COMPUTER SCIENCE REVIEW 19 (2016) 56–77

Available online at www.sciencedirect.com

ScienceDirect

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

Survey

Citations, research topics and active countries in software engineering: A bibliometrics study

Vahid Garousi[a,*], *Mika V. Mäntylä*[b]

[a] *Software Engineering Research Group, Department of Computer Engineering, Hacettepe University, Ankara, Turkey*

[b] *M3S, Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland*

ARTICLE INFO

Article history:
Received 15 December 2015
Accepted 22 December 2015
Published online 2 March 2016

Keywords:
Software engineering
Research literature
Citation analysis
Thematic and topic analysis
Bibliometrics

ABSTRACT

Context: An enormous number of papers (more than 70,000) have been published in the area of Software Engineering (SE) since its inception in 1968. To better characterize and understand this massive research literature, there is a need for comprehensive bibliometrics assessments in this vibrant field.

Objective: The objective of this study is to utilize automated citation and topic analysis to characterize the software engineering research literature over the years. While a few bibliometrics studies have appeared in the field of SE, this article aims to be the most comprehensive bibliometrics assessments in this vibrant field.

Method: To achieve the above objective, we report in this paper a bibliometrics study with data collected from Scopus database consisting of over 70,000 articles. For thematic analysis, we used topic modeling to automatically generate the most probable topic distributions given the data.

Results: We found that number of papers published per year has grown tremendously and currently 6000–7000 papers are published every year. At the same time, nearly half of the papers are not cited at all. Using text mining of articles titles, we found that currently the hot research topics in software engineering are: (1) web services, (2) mobile and cloud computing, (3) industrial (case) studies, (4) source code and (5) test generation. Finally, we found that a small share of large countries produce the majority of the papers in SE while small European countries are proportionally the most active in the area of SE, based on the number of papers.

Conclusion: Due to large volumes of research in SE, we suggest using the automated analysis of bibliometrics as we have done in this paper. By picking out the most cited papers, we can present the land marks of SE and, with thematic analysis, we can characterize the entire field. This can be useful for students and other new comers to SE

* *Corresponding author.*
E-mail addresses: vahid.garousi@hacettepe.edu.tr (V. Garousi), mika.mantyla@oulu.fi (M.V. Mäntylä).

http://dx.doi.org/10.1016/j.cosrev.2015.12.002

and for presenting our achievements to other disciplines. In particular, we see and report the value of such an analysis in situations where performing a full scale SLR is not feasible due to restrictions on time or to lack of exact research questions.

Contents

1. Introduction

According to the data from the Scopus publication database, more than 70,000 papers have been published in the area of Software Engineering (SE) since its inception in 1968. As the SE research literature has grown tremendously, there is a need for bibliometrics studies in this area. Bibliometrics is a set of methods to quantitatively analyze research literature.

Bibliometrics studies in SE have focused in the following areas; (a) generating ranking lists of top performing institutions and scholars [1–9], (b) citation analysis to identify the most popular articles [10–13], and (c) content analysis of SE research [14–16]. Papers in area (a) can mainly be used internally within the SE research community. Papers on areas (b) and (c) can be used to explain our science to outsiders, e.g. to funding authorities or to scientists representing other disciplines. Additionally, such works can be helpful in teaching students about software engineering research or to highlight the top areas under study to industry, and help outsider to get acquainted with the latest research trends. Thus, bibliometrics papers can be important aid in distributing knowledge beyond the software engineering community.

New bibliometrics studies are needed regularly to keep up with the most recent research developments. Furthermore, this study contributes beyond the past works in the following ways. First, this study covers the largest pool of software engineering papers so far 72,787 papers, for example this is over two times more than in prior work that analyzed 26,624 papers [17]. Second, we analyze the citations in the SE research literature. The past series of work by Wohlin [10–13] in this area covers only papers published in selected SE journals and analyzes papers on individual years only, whereas we cover far greater area of publication forums. Furthermore, Wohlin does not consider the citations landscape beyond individual papers. Third, we present automated topic analysis to identify software engineering research themes and the hot and cold research topics in SE. Past work in this area has manually analyzed a rather small set of articles, e.g., Glass et al. [14] manually analyzed a small set of papers ($n = 369$) from six leading SE journals. Cai and Card [15] analyzed 691 papers from 7 leading journals

SE and 7 leading conferences SE. To our knowledge, the only automated thematic analysis of SE literature is by Coulter et al. [16] who in 1998 performed co-word analysis using ACM Computing Classification System. Our study on research topics is automated, focuses on our entire corpus and follows the approach by Griffiths and Steyvers [18]. In summary, the contributions of this paper are four-fold:

- The most comprehensive citation analysis reported to date on the entire SE research literature (Section 4.2).
- Topics and thematic analysis of the entire SE research literature (Section 4.3).
- Ranking of the world nations by the number of SE papers contributed by each country (Section 4.4).
- To enable other researchers to conduct similar types of analyses, the entire raw dataset (including 71,668 papers) has been made available as an Excel file which can be downloaded online [19].

Section 2 discusses the related work in which we briefly review the existing bibliometrics studies in SE. We then present in Section 3 the research methodology, the data source and data extraction process which we used to prepare the pool of all SE papers used later for analysis. Section 4 presents the results of the study. Section 5 summarizes the findings, implications, and discusses the potential threats to validity of our study. Finally, Section 6 concludes this study and states the future work directions.

2. Related work: existing bibliometrics studies in SE

A number of bibliometrics studies have been published in SE, several of which are discussed next. Table 1 lists a few representative studies along with their notable findings.

The sequential series of four papers by Wohlin [10–13] analyzes the most cited papers in SE journals between 1999 and 2002. As discussed by Wohlin, the intention of the analysis in those four papers was twofold: (1) first, to identify the most cited papers, and (2) second, to invite the authors of the most cited papers to contribute to a special section of the Information and Software Technology journal.

Cai and Card [15] analyzed 691 papers from 7 leading journals SE and 7 leading conferences SE. Among their findings was that 73% of journal papers focus on 20% of subjects in SE, including testing and debugging, management, and software/program verification.

The series of 12 papers by Glass et al., three of which are cited in Table 1 [4,5,20], was an ongoing, annual event that identified the top-15 SE scholars and institutions for the five-year period in systems and software engineering between 1995 and 2006. The rankings were based on the number of papers published in a selected set of leading SE journals.

The study reported in [21] presented a bibliometric assessment of Canadian SE scholars and institutions. Additional findings reported in [21] included correlation analysis of the SE research productivity (output in terms of number of papers) of Canadian provinces versus their national research grant amounts.

Focusing on specific sub-areas under SE, the study reported in [22] presented a bibliometric analysis of ten years of search-based SE. Some recent systematic mapping (SM) have included bibliometric analyses of SE sub-areas, e.g., development of scientific software in [23]. Among the findings reported in [23] was that the most active authors in the area of development of scientific software were mostly located in the US (approximately 50%), followed by the Canadian and British researchers.

Ren and Taylor's developed a Java tool [24] in 2007 and used it for automatic publication ranking of research institutions and scholars [24] presented a proof of concept of that tool in ranking SE institutions and scholars. The tool incorporates the impact factors of publication venues. Again, similar to works of Glass et al. [5,6], instead of covering the entire SE research literature landscape, only a selected subset of SE journals were considered. In a previous work [21], the first author and a colleague used Ren and Taylor's tool in 2010 and presented a bibliometric ranking and assessment of the Canadian SE scholars and institutions with data covering the time window of 1996–2006.

More recently, in a 2013 paper [17], Garousi and Ruhe conducted and reported a bibliometric/geographic assessment of the entire SE research landscape covering the papers published between 1969 and 2009. Among the most interesting findings of [17] are: (1) Over the 40 years, in total about 60% of the SE literature has been contributed by only 7% of all countries, (2) the SE research output of different countries does not necessarily correlate with their GDPs, (3) the share of contributions to the SE discipline by the American researchers has declined from 71.4% (in 1980) to 14.9% (in 2008), and (4) China is the country with the biggest share growth in the number of publications, from 0.8% of the entire SE publications in 1991 to 13.8% in 2009.

While [17] reported interesting findings as discussed above, the dataset used in that study lacked the citation data of the papers and thus it was impossible to conduct citation analysis in the context of the SE literature. The current study intends to fill those gaps by extracting and analyzing the citation landscape for the SE literature. Furthermore, in this paper we also study the search for SE research with topic modeling by partially replicating a popular paper by Griffiths and Steyvers [18] who applied topic modeling (text-mining technique) to discover scientific topics. Also, the current study widens the analysis time window of [17] (1969–2009) by including the latest papers in the study pool as well, i.e., considering the publication time window of 1969–2014. Finally, the number of papers analyzed is larger 72,787 versus 26,624.

The paper entitled "*Trends in computer science research*" [25] is related since CS is closely related to SE. This paper identified trends, bursty topics, and interesting inter-relationships between the American National Science Foundation (NSF) awards and CS publications, finding, for example, that if an uncommonly high frequency of a specific topic is observed in publications, the funding for this topic is usually increased.

Fernandes reports a bibliometric study [26] which focuses on authorship trends in SE. The researcher collected around 70,000 entries from the DBLP (a well-known online computer science bibliography website) for 122 conferences and

Table 1 – A few selected bibliometrics studies in SE (sorted by years of publications).

Ref.	Year	Topic	Notable findings
[10]	2005	An analysis of the most cited papers in software engineering journals—1999	• An analysis of the 20 most cited SE journal papers in the 20 year period of 1979–1999 is presented. • Most cited papers are ranked using two metrics: absolute numbers of citations and the average number of citations per year. • The research topics and methods of the most cited papers in 1999 are compared with those from the most cited papers in 1994 to provide a picture of similarities and differences between the years. • The top cited paper is "use case maps as architectural entities for complex systems" [38] with only 25 citations.
[11]	2007	An analysis of the most cited papers in software engineering journals—2000	• The paper describing the SPIN model checker [39] by G.J. Holzmann published in 1997 is the first using both metrics.
[12]	2008	An analysis of the most cited papers in software engineering journals—2001	• The most productive author in the 20-year period of 1981–2001 is Victor Basili.
[13]	2009	An analysis of the most cited papers in software engineering journals—2002	• The top cited paper is "Preliminary guidelines for empirical research in software engineering" with 64 citations.
[15]	2008	An analysis of research topics in software engineering—2006	• The paper examines all the 691 papers published in a selected list of venues in 2006. • 73% of journal papers focus on 20% of subjects in SE, including testing and debugging, management, and software/program verification. • 89% of conference papers focus on 20% of subjects in SE, including software/program verification, testing and debugging, and design tools and techniques. • The average number of 7 top journals and 7 top international conferences in SE references cited by a journal paper is about 33, whereas this number becomes around 24 for a conference paper.
[4]	2008	Assessment of systems and software engineering scholars and institutions (2001–2005)	• The rankings are calculated based on the number of papers published in journals: IEEE TSE, TOSEM, JSS, SPE, EMSE, IST, and IEEE Software. • The top scholar is Magne Jørgensen of Simula Research Laboratory, Norway. • The top institution is Korea Advanced Institute of Science and Technology, Korea.
[5]	2009	Assessment of systems and software engineering scholars and institutions (2002–2006)	• The top-ranked scholar is Magne Jørgensen of Simula Research Laboratory, Norway. • The top-ranked institution is Korea Advanced Institute of Science and Technology, Korea.
[21]	2010	Bibliometric assessment of Canadian software engineering scholars and institutions (1996–2006)	• The study used two metrics: impact factors, and h-index, based on papers published in top 12 selected software engineering journals and conferences. • The top-ranked institution is Carleton University. • The top-ranked scholars (by each of the two metrics) are Lionel Briand (formerly with Carleton University) and Gail Murphy from UBC.
[22]	2011	Ten years of search-based software engineering: a bibliometric analysis	• The study covered 740 publications of the SBSE community from 2001 through 2010. • The performed bibliometric analysis concerned mainly in four categories: publication, sources, authorship, and collaboration. The study also analyzed the applicability of bibliometric laws in SBSE, such as Bradfords and Lotka.

