Information and Software Technology 72 (2016) 1–15

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

A bibliometric analysis of 20 years of research on software product lines

Ruben Heradio[a,*], Hector Perez-Morago[a], David Fernandez-Amoros[a], Francisco Javier Cabrerizo[a], Enrique Herrera-Viedma[b]

[a] Department of Software Engineering and Computer Systems, Universidad Nacional de Educacion a Distancia (UNED), Madrid 28040, Spain
[b] Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain

ARTICLE INFO

Article history:
Received 11 May 2015
Revised 11 November 2015
Accepted 12 November 2015
Available online 2 December 2015

Keywords:
Software product lines
Bibliometrics
Science mapping
Performance analysis

ABSTRACT

Context: Software product line engineering has proven to be an efficient paradigm to developing families of similar software systems at lower costs, in shorter time, and with higher quality.

Objective: This paper analyzes the literature on product lines from 1995 to 2014, identifying the most influential publications, the most researched topics, and how the interest in those topics has evolved along the way.

Method: Bibliographic data have been gathered from ISI Web of Science and Scopus. The data have been examined using two prominent bibliometric approaches: science mapping and performance analysis.

Results: According to the study carried out, (i) software architecture was the initial motor of research in SPL; (ii) work on systematic software reuse has been essential for the development of the area; and (iii) feature modeling has been the most important topic for the last fifteen years, having the best evolution behavior in terms of number of published papers and received citations.

Conclusion: Science mapping has been used to identify the main researched topics, the evolution of the interest in those topics and the relationships among topics. Performance analysis has been used to recognize the most influential papers, the journals and conferences that have published most papers, how numerous is the literature on product lines and what is its distribution over time.

1. Introduction

A *Software Product Line* (SPL) is an engineering approach to efficient development of whole portfolios of software products [1]. The basis of the approach is that products, instead of being developed from scratch one by one, are built from a core asset base, i.e., a collection of artifacts that have been designed specifically for use across the portfolio.

The SPL approach brings the benefits of economies of scope to software engineering, since less time and effort are needed to produce a greater variety of products. Many companies have exploited the concept of software product lines to increase the resources that focus on highly differentiating functionality and thus improve their competitiveness with higher quality and reusable products while decreasing the time-to-market condition. For instance, van der Linden et al. [2] summarize experience reports from ten different companies working on diverse domains (e.g., Bosch on Gasoline Systems, Nokia on Mobile Phones, Philips on Consumer Electronics Software for Televisions, Siemens on Medical Solutions, etc.).

The goal of this paper is to analyze, using *bibliometric* techniques, the literature on SPLs for the last twenty years in order to determine the main topics and trends of this research area. The outcomes of our analysis provide information regarding the following issues:

1. What are the most influential papers on SPL literature?
2. Who are the most prolific authors?
3. What journals, conferences, etc. have published the majority of the papers?
4. How numerous is the SPL literature? How has paper publication been distributed over time?
5. What are the main topics studied in the area? How has the interest in those topics evolved with time?
6. What are the most impacting papers for a given a topic along a certain period of time?

We have processed 2845 records retrieved from ISI Web of Science (ISIWoS) and Scopus by using two approaches to examine bibliographic data: *performance analysis* [3] and *science mapping* [4].

* Corresponding author. Tel.: +34 913988242.
E-mail addresses: rheradio@issi.uned.es, rheradio@gmail.com (R. Heradio), hperez@issi.uned.es (H. Perez-Morago), david@issi.uned.es (D. Fernandez-Amoros), cabrerizo@issi.uned.es (F. Javier Cabrerizo), viedma@decsai.ugr.es (E. Herrera-Viedma).

http://dx.doi.org/10.1016/j.infsof.2015.11.004
0950-5849/

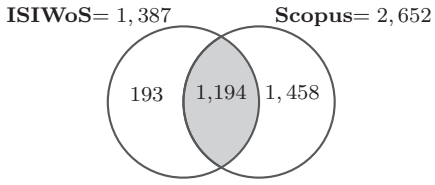

Fig. 1. Number of records retrieved from ISIWoS and Scopus.

Performance analysis tries to quantify the impact of scientific actors (researchers, journals, etc.) on a research field by measuring how often a paper is cited. In particular, this paper uses *H-index* [5] to measure publication impact, which is one of the most popular indicators for citation analysis.

Science mapping attempts to display the structural and dynamic aspects of scientific research, delimiting a research field, and identifying, quantifying and visualizing its thematic subfields. To map the SPL research area, we have used a technique called *co-word analysis*, that measures the association strengths of publication keywords [6].

