On extracting relevant and complex variability information from software descriptions with pattern structures
Jessie Carbonnel, Marianne Huchard, Clémentine Nebut

To cite this version:

HAL Id: lirmm-01872807
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01872807
Submitted on 12 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License
Poster: On Extracting Relevant and Complex Variability Information from Software Descriptions with Pattern Structures

Jessie Carbonnel, Marianne Huchard, Clémentine Nebut
LIRMM, Université de Montpellier, CNRS, Montpellier, France
{jcarbonnel,huchard,nebut}@lirmm.fr

ABSTRACT
The migration from existing software variants to a software product line is an arduous task that necessitates to synthesise a variability model based on already developed softwares. Nowadays, the increasing complexity of software product lines compels practitioners to design more complex variability models that represent other information than binary features, e.g., multi-valued attributes. Assisting the extraction of complex variability models from variant descriptions is a key task to help the migration towards complex software product lines. In this paper, we address the problem of extracting complex variability information from software descriptions, as a part of the process of complex variability model synthesis. We propose an approach based on Pattern Structures to extract variability information, in the form of logical relationships involving both binary features and multi-valued attributes.

CCS CONCEPTS
- Information systems → Information extraction; - Software and its engineering → Software product lines; Software reverse engineering;

KEYWORDS
Software Product Line, Reverse Engineering, Variability Extraction

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ACM Reference format:
https://doi.org/10.1145/3183440.3194982

2 METHOD OVERVIEW

The proposed extraction method is based on Pattern Structures (PS) [2], a mathematical framework used for knowledge discovery on a set of objects described by complex data, e.g., numerical values, graphs, sets of values. PS describe each object of a set \( O \) according to a "pattern", that can be of any type of data on which one can define a similarity operator \( \cap \). Applied to a set of patterns, this operator returns a pattern representing their most specific common generalisation. We applied our method on existing multi-valued matrices taken from the software comparison category of wikipedia. Here, the objects represent the software variants, and the patterns represent the variant descriptions. Table 1 presents an excerpt of a multi-valued matrix depicting softwares and their characteristics. It describes 5 software variants depending on 5 characteristics: Open Source, Corporate and Personal are binary characteristics, First Release is a characteristic having 4 possible Date-typed values, and Language is a characteristic with 3 possible String-typed values. Our method can be decomposed in the 4 steps of Figure 1.

Figure 1: Extracting complex variability information from variant descriptions using Pattern Structures.
Identifying patterns. Patterns can be values of atomic types, e.g., a Date, a String or a number. But, one can choose to compose several patterns to form a vector of patterns, that can be considered a pattern as well. In our case, a variant description may be represented as a pattern vector depicting the values of each characteristic of the matrix. This first step consists in identifying in the variant descriptions the set of component patterns that will form the pattern vectors. As binary characteristics have a $\{true, false\}$ domain, the set of binary characteristics having a true value of a variant can be considered as one multi-valued characteristic. This representation choice does not alter the extracted information and permits to simplify the processed data.

The pattern vectors for the software variants of Table 1 are thus of the form $v = \langle \{f_1, f_2, \ldots, f_n\} \rangle$, where $P_f$ is the set of features (i.e., binary characteristics), followed by the value of the attribute First Release (FR) ($P_f$) and then the value of the attribute Language ($P_l$). The two following pattern vectors respectively represent the softwares Confluence and GrokOla:

$$v_1 = \langle \{Corporate, Personal\}, FR = 2004, Language = Java EE \rangle$$
$$v_2 = \langle \{Corporate\}, FR = 2014, Language = JavaScript \rangle$$

Defining taxonomies. Now that we have specified the pattern vectors, we have to define similarities between the elements of each component pattern. A similarity operator is associated to a subsumption relation $\subseteq$, allowing the set of patterns $P$ to be partially ordered in a taxonomy, in which each pair of elements has a unique most specific common generalisation pattern. In other words, it structures the patterns by specialisation, and $p_1 \subseteq p_2$ if and only if $O \subseteq O'$ and $p_1 \subseteq p_2$. The set of all pattern concepts of $PS$ provided with the order $\leq_{ps}$ forms a lattice structure called a pattern concept lattice.

Building the lattice. The set of objects $O$, the pattern taxonomy $(P, \sqcap)$ and the mapping $\delta : O \rightarrow P$ associating each object of $O$ to its pattern in $P$, form a Pattern Structure. Given a Pattern Structure $PS = \langle O, (P, \sqcap), \delta \rangle$, a pattern concept of $PS$ is a tuple $(O', p)$ such that $O' \subseteq O$ and $p \in (P, \sqcap)$, which represents a maximal set of objects and the most specific pattern corresponding to these objects. This means that there is no other object than the ones in $O'$ that corresponds to the pattern $p$, and that there is no other pattern more specific than $p$ which corresponds to all the objects from $O'$. These pattern concepts can be partially ordered: given two pattern concepts $C_1 = (O_1, p_1)$ and $C_2 = (O_2, p_2)$ of $PS$, $C_1 \leq_{ps} C_2$ if and only if $O_1 \subseteq O_2$ and $p_1 \subseteq p_2$. The set of all pattern concepts of $PS$ provided with the order $\leq_{ps}$ forms a lattice structure called a pattern concept lattice.

### Table 1: Excerpt of a multi-valued matrix representing software variant descriptions taken from wikipedia

<table>
<thead>
<tr>
<th></th>
<th>Open Source</th>
<th>Corporate</th>
<th>Personal</th>
<th>First release</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Confluence</td>
<td>X</td>
<td>X</td>
<td>2004</td>
<td>Java EE</td>
</tr>
<tr>
<td>2</td>
<td>GrokOla</td>
<td>X</td>
<td></td>
<td>2014</td>
<td>JavaScript</td>
</tr>
<tr>
<td>3</td>
<td>Jive</td>
<td></td>
<td></td>
<td>2006</td>
<td>Java EE</td>
</tr>
<tr>
<td>4</td>
<td>Wagner</td>
<td>X</td>
<td>X</td>
<td>2006</td>
<td>Ruby</td>
</tr>
<tr>
<td>5</td>
<td>Wiki.js</td>
<td>X</td>
<td>X</td>
<td>2017</td>
<td>JavaScript</td>
</tr>
</tbody>
</table>

3 CONCLUSION AND PERSPECTIVES

We presented an approach to extract complex variability information from software variant descriptions based on Pattern Structures. It allowed us to extract variability information by considering not only the software variants’ features, but also multi-valued attributes. This work is a first step toward a more generic approach to assist practitioners into extracting complex variability models.

Our extraction method is complete and therefore extracts an important number of relationships. It appears that a significant part of them are “accidental”, i.e., true for the considered set of variants but not regarding the domain. In future work, we plan to study existing techniques to lower the number of considered extracted relationships by deepening the separation of the meaningful ones from the accidental ones. Another future work will be to consider the extraction of other kinds of variability information that could be useful to synthesise complex variability models from software descriptions. Particularly, relationships between several independent but connected software families are to be studied, allowing applications in the field of multiple software product lines.

## REFERENCES