(*continued on next page*)

journals, for the period 1971–2012. Interestingly enough, the author indicated that the number of authors of articles in SE is increasing on average around 0.40 authors/decade. Also, the results indicate that until 1980, the majority of the articles have one author, while articles from 90s until today with 3 or 4 authors represent almost half of the total number of papers. Since the average number of authors of scientific articles is increasing, it was the opinion of the researcher that the system of authorship is consequently becoming inappropriate, in the sense that it becomes more difficult to credit all the authors for the specific contributions they made to each article. Therefore, the researcher suggests that the

Table 1 (*continued*)

Ref.	Year	Topic	Notable findings
[20]	2011	Assessment of systems and software engineering scholars and institutions (2003–2007 and 2004–2008)	• The top-ranked institution is Korea Advanced Institute of Science and Technology, Korea for 2003–2007, and Simula Research Laboratory, Norway for 2004–2008. • Magne Jørgensen is the top-ranked scholar for both periods.
[23]	2011	Development of scientific software: a systematic mapping, bibliometrics study and a paper repository	• 17 out of 130 publications in the pool were cited more than 25 times. • The most active author in the field is Diane Kelly, with Royal Military Collage of Canada, with a total of ten (co-authored) publications. • The authors' most frequent affiliations are located in the US (approximately 50%), followed with a large distance by Canada and the UK.
[17]	2013	Bibliometric/geographic assessment of 40 years of software engineering research (1969–2009)	• The first bibliometric quantitative analysis of publications in SE, including relative and absolute growth in the number of all SE publications as well as an analysis among countries. • Over the 40 year period (1969–2009), in total about 60% of the SE literature has been contributed by only 7% of all countries. • The US is the clear leader, followed by UK and China. • The SE research output of different countries does not necessarily correlate with their GDPs. • The share of contributions to the SE discipline by the American researchers has declined from 71.43% (in 1980) to 14.90% (in 2008). • China is the country with the biggest share growth in the number of SE publications (from 0.82% of the entire SE publications in 1991 to 13.82% in 2009).
[25]	2013	Trends in computer science research	• Only a small fraction of authors attribute their work to the same research area for a long period of time, reflecting for instance the emphasis on novelty (use of new keywords) and typical academic research teams. • Highlighted the dynamic research landscape in CS, with its focus constantly moving to new challenges arising from new technological developments. • Computer science is atypical science in that its universe evolves quickly, with a speed that is unprecedented even for engineers.
[26]	2014	Authorship trends in SE	• Around 70,000 entries from the DBLP for 122 conferences and journals, for the period 1971–2012, were collected. • The number of authors of articles in SE is increasing on average around 0.40 authors/decade. • Until 1980, the majority of the articles have one author, while articles from 90s until today with 3 or 4 authors represent almost half of the total number of papers.

(*continued on next page*)

SE community must establish an agreed publishing standard to define how to assign the academic contribution to all collaborators of a research project.

Garousi (the first author of the current paper) recently conducted and published a bibliometric assessment [27] of Turkish software engineering scholars and institutions covering years 1992–2014. Among the results were that: (1) Turkey produces only about 0.49% of the world-wide SE knowledge, as measured by the number of papers in Scopus, which is very negligible unfortunately. (2) There is a lack of diversity in the general SE spectrum in Turkey, e.g., we noticed very little focus on requirements engineering, software maintenance and evolution, and architecture. This denotes the need to further diversification in SE research topics in Turkey, and (3) In total, 89 papers in the pool (30.8% of the total) are internationally-authored SE papers. Having a good level of international collaborations is a good sign for the Turkish SE community. The current article follows the same bibliometric approach as was conducted in [27] (details are discussed in Section 3).

Garousi and Fernandes conducted and reported a recent bibliometric assessment [28] to identify the top-100 highly-cited papers in SE in terms of two metrics: total number of citations and average annual number of citations. These two researchers argued that, as the subject of research excellence has received increasing attention (in science policy) over the last few decades, increasing numbers of bibliometric studies have been published dealing with characterizing and ranking highly-cited papers [29]. For example, the cover story of the October 2014 issue of the prestigious *Nature* magazine was "*The top 100 papers*" [30]. That Nature issue includes several papers (e.g., [31]) on the issue of highly-cited papers

Table 1 (continued)

Ref.	Year	Topic	Notable findings
[27]	2015	Bibliometric assessment of Turkish software engineering scholars and institutions (1992–2014)	• Turkey produces only about 0.49% of the world-wide SE knowledge, as measured by the number of papers in Scopus, which is very negligible unfortunately. • There is a lack of diversity in the general SE spectrum in Turkey, e.g., we noticed very little focus on requirements engineering, software maintenance and evolution, and architecture. This denotes the need to further diversification in SE research topics in Turkey. • In total, 89 papers in the pool (30.8% of the total) are internationally-authored SE papers. Having a good level of international collaborations is a good sign for the Turkish SE community.
[28]	2016	Highly-cited papers in software engineering: The top-100	• A study, comprised of five research questions, to identify and classify the top-100 highly-cited SE papers in terms of two metrics: total number of citations and average annual number of citations. • By total number of citations, the top paper is "A metrics suite for object-oriented design", cited 1817 times and published in 1994. By average annual number of citations, the top paper is "QoS-aware middleware for Web services composition", cited 154.2 times on average annually and published in 2004. • It was concluded that it is important to identify the highly-cited SE papers and also to characterize the overall citation landscape in the SE field. It was hope that this paper would encourage further discussions in the SE community towards further analysis and formal characterization of the highly-cited SE papers, as it has been done in other fields.

in various scientific disciplines. Garousi and Fernandes [28] report, among other things, that: by total number of citations, the top paper is "*A metrics suite for object-oriented design*", cited 1817 times and published in 1994. By average annual number of citations, the top paper is "*QoS-aware middleware for Web services composition*", cited 154.2 times on average annually and published in 2004. Garousi and Fernandes [28] also identified works pointing out possible determinants of the likelihood of high citations, e.g., based on a paper entitled "*Highly-cited works in neurosurgery*" [32], the determinants are: the time of publication, field of study, nature of the work, and the journal in which the work appears. One would wonder if those determinants are also applicable in the SE domain.

3. Research method and data extraction

In the following, the goal, research questions of our study and the metrics we have used are presented. We then present the data extraction phase of our study.

3.1. Goal and research questions

The goal of this study is to conduct a bibliometrics assessment in SE, focusing on citations and topics, to better characterize and understand the research literature in this field from the point of view of researchers. Based on the above goal, the following research questions (RQs) were raised (grouped under four categories). The goal and RQs of the study are *exploratory* and *descriptive* in nature [33].

- **RQ 1**: **Volume of papers**: How many SE papers have been published each year since the field's inception in 1968?
- **RQ 2**: **Citation landscape**: What is the citation landscape of the SE literature? This RQ has been divided into five sub-RQs.
  - **RQ 2.1**: What is the distribution of citations for the SE papers? For example, what ratio of SE papers has had no citations?
  - **RQ 2.2**: What are the highly-cited papers in SE?
  - **RQ 2.3**: What are the citation trends of different venue types? For example, do journal papers get more citations, on average, than conference papers?
  - **RQ 2.4**: What are the annual trends of citations in SE? For example, do older papers get more citations on average compared to newer papers?
  - **RQ 2.5**: How have the volume of and citations for papers in different SE sub-areas evolved over the years?
- **RQ 3**: **Topics and thematic analysis**: This RQ has been divided into three sub-RQs.
  - **RQ 3.1**: How have focus areas of the papers changed over the years?
  - **RQ 3.2**: What research topics have increased/decreased in popularity (hot and cold topics)?
- **RQ 4**: **the most active countries in SE**: How do different countries rank in terms of number of contributed papers?

3.2. Data source and data extraction

3.2.1. Selection of the publication database

To identify the list of all SE papers, we had to select a suitable publication database. For systematic selection of such a database, by reviewing the related review studies (discussed in Section 3), we devised three important selection criteria:

1. The publication database should provide the highest quality and reliability in terms of coverage of the SE literature, i.e., including all the SE papers.
2. The publication database should include the citation data for papers.

Table 2 – Rating of the three candidate publication databases in terms of the three selection criteria.

Criteria	Publication databases		
	Scopus	Web of science	Google scholar
1-Quality and reliability in terms of coverage of the SE literature	Since Scopus has the feature to search by "Source name" (venue names), quality and reliability of search results in terms of complete coverage can be achieved to a great extent.	Given the nature of SE papers, quality and reliability of search results in terms of complete coverage cannot be guaranteed.	Given the nature of SE papers, quality and reliability of search results in terms of complete coverage cannot be guaranteed.
2-Including citation data	Yes	Yes	Yes
3-Convenient/usable interface for searching and data extraction	Allows saving the list of all extracted papers into CSV files.	Only allows saving the list of extracted papers into CSV files on a page by page basis.	Exporting the list of extracted papers to files is not automatically possible. We were not able to find any API for it.

3. The publication database should provide a convenient/usable interface to search and extract the citation data.

To find the candidate publication databases, we reviewed a large number of bibliometrics studies, in SE (e.g., [5,6,17,21,22]), and fields other than SE (e.g., [34–37]). We short-listed the candidate publication databases as follows: DBLP (www.dblp.org), Scopus (www.scopus.com), Web of Science (www.webofknowledge.com) and Google Scholar (scholar.google.com). These databases are among the most popular databases that researchers regularly use in various bibliometrics studies. DBLP was not further considered, since it does include citation data. In Table 2, we discuss how the remaining three candidate publication databases rate in terms of the selection criteria discussed above.

Regarding criterion #3, as we discuss in Table 1, Google Scholar became ineligible for our selection, since exporting the list of extracted papers to files is not automatically possible in a convenient manner (except that one has to write complex scripts), and we were not able to find any API for it. One can easily imagine that manual analysis of huge number of SE papers using Google Scholar would be very time consuming. Web of Science only allows saving the list of extracted papers into CSV files on a page by page basis, e.g., if the paper search results returns 100 pages of papers, exporting the data would be very tedious. Only Scopus allows saving the list of all extracted papers into CSV files. Thus, this is an advantage of Scopus over Web of Science.

Regarding criterion #1, as we discuss in Table 1, Scopus scores better than Web of Science, since Scopus has the feature to search by "Source name" (venue names). Thus, using Scopus, quality and reliability of paper search results in terms of complete coverage of the SE domain can be achieved to a great extent, i.e., as we discuss in the following, we included in the search query the phrase "software" in venue names which we found to be a suitable approach to ensure including almost all major SE journals and conferences in the search approach. Given the nature of SE papers, quality and reliability of search results in terms of complete coverage cannot be guaranteed using Web of Science, since searching by paper title having the phrase "software engineering" does not guarantee including all the SE papers as many SE paper do not explicitly include that phrase in their title, nor in the abstract, nor in the keywords. The first author actually experienced this challenge in a recent bibliometrics study [17] in which a bibliometric/geographic assessment of 40 years of SE research (1969–2009) was reported. All the major SE venues including the top SE conferences and journals, e.g., ICSE, ICSM, ICST, IEEE TSE, ACM TOSEM, were included in the results returned by Scopus when the search via source name including 'software' was conducted.

Regarding criterion #2, all three candidate publication databases include citation data (i.e., the number of times a given paper has been cited).

In conclusion, by summarizing the outcomes with respect to our three selection criteria, the Scopus publication database was chosen as the publication database from which the set of SE papers would be identified. A recent paper published in the Nature magazine, titled "The top 100 papers" [30], which was discussed in Section 2, also used Scopus. There have been empirical studies, e.g., [34–37], which have compared the performance and coverage of Web of Science versus Scopus in several fields, e.g., social sciences. Some studies, e.g., [36], have found empirically that Scopus is better than Web of Science in certain aspects, e.g., "larger coverage of titles" [36].

3.2.2. *Extraction of all SE papers from Scopus*

Having selected Scopus as the publication database to conduct the search for the SE papers, the next step was to actually conduct the search for those papers.