Our work uses both performance analysis and science mapping in a complementary way. Bibliometric maps supports visualizing the main topics in the literature, their evolutions and inter-relationships. Performance analysis helps to identify the most productive topics (in terms of number of published papers), and the most impacting ones (according to received citations).

The remainder of the paper is structured as follows: Section 2 summarizes the methodology and techniques used to carry out our work. Section 3 reports the performance analysis of the whole SPL research area, identifying those papers that due to their number of citations should be considered as *classics*. Section 4 applies science mapping techniques to identify the cognitive structure and evolution of the SPL literature. Section 5 discusses threats to the validity of our work. Finally, some concluding remarks are provided in Section 6.

2. Materials and methods

To perform the bibliometric analysis described in this paper, the workflow proposed in [7,8] has been followed, which is composed of the following steps:

(1) **Data retrieval.** The data processed in this paper come from ISIWoS and Scopus, which are the most reliable bibliographic databases at the moment [9,10]. On September 2015, the following query[1] was made on Scopus and the ISIWoS Core Collection for the time span 1995–2014:

```
TOPIC =
''software product line*'' or
(
(
''product line*'' or ''mass customization'' or ''product famil*'' or
''program famil*'' or ''software factor*'' or ''product platform*''
) and
(
''domain engineering'' or ''application engineering'' or
''feature model*'' or ''feature diagram*'' or
''decision model*'' or ''decision diagram*'' or
(software and variabilit*)
)
)
```

Venn diagram in Fig. 1 depicts the number of publication records provided by each database: 1387 records from ISIWoS and 2652 records from Scopus. 1194 of those records were common for both Scopus and ISIWos. Thus, the total number of records processed in this paper is 2845.

[1] The asterisk pattern character means zero to many characters; it is used in our query to catch the noun plurals.

(2) **Data aggregation.** The scatter plot in Fig. 2 shows the number of citations of the records common to ISIWoS and Scopus, including the corresponding regression line as well. ISIWoS has a more selective procedure to include bibliographic references than Scopus. As a result, Scopus provides more records than ISIWoS, and the citations tend to be higher as well.
For the query we performed, the Pearson's correlation coefficient of the citation counts is 0.87. So the information provided by both databases is rather consistent.
To combine the records, the citation count for the common records was computed as the maximum of the citations given by ISIWoS and Scopus.

(3) **Preprocessing.** The data retrieved from bibliographic databases usually have errors. For instance, references may be duplicated, authors' names may appear in different ways, etc. So, it is necessary to preprocess the data before carrying out any analysis.
To track the evolution of the SPL research area and measure its performance, we have used two approaches that require analyzing publication keywords and citations: Co-Word Analysis and H-index. Hence, we have performed a laborious preprocessing procedure to
   (1) *Correct invalid citations*; e.g., the technical report [11] appears cited in the raw data gathered from ISIWoS as "Bachman F., 2005, CMUSEI2005 TR012" and "Bachman**n** F., 2005, CMUSEI2005TR012".
   (2) *Standardize keywords*. From the ISIWoS records, a set of 2000 keywords was available, 1667 were authors' keywords and 333 were words provided by ISIWoS *KeyWords Plus* (index terms created by Thomson Reuters from significant, frequently occurring words in the titles of an article cited references). The Scopus records included a set of 9308 keywords. As Fig. 3 summarizes, the initial aggregated set of 11,308 keywords was progressively reduced by applying the following steps:
      (a) Keywords were converted to uppercase, leading and trailing white-spaces were removed, and inner white-spaces were replaced by the character '-'. After that, the repeated keywords were removed and plurals were grouped.
      (b) Keywords useless to identify research topics inside the SPL area were discarded. For example, SOFTWARE-PRODUCT-LINE, PRODUCT-FAMILY, SOFTWARE-ENGINEERING, etc. are applicable to all the records and thus they cannot be used to distinguish particular topics. Therefore, those general keywords were removed.
      (c) Keywords were grouped according to their meaning. To improve the interpretability of the co-word analysis results, the set of keywords was reduced by grouping those words that refer to the same topic. For instance, STAGED-CONFIGURATION, AUTOMATED-CONFIGURATION, FEATURE-BASED-CONFIGURATION, PRODUCT-DERIVATION-TOOL, etc. were grouped as PRODUCT-DERIVATION.

(4) **Analysis.** There are two main approaches to examine bibliographic data [12]:
   (1) *Performance analysis* tries to quantify the impact of scientific actors (researchers, journals, etc.). In particular, the performance analysis we have carried out is based on H-index, which is introduced in Section 2.1.
   (2) *Science Mapping* looks for identifying the cognitive structure and evolution of a research field. Section 2.2 summarizes the techniques used in this paper for mapping the SPL research area.