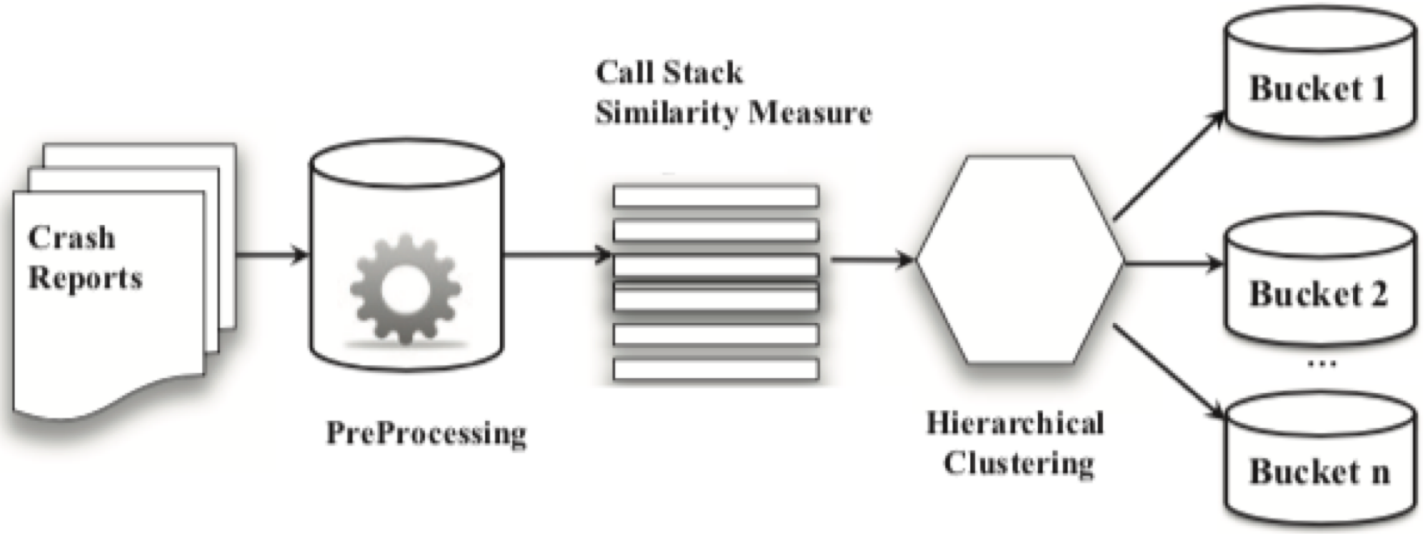
Cluster Duplicate Bug Reports

第二组

龚玥 张兆旭 周兴友 吴江宁 唐德轩

11611908@sustc.mail.edu.cn

基本流程



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/rebucket-icse2012.pdf

数据集构建

1. 提取Traceback信息： 正则表达式匹配

2. 预处理： 去除Immune functions. E.g. Native Method in Java library

3. 预处理： 去除Recursive function.

