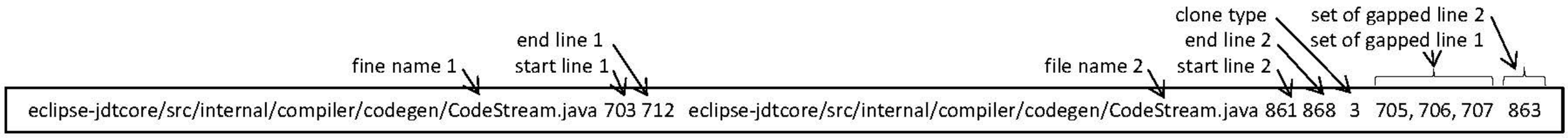
A Dataset of Clone References with Gaps

Outline

Bellon's dataset [1][2] is one of widely used clone datasets. The dataset contains many clone references, which are correct clones consisting of their locational information. Thus, the dataset is useful for comparing accuracies among clone detectors. However, Bellon's dataset does not have locational information of gapped lines. For this reason, Bellon's dataset does not evaluate some Type-3 clones correctly. In order to resolve the problem, we added locational information of gapped lines to Bellon's dataset and made it public.

The following figure shows an example of our clone reference and the correspondent source files. The left source file has three gapped lines (lines 705, 706 and 707), and the right one has a single gapped line (line 863).




```
   ⋮
703: final public void dstore_2() {
704:   countLabels = 0;
*  705:   stackDepth -= 2;
*  706:      if (maxLocals < 4) {
*  707:         maxLocals = 4; }
708:      try {
709:         position++;
710:         bCodeStream[classFileOffset++] = OPC_dstore_2;
711:      } catch (IndexOutOfBoundsException e) {
712:         resizeByteArray(OPC_dstore_2); } }
   ⋮
```

* gapped line

```
   ⋮
861: final public void fcmpl() {
862:   countLabels = 0;
*  863:   stackDepth--;
864:   try {
865:      position++;
866:      bCodeStream[classFileOffset++] = OPC_fcmpl;
867:   } catch (IndexOutOfBoundsException e) {
868:      resizeByteArray(OPC_fcmpl); } }
   ⋮
```


[1] S. Bellon, R. Koschke, G. Antniol, J. Krinke, and E. Merlo, ``Comparison and evaluation of clone detection tools", IEEE TSE, 33(9):577-591, 2007.
[2] Detection of Software Clones <http://www.bauhaus-stuttgart.de/clones/>

Datasets

Project name	Language
weltab	C
cook	C
snns	C
postgresql	C
netbeans-javadoc	Java
eclipse-ant	Java
eclipse-jdtcore	Java
j2sdk1.4.0-javax-swing	Java

In each of above datasets, first two elements in a single line represent identifiers.
First one is serial number of clone references, and second one is that of software systems.
These elements are not used for evaluations directly.

We used only Java datasets for the evaluation in our paper.

Data Format

Each of datasets has following BNF grammar:

```
lineend := '\n'
separator := '\t'
comma := ','
hyphen := '-'
digit := '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9'
digit0 := '0' | digit
number := digit [digit0*]
filename1 := valid-filename
fromline1 := number
toline1 := number
```

gapline1 := number
gapset1 := hyphen | gapline1 [comma gapline1*]
filename2 := valid-filename
fromline2 := number
toline2 := number
gapline2 := number
gapset2 := hyphen | gapline2 [comma gapline2*]
clonetype := '0' | '1' | '2' | '3'
line := filename1 separator fromline1 separator toline1 separator filename2 separator fromline2 separator toline2 separator clonetype separator gapset1 separator gapset2 lineend
file := line [line*]

Contact

Hiroaki Murakami <h-murakm at ist.osaka-u.ac.jp> (please replace "at" with "@")
Software Design Laboratory
Graduate School of Information Science and Technology, Osaka University