Course Organization

Preface

Rui ZHANG

Jilin University, China

September 19, 2023

Contents

Contents

Faculty Members

- Fausto Giunchiglia
  Homepage

Faculty Members

- Fausto Giunchiglia
  Homepage
- Simone Bocca

Faculty Members

- Fausto Giunchiglia
  Homepage
- Simone Bocca
- Mayukh Bagchi

Faculty Members

- Fausto Giunchiglia
  Homepage
- Simone Bocca
- Mayukh Bagchi
- Amarsanaa Ganbold

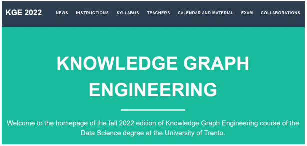
Faculty Members

- Fausto Giunchiglia
  Homepage
- Simone Bocca
- Mayukh Bagchi
- Amarsanaa Ganbold
- Rui ZHANG

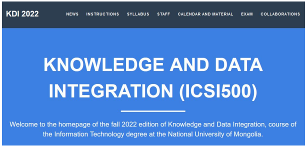
Faculty Members

- Fausto Giunchiglia Homepage
- Simone Bocca
- Mayukh Bagchi
- Amarsanaa Ganbold
- Rui ZHANG

Course Web Site Unitn



Course Web Site NUM



KDI JLU Resources

邀请码: 95716965

APP首页右上角输入

该邀请码2024年03月17日前有效

KDI-2023

- Course site at https://mooc1.chaoxing.com/course/228885246.html

KDI JLU Resources

群聊：KDI-JLU-2003

该二维码7天内(9月26日前)有效，重新进入将更新

- Course site at https://mooc1.chaoxing.com/course/228885246.html
- Wechat group by barcode...

Contents

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...
- Data is fundamental in real-world applications...

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...
- Data is fundamental in real-world applications...
  - Why cannot we publish a paper entitled as '*A Library Management System*'?

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...
- Data is fundamental in real-world applications...
  - Why cannot we publish a paper entitled as '*A Library Management System*'?
  - What if buidu.com does not answer your query properly?

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...
- Data is fundamental in real-world applications...
  - Why cannot we publish a paper entitled as '*A Library Management System*'?
  - What if buidu.com does not answer your query properly?
- Data diversity is everywhere...

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...
- Data is fundamental in real-world applications...
  - Why cannot we publish a paper entitled as '*A Library Management System*'?
  - What if buidu.com does not answer your query properly?
- Data diversity is everywhere...

Q: Can we build an application on a SINGLE data source?

Ubiquitous Data Diversity

- Big Data is already there...
  - 5 V's from domains like finance (e-Business), biomedicine, transportation, search engine...
  - Heterogeneity even for data sources from the same domain...
- Data is fundamental in real-world applications...
  - Why cannot we publish a paper entitled as '*A Library Management System*'?
  - What if buidu.com does not answer your query properly?
- Data diversity is everywhere...

Q: Can we build an application on a SINGLE data source?
A: Yes so far, but soon unlikely, if not impossible.

Feature or Bug?

It's Not a Bug, It's a Feature: How Misclassification Impacts Bug Prediction

Kim Herzig
Saarland University
Saarbrücken, Germany
herzig@cs.uni-saarland.de

Sascha Just
Saarland University
Saarbrücken, Germany
just@st.cs.uni-saarland.de

Andreas Zeller
Saarland University
Saarbrücken, Germany
zeller@cs.uni-saarland.de

***Abstract*—In a manual examination of more than 7,000 issue reports from the bug databases of five open-source projects, we found 33.8% of all bug reports to be *misclassified*—that is, rather than referring to a code fix, they resulted in a new feature, an update to documentation, or an internal refactoring. This misclassification introduces *bias* in bug prediction models, confusing bugs and features: On average, 39% of files marked as defective actually never had a bug. We discuss the impact of this misclassification on earlier studies and recommend manual data validation for future studies.**

***Index Terms*—Mining software repositories, bug reports, data quality, noise, bias**

I. Introduction

In empirical software engineering, it has become commonplace to mine data from change and bug databases to detect where bugs have occurred in the past, or to predict where they will occur in the future. The accuracy of such measurements and predictions depends on the *quality of the data.* Therefore, mining software archives must take appropriate steps to assure data quality.