难点

1. 缺乏有标注的新数据

2. Duplicate的定义？如何定义bug的颗粒度：
   同一行？ 同一个函数？同一个包？

3. 人工标注duplicate的准确率有多少？难以review
模型训练出来的cluster的准确度

衡量相似度

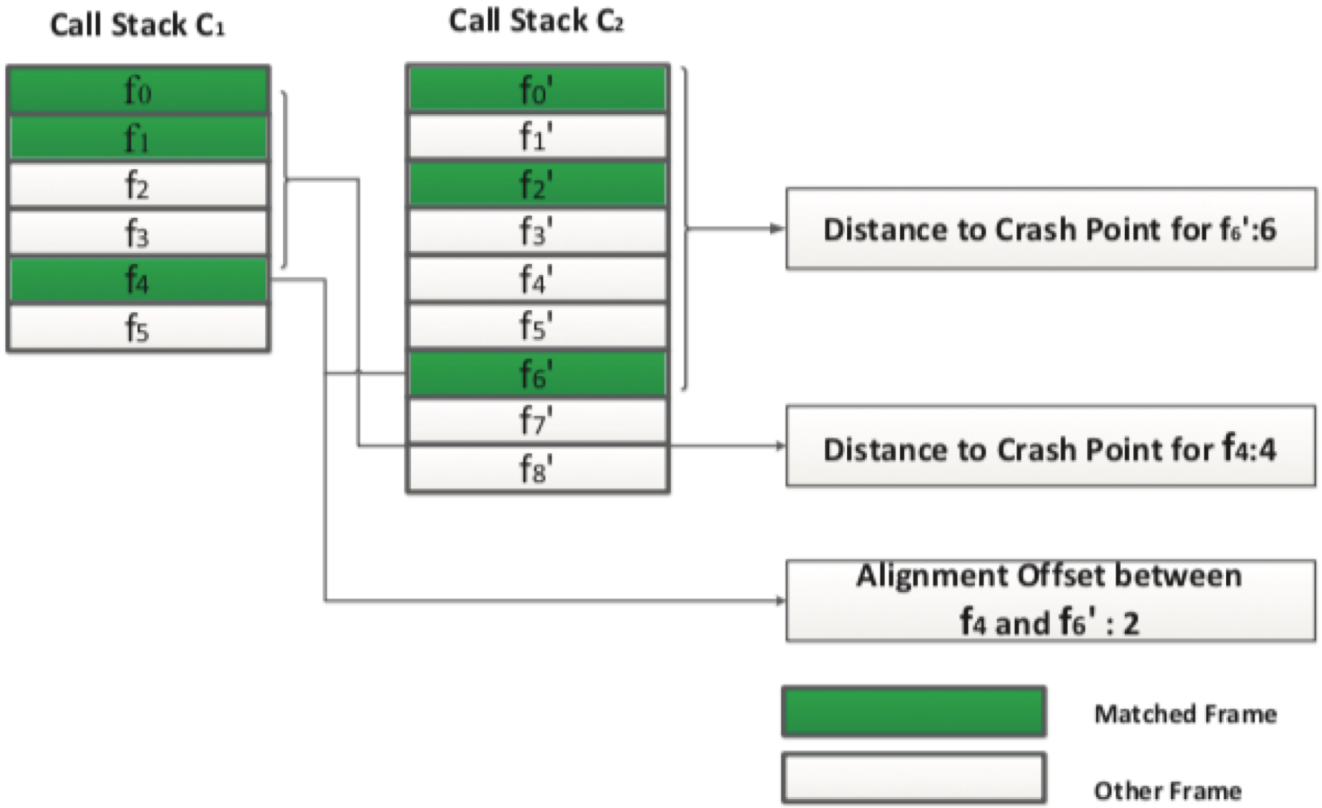
- DURFEX: A Feature Extraction Technique for Efficient Detection of Duplicate Bug Reports
- ReBucket – A Method for Clustering Duplicate Crash Reports based on Call Stack Similarity

难点

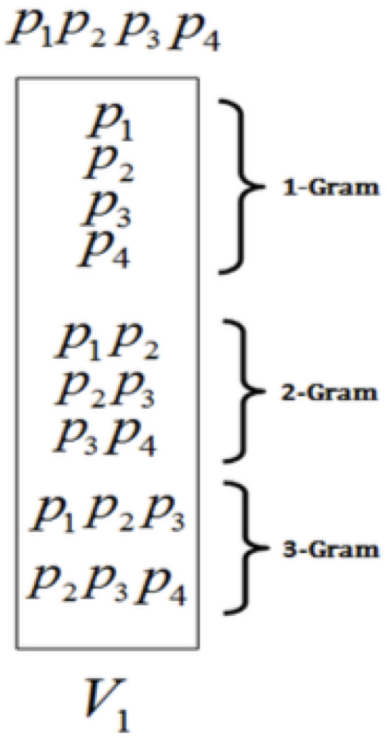
1. Scalability

2. 参数训练

Rebucket


Call Stack C1
f0
f1
f2
f3
f4
f5
Call Stack C2
f0'
f1'
f2'
f3'
f4'
f5'
f6'
f7'
f8'
Distance to Crash Point for f6':6
Distance to Crash Point for f4:4
Alignment Offset between
f4 and f6' : 2
Matched Frame
Other Frame