We found that, when conducting searches in Scopus, including the phrase "software" in "source title" (a term used in Scopus interface meaning the conference or journal where a paper has been published) is a suitable approach to ensure targeting the entire SE literature with a high precision (coverage). By experimentation, we found that this approach is indeed quite reliable in terms of coverage of the SE literature and has been used in other disciplines as well [29–32,40–53]. We should further note that the same approach has showed to be effective and it has also been used in two other recent bibliometric studies by the first author of the current article: (1) in a recent bibliometric assessment [27] of the Turkish SE scholars and institutions by extracting the list of all SE papers which have originated from Turkey (authored

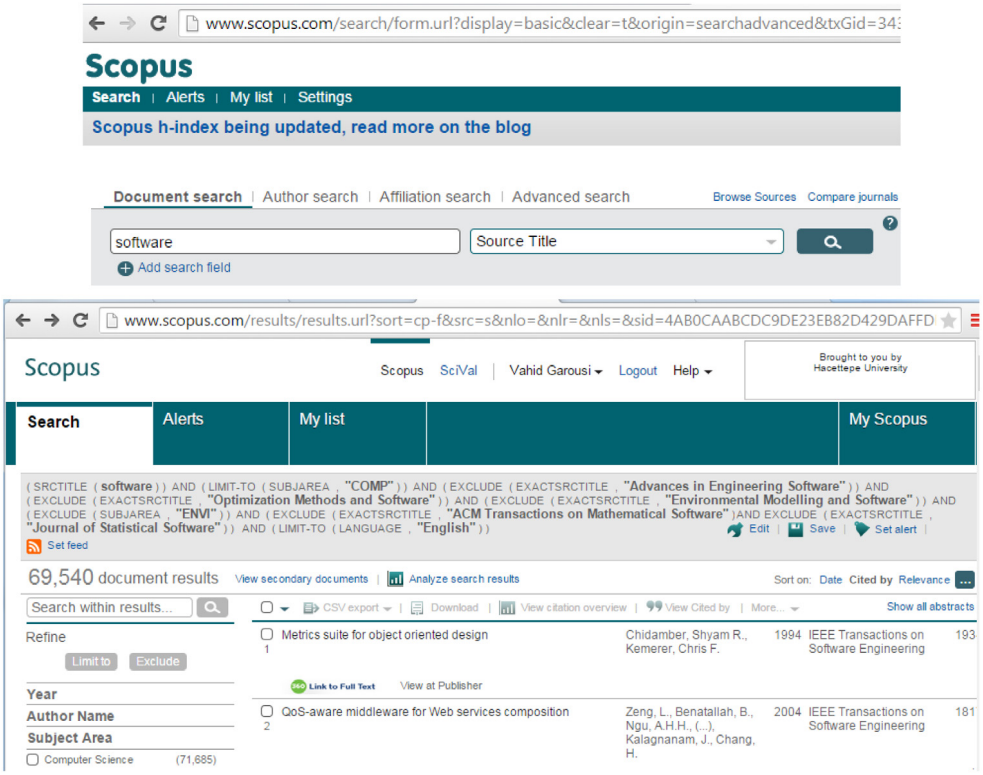

Fig. 1 – Two screenshots showing the method used to identify the top papers in the Scopus publication database (www.scopus.com).

or co-authored by Turkish authors) using the same approach, (2) in a recent bibliometric assessment to identify the top-100 highly-cited papers in SE [28].

In the Scopus search interface, we included the phrase "software" under "source title" as shown in Fig. 1. The exact search query that was developed to extract all SE papers from Scopus is shown in Table 3 along with explanations for each phrase in the query. We conducted several rounds of iterative review and excluded unrelated venues (such as, Journal of Optimization Methods and Software) and also non-English papers.

We should also note that the data extraction phase of this study was conducted on Dec. 25, 2014. Even if the analysis was done at the end of 2014, as per our analysis, we found that it takes a while for the Scopus database engine to record/import all the data from other sources (it seems that there is some sort of a batch processing scheme in place). Thus, the data for 2014 were partial. Furthermore, the citations for papers in 2014 were relatively very low since they were either "In Press" or recently published. For instance, our analysis showed that the 2443 papers (partial count as per the Scopus approach discussed above) published in 2014 had 203 citations, while for 6403 papers published in 2013, there were 3365 citations. Due to the partial situation of the 2014 dataset, we decided to not include the 2014 papers altogether in our dataset and used 2013 as the last publication year.

As a result of applying the above approach, we had an initial dataset of 69,540 papers. Obviously, all the major SE venues including the top SE conferences and journals such as ICSE, ICSM, ICST, IEEE TSE, ACM TOSEM, were included in the results returned by Scopus since all the names include the word 'software'.

Furthermore, we were also aware that a number of SE-related venues do not have the term 'software' in their titles, such as the following ones:

- Venues on requirements engineering: Springer Journal on Requirements Engineering and the International Requirements Engineering Conference (RE).
- Venues including the "Formal Methods" phrase: Formal Methods in System Design (journal), and the International Symposium on Formal Methods (FM).
- International Conference on Program Comprehension (ICPC).
- Working Conference on Reverse Engineering (WCRE).
- International Conference on Model-Driven Engineering Languages and Systems (MoDELS).
- International Conference Technology of Object-Oriented Languages and Systems (TOOLS).
- European Conference on Object-Oriented Programming (ECOOP).
- Object-Oriented Programming, Systems, Languages & Applications (OOPSLA).

We should mention that, at some point, the line between SE and other related disciplines such as the programming language community often seems "gray". Thus, for the purpose of this study, we had to draw the border somewhere. As we have listed in the above additional list of venues not including the term 'software', we included those that have a focus on object-oriented concepts and thus related to the design phase of SE.

Table 3 – The search query that was developed to extract SE papers from Scopus.

Search query:	Explanations:
(SRCTITLE (software)) AND	Only venues with the "software" phrase
(LIMIT-TO (SUBJAREA, "COMP")) AND	Only the sub-area of "Computer Science"
(EXCLUDE (EXACTSRCTITLE, "Advances in Engineering Software")) AND	Excluding this particular journal
(EXCLUDE (EXACTSRCTITLE, "Optimization Methods and Software")) AND	Excluding this particular journal
(EXCLUDE (EXACTSRCTITLE, "Environmental Modelling and Software")) AND	Excluding this particular journal
(EXCLUDE (SUBJAREA, "ENVI")) AND	Excluding the sub-area of environmental science
(EXCLUDE (EXACTSRCTITLE, "ACM Transactions on Mathematical Software") OR	Excluding this particular journal
EXCLUDE (EXACTSRCTITLE, "Journal of Statistical Software"))AND	Excluding this particular journal
(LIMIT-TO (LANGUAGE, "English"))	Only including papers written in English

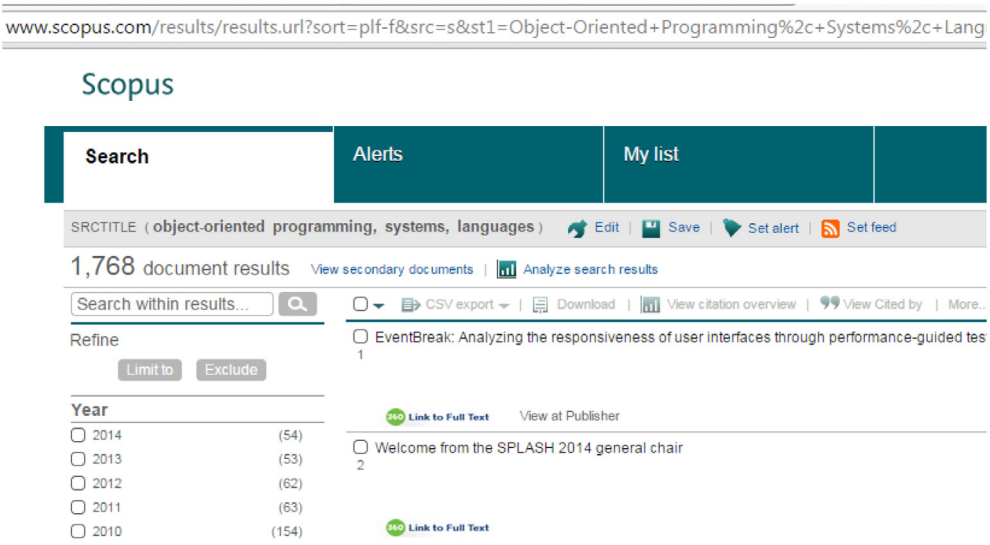

Fig. 2 – Screenshot showing the query used to identify papers published in the proceedings of the Conference on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA).

We conducted searches for the above venues separately (in the first week of May 2015), and as a result, 3240 additional papers were found and added to the pool. As an example, Fig. 2 shows the query used to extract the list of papers published in the proceedings of the Conference on Object-Oriented Programming, Systems, Languages and Applications (OOPSLA).

We should add that Scopus stores the following 12 document (resource) types: article, article in press, book, book chapter, conference paper, conference review, editorial, erratum, letter, note, review and short survey. We only wanted to include scientific papers, thus we included records of the following types only: articles, articles in press, book chapters, conference papers and review papers (e.g., survey and systematic review papers), and excluded the rest.

Once we had the pool of papers, we reviewed the records to ensure its integrity, e.g., not having duplicate records of a given paper. It was somewhat surprising that data exported from Scopus had some duplicates. We cleaned up the dataset and after applying all the above steps, the final paper pool was finalized with 71,668 papers. To ensure transparency and replicability of our analysis, and also to enable other researchers to conduct other types of analyses, the entire raw dataset for all the papers is available as an Excel file which can be downloaded online [19].

4. Results

4.1. RQ 1: *annual volume of papers over years*

In terms of the growth of the SE literature, Fig. 3 shows the number of SE papers included in Scopus by their publication year. The earliest publication year was 1972 from which 29 papers were included in Scopus. The annual number of papers has grown and reached 6317 papers in 2013. A major growth after year 2004 is visible.

4.2. RQ 2: *citation analysis*

4.2.1. *RQ 2.1: citation landscape*

Citations are crucial in any research to position the work and to build on the work of others. A high citation count is usually considered an indication of the influence and impact of a given paper [41].

Based on the data extracted from Scopus, Fig. 4 shows an overview of the SE citation landscape as a scatter plot of all the papers' citation counts versus publication years, along with the corresponding box-plots (in top and right side of Fig. 4). Note that there are 71,668 points on this scatter plot, corresponding to all papers in the pool.

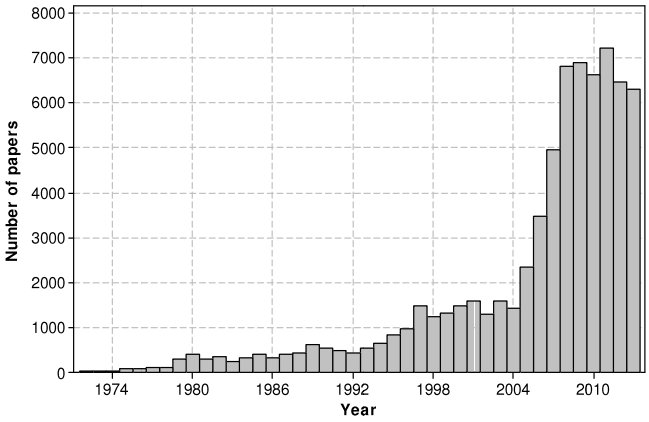

Fig. 3 – Number of SE papers included in Scopus by their publication year.

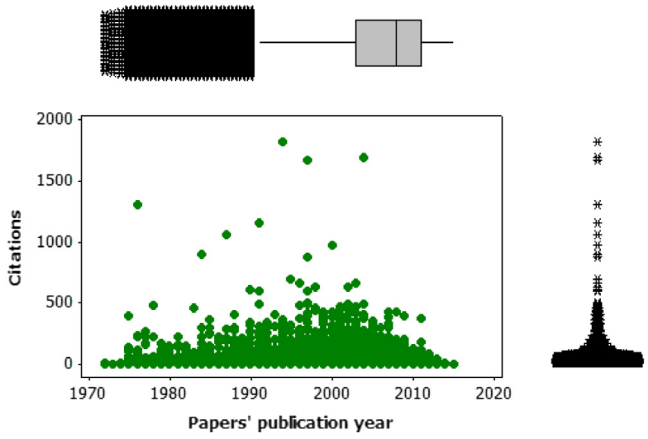

Fig. 4 – Scatter plot of citation counts versus publication years of all the SE papers (also including box-plots).

