

## Automatically Inferring Concern Code from Program Investigation Activities

Martin P. Robillard and Gail C. Murphy

## Outline

- Motivation
- Investigation Transcripts
- Inference Algorithm
- Evaluation and Results
- Conclusions

## Motivation: An Example

### Bob

- Modify a text editor to allow explicit disabling of the auto-save feature through the GUI
- Investigate code to answer questions and make the change

### Alice

- Modify the auto-save feature to use a new library
- Investigate code to answer questions and make the change

## Motivation: An Example

### Bob

- Modify a text editor to allow explicit disabling of the auto-save feature through the GUI
- Investigate code
- Make the change
- Save concern in a database

### Alice

- Modify the auto-save feature to use a new library
- Find the concern in the database to learn what the relevant code it
- Make the change

## Motivation: Problems

- Programmers spend a lot of time investigating source code in order to perform a program evolution task.
- Determining which pieces of code to examine is a complex problem.
- Investigated concerns are rarely documented.

## Solution

- Automatically infer the meaning of a program investigation session.
- The results of the inference algorithm are a set of concern descriptions that can serve as documentation.

## Investigation Transcripts

- Informally, a transcript records all source code visible during a program investigation session.
- Granularity: method/field declaration
- Formally, a transcript is an ordered set of *investigation events*,  
 $E = \{e_1, \dots, e_n\}$

7

## Investigation Events

- An event is a change in what is visible.
- A method declaration is visible if it is completely or partially visible in the active window.
- An event  $e$  consists of a tuple  $(D, c, X)$ 
  - $D$  is a set of identifiers visible immediately after the event
  - $c$  is the category of the event
  - $X$  is an order set of extra information

8

## Event Categories & Extra Info

Event Category ( $c$ )	Visible code changed as a result of...	Extra Info ( $X$ )
<b>B</b>	Selecting an element in a code browser	Declaration accessed through browser
<b>C</b>	Following a cross-reference	Domain and range of cross-reference
<b>R</b>	Editor window recalled	None
<b>L</b>	Scrolling up or down	None
<b>K</b>	Keyword search	Declaration in which keyword was found

9

## Example Investigation Transcript

```

B137           K   B137
F29, F30, F31 C   B137, F30
B137           R
B24, B167      B   B24
B167, B168    L
    
```

10

## Inference Algorithm

- Input: Investigation transcript
- Output: Possible concern descriptions
- Concern description: a group of program elements that are related in the investigation
- Idea is to extract potential *concern descriptions* from a transcript

11

## Inference Algorithm: Phase 1

- For each event  $e_i$ , calculate the probability  $a_{i,j}$  that each declaration in the set  $D_i$  is relevant.
- Each event is associated with a weight,  $w_{i,j}$  that is determined by the category of the event and the extra info set ( $x^a$ )

$$p(d_{i,j}) = \frac{w_{i,j}}{\sum_{k=1}^n w_{i,k}}$$

12

## Calculating Probabilities

```

for all  $e_i = \{D_i, c_i, X_i\}$  in E do
  for all  $d_{i,j}$  in  $D_i$  do
     $w_{i,j} = 1$ 
    if ( $c_i = B$  ||  $c_i = K$ ) &&  $d_{i,j} = x_{i,1}$  then
       $w_{i,j} = w_{i,j} + \alpha$ 
    else if  $c_i = C$  &&  $d_{i,j} = x_{i,2}$  then
       $w_{i,j} = w_{i,j} + \alpha$ 
    if  $c_{i+1} = C$  &&  $d_{i,j} = x_{i+1,1}$  then
       $w_{i,j} = w_{i,j} + \alpha$ 
  
```

**F29, F30, F31    C    B137, F30**

$\alpha = 5$                        $p(F29) = 0.125$   
 $p(F30) = 0.75$              $p(F31) = 0.125$