DURFEX



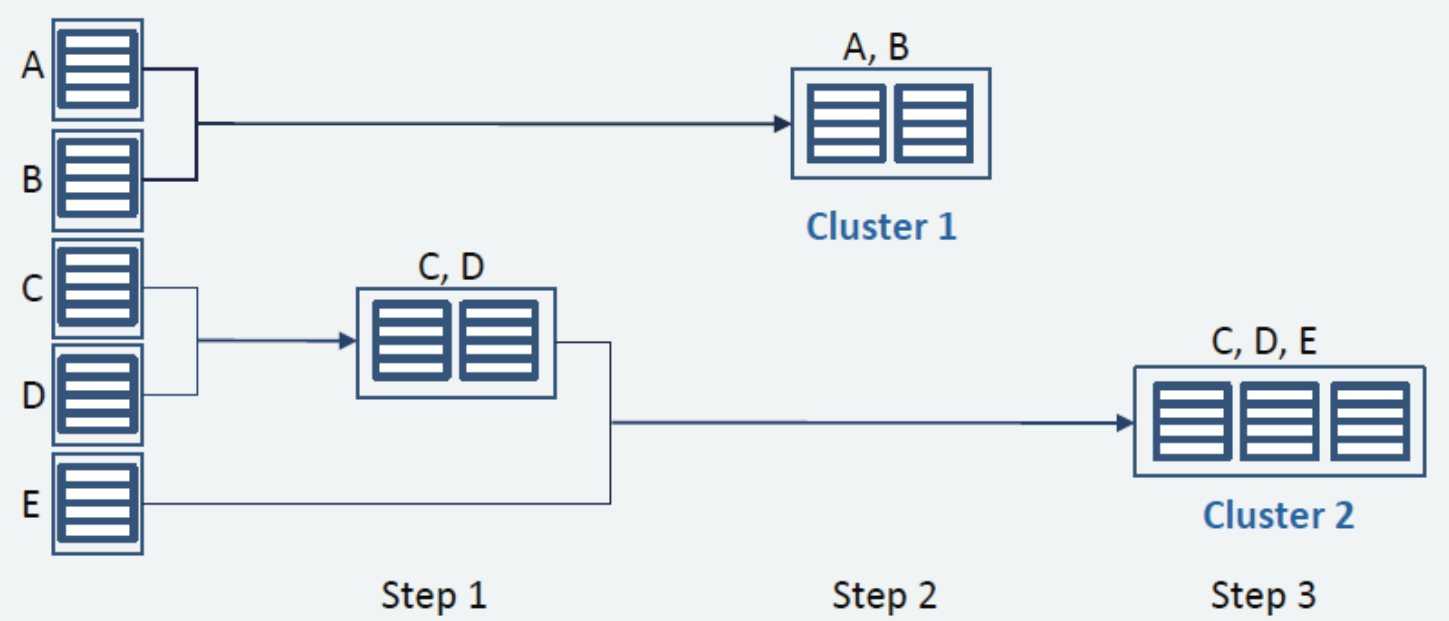
Typically, given $V_1 =< w_{11}, w_{12}, \ldots\ldots w_{1n} >$ and $V_2 =< w_{21}, w_{22}, \ldots\ldots w_{2n}$, the cosine similarity is calculated as in [9]:

$$Cos(\theta) = \frac{V_1.V_2}{|V_1|.|V_2|} \quad (5)$$

$$\phi_{tf.idf}(p,T,\Gamma) = \frac{\mathrm{K}}{df(p_i)} freq(p_i); i = 1, \ldots, m$$

Cluster

Agglomerative Hierarchical clustering technique



分工(暂定）

1. 数据集构建：吴江宁

2. 复现DUPFEX: 龚玥 唐德轩 周兴友

3. 优化Rebucket: 张兆旭

Thank you!
Q & A