$getPaths = 1;

next;

#print STDERR "getHeader $rev, $login, $date, $line\n";

}

if ($getPaths && /^$/){

next;

}

if ($getPaths && /^[0-9]/){

/(\d+)\s+(\d+)\s+(.\*)$/;

my ($nadd, $ndel, $path) = ($1, $2, $3);

$paths{$path}="$nadd:$ndel";

}

}

&output ();

#########################################################

（3）Linux kernel的patch data:

<https://zenodo.org/record/1165576#.WsGWPkxuKUk>

论文参见：Yulin Xu and Minghui Zhou. A Multi-level Dataset of Linux Kernel Patchwork. In MSR '18: MSR '18: 15th International Conference on Mining Software Repositories , May 28–29, 2018, Gothenburg, Sweden. ACM, New York, NY, USA.

4，关于开源相关的资源。

* 北卡罗来纳大学维护的开源资源的主页（https://www.ncsu.edu/it/open\_source/），分门别类地索引了开源操作系统、开源应用软件、开源中间件、开源硬件、开源课程、开源出版等资源，为学生、科研与教育工作者们提供了一个利用开源的入口。
* Schoolforge联盟是一个旨在联合开发、推广开源教育资源的组织的联合体，其所搭建的schoolforge.net（<https://schoolforge.net/）平台集合了面向教育的各类开源软件、开源课程、开源教材等资源，为开源教育提供了丰富资源>。
* Datahouse.trustie.org是我们提供数据资源的一个平台，以后还会陆续放置相关的开源资源和服务。

1. 关于R语言和Shell脚本。

我们的论文基本都提供数据集和脚本，对于数据前期处理的脚本基本一样。

例如，git数据的前期处理可参见：https://github.com/minghuizhou/maintainerAnalysis，issue data的前期处理可参见：<https://zenodo.org/record/971162#.Wcb12BfadCU>

====================

ZhouMinghui 2018-07-26