A general challenge in mining is to separate *bugs from non-bugs*. In a bug database, the majority of issue reports are classified as *bugs*—that is, requests for corrective code maintenance. However, an issue report may refer to "perfective and adaptive maintenance, refactoring, discussions, requests for help, and so on" [1]—that is, activities that are unrelated to errors in the code, and would therefore be classified in a non-bug category. If one wants to mine code history to locate or predict error prone code regions, one would therefore only consider issue reports classified as bugs. Such filtering needs nothing more than a simple database query.

However, all this assumes that the category of the issue report is accurate. In 2008, Antoniol et al. [1] raised the problem of *misclassified* issue reports—that is, reports classified as *bugs*, but actually referring to *non-bug issues*. If such mix-ups (which mostly stem from issue reporters and developers interpreting "bug" differently) occurred frequently and systematically they would introduce *bias* in data mining models threatening the external validity of any study that builds on such data: Predicting the most error-prone files, for instance, may actually yield files most prone to new features. But how often does such misclassification occur? And does it actually bias analysis and prediction?

TABLE I
Project details.

	Maintainer	Tracker type	# reports
HTTPClient	APACHE	Jira	746
Jackrabbit	APACHE	Jira	2,402
Lucene-Java	APACHE	Jira	2,443
Rhino	MOZILLA	Bugzilla	1,226
Tomcat5	APACHE	Bugzilla	584

These are the questions we address in this paper. From five open source projects (Section II), we manually classified more than 7,000 issue reports into a fixed set of issue report categories clearly distinguishing the kind of maintenance work required to resolve the task (Section III). Our findings indicate substantial data quality issues:

Issue report classifications are unreliable. In the five bug databases investigated, more than 40% of issue reports are inaccurately classified (Section IV)

Every third bug is not a bug. 33.8% of all bug reports do not refer to corrective code maintenance (Section V).

After discussing the possible sources of these misclassifications (Section VI), we turn to the consequences. We find that the validity of studies regarding the distribution and prediction of bugs in code is threatened:

Files are wrongly marked as fixed. Due to misclassifications, 39% of files marked as defective actually have never had a bug (Section VII).

Files are wrongly marked to be error-prone. Between 16% and 40% of the top 10% most defect-prone files do not belong in this category after reclassification (Section VIII).

Section IX details studies affected and unaffected by these issues. After discussing related work (Section X) and threats to validity (Section XI), we close with conclusion and consequences (Section XII).

II. Study Subjects

We conducted our study on five open-source JAVA projects described in Table I. We aimed to select projects that were under active development and were developed by teams that follow strict commit and bug fixing procedures similar to industry. We also aimed to have a more or less homogeneous

978-1-4673-3074-9/13/$31.00 © 2013 IEEE ICSE 2013, San Francisco, CA, USA

Feature or Bug?

- Data diversity is a feature of the Big Data era, but NOT a bug harmful only.

Feature or Bug?

- Data diversity is a feature of the Big Data era, but NOT a bug harmful only.
- Therefore, the proper attitude is to take the advantages.

Feature or Bug?

- Data diversity is a feature of the Big Data era, but NOT a bug harmful only.
- Therefore, the proper attitude is to take the advantages.
- HOW?

What to get...

- What are Knowledge Graphs (KGs).
- What Knowledge Graphs can be used for, and example of already used KGs.
- What does it means to build a KG.
- How to solve the different problems involved in KG construction, using the iTelos KGE methodology.
- How to use new and existing tools and libraries to address the problems encounterd in KGs construction.
- How to develop an entire project of KGE on real-world case studies.

Contents

Abilities preferred...

- Open mind
  - Motivated
  - Collaboration
  - Internationalization

Abilities preferred...

