RWTH Aachen University
Software Engineering Group

Feature Location Techniques

Seminar Paper

presented by
Bergerbusch, Timo

1st Examiner: Prof. Dr. B. Rumpe

2nd Examiner: Dipl.-Inform. C. Schulze

Advisor: Dipl.-Inform. C. Schulze

The present work was submitted to the Chair of Software Engineering

Aachen, December 16, 2016

Eidesstattliche Versicherung

Name, Vorname

Matrikelnummer (freiwillige Angabe)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel

selbständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum

Unterschrift

*Nichtzutreffendes bitte streichen

Belehrung:

§ 156 StGB: Falsche Versicherung an Eides Statt

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:

Ort, Datum

Unterschrift

Abstract

Locating software artifacts that implement a specific program functionality, whether it's functional or non-functional, are called a feature. Detecting features in a program is the main goal of Feature Location Techniques(FLT). It assists software developers during the maintenance and refactoring of the code. But also the software product line engineering(SPLE), which specifies, designed and implements different products by managing features, uses these techniques to create a product without copying code unstructured but by systematic reuse of the artifacts the FLT's locate [PBvDL05].

Therefor my seminar paper deals with different feature location techniques from very fundamental methods to some of today's newest research fields. In this paper I introduce a real use case example, to show the real utility of the techniques, of the Freemind mind mapping software [www16b].

In this paper we continue to get to know to the basics of FLT's to understand how they are able to define artifacts, the classification of FLT's considering their approach strategy, explaining different techniques of different previously mentioned classes, regarding their strengths and weaknesses, on a realistic use case of a real software segment. At the end will be an outlook to leveraging SPLE architectures and possible improvements of the existing techniques[ZZL+06].

Contents

Chapter 1

Introduction

A feature location technique is aiming at the locating of software artifacts as a realization of a system requirement. It could be *functional*, like the ability of doing a special kind of computation for example counting elements, or it could be *non-functional* like doing a functional requirement in a given time. To be able to understand what a feature location technique in detail should be it is necessary to have a basic knowledge about two aspects of modern software engineering. Without either one of the following two underling definitions it's is not clearly definable what a feature location technique should be capable of and there is also no way to rate if a technique is efficient and correct.

On the one hand there a the features. As defined by the Institute of Electrical and electronics Engineers (IEEE) a feature is defined as 'A distinguishing characteristic of a software item (e.g., performance, portability, or functionality)'.[Wik04a] For us simplified a feature is a software artifact implementing a given requirement. Features are often described by the definition of *Rajlich and Chen*, who describe a feature or concept as a triple of *name*, the name of the feature, *intension*, a short precise description, and *extension*, the artifacts implementing the feature.[KC00]

The the other hand there is the software product line engineering (SPLE). A product line is a variety of products, which in our case are software products, which 'share a common, managed set of features satisfying the specific needs a particular market segment or mission and that are developed from a common set of core assets in a prescribed way.' [www16a]. A good example are the products of SAP like the *Business One*, *Business All-In-One* and *Business ByDesign*, which share a basic set of functionality, build up on each other and often are modified to fit the needs of a customer. The SPLE promotes *systematic* software reuse being base on the knowledge about the set of available features, relationships among the features and the relationship between features and their artifacts. The most essential step for unfolding the complexity of existing implementations to be able to transform it into a SPLE includes the identifying of the implemented features and their corresponding artifacts.

This, the locating and defining of a feature, is the problem a feature location technique should solve, so that developers of software product lines are supported during the maintenance and the aspect-/feature oriented refactoring of software.

Chapter 2

Freemind Example

The example used for this paper is the *automatic save file* feature of Freemind. Freemind is an open source mind-mapping tool. The *automatic save file* feature is a good example, because of it's name. Parts of the name are also mentioned in other features, which makes it slightly more difficult to only locate this specific feature. A representiv callgraph of the important parts is shown in Fig 2.1.