The cross black points in the two box-plots in the top (for publication years) and the right side of the chart (for citation values) are 'outliers' and, as the two box-plots depict, the data in both X and Y axes are somewhat (for the case of publication years) to extremely skewed (for the case of number of citations). This denotes that, for the case of publication years, most of the papers have been published in later years. For instance, 81.8% of the papers were published in the last 15 years (2000–2014), while the remaining 18.2% were published in the first 28 years (1968–1999). This shows that the volume of SE papers is experiencing a major growth lately. Note that the right box-plot in Fig. 4 is hidden under the numerous outlier points since there are many of such points. Let us recall that, as per notational rules of box-plots, a box-plot shows 25%–75% quartile of data in a 'box' notation and that quartile is quite tiny in the case of the right box-plot in Fig. 4, since half of the citation values are simply zero and other are quite small, as discussed next.

Out of all the 71,668 SE papers in the pool indexed in the Scopus publication database, 30,958 papers (~43% of the pool) had no citations at all, 10,095 papers (~14% of the pool) had only one citation. In total, 30,615 papers (~43% of the pool) had received more than one citation. The sum of all the citation numbers is 448,050. Thus, the average citation value is 6.82 per paper. The highest cited paper was cited 1817 times (to be discussed in further detail in Section 4.2.2). Fig. 5 shows the histogram of the citation data for all the SE papers.

Focusing on the issue of inequality in citation distributions, there are many studies in the scientometrics and bibliometrics literature, from as early as in the 1960s, e.g., [54–58]. In a classical book titled "*Little Science, Big Science*" and written in 1963 [54], the author observed that only about six percent of publishing scientists produce one-half of all papers published. Allison and Stewart [55] demonstrated that counts of citations to scientists' work are even more unequally distributed than counts of publications.

More recently, a 2014 paper [58] adopted the well-known Gini index, from the economy literature, to quantitatively measure inequality in academic institutions and science journals. The study showed a universal nature of academic inequalities in terms of citations. In economy and social sciences, the Gini coefficient (also known as the Gini index or Gini ratio) is a measure of statistical dispersion intended to represent the income distribution of a nation's residents, and is the most commonly used measure of inequality.

While we showed an initial view of the citation inequality in the SE literature in the histogram of Fig. 5, it would be interesting to explore this issue in further depth in

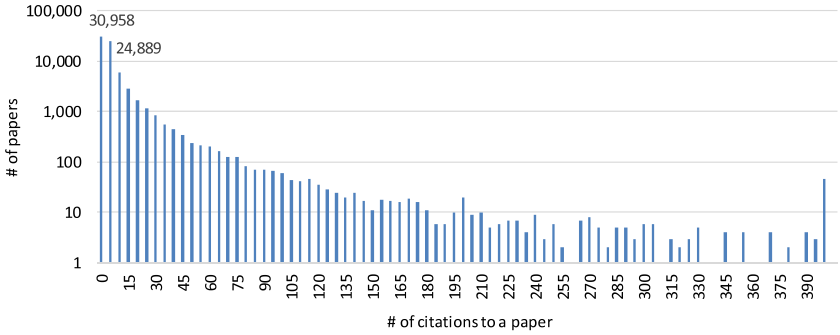

Fig. 5 – Histogram of citation data for all the SE papers included in Scopus.

future studies by adopting rigorous approaches from the scientometrics literature, e.g., [54–58].

4.2.2. RQ 2.2: *highest-cited papers*

This RQ was the main RQ of another recent bibliometric study in which the first author was involved in [28]. We thus do not intend to duplicate those results here, but only would like to report brief results to establish the linkage between the two studies and to invite the reader to review that paper [28] for in-depth analyses of highest-cited papers in SE.

To identify the highest-cited papers, we used two metrics: absolute numbers of citations and the average annual number of citations to a given paper, since its publication year until 2014. The latter metric normalizes the effect of publication year (age) on the total numbers of citations and has been used in many bibliometrics studies. The top five papers using each of the two metrics are shown in Tables 4 and 5. For the list of top-100 papers and more comprehensive discussions, refer to [28].

Two of the top five papers appear in both rankings. We can see that both old and new papers are appearing in the top lists, e.g., the paper titled "*Complexity measure*" from 1976 and "*Guidelines for conducting and reporting case study research*" from 2009.

Identification and classification of highly-cited papers are common and are regularly reported in various disciplines, e.g., biology, medicine, ecology, and social sciences. More recently, the cover story of the October 2014 issue of the prestigious *Nature* magazine was "*The top 100 papers*" [30] which ranked the top-100 papers of all areas of science. The study reported that only 14,499 papers out of 58 million items indexed in the Thomson Reuter's Web of Science have more than 1000 citations. The top three papers identified in [30] were cited 305,148; 213,005 and 155,530 times and all three were "biological lab techniques".

4.2.3. RQ 2.3: *volume and citation statistics for different publication types*

As discussed in Section 3.2, Scopus stores the following 12 document (resource) types in its database: article, article in press, book, book chapter, conference paper, conference review, editorial, erratum, letter, note, review and short survey. We only wanted to include scientific papers, thus we included records of the following five types only: articles, articles in press, book chapters, conference papers and review papers (e.g., survey and systematic review papers), and excluded the records of the other types.

We calculated six types of statistics for different documents types, as shown in Table 6. In terms of the ratio of the papers, journal and conferences papers, by covering 31.4% and 66.0% of the pool, are in the majority. In terms average number of citations per document type, review papers (e.g., surveys and systematic reviews) and journal articles, with averages of 18.4 and 12.6, are the top two. Thus, it seems that, as one would expect, review papers are quite popular and receive relatively high citations compared to all other paper types.

In terms of median citation values, only journal and review articles have non-zero values, denoting that for the other types, the data is highly skewed towards zero. In terms of % of documents with no citations, about 61% of book chapters and 55% of conference papers have not received any citations. Understandably, a high ratio of articles in press also have no citations.

4.2.4. RQ 2.4: *annual analysis of citations*

Fig. 6 shows the annual number of papers and citations to papers published in different years. Both yearly and also cumulative values are shown. The citations to more recent papers (after 2008) are in a decreasing order, since as it is well known, more time is needed for the recent papers to get enough exposure and thus citations.

Next, we wanted to know how different is the number of citations to papers published in different years. Fig. 7 shows the trend of average citations to papers in different years, which is essentially the result of division of the values in Fig. 6. Also, a scatterplot of all the individual data points is shown.

In the first glance, the trend of Fig. 7 looks like the "hype cycle" (the trend form of which has been shown in Fig. 7 as well). However, as discussed next, we do not think the SE literature, as a whole, has such a characteristics. By a closer analysis of the papers published in earlier years of 1975–77 where a high peak is visible, we found that relatively small number of papers were published in those years but they have been quite influential in the area, and thus have received relatively high citations, which have led to high average values seen in Fig. 7. The citations to more recent papers (after 2005) are quite low, since as it is well known, again, more time is needed for recent papers to get enough exposure.

Table 4 – Top-five papers based on total number of citations.

Rank	Paper title	Publication year	Times cited
1	A metrics suite for object-oriented design	1994	1817
2	QoS-aware middleware for Web services composition	2004	1696
3	The model checker SPIN	1997	1669
4	Complexity measure	1976	1304
5	Graph drawing by force-directed placement	1991	1162

Table 5 – Top-five papers based on average annual number of citations.

Rank	Paper title	Publication year	Average citations	Total citations
1	QoS-aware middleware for Web services composition	2004	154.2	1696
2	CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms	2011	92.8	371
3	The model checker SPIN	1997	92.7	1669
4	A Metrics suite for object oriented design	1994	86.5	1817
5	Guidelines for conducting and reporting case study research in software engineering	2009	65.3	392

Table 6 – Volume and citation statistics by document types.

Statistics	Document types				
	Article	Article in press	Book chapter	Conference paper	Review
Total # in the pool	22,523	214	985	47,275	671
% of the pool	31.4%	0.3%	1.4%	66.0%	0.9%
Times cited (average)	12.6	0.3	2.5	3.6	18.4
Times cited (median)	2	0	0	0	4
% with no citations	33.2%	59.3%	61.3%	54.8%	27.7%
% with at least one citation	66.8%	40.7%	38.7%	45.2%	72.3%

4.2.5. RQ 2.5: *volume of and citations for papers in different SE sub-areas*

Our dataset (which is also available online [19]) is quite rich since, in addition to the analyses conducted above, it enables other types of analyses too. As the next analysis (to address RQ 2.5), we grouped papers by different SE sub-areas. To do this, our approach was to calculate the volume of papers in five representative SE sub-areas by searching in the paper titles. The five sub-areas are: 'requirement', 'test', 'maintenance', 'verification' and 'validation'. We additionally included V&V to complement the testing sub-area. Fig. 8 shows the trends. We should note that of course, there are limitations to this simple textual analysis and phrases with similar meanings to a topic have not been included, e.g., 'program comprehension' which is a topic under 'maintenance' has not been included. Recently after year 2004, there has been a major increase in the number of papers on testing compared to research focus on maintenance.

As the next analysis, since we had the citation data as well, we calculated the average number of citations to papers with 'requirement', 'test' and 'maintenance' in their titles and the results are shown in Fig. 9. As we can see, the trends in early years (from 1970 to 1990) for all three series were quite similar. Quite an abnormal situation occurs around years 1990–1992, in which a sudden increase in average number of citations to papers occurs. The trends in years after 1995 to date are quite similar among all three series, however, citations to testing papers are slightly higher than the other two.

4.3. RQ 3: *topics and thematic analysis*

To address RQ 3, we conducted two types of topics and thematic analysis: (1) by word cloud visualization of paper titles in different decades, and (2) topics analysis based on text-mining, which we report next.

4.3.1. RQ 3.1: *focus areas of the papers through each decade*

Research trends of every field change by time. We used word cloud analyses to see how the focus areas of SE papers have been changing by time. Fig. 10 shows the word cloud of subsets of paper titles, grouped by the decades of their publications years, e.g., 1980–1989. An online tool named Wordle (www.wordle.net) was used to generate these word clouds. For brevity, common words such as "software", "using" and "of" have been removed. As we can see, in earlier decades, e.g., 1970s, phrases such as "program" and "implementation" were the most common, while the focus areas have shifted to topics such as "analysis" and "design" in 1980s, to "process" and "engineering" in 1990s, and to different topics such as "model", "testing" and "web" in 2000s and afterwards.

4.3.2. RQ 3.2: *topics analysis based on text-mining*

We conducted a systematic trend analysis of SE research topics with text mining. More specifically, we used topic modeling and Latent Dirichlet Allocation (LDA) [18]. Topic models are statistical models for discovering abstract topics that

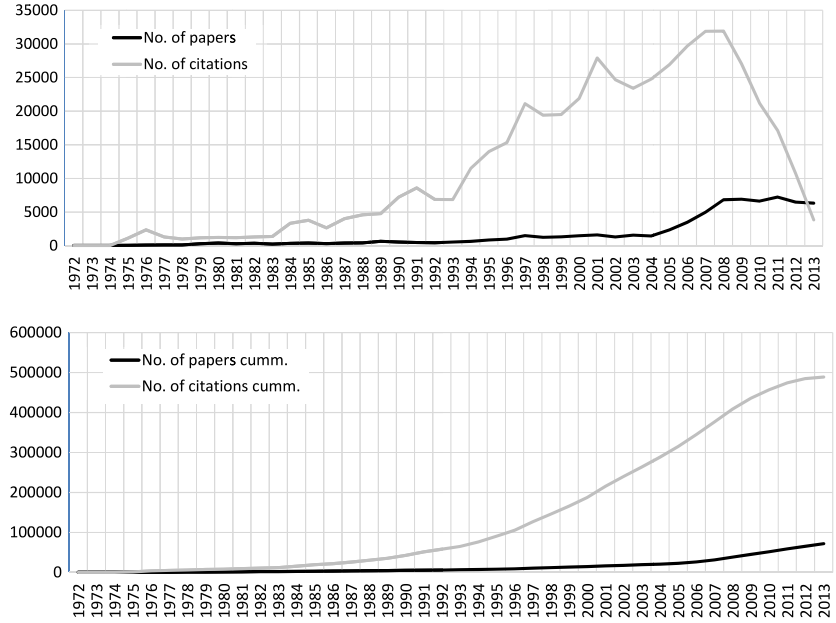

Fig. 6 – Annual number of papers and citations (top: yearly values, bottom: cumulative trend).