13

## Inference Algorithm: Phase 2

- Infer concerns by analyzing the correlation between pairs of elements.
- Correlation metric
  - How close are two elements in the investigation sequence?
  - Category of the two elements
  - Are the elements directly related?
  - Calculated probabilities

14

## Correlation Metric

- Nine parameters
  - $\beta_0$ : importance of two elements being displayed consecutively
  - $\beta_1, \beta_2$ : importance of two elements being displayed separated by one or two elements
  - $\beta_B, \beta_C, \beta_R, \beta_L, \beta_K$ : importance of the event categories
  - $\beta_S$ : importance that two elements are actually related

15

## Inference Algorithm: Phase 3

- Concern generation
  - $\eta$  is desired number of elements in all concern
- Example,  $\eta = 5$ 
  - Pairs: [A, B][B, C][D, E]
  - Concerns: [A, B, C][D, E]

16

## Evaluation

- Two case studies
  - Investigate a program to do an evolution task using Eclipse.
    - jEdit – make a change, given a hint
    - jHotDraw – make a change, not given a hint
- Five configurations
  - Basic (1) - Intuition, linear progression
  - Neighbors (2) - Consecutive events only
  - No structure (3) - Emphasis on developer
  - Structure (4) - Emphasis on structure
  - Guesses (5) - Emphasis on guessing and browsing

17

## Subject C2

Id	Concern	1	2	3	4	5
1	A, B	X	X	X	X	
2	A, B, C		X			
3	D, E	X	X	X		X
4	D, E, M		X			
5	D, E, M, P, Q, R				X	
6	F, G		X			
7	F, G, H	X			X	
8	F, G, H, K			X		
9	F, G, H, K, L					X
10	I, J	X	X	X	X	
11	K, L	X				
12	M, N	X	X	X		X
13	K, O		X			

18

### Subject C3

Id	Concern	1	2	3	4	5
1	D, E, M, N, P, R, S, T	X		X		
2	D, E, M, N, P, R, S, T,		X			
3	D, E, M, N, R, S, T					X
4	D, E, M, P, R, S, T, V				X	
5	G, H	X	X	X	X	X
6	F, U	X	X	X	X	
7	K, W, X					X

19

### Subject C4

Id	Concern	1	2	3	4	5
1	I, J, Q, Y	X	X			
2	I, J, K, M, W, X, Y,			X		
3	I, J, M, Q, Y, BB				X	
4	I, J, M, Y, CC					X
5	K, X	X	X			
6	K, W, X					X
7	F, AA		X	X		
8	F, Z, AA	X				
9	F, Z, AA, DD				X	
10	G, H	X	X	X	X	X
11	M, BB	X	X			

20

### Subject J1

Id	Concern	1	2	3	4	5
1	A, B, C	X	X	X	X	
2	D, E	X	X			
3	D, E, F, G		X			
4	D, E, F, G, O				X	
5	D, E, F, G, P,					X
6	P, G	X				
7	F, G, M		X			
8	H, I	X	X			
9	H, I, J					X
10	J, K			X	X	
11	J, K, L	X	X			
12	M, N		X	X		

21

### Subject J2

Id	Concern	1	2	3	4	5
1	R, S, T, U, V, W, X, Y, Z	X				
2	R, S, T, U, V, W, Z, FF, HH					X
3	R, S, T, U, W, X, Y, Z, DD, EE, FF,			X		
4	R, S, T, U, W, X, Z, FF, II				X	
5	R, T, U, V, W, Y, Z, FF, HH			X		
6	AA, BB, CC	X	X	X		

22

### Results

- Basic configuration gave best results
  - Browser, cross-reference, and keyword search transitions are more important
- Scrolling
  - Increases the likeliness that an element is selected as relevant
- Transcript boundaries
  - Turn transcript on and off as needed

23

### Conclusions

- Feasibility of inferring concerns from investigation transcripts
- Results of the algorithm depend on parameters that can be tweaked
- Inferred concern descriptions can be used as documentation

24