As you can see here only the relevant constructors and methodsare shown and numbered with indices from 1 to 8. We reference them by using the number sign # and then the corresponding number. Also the feature of the regarded function are highlighted with a blue background color. These are the methods which should be located if the *automatic save file* function is the wanted feature. Note that the all the methods of different classes can in addition call other methods and constructors, which are irrelevant to the feature. So as we can see the feature is mainly implemented by two methods of a subclass of *MindMapMapModel* so called *doAutomaticSave*:

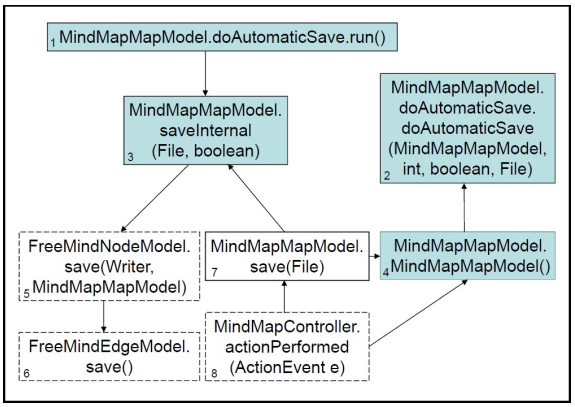

Figure 2.1: The Freemind callgraph [www16b] [RC13]

- the constructor, which is #2. This constructor gets a few parameters to configurate the *doAutomaticSave*-function and registers the class in the sheduling queue, so that it gets called.
- the *run()*-function #1. This Methods gets called after the class is registered in the sheduling queue and everytime a special event is occurs. That can be different, like a period of time to shedule an automatic save or a preset number of actions within the main-programm. It calles the *saveInternal*-method to do actual saveoperation.

Regarding the previously mentioned definition of a feature by Rajlich and Chen 1, we can now define the regarded feature as the following:

name: *automatic save file*
intesion: saves a file automaticly afer the occuring of an event The methods #5 to
extension: #1, #2, #3 and #4
#8 aren't in the extension of the *automaticSaveFile* feature. Mainly # 5 and #6 are called by methods of the *automaticSaveFile* feature, but aren't relevant to the specifies of this function. #7 and #8 in fact call #3 and #4, but they handle a user triggered save-event, which obviously isn't important to the *automaticSaveFile* feature.

While all feature location techniques try to achive the same goal, which is the locating the feature extension to a given feature intension, they differate in the underlying base of assumptions they make to be able to get the tracebility. It will be declared more specific in chapter 4.

Chapter 3

Basic Underlying Techniques

To understand how feature location techniques work it is important to understand a few basic techniques that are commonly used to create or improve feature location. All the basic techniques will be exemplary executed on the previously introduced Freemind-example in chapter 2.

3.1 Formal Concept Analysis (FCA)

Formal Concept Analysis (short: *FCA*) is a predominantly mathematical approach to identify groups of classes and methods compared by the sharing of attributes. Therefor the *FCA* regards the binary relation between all objects and attributes and therefor can also provide a model to analyze hierarchy, because hierarchy structures often have similar relations.

The *FCA*'s goal is to define so called *concepts*. A *concept* is a tuple of extension, the objects that belong to a concept, and intension, all the attributes that every object of the extension has. In order to be able to derive such a *concept* the *FCA* creates an incidence table. The table can be derived in 3 steps as seen in Fig. 3.1:

1. declaring every word in the objects and methods as w_i to a new i if the word isn't already defined
2. decapitalizing every w_i
3. creating the table with every decapitalize word as a row and every σ as a column. The cells c_ij are checked if σ contains the word w_i