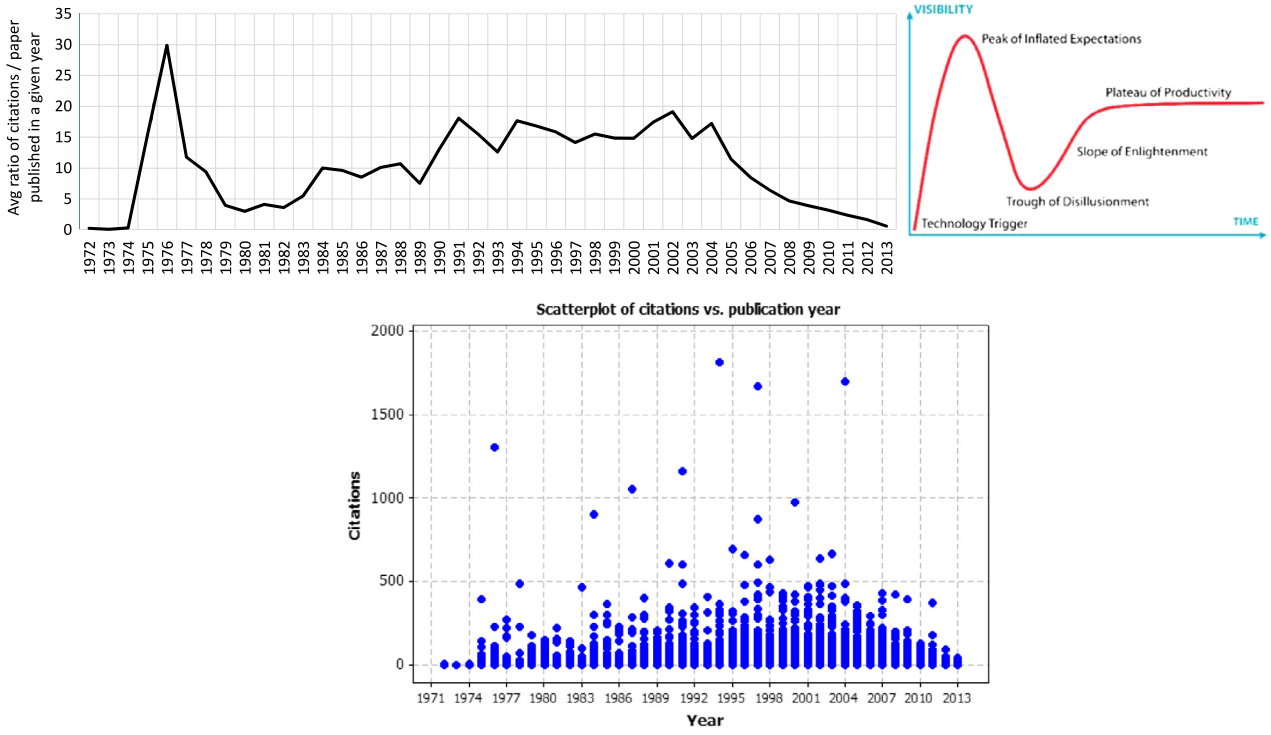

Fig. 7 – Citations to papers published in different years.
Source: **The top-right figure has been taken from: en.wikipedia.org/wiki/Hype_cycle.**

appear in a collection of documents. Our approach is a partial reproduction to the one by Griffiths and Steyvers [18] who used it to discover scientific topics appearing in the papers in the Proceedings of the US National Academy of Sciences (PNAS). We used the R statistical analysis program and utilized the R scripts provided by Ponweiser [59] who performed an exact replication of the work by Griffiths and Steyvers.

The automated thematic analysis of the SE research literature has been done in the past by Coulter et al. [16] who in their 1998 paper used co-word analysis and relied on the fixed set terms from ACM's taxonomy. Co-word analysis is an older method in scientometrics and has lost its popularity to LDA as it cannot handle synonym terms very well for example. Recent, studies also suggest that LDA produces

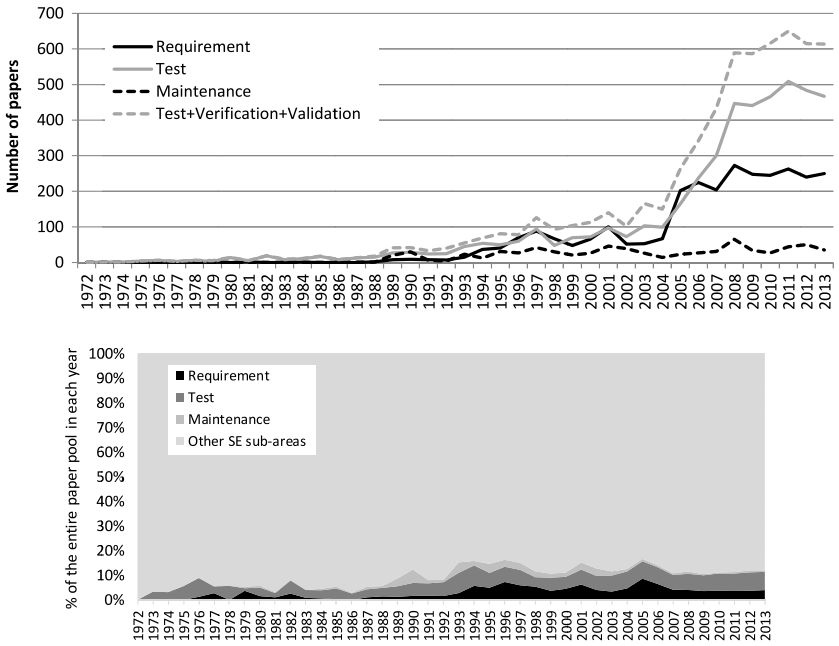

Fig. 8 – Top: Annual trends for number of papers with four different phrases in their titles. Bottom: Annual ratios of papers in four different sub-areas in the entire pool.

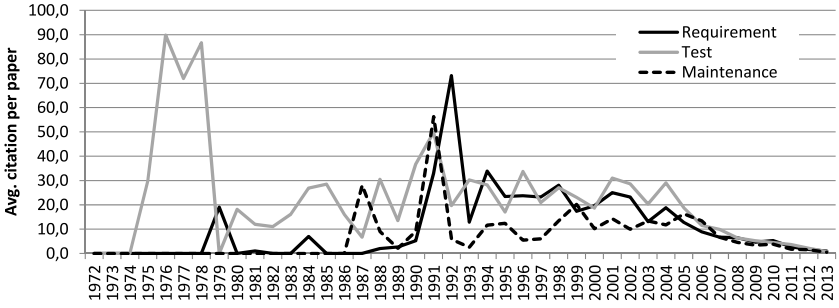

Fig. 9 – Average number of citations to papers with 'requirement', 'test' and 'maintenance' in their titles.

better results [60,61]. In our approach, we first created a document term matrix using the package '*tm*' of the R toolset by issuing the following command:

- `dtm = DocumentTermMatrix (corpus, control = list (tolower = TRUE, stopwords = TRUE, stemming = TRUE, minwordLength = 3, removeNumbers = TRUE, removePunctuation = TRUE, bounds = list (global = c(5,Inf))))`

We decided to remove standard stop words included in R package *tm*. Following the original paper [18], we removed words with less than three letters and included only terms that appear at least in five documents. The original paper [18] does not mention the use of stemmer, i.e. a tool to find the root form of a word, which hints that stemmer was not used. We created models both with and without the stemmer and in general we found no big differences. In some cases, the topics created with the stemmer made more sense and sometime less sense than the ones without stemmer, e.g. when using the stemmer, the algorithm created a topic with the top stemmed words of "require", "review", "systemat" "engin", "map". In that case, it is obvious that the topic modeling algorithm has categorized requirements engineering and systematic reviews under a single topic. Thus, as the stemmer seemed to produce no obvious benefits, we decided not to use it.

We applied the topic modeling only to document titles as the abstracts were not available in our dataset due to restrictions in the amount of paper abstracts one could export from Scopus. If one chooses to do similar analysis with a smaller number of document, e.g. with only papers about software testing then we recommend also downloading the abstracts. We ended up with vocabulary of 6681 words while the work by Griffiths and Steyvers reports a vocabulary of 20,551. A possible explanation may be that Griffiths and Steyvers [18] analyzed the articles from the Proceedings of the US National Academy of Sciences (PNAS) that consists of several science areas, such as astronomy, chemistry, statistics, and geology, where as our articles purely focused on various forms of SE. Another possible explanation is that they used papers abstracts while we were limited to paper titles in our analysis.

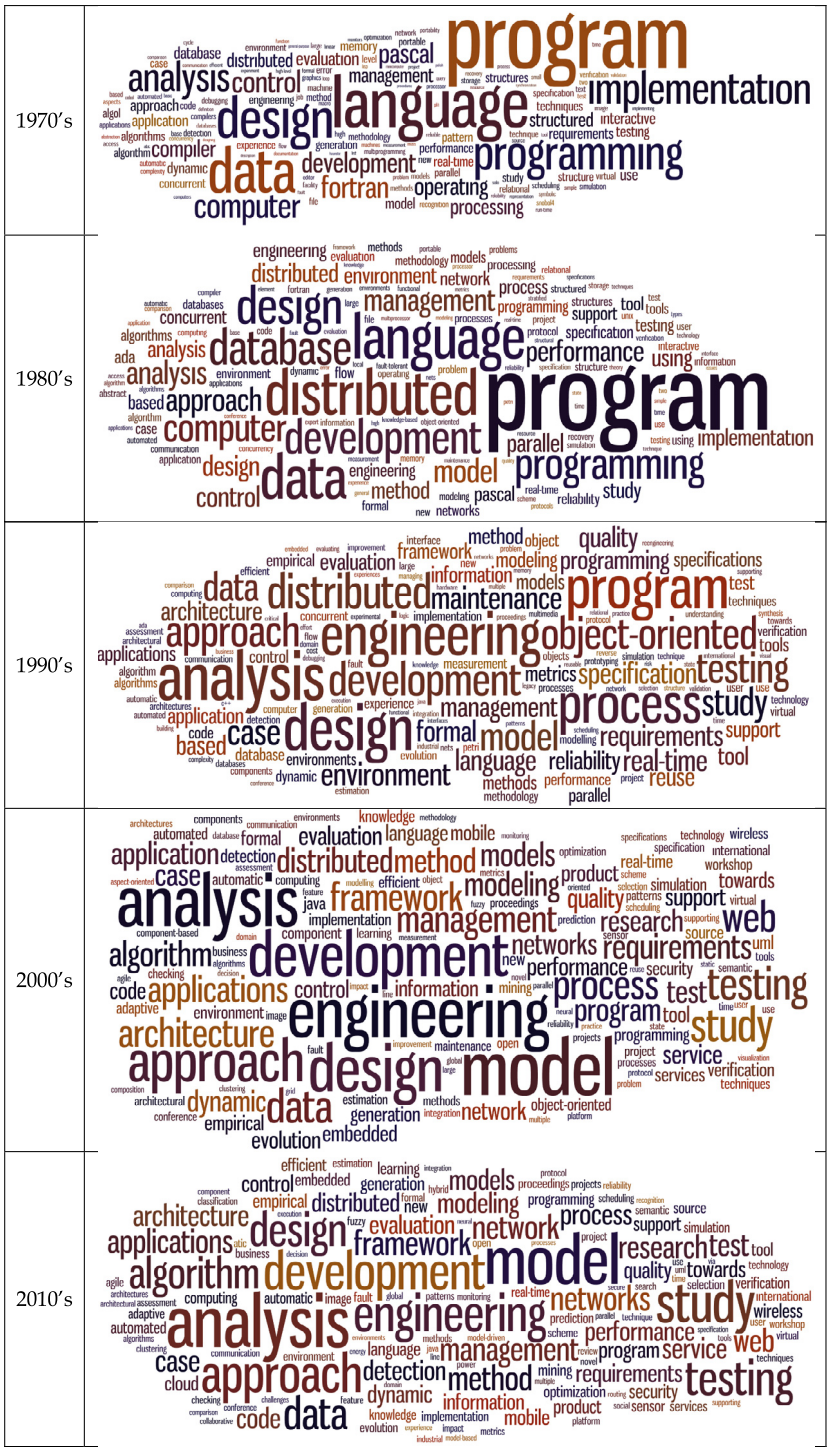

Fig. 10 – Focus areas of SE papers in each decade.

