Mining Software Engineering Repositories

Vincent Leung
Pat McGowan
Pascal Turmel

What is Repository Mining?

- Gaining valuable data from software repositories

- A form of mass data mining

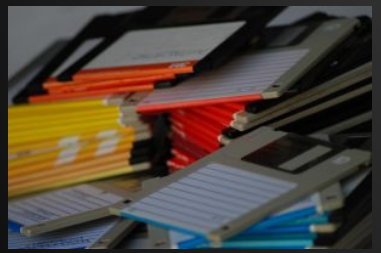
Forms of Repositories

- Source Control Repositories
  - Version Control Systems
- Issue/Bug repositories
  - Systems that are used to keep track of reported issues/bugs
- Code repositories
  - Where source files are usually stored
  - Open-sourced project hosting servers

What is data mining?

- Gaining useful information through analyzing data by applying algorithms

- Used to detect fraud, minimize risk, anticipate demand etc.



Why do it?

- By mining data, it is possible to gain new information about large systems
- By applying algorithms to the data, it is possible to gain or predict information about a specific project or area of research
- Can be used to predict source code changes by mining change history
- Can facilitate issue resolution by comparing multiple software solutions

Benefits

- Can gain knowledge about a system by analyzing it
- May find useful information in the bug/issue tracking repositories
- Analyzing information about the software environment and placing more personnel to work on specific parts of a project

How do we do it?

- **Manually Farm Repos**
  - Ex. going into individual github acounts.
- **Automated Scripts**
  - Ex. The script on the right-->
- **Pre-compiled Programs**
  - Ex. the next slide

```
#!/bin/bash

while read user; do
 mkdir $user
 curl https://api.github.com/users/$user/repos >$user/repos
 for user_repo in $(grep \"name\" $user/repos |awk -F "" '{print $4}'); do
  mkdir -p $user/$user_repo &&\
  git clone git@github.com:$user/$user_repo $user/$user_repo
 done
 rm $user/repos
done <users
```

Pre-Compiled Programs

- Pre-Packaged Programs for Data Mining
  - weka, rapidminer, Columbus, MetricsGrimoire, LibreSoft
- Software Analytics
  - PMD, XIAO, Hadoop, Vertica

Data Discovery

Discovering data comes in 2 forms:

Description and Prediction.

Some methods used to detect patterns are:

- Anomaly detection
  - Check for notable differences. Data that differ from known patterns.
- Association learning
  - ex. Ebay and Amazon takes browsing data and uses it to recommend you products
- Cluster detection
  - used to group data findings into categories
- Classification
  - if an existing structure is already defined, it can be used to categorize them into the pre-determined categories.
- Regression
  - Used to create a model with data to be able to predict future behaviours

References

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388/

http://www.slideshare.net/herraiz/20100618-daniel-uah

http://dailytruthbase.blogspot.ca/2011/09/genesis-29-31-family-friction-with.html

http://github.com

https://soundcloud.com/snare-fetischisten-radio/snare-fetisch