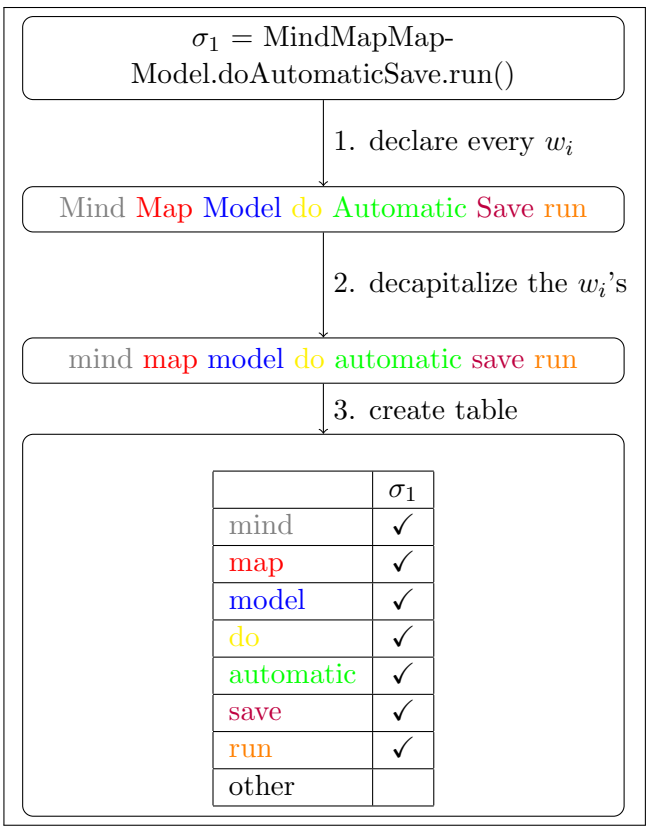

Figure 3.1: #1 of the Freemind Example as example

objects ↓	σ_1 ↓	σ_2 ↓	σ_3 ↓	σ_4 ↓	σ_5 ↓	σ_6 ↓	σ_7 ↓	σ_8 ↓
action								✓
automatic	✓	✓						
controller								✓
do	✓	✓						
file								
free					✓	✓		
internal			✓					
map	✓	✓	✓	✓			✓	✓
mind	✓	✓	✓	✓	✓	✓	✓	✓
model	✓	✓	✓	✓	✓	✓	✓	✓
node						✓	✓	
performed								✓
run	✓							
save	✓	✓	✓		✓	✓	✓	

Figure 3.2: The complete incidence table of the Freemind Example

Keeping the methods numbers as we did we get Figure 3.2 as a result. Mathematically it leads us to defining O as a set of objects, A as a set of attributes and R as the set of relations $r = (o, a) \quad o \in O, a \in A$ as derivable of the table. Also we define that
$\sigma(O) = \{a \in A | (o, a) \in R, \forall o \in O\}$ "all attributes that every $o \in O$ has"
$\rho(A) = \{o \in O | (o, a) \in R, \forall a \in A\}$ "all objects that every $a \in A$ has"
So a concept can be declared as a tuple $c = (O, A)$ so that $A = \rho(0)$ and $O = \sigma(A)$. So O is the extension and A is the intension.

From there it is very easy to see, that the set of all concepts C is a partial order (*superconcept - subconcept*) defined as:

$$(O_1, A_1) \leq (O_2, A_2) \quad \Leftrightarrow \quad O_1 \subset O_2 \text{ or } A_1 \subset A_2.$$

Which leads to the definition that $C, \leq$ form a concept lattice and in our example it's a taxonomy of name tokens.

3.2 Latent Semantic Indexing (LSI)

$$A = \begin{array}{l|cccccccc} \text{Documents/ Terms} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 \\ \hline \texttt{action} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \texttt{automatic} & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \texttt{controller} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \texttt{do} & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \texttt{file} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \texttt{free} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ \texttt{internal} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \texttt{map} & 2 & 2 & 2 & 4 & 0 & 0 & 2 & 1 \\ \texttt{mind} & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \\ \texttt{model} & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 0 \\ \texttt{node} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ \texttt{performed} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \texttt{run} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \texttt{save} & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 0 \end{array} \qquad q = \begin{array}{c} q \\ \hline 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{array}$$

Figure 3.3: The term-document matrix

The *Latent Semantic Indexing*(short: *LSI*) is an automatic statistical technique. It derives to a given document a vector representation of the query and the corpus by creating a term-document matrix of co-occurring terms. A term t_i is a word, as a tokenized and decapitalized word of the methods ordered alphabetically and is represented in a row of the matrix. A document d_j, which are in our example the different method- and class names, are represented as the columns of the matrix. So the matrix, shown in Fig. **??**, looks very similar to the table of FCA (Fig. 3.2) with the difference of an unsigned integer value v_ij, representing how often a document $d_j = MindMapMapModel.doAutomaticSave.run()$ contains token t_i, i.e. d_1 contains the token $t_7 = map$ twice, but the token $t_2 = automatic$ only once and doesn't contain $t_1 = action$ at all. Also a query q is given, which has a *1* at the terms *automatic*, *save* and *file* representing the feature that should be analyzed.