We fitted LDA with Gibbs sampling. In the R tool-set, we used the following parameters that are typical [62]:

- `alpha = 50/k, estimate.beta = TRUE, verbose = 0, save = 0, keep = 50, seed = as.integer (Sys.time()), nstart = 1, best = TRUE, delta = 0.1, iter = 1000, burnin = 1000, thin = 2000.`

In topic modeling, finding the optimal number of topics (k) that describes the corpus needs to be tested. Similar to the work of Griffiths and Steyvers [18], we used standard Bayesian Methods and computed maximum likelihood of the data given the model, using harmonic mean method, for models with different number of topics. Fig. 11-(a) shows how the log-likelihood measure performs first with topic numbers ranging from 100 to 600 with an increment of 100. Fig. 11-(b) shows the number of topics ranging from 20 to 150 with an increment of 10. Finally, we computed topics ranging from 50 to 80 with

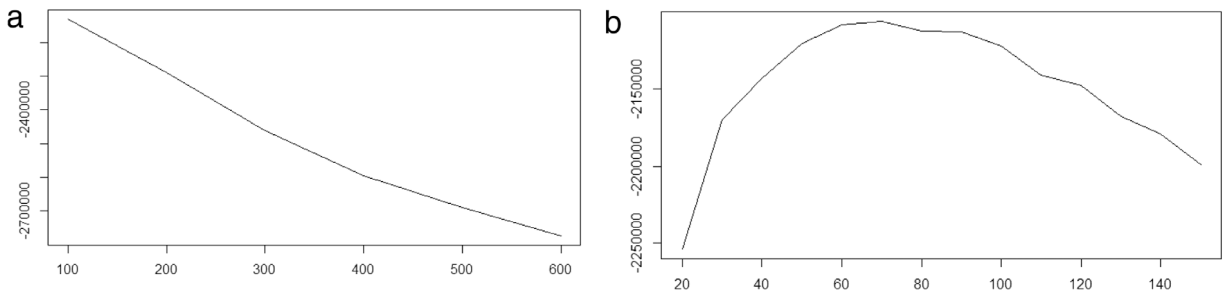

Fig. 11 – Number of topics (*x*-axis) and goodness of the model measured with log likelihood measure of the data given the model (*y*-axis), see [18] for details.

Table 7 – Number of topics with significant trends.

	p-level values			
	0.05	0.01	0.001	0.0001
Negative trend	16	14	12	8
Positive trend	29	23	16	12
Total	45	37	28	20

an increment of 1 (not plotted) and found that, for our case, the number of topics that produced maximal likelihood was 67. The work by Griffiths and Steyvers [18] tested number of topics with 8 samples (50, 100, 200, 300, 400, 500, 600 and 1000) and found that 300 topics produced the maximal likelihood.

We analyzed mean Theta, i.e. topic distribution over all documents describing the popularity of that topic, by years to discover the hot and cold topics within our dataset. We fitted linear regression model in order to study whether the increase in time (independent variable) could explain the changes in the mean of Theta (dependent variable) for each topic per year. For the trend analysis, we discarded all years prior to 1979, i.e., from 1972 to 1978 as then less 250 papers per year were published. In the early years, the year to year variation between the topic popularity was much larger than in the later years due to small number of papers. Similarly all 7 papers for year 2015 were discarded. Performing these data removals only removed 463 papers from our dataset.

Table 6 shows the number of topics with significant trends on various *p*-levels. Overall in our dataset of the SE literature, 29% of the topics (20/67) showed to have significant ($p = 0.0001$) positive or negative trend. In the previous works on the PNAS dataset, percentage values quite comparable to the values we calculated were reported: 35% (104/300) by Griffiths and Steyvers [18] and 17% (50/300) by Ponweiser [59]. Based on this and further comparison with Ponweiser's findings, who provides an identical data as in our Table 7, we conclude that SE is not a more or less trend-oriented than the other sciences represented in the PNAS dataset.

Fig. 10 shows the five hottest and coldest topics for SE. In Tables 8 and 9, we see the ten most probable terms for the hottest and coldest topics. The coldest topics are on programs, databases and systems. We have given each topic a descriptive name, see Table 8, and the five coldest topics can be listed as: (1) programming languages, (2) program execution and debugging, (3) databases, (4) system design, and (5) distributed systems. However, Fig. 12 shows that for the coldest topics, the mean Theta is in recent years just a little less than it is for the hottest topics. So the label "cold topic" does not mean that the topic would be dead. It means that focus on those topics has decreased by time. The coldest topics are the ones that used to be hot when SE was an emerging discipline, i.e. in order to become a cold topic, one has to achieve a decreasing trend.

In Table 9, we have titled the five hottest topics as: (1) web services, (2) mobile and cloud computing, (3) industrial studies, (4) source code and (5) test generation. The third topic (i.e., industrial studies) is different from the other hot and cold topics, as it describes the context in which the particular study has been made, i.e. an industrial case study. All the other topics describe an artifact or a phenomena rather than the context.

Comparing the hottest topics and the highest cited papers (per year), from Table 5 shows that there is a relationship between highly-cited papers and hot topics. Web-services is one of the hot topics and "*QoS-aware middleware for Web services composition*" is the highest cited paper in our pool. Similarly, mobile and cloud computing is the second hottest topic and the second highest cited paper is "*CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms*". Finally, industrial studies are the third hottest topic and its matching paper is the fifth highest cited paper titled "*Guidelines for conducting and reporting case study research in software engineering*".

We also analyzed which topics are the most popular based on the number of papers that are primarily assigned to each topic. We found that requirements engineering has the highest number of papers (3096), followed by databases (2601 papers), software effort estimation (2601), and design patterns (2172). In Section 4.2.5, we compared the popularity of different SE sub-domains based on word frequency. In order to do that, we have to know what the different sub-domains are, e.g. requirements engineering and testing. The topic analysis in this section is different as the topics presented rise from the data based on the fitted probabilistic model. This is completely automated and it does not require any prior knowledge of SE sub-domains. At the same time, the automation sometimes makes mistakes as we shall next see.

At fifth place (column) in Table 10, we have several topics under the 'Incoherent' category. The most probable terms for such topics are "using"; "analysis", "time" and "model. Those topics are incoherent and would not make that much sense to readers that know the domain of SE. Having an incoherent topic is a well-known problem when using topic

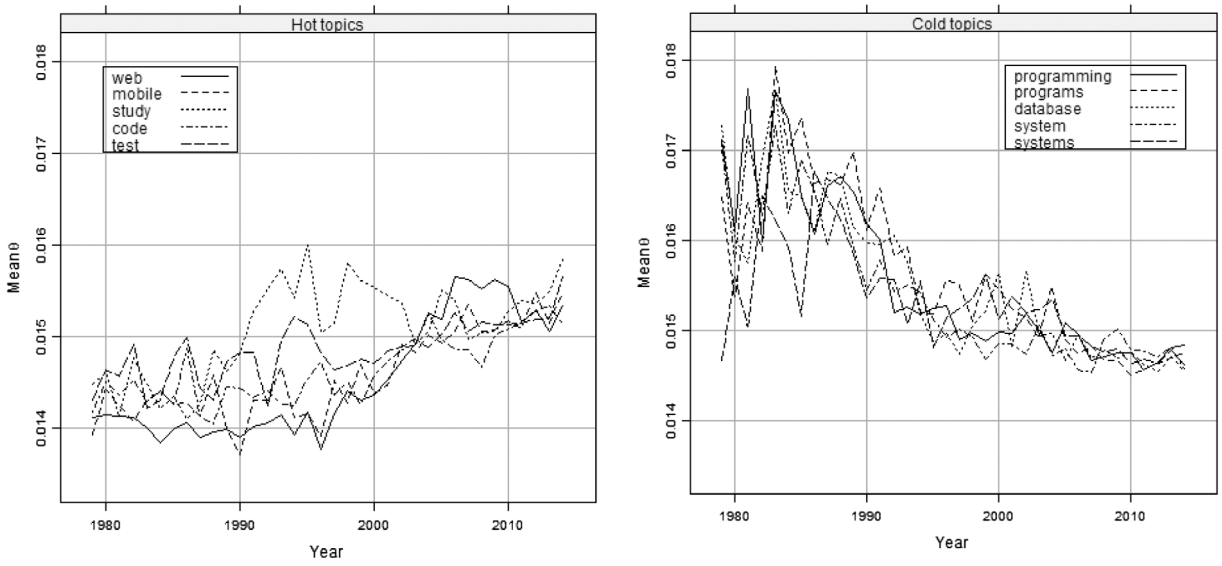

Fig. 12 – The five hottest and coldest topics.

Table 8 – Ten most probable terms in the five coldest topics.

Given topic name	Programming languages ($n = 1145$)	Program execution and debugging ($n = 701$)	Databases ($n = 2601$)	System design ($n = 1192$)	Distributed system ($n = 891$)
1	"programming"	"programs"	"database"	"system"	"systems"
2	"language"	"java"	"data"	"design"	"distributed"
3	"languages"	"execution"	"processing"	"implementation"	"multi agent"
4	"visual"	"concurrent"	"databases"	"operating"	"complex"
5	"aspect oriented"	"parallel"	"xml"	"file"	"reactive"
6	"natural"	"symbolic"	"relational"	"intelligent"	"large scale"
7	"description"	"debugging"	"query"	"expert"	"self adaptive"
8	"domain specific"	"Ada"	"objects"	"generator"	"cooperative"
9	"structured"	"traces"	"efficient"	"realization"	"fault tolerant"
10	"high level"	"multithreaded"	"spatial"	"Unix"	"timing"

Table 9 – Ten most probable terms in the five hottest topics.

Given topic name	Web services ($n = 915$)	Mobile and Cloud computing ($n = 862$)	Industrial studies ($n = 1376$)	Source code ($n = 988$)	Test generation ($n = 923$)
1	"web"	"mobile"	"study"	"code"	"test"
2	"service"	"computing"	"case"	"source"	"generation"
3	"services"	"cloud"	"empirical"	"open"	"automatic"
4	"applications"	"environments"	"industrial"	"projects"	"coverage"
5	"composition"	"agent"	"studies"	"changes"	"automated"
6	"semantic"	"middleware"	"comparative"	"usage"	"selection"
7	"discovery"	"grid"	"use"	"documentation"	"generating"
8	"composite"	"devices"	"exploratory"	"API"	"cases"
9	"QoS"	"smart"	"pilot"	"detecting"	"suite"
10	"BPEL"	"trust"	"importance"	"clones"	"tests"

modeling algorithms and they are caused by the mismatch between the topic modeling algorithms that are based on probabilistic word distributions and humans who know what words together form semantically meaningful topics for a particular domain.

Finally, we studied the most-cited topics when normalized by topic age and paper count (see Table 11). As all papers can be assigned to a single topic, we computed the number of paper citations for each topic and divided this by the sum of the age of the papers for that topic. As the results of computations showed, 'model checking' was the most cited topic with 1.78 cites per paper per year. It was followed by test generation (1.60), source code (1.47), automated testing (1.40) and systematic review (1.29). The fifth most cited topic is

Table 10 – Ten most probable terms in the topics with the highest number of papers.