- Open mind
  - Motivated
  - Collaboration
  - Internationalization
- Skills
  - Data management: basic coding skills in python and/or java/javascript
  - Databases modeling: ER modeling, (Ontology modeling if possible, Ontology definition desirable via web languages mainly as RDF and OWL)

International Communication

- What is communication?

International Communication

- What is communication?
- Why to communicate internationally?

International Communication

- What is communication?
- Why to communicate internationally?
- Who speaks louder?

International Communication

- What is communication?
- Why to communicate internationally?
- Who speaks louder?

It is vital to understand what is the top and to promote collaboration.

Coding Skills

> "I will do research but NOT coding."
>
> —Someone

Coding Skills

Coding Skills

Code eases the life as it...

- checks results...
- verifies ideas...
- explores assumptions...

Coding Skills

"Coding builds confidence."

5 Excellent Ways to Improve ...

Motivation

- Interest makes good motivation...

Motivation

- Interest makes good motivation...
- Practical requirements...

Motivation

- Interest makes good motivation...
- Practical requirements...
- Objective necessities...

Contents

Theory

1. Diversity in Data (and Knowledge)

2. Knowledge Graph for Modeling

3. Purpose Oriented Data Integration Pipeline

Theory

1. Diversity in Data (and Knowledge)
    - Different levels of diversity...
    - Strategies to handle diversities
2. Knowledge Graph for Modeling

3. Purpose Oriented Data Integration Pipeline

Theory

1. Diversity in Data (and Knowledge)
   - Different levels of diversity...
   - Strategies to handle diversities
2. Knowledge Graph for Modeling
   - Conceptual modeling
   - Necessity of reuse
3. Purpose Oriented Data Integration Pipeline

Theory

1. Diversity in Data (and Knowledge)
   - Different levels of diversity...
   - Strategies to handle diversities
2. Knowledge Graph for Modeling
   - Conceptual modeling
   - Necessity of reuse
3. Purpose Oriented Data Integration Pipeline
   - Purpose clarification
   - Entity relationship modeling
   - Schema modeling
   - Mapping from data to knowledge

Practice

1. Data Preparation
2. Common Knowledge Reuse
3. Modeling
4. Integration

Practice

1. Data Preparation
   - Collection
   - Laundry
2. Common Knowledge Reuse
3. Modeling
4. Integration

Practice

1. Data Preparation
   - Collection
   - Laundry
2. Common Knowledge Reuse
   - Teleology
   - Schema Overlap
3. Modeling
4. Integration

Practice

1. Data Preparation
   - Collection
   - Laundry
2. Common Knowledge Reuse
   - Teleology
   - Schema Overlap
3. Modeling
   - Informal and Formal modeling
   - Purpose, Data source, and Knowledge source.
4. Integration

Practice

1. Data Preparation
   - Collection
   - Laundry
2. Common Knowledge Reuse
   - Teleology
   - Schema Overlap
3. Modeling
   - Informal and Formal modeling
   - Purpose, Data source, and Knowledge source.
4. Integration
   - Semantic Matching/Mapping
   - Individual Population into Knowledge Graph

Contents

Teamwork

We need to build teams of 3 students to complete the following...

Teamwork

We need to build teams of 3 students to complete the following...

1. Organization
   - Roles: project manager, knowledge engineer, data scientist...
   - Hot backups

Teamwork

We need to build teams of 3 students to complete the following...

1. Organization
   - Roles: project manager, knowledge engineer, data scientist...
   - Hot backups
2. Work
   - Weekly evaluation
   - Stage document

Teamwork

We need to build teams of 3 students to complete the following...

1. Organization
   - Roles: project manager, knowledge engineer, data scientist...
   - Hot backups
2. Work
   - Weekly evaluation
   - Stage document
3. Presentation
   - Result
   - Style

Personal Assignment

- to read ≥ 1 related article (and reference it in the final presentation).
- to complete ≥ 1 share (role) of the project work.
- to take charge of ≥ 1 related document.

Scale

50% Midterm Presentation

50% Final Presentation

Contents

Summary

In this lecture we discussed:

- The preparations for the course.
- The emphasis of teamwork in KDI.
- The expected output and gain of KDI.

Thanks!