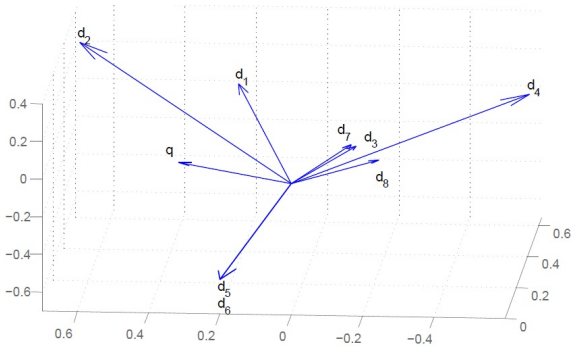

Figure 3.4: The vector representation of the documents d_j and the query q from the Freemind Example 2

By normalizing and decomposing using a singular value decomposition the documents can be put into vector representation so that every document has a vector representing their equality to the query q. Taking the $cosine()$ of the query q and a document d_j it is possible to measure the similarity. If a document and a query are equal the spreading angle would be 0 and therefor the best possible similarity is given by $cosine(0) = 1$. Also the worst possible angle is 180, which is equal to document "pointing in the opposite direction", and therefor the worst similarity is given by $cosine(180) = -1$. The common interpretation of the values, regarding that D is the set of all documents, is, that the set $\{d_i \in D | cosine(d_i, q) \geq 0\} \subseteq D$ are considered to be a related to the query of interest, hence every other document is not. It's simple to see that a document is more similar if it points in the same general direction as the query, because of the shared terms. In the Freemind Example the document $d_2 = MindMapMapModel.doAutomaticSave.doAutomaticSave$ is the most similar to the query $q = automaticSaveFile$, while $d_8 = MindMapController.actionPerformed$ is the least similar.

d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
0.6319	0.8897	-0.2034	-0.5491	0.2099	0.2099	-0.1739	-0.6852

Like previously mentioned d_2 is the most similar to q, because of the "pointing in the same general direction", which it now proven by having the highest value $cosine(d_2, q) = 0.8897 = max\{cosine(d_i, q) | d_i \in D\}$ and also d_8 is the least similar with a value $cosine(d_8, q) = -0.6852 = min\{cosine(d_i, q) | d_i \in D\}$.

3.3 Term Frequency - Inverse Document Frequency (tf-idf)

The *term frequency - inverse document frequency* technique is a statistical technique to derive a feature to a given intension. It measures the importance of a term or multiple terms to documents by its frequency of appearing. The terms are terms of the intension of the feature that is wanted to be analyzed. In a simple way it can be described as: "the more frequent a term occurs in the document, the more relevant the document is to the term".
This is mathematically described as the $documentfrequency tf = (t, d)$, counting how often the term t is contained in the document d. In our example the term $t_2 = save$ appears in d_3 once and the term $t_1 = automatic$ doesn't appear at all so: $tf(t_2, d_3) = 1$ and $tf(t_1, d_3) = 0$.
Doing that for the terms $t_1 = automatic, t_2 = save$ and $t_3 = file$ and the documents d_1 to d_8 we get the matrix shown in Fig **??**. The main problem of this technique is, that uninformative terms appearing within a document-set, often referred as *corpus* and shortened by D, maybe even multiple times can distract from terms, which are mentioned

less frequent but are more relevant. To compensate that, the technique relativizes by calculating how many documents contain the term and normalizing it. If it's a commonly used term shared by many documents this term can't be taken as a measurement to differentiate between documents. Or colloquially "the more documents include a term, the less this term discriminates between documents".
So the so-called *inverse document frequency (idf(t))* is calculated as

$$idf(t) = log((|D|)/|\{d \in D|t \in d\}|)$$

with D still being the set of documents. And the final *term frequency - inverse document frequency* is the multiplication of both scores, so:

$$tf\text{-}idf(t,d) = tf(t,d) * idf(t)$$

Regarding our example we can compute the *idf* of our terms:

$$t_1 = automatic : idf(t_1) = log(8/2)$$
$$t_2 = save : idf(t_2) = log(8/6)$$
$$t_3 = file : idf(t_3) = 0$$

Like in the example if the focus isn't on one term but on a set of terms the *tf-idf(t,d)* values to a document d are added up. So finally the matrix can be derived as it is shown in Table **??**.