Given topic name	Requirements engineering ($n = 3096$)	Database ($n = 2601$)	Software effort estimation ($n = 2601$)	Design patterns ($n = 2172$)	Incoherent topic ($n = 2155$)
1	"requirements"	"database"	"software"	"design"	"using"
2	"engineering"	"data"	"estimation"	"patterns"	"analysis"
3	"elicitation"	"processing"	"effort"	"implementation"	"time"
4	"nonfunctional"	"databases"	"cost"	"architectural"	"model"
5	"role"	"xml"	"measurement"	"pattern"	"real"
6	"tracing"	"relational"	"models"	"matching"	"behavior"
7	"goals"	"query"	"functional"	"style"	"functions"
8	"managing"	"objects"	"function"	"decisions"	"world"
9	"legal"	"efficient"	"size"	"principles"	"structure"
10	"goal-oriented"	"spatial"	"point"	"rationale"	"paper"

Table 11 – Ten most probable terms in the topics with the highest amount of citations per paper per year.

Given topic name	Model checking ($n = 686$)	Test generation (=923)	Source code ($n = 988$)	Automated testing ($n = 785$)	Systematic reviews ($n = 653$)
1	"model"	"test"	"code"	"testing"	"software"
2	"checking"	"generation"	"source"	"automated"	"systematic"
3	"transformation"	"automatic"	"open"	"model based"	"review"
4	"driven"	"coverage"	"projects"	"regression"	"challenges"
5	"transformations"	"automated"	"changes"	"mutation"	"survey"
6	"probabilistic"	"selection"	"usage"	"random"	"future"
7	"Markov"	"generating"	"documentation"	"strategies"	"issues"
8	"bounded"	"cases"	"API"	"GUI"	"mapping"
9	"graph"	"suite"	"detecting"	"conformance"	"results"
10	"properties"	"tests"	"clones"	"techniques"	"approaches"

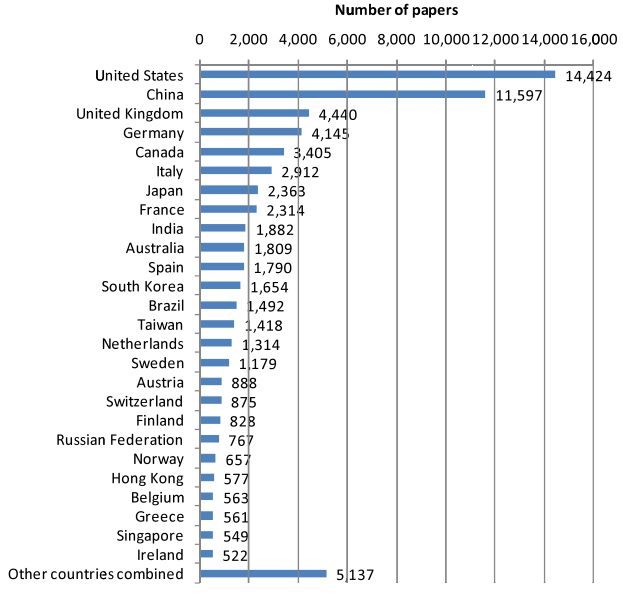

Fig. 13 – Ranking of the countries with more than 500 contributions.

'systematic reviews' providing evidence for the widely quoted idea that literature reviews get many citations. We can also notice that, in the top-5 list, we have two topics related to software testing (i.e., test generation and automated testing). Had we used the stemmer, the two testing topics would have been merged. However, as previously pointed out, the use of the stemmer also created problems when two unrelated topics of requirements engineering and systematic reviews became merged into a single topic.

4.4. RQ 4: *ranking of countries by number of contributed papers*

For each search query, Scopus provides statistics of countries based on author affiliations. Thus, our pool included that information. Ranking of the countries with more than 500 research contributions in the pool is shown in Fig. 13. For papers with multiple country affiliations, all the involved countries are considered in the Scopus metric. Thus, the sum of the values in Fig. 13 is larger than the pool size. We can see that the top three countries (US, China and the UK) have generated almost half (~44%) of all the SE research contributions, which makes the contributions pool very non-normal across countries.

To consider country populations in the ranking, we normalized the country contributions by population values and Fig. 14 shows the results. We can see that the top six countries are all small Nordic or middle European countries in respective order: Finland, Norway, Sweden, Ireland, Switzerland, and Austria. Proportionally, the relative contributions of these small countries are far greater than the countries with larger populations (e.g., US and China).

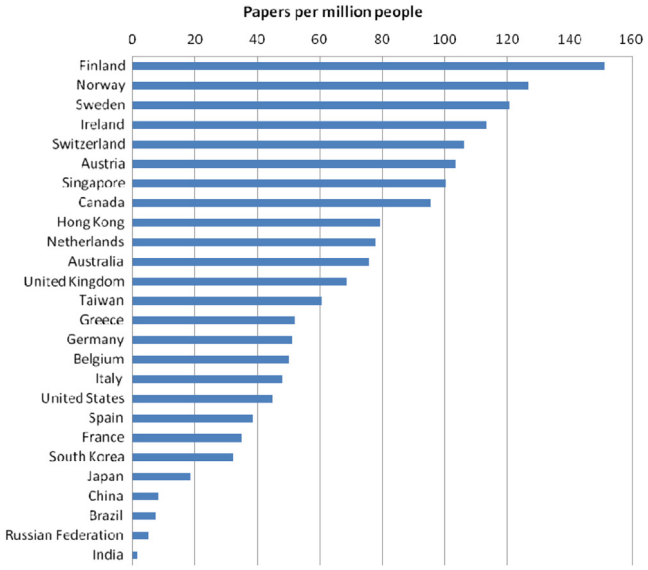

Fig. 14 – Ranking of the countries with more than 500 contributions normalized by country population.

5. Discussions

We summarize the findings, and then discuss potential threats to validity in this section.

5.1. *Summary of findings, trends, and implications*

Our study showed that the volume of papers on SE has grown rapidly since its inception in 1968. During 1990s, less than 1000 SE papers were published annually. However, in early 2000s, that number was between 1000 and 2000 annually, and lately (as of 2015), the number of SE papers published yearly is between 6000 and 7000 papers. We do not know whether this number reflects the "publish or perish" academic culture or a true increase in the progress. However, it clearly suggests that performing automated bibliometrics analysis as we have done in this paper is preferable, as otherwise it is impossible to form an objective and holistic view of the landmarks of our science and the research topics that are investigated.

The citation analysis showed that almost half of the papers in SE (~43% of the pool) are not cited at all. This might suggests that the work is of poor quality with respect to research or reporting. An alternative explanation is that the topics studied in papers without citations are of little interest to the research community. Finally, also this could be due to "publish or perish" academic culture that pushes academics to "flush" out even the less successful and low-quality papers, e.g., quality versus quantity.

The citation landscape presented a power-law (Pareto principle)-like distributions where only a small share of studies gather the majority of the citations. Looking at the top cited papers (Section 4.2.2), we found that they represent broad topics from technological advancement to research methodology and guideline papers. We are aware that citations have also drawn a lot of criticisms. For example, a recent column in the Nature magazine [63] heavily criticizes citation metrics: they drive people to work in highly fashionable subfields, to submit shorter paper to popular forums instead of publishing an in-depth research in a focused journal, and that reasons for citations are not always ethical, "*to satisfy a potential big-shot referee*", "*to give the impression that there is a community interested in the topic by stuffing the introduction with irrelevant citations*".

Analysis of paper types showed that roughly two third of the papers are published in conferences while only one third is published in journals. This is typical in computer science but different from natural sciences where the majority of the papers are published in journals. When comparing citations, we found that journal papers get more citations than conference papers.

Thematic analysis showed shifting trends in the SE research. Current hot topics are (1) web services, (2) mobile and cloud computing, (3) industrial studies, (4) source code and (5) test generation. Similarly, five cold topics that have been decreasing with interest are (1) programming languages, (2) program execution and debugging, (3) databases, (4) system design, and (5) distributed systems. These results produced by automated quantitative analysis of paper titles match quite nicely with the authors' intuitive feeling of the trends of SE.

Finally, the analysis of countries showed that top three countries (US, China and UK) produce almost half of all the papers. At the same time, it was found that small northern and middle European countries are the most productive in producing papers when normalized to the size of the country populations.

5.2. *Potential limitations and threats to validity*

In this section, the potential threats to the validity of the sturdy are discussed in the context of the four types of threats to validity based on a standard checklist presented in [64]. We also discuss the steps that we have taken to minimize or mitigate those potential threats.

5.2.1. *Internal validity*

Internal validity reflects the extent to which a causal conclusion based on a study is warranted [64]. The systematic approach that has been utilized for the selection of publication database and of SE papers was described in Section 2. In order to make sure that this study and ranking are repeatable, search engines, search terms were carefully defined and reported. Also, to ensure transparency and replicability of our analysis, the entire raw and ranking data for all the 71,668 papers is available as an Excel file which can be downloaded online [19].

Limitation of search terms and search engines can lead to incomplete set of papers in the pool. We empirically found that, when conducting searches in Scopus, including the phrase "software" in venue names is an effective way to ensure targeting the entire SE literature.

5.2.2. *Construct validity*

Construct validities are concerned with issues that to what extent the object of study truly represents theory behind the study [64]. Threats related to this type of validity in this study were suitability of RQs and categorization scheme used for the data extraction.

To limit potential construct threats in this study, the GQM approach was used to preserve the tractability between research goal, questions and measurements. RQs were designed to cover our goal and different aspects of the top papers. For designing a good categorization scheme for the systematic mapping, we adapted standard classifications from [65] and also have finalized the used schema through an iterative improvement process.

As a limitation w.r.t construct validity, we should note that we assumed that all the papers published by the venues (e.g., journals) that we agreed to include in the pool are SE-related. However, we have seen then in a small proportion that this is not true. For example, Journal of Software and Systems sometimes publishes paper on non SE-related topics, e.g., paper [66] published in this journal which proposes a concurrency control protocol for real-time databases. Unfortunately, filtering out such papers manually would not be feasible and automating such as task would need a sophistical intelligent machine-learning algorithm. Similar to many other automated analyses, the authors believe that the impact of such false-positive should be minimal on the results of this study. This belief was also found to be true in a case where the second author studied the top cited papers of the five hottest topics, see Table 9, for teaching purposes. At least among those papers all were related to SE. However, we also recognize that what SE is can be ill-defined. For example, the second highest ranked paper based on citations per year, see Table 5, titled "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms" might not be at first considered as a SE paper since it relates to cloud computing environments. However, once one realizes that the experienced performance of cloud applications, i.e. the software running on the cloud, heavily depends on the cloud computing setup then the value of such cloud simulators becomes evident to software engineers that try to deploy the best possible performance to their customers.

5.2.3. *Conclusion validity*

Conclusion validity of a study deals with whether correct conclusions are reached through rigorous and repeatable treatments [64].

Conclusions that we discussed throughout the paper were based on actual quantitative measures and statistics on the data extracted from the top papers. Following the systematic approach that we used to identify and map the top papers assured that, if the study is conducted by other researchers, it is expected that results will not have major deviations from our results.

5.2.4. *External validity*

External validity is concerned with to what extent the results of this secondary study can be generalized [64]. The results of this study are not meant to be generalized to fields outside SE. However, we believe that given the rigor of the systematic approach that we used to identify and map the top papers, the results highlight the citation landscape of the SE and the top papers in this area.

6. Conclusions and future work

This review presented an exploratory bibliometrics assessment of the SE research literature. As the trends in this paper showed, the SE literature is very active and the number of papers in this area is increasing each year. However, about half (43%) of the papers in this area so far have received no citations at all. This raises the question that: why is there such a large ratio of non-cited papers in this area? How does this trend compare to other areas of science? Is it because we have too many less-known venues which publish papers not seen or read by others? Does this have anything to do with papers' or venues' quality? As discussed in a bibliometrics study entitled "*Characteristics of highly-cited papers*" [41], quality dynamics and visibility dynamics increase the number of citation to a given paper. Perhaps those dynamics should be studied further in depth in the SE literature.

The pool that we have made publicly available [19] can be used to conduct other thematic and demographic analysis in SE and its sub-domains. We for example demonstrated the hot and cold topics as well as the most cited research topics in our field. Also, this bibliometrics approach can be repeated periodically to analyze the growth and trends in the field in upcoming years and compare the future trends with the findings of this study.