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
$t_1 = automatic$	0.6021	1.2041	0	0	0	0	0	0
$t_2 = save$	0.1249	0.2499	0.1249	0	0.1249	0.1249	0.1249	0
$t_3 = file$	0	0	0	0	0	0	0	0
$\sum_{i=1}^{3} tf\text{-}idf(t_i, d_j)$	0.727	1.454	0.1249	0	0.1249	0.1249	0.1249	0

Table 3.1: Term Frequency - Inverse Document Frequency

3.4 Hyper Link Induced Topic Search (HITS)

The *Hyper Link Induced Topic Search* (short: *HITS*) is a page ranking algorithm for web mining[1], which is the counterpart of the famous *Google Page Rank*-algorithm and is currently used by the *Ask Search Engine* [Wik04b]. Its basically used to get websites that correspond best to a given input, like every search engine. The *HITS*-algorithm distinguishes between two forms of web pages, which aren't necessarily disjoint:

1. hub
   A hub is a web page pointing towards other web pages , which can be a hub, an authority or even both. A pragmatism is to say: "a good hub points to many authorities."

[1] web mining is the analysis step of the knowledge discovery in databases process within the World Wide Web CITE

2. authority
   An authority is a web page, that other pages point to in order to cite or prove. The rule of thumb is: "a good authority is pointed by many good hubs."

Regarding the definition of hubs and authorities it seems quite natural to define a directed graph $G = (V, E)$ with vertices V = web pages and edges $E = \{(v, w)|v \text{ refers to } w\}$ (also called *links*) . A hubscore is the number of authorities the hub refers to. An authorityscore is a number of good links that refer towards this authority. Both are initialized with 1. Keeping the graph G in mind the hub- and authority scores can be defined as the following.

$$\begin{aligned} \text{authority score of page } p \qquad & A_p = \{\textstyle\sum_{\{q|(q,p)\in E\}} H_q\} \\ \text{hub score of page } p \qquad & H_p = \{\textstyle\sum_{\{q|(p,q)\in E\}} A_q\} \end{aligned}$$

Given the two values the graph can be rewritten as $G' = (V', E')$ with $V' = \{(p, H_p, A_p)|\forall p \in V\}$ and $E' = E$. By iterating over the graph the values of H_p and A_p are calculated for every page p. In order to don't just count up to infinity the values have to be normalized like the following:

$$\begin{aligned} \text{normalizing the authority score of page } p \qquad & A_p = A_p/\sqrt{\textstyle\sum_{(q,H_q,A_q)\in V'} A_q^2} \\ \text{normalizing the hub score of page } p \qquad & H_p = H_p/\sqrt{\textstyle\sum_{\{(q,H_q,A_q)\}\in V'} H_q^2} \end{aligned}$$

The normalized values satisfy the condition, that $\sum_{\{(p,H_p,A_p)\}\in V'} H_p^2 = \sum_{\{(p,H_p,A_p)\}\in V'} A_p^2 = 1$.

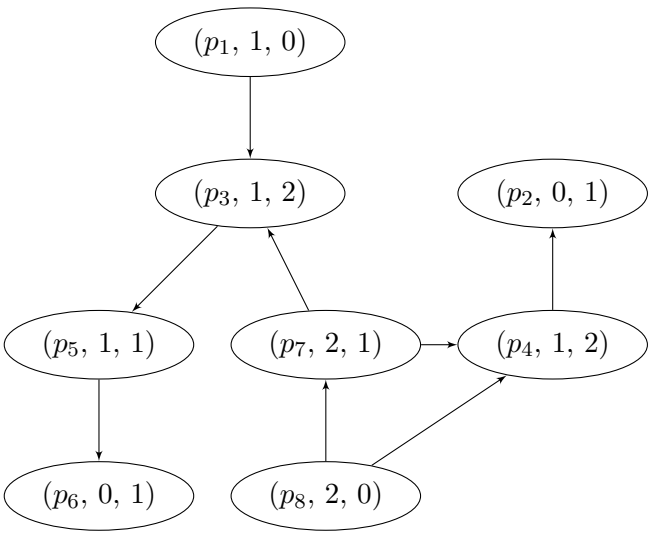

Figure 3.5: The graph G

Applying the *HITS*-algorithm to program code hubs can be colloquially described as methods, that call many other methods, and authority's can be described as methods, that implement a function.
In the Freemind Example the first graph will look very similar to the class diagram, as it is shown in Fig 3.5. The class #i will refer to page p_i. After transferring it into the graph of the form of G' and after the first iteration the graph looks like Fig. 3.5.
Including the normalization the graph G' looks like Fig. 3.6. The normalization was done by calculating for every H_p and A_p of a page p as:

$$\begin{aligned} H_p &= H_p/\sqrt{1^2+0^2+1^2+1^2+1^2+2^2+0^2+2^2} = H_p/\sqrt{12} \text{ and} \\ A_p &= A_p/\sqrt{0^2+1^2+2^2+2^2+1^2+1^2+1^2+0^2} = A_p/\sqrt{12}. \end{aligned}$$

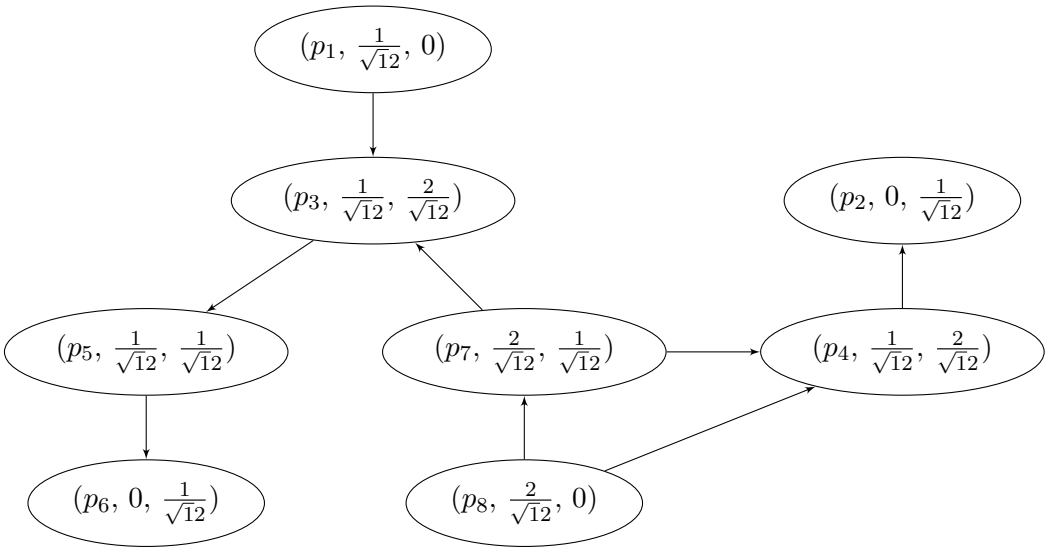

Figure 3.6: G' after the first iteration with normalizing the scores

Chapter 4

Classification and Methodology

The classification of feature location techniques is very important, because of the different special demands of some classes of techniques and their assumptions they have towards special parts of the system or code. The first big distinction is the difference of dynamic and static techniques.