The process described here is replicable, and we think that it can be used to study the subareas of SE, or any other field for that matter. To demonstrate the replicability of the process, we have published a recent short paper focusing on the ESEM conference papers only [67]. For that paper a freshly hired Ph.D. student, the first author of the ESEM paper, performed identical analysis as in this paper, and documented it. For the new Ph.D. student, performing the analysis took roughly 10 full working days. This suggests that an analysis described in this paper where one looks at the top-cited papers and performs text-mining to discover the research topics of a certain area can perhaps be used in situations where the lack of exact research questions and lack of time prevent utilizing

the rather laborious SLR method. For example, following our process could make the literature reviews of Master's or Bachelor's thesis more structured. Currently, the second author is utilizing this process as an exercise work in a master's thesis course where students replicate the process for the sub-areas of SE, such as testing. Furthermore, the second author has already used the hot topics of Table 9 for selecting contents for a course titled Emerging Trends in Software Engineering at the University of Oulu. We think that such topic selection for a course is evidence-based when the usually used methods are either subjective, the professor selects the topics based on his/her prior knowledge, or convenience based, where course contents are the ones covered in a particular textbook.

The last but not the least, it is also important for researchers in the SE community to pay more attention to the quality of publications versus their quantity, which was put nicely by David Parnas as "*Stop the numbers game*" [68].

Acknowledgments

Vahid Garousi was partially supported by several internal grants provided by the Hacettepe University.

REFERENCES

[1] R.L. Glass, An assessment of systems and software engineering scholars and institutions, J. Syst. Softw. 27 (1994) 63–67.
[2] R.L. Glass, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (1999–2003), J. Syst. Softw. 76 (2005) 91–97.
[3] T.H. Tse, T.Y. Chen, R.L. Glass, An assessment of systems and software engineering scholars and institutions (2000–2004), J. Syst. Softw. 79 (2006) 816–819.
[4] W.E. Wong, T.H. Tse, R.L. Glass, V.R. Basili, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (2001–2005), J. Syst. Softw. 81 (2008) 1059–1062.
[5] W.E. Wong, T.H. Tse, R.L. Glass, V.R. Basili, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (2002–2006), J. Syst. Softw. 82 (2009) 1370–1373.
[6] W.E. Wong, T.H. Tse, R.L. Glass, V.R. Basili, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (2003–2007 and 2004–2008), J. Syst. Softw. 84 (2011) 162–168.
[7] R.L. Glass, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (1997–2001), J. Syst. Softw. 64 (2002) 79–86.
[8] R.L. Glass, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (1996–2000), J. Syst. Softw. 59 (2001) 107–113.
[9] R.L. Glass, T.Y. Chen, An assessment of systems and software engineering scholars and institutions (1998–2002), J. Syst. Softw. 68 (2003) 77–84.
[10] C. Wohlin, An analysis of the most cited articles in software engineering journals—1999, Inf. Softw. Technol. 47 (2005) 957–964.
[11] C. Wohlin, An analysis of the most cited articles in software engineering journals—2000, Inf. Softw. Technol. 49 (2007) 2–11.
[12] C. Wohlin, An analysis of the most cited articles in software engineering journals—2001, Inf. Softw. Technol. 50 (2008) 3–9.
[13] C. Wohlin, An analysis of the most cited articles in software engineering journals—2002, Inf. Softw. Technol. 51 (2009) 2–6.
[14] R.L. Glass, I. Vessey, V. Ramesh, Research in software engineering: an analysis of the literature, Inf. Softw. Technol. 44 (2002) 491–506.
[15] K.-Y. Cai, D. Card, An analysis of research topics in software engineering—2006, J. Syst. Softw. 81 (2008) 1051–1058.
[16] N. Coulter, I. Monarch, S. Konda, Software engineering as seen through its research literature: A study in co-word analysis, J. Am. Soc. Inf. Sci. 49 (1998) 1206–1223.
[17] V. Garousi, G. Ruhe, A bibliometric/geographic assessment of 40 years of software engineering research (1969–2009), Int. J. Softw. Eng. Knowl. Eng. 23 (2013) 1343–1366.
[18] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (2004) 5228–5235.
[19] V. Garousi, M.V. Mäntylä, All source data for bibliometrics study of the software engineering community, in https://goo.gl/G8f0M0, 2015 (accessed: December 2015).
[20] W.E. Eric, T.H. Tse, Robert L. Glass, Victor R. Basilid, T.Y. Chene, An assessment of systems and software engineering scholars and institutions (2003–2007 and 2004–2008), J. Syst. Softw. 84 (2011) 162–168.
[21] V. Garousi, T. Varma, A bibliometric assessment of canadian software engineering scholars and institutions (1996–2006), Canad. J. Comput. Inform. Sci. 3 (2010) 19–29.
[22] F. de Freitas, J. de Souza, Ten years of search based software engineering: A bibliometric analysis, in: M. Cohen, M. Ó Cinnéide (Eds.), in: Search Based Software Engineering, 6956, Springer, Berlin, Heidelberg, 2011, pp. 18–32.
[23] R. Farhoodi, V. Garousi, D. Pfahl, J.P. Sillito, Development of scientific software: A systematic mapping, bibliometrics study and a paper repository, Int. J. Softw. Eng. Knowl. Eng. 23 (2013) 463–506.
[24] J. Ren, R.N. Taylor, Automatic and versatile publications ranking for research institutions and scholars, Commun. ACM 50 (2007) 81–85.
[25] A. Hoonlor, B.K. Szymanski, M.J. Zaki, Trends in computer science research, Commun. ACM 56 (2013) 74–83.
[26] J. Fernandes, Authorship trends in software engineering, Scientometrics 101 (2014) 257–271.
[27] V. Garousi, A bibliometric analysis of the Turkish software engineering research community, Springer J. Scientometrics 105 (2015) 23–49.
[28] V. Garousi, J.M. Fernandes, Highly-cited papers in software engineering: The top-100, Inf. Softw. Technol. 71 (2016) 108–128.
[29] L. Bornmann, How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature, Res. Eval. 23 (2014) 166–173.
[30] R.V. Noorden, B. Maher, R. Nuzzo, The top 100 papers, Nature 514 (2014) 550–553.
[31] J.P.A. Ioannidis, K.W. Boyack, H. Small, A.A. Sorensen, R. Klavan, Is your most cited work your best? Nature 514 (2014) 561–562.
[32] F.A. Ponce, A.M. Lozano, Highly cited works in neurosurgery. Part I: the 100 top-cited papers in neurosurgical journals, J. Neurosurg. 112 (2010) 223–232.
[33] S. Easterbrook, J. Singer, M.-A. Storey, D. Damian, Selecting empirical methods for software engineering research, in: F. Shull, J. Singer, D.K. Sjøberg (Eds.), Guide to Advanced Empirical Software Engineering, Springer, London, 2008, pp. 285–311.
[34] É. Archambault, D. Campbell, Y. Gingras, V. Larivière, Comparing bibliometric statistics obtained from the web of science and scopus, J. Am. Soc. Inf. Sci. 60 (2009) 1320–1326.

[35] M.E. Falagas, E.I. Pitsouni, G.A. Malietzis, G. Pappas, Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses, FASEB J. 22 (2008) 338–342.
[36] A. Abrizah, A.N. Zainab, K. Kiran, R.G. Raj, LIS journals scientific impact and subject categorization: a comparison between Web of Science and Scopus, Scientometrics 94 (2013) 721–740.
[37] A.A. Chadegani, H. Salehi, M.M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, et al., A comparison between two main academic literature collections: Web of science and scopus databases, Asian Soc. Sci. 9 (2013) 18–26.
[38] R.J.A. Buhr, Use case maps as architectural entities for complex systems, IEEE Trans. Softw. Eng. 24 (1998) 1131–1155.
[39] G.J. Holzmann, The model checker SPIN, IEEE Trans. Softw. Eng. 23 (1997) 279–295.
[40] R. Tijssen, M. Visser, T. van Leeuwen, Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? Scientometrics 54 (2002) 381–397.
[41] D.W. Aksnes, Characteristics of highly cited papers, Res. Eval. 12 (2003) 159–170.
[42] P. Pyšek, D.M. Richardson, J. V, Who cites who in the invasion zoo: insights from an analysis of the most highly cited papers in invasion ecology, Preslia 78 (2006) 437–468.
[43] L. Bornmann, F. de Moya Anegón, L. Leydesdorff, Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the ortega hypothesis, PLoS One 5 (2010).
[44] Z. Corby, To be the best, cite the best, Nature News (2010). http://dx.doi.org/10.1038/news.2010.539.
[45] T.A. Hamrick, R.D. Fricker, G.G. Brown, Assessing what distinguishes highly cited from less-cited papers published in interfaces, Interfaces 40 (2010) 454–464.
[46] O. Persson, Are highly cited papers more international? Scientometrics 83 (2010) 397–401.
[47] M. Wang, G. Yu, D. Yu, Mining typical features for highly cited papers, Scientometrics 87 (2011) 695–706.
[48] N. Miyairi, H.-W. Chang, Bibliometric characteristics of highly cited papers from Taiwan, 2000–2009, Scientometrics 92 (2012) 197–205.
[49] G. Abramo, T. Cicero, C.A. D'Angelo, Are the authors of highly cited articles also the most productive ones? J. Informetr. 8 (2014) 89–97.
[50] J. Antonakis, N. Bastardoz, Y.H. Liu, C.A. Schriesheim, What makes articles highly cited? Leadersh. Quart. 25 (2014) 152–179.
[51] M.E.J. Newman, Prediction of highly cited papers, Europhys. Lett. (EPL) 105 (2014) 6.
[52] D.J. Eaton, Highly cited papers in Medical Physics, Med. Phys. 41 (2014) 43–44.
[53] E. Aversa, Citation patterns of highly cited papers and their relationship to literature ageing: A study of the working literature, Scientometrics 7 (1985) 383–389.
[54] D. Price, Little Science, Big Science, Columbia University Press, 1963.
[55] P.D. Allison, J.A. Stewart, Productivity differences among scientists: Evidence for accumulative advantage, Amer. Sociol. Rev. 39 (1974) 596–606.
[56] W. Okrasa, Differences in scientific productivity of research units: Measurement and analysis of output inequality, Scientometrics 12 (1987) 221–239.
[57] P.D. Allison, Inequality and scientific productivity, Soc. Stud. Sci. 10 (1980) 163–179.
[58] A. Ghosh, N. Chattopadhyay, B.K. Chakrabarti, Inequality in societies, academic institutions and science journals: Gini and -indices, Physica A 410 (2014) 30–34.
[59] M. Ponweiser, Latent Dirichlet allocation in R (Master's thesis), University of Economics and Business, Vienna, 2012.
[60] W. Ding, C. Chen, Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods, J. Assoc. Informat. Sci. Technol. 65 (2014) 2084–2097.
[61] K. Lu, D. Wolfram, Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches, J. Am. Soc. Inf. Sci. Technol. 63 (2012) 1973–1986.
[62] D. Binkley, D. Heinz, D. Lawrie, J. Overfelt, Understanding LDA in source code analysis, in: Presented at the Proceedings of the 22nd International Conference on Program Comprehension, Hyderabad, India, 2014.
[63] R. Werner, The focus on bibliometrics makes papers less useful, Nature 517 (2015) 245.
[64] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering: An Introduction, Kluwer Academic Publishers, 2000.
[65] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, in: Presented at the 12th International Conference on Evaluation and Assessment in Software Engineering, EASE, 2008.
[66] K.-w. Lam, K.-y. Lam, S.-l. Hung, Optimistic concurrency control protocol for real-time databases, J. Syst. Softw. 38 (1997) 119–131. 8/.
[67] P. Raulamo-Jurvanen, M.V. Mäntylä, V. Garousi, Citation and topic analysis of the ESEM papers, in: International Symposium on Empirical Software Engineering and Measurement, 2015.
[68] D.L. Parnas, Stop the numbers game, Commun. ACM 50 (2007) 19–21.