dynamic:
Dynamic approaches collect information about the program at runtime. They do so by using program dependency analysis, information retrieval, latent semantic indexing (3.2 or the term frequency - inverse document frequency **??**, which only consider the methods and classes, which are involved during the current execution of the program. This is a big advantage, because by knowing that looking for a feature knowing roughly the part of the program where it could be used the user is able to steer the program into the direction. In our example the *automaticSaveFile*-function wouldn't be in the main-menu or the settings, but is more likely to be involved if the user creates a mind map and waits till the *automaticSaveFile*-function is triggered. But that advantage has also it's flipside. By only analyzing the involved parts of the program the whole information retrieval is based on the input the program gets and has to generalize from that, which may not be the right thing to do. Also collecting information on test-cases can only derive *functional* requirements, but isn't able to derive non-functional requirements. In general the dynamic approaches under approximate.
static:
Static approaches don't need the program to be executed. They collect information directly out of the source, which has one big disadvantage. A static approach would look at every single part of the code to derive information about the feature, the user want's to locate, which can be very costly. Imagine a program which is very very complex and big and the user wants to locate a very small special feature which is contained in very little of the code for example in only 0.01% . The static approach will look through the whole 100% of code, of which 99.99% are not related to the feature. The big advantage of the static approach is that the information it reveals are safe, which means it doesn't has to generalize out of a case but can validate on the whole information. This results in the ability to derive functional and non-functional requirements. This whole information can on the other side lead again to problems. Knowing every little detail can lead to situations in which the information's are undecidable in the matter of affiliation to the feature. So the technique has to approximate a solution, which may be to imprecise. In general the static approaches over approximate.

The techniques can also be splitted within the *static*/*dynamic*-groups due to the form of output the methods give.

plain:
The plain-output techniques present an unsorted list of artifacts, which are considered by the technique to be relevant to the feature. They leave the interpreting of the output to the user.
guided:
The guided-output techniques present the collected artifacts in a special arrangement to build an interpretation, like ordering the artifacts based on the relevance it is considered to have. Also often a so called *Program Dependency Graph* is given to not only show relevant artifacts, but also give a dependency of these artifacts. This topic is further explained in "Case Study of Feature Location Using Dependence Graph" by K. Chen and V. Rajlich [CR00].

Also the different techniques make assumptions. For example the in chapter 3.2 mentioned *Latent Semantic Indexing* does the assumption that the classes and methods of the code are named like the function they implement. The same technique can be useful on one code fragment, which fits the assumptions, but completely useless on an other one, which doesn't fulfill the assumptions. [RC13] [DRGP13]

Chapter 5

Feature Location Techniques

In this chapter we want to look at four different feature location techniques in detail. We choose two static and two dynamic techniques with each one technique giving plain and one giving guided output.

5.1 Static - Plain

This could be one of the following techniques:

- CLDS - Chen et al.
- Find-Concept - Wlakinshaw at al.
- SNIAFL - Zhao et al. (recommended)

5.2 Static - Guided

- Suade - Robbilard
- Dora - Hill
- CVSSearch - Chen

5.3 Dynamic - Plain

- Sw. Reconnaissance - Wilde
- Koschke
- Asadi

5.4 Dynamic - Guided

- SITIR - Liu
- Cerberus - Eaddy

Bibliography

[CR00] Kunrong Chen and Václav Rajlich. Case study of feature location using dependence graph. In *IWPC*, pages 241–247. Citeseer, 2000.

[DRGP13] Bogdan Dit, Meghan Revelle, Malcom Gethers, and Denys Poshyvanyk. Feature location in source code: a taxonomy and survey. *Journal of Software: Evolution and Process*, 25(1):53–95, 2013.

[KC00] Václav Rajlich Kunrong Chen. *Case Study of Feature Location Using Dependence Graph.* IWPC, 2000.

[PBvDL05] Klaus Pohl, Günter Böckle, and Frank J van Der Linden. *Software product line engineering: foundations, principles and techniques.* Springer Science & Business Media, 2005.

[RC13] Julia Rubin and Marsha Chechik. A survey of feature location techniques. In *Domain Engineering*, pages 29–58. Springer, 2013.

[Wik04a] Wikipedia. Plagiarism — Wikipedia, the free encyclopedia, 2004. [Online; accessed 22-July-2004].

[Wik04b] Wikipedia. Plagiarism — Wikipedia, the free encyclopedia, 2004. [Online; accessed 13-December-2016].

[www16a] Software Product Lines `http://www.sei.cmu.edu/productlines/`, november 2016.

[www16b] Freemind website `http://freemind.sourceforge.net/`, october 2016.

[ZZL+06] Wei Zhao, Lu Zhang, Yin Liu, Jiasu Sun, and Fuqing Yang. Sniafl: Towards a static noninteractive approach to feature location. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 15(2):195–226, 2006.