Citation and Topic Analysis of the ESEM papers

Päivi Raulamo-Jurvanen, Mika V. Mäntylä
M-Group
Department of Information Processing Science
University of Oulu, Oulu, Finland
{paivi.raulamo-jurvanen, mika.mantyla}@oulu.fi

Vahid Garousi
Software Engineering Research Group
Department of Computer Engineering
Hacettepe University, Ankara, Turkey
vahid.garousi@hacettepe.edu.tr

Abstract—Context: **The pool of papers published in ESEM.** ***Objective*****: To utilize citation analysis and automated topic analysis to characterize the SE research literature over the years focusing on those papers published in ESEM.** ***Method*****: We collected data from Scopus database consisting of 513 ESEM papers. For thematic analysis, we used topic modeling to automatically generate the most probable topic distributions given the data.** ***Results*****: Nearly 42% of the papers have not been cited at all but the effect seems to wear off as time passes. Using text mining of article titles and abstracts, we found that currently the most popular research topics in the ESEM community are: systematic reviews, testing, defects, cost estimation, and team work.** ***Conclusions*****: While this study analyzes the paper pool of the ESEM symposium, the approach can easily be applied to any other sub-set of SE papers to conduct large scale studies. Due to large volumes of research in SE, we suggest using the automated analysis of bibliometrics as we have done in this paper.**

Keywords— Software engineering; research literature; citation analysis; thematic and topic analysis; bibliometrics; International Symposium on Empirical Software Engineering and Measurement (ESEM)

I. INTRODUCTION

The purpose of this study is to provide an overview of the literature published in the ESEM symposium since it started in 2007, by bringing together the ISESE and METRICS events. This paper covers three aspects: (1) the citation landscape and the most cited works of ESEM, (2) the most active authors and institutions, and (3) the hot research topics in ESEM publications.

Citations are a representation of connection between two publications. Citations are defined to be the backbone of a research: a way of building new studies on existing research results and a way to judge influential work in different fields [1-2]. The findings of this study visualize the skeleton of citations of the papers and discover trends in topics of those papers in ESEM.

Bibliometrics-based identification of active authors and institutions could serve for different reasons, e.g. allow graduate students and researchers to better identify where they want to study or with whom to work, or enable employers to recruit the most qualified potential researchers. The series of 12 papers by Wong, Glass et al., e.g. [3-4] was an ongoing, annual effort that periodically identified the top-15 SE scholars and institutions in systems and software engineering between 1995 and 2006. Also, some recent systematic mapping studies have included bibliometrics analyses of active authors and institutions in SE sub-areas [5-6], e.g. development of scientific software. Among the findings of one such study was that the most active authors in the area of development of scientific software were mostly located in the US, followed by the Canadian and British researchers [6].

The third and last goal of our study, identifying hot (active) research topics, is important as it can help in mapping the given area to outsiders such as students or industry. Such mappings of science have been done in computer science and other disciplines before, e.g. [7-10]. For example, Hoonlor et al. [9], mention that "*Keywords in the ACM Digital Library and IEEE Xplore digital library and in NSF grants anticipate future CS research [directions]*".

II. METHOD

The Scopus publication database was selected for our study due to its coverage of all ESEM papers whether originally published by ACM or IEEE. It provides titles, abstracts and citations that we used in the analysis. The data from Scopus was collected on May 22nd 2015 with the search string `SRCTITLE (international symposium on empirical software engineering and measurement)`. The search from Scopus resulted in total of 513 papers written by a total of 155 individual authors (either as a single or co-author(s)). Data was exported to Excel and R for conducting further analysis. For topic analysis, we used text-mining of paper titles and abstracts. The topics were created in R with package "topicmodels" using the LDA topic modeling with Gibbs sampling. Similar approach have been utilized for other disciplines previously, e.g. [10]. The data used is available at https://goo.gl/zTbFSt.

III. RESULTS

A. Citation landscape and the most cited works of ESEM

Out of the total 513 ESEM papers (published from 2007-2015), 295 have been cited at least once, leaving 218 un-cited papers. The 295 cited papers have 1,897 citations in total, which gives an average of 6.43 and median of 2 citations per paper for those cited papers. For total number of papers (including the un-cited ones), the average citations per paper is 3.70 and median is 1. The numbers of citations for papers resemble a power law distribution, i.e. a small share of papers has a large number of citations while many papers have only a few or no citations at all, as observable in Figure 1. It is interesting that such a large number of papers have no citations at all (218 in the given observation period 2007-2015). However, Figure 2 shows that 81% of papers prior to 2010 (173/211) have been cited. Thus the lack of citations in later papers may be explained by the fact that the longer a paper has been available the better the chances for it to be cited.

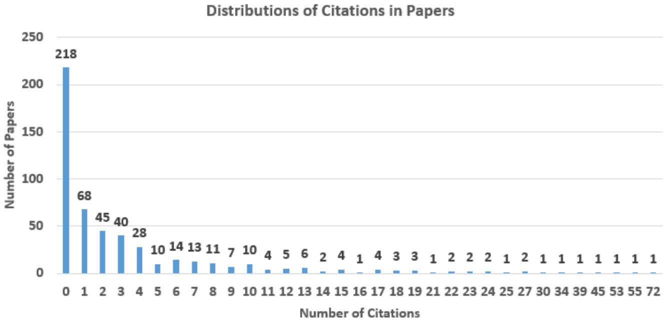
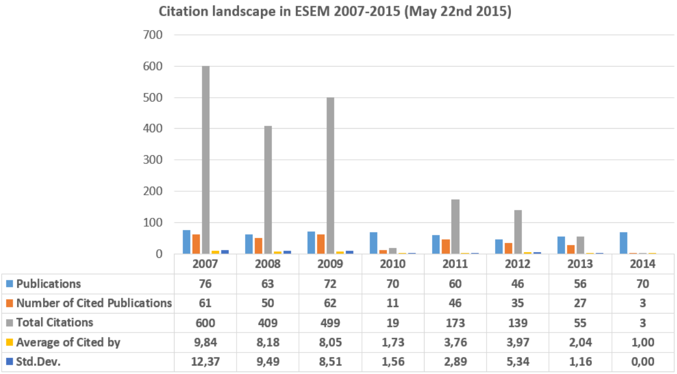


Fig. 1 Distribution of citations in papers (2007-2015).

The overall landscape of citations for papers for the observation period 2007-2015 is shown in Figure 2. Notably, papers from 2011 and 2012 already are more cited than the papers in 2010 even though they are more recent. The issue that there are so many un-cited papers seems exceptional due to the fact that many times authors prefer to give credit to their own preceding work worth the effort (self-citation).



	2007	2008	2009	2010	2011	2012	2013	2014
Publications	76	63	72	70	60	46	56	70
Number of Cited Publications	61	50	62	11	46	35	27	3
Total Citations	600	409	499	19	173	139	55	3
Average of Cited by	9,84	8,18	8,05	1,73	3,76	3,97	2,04	1,00
Std.Dev.	12,37	9,49	8,51	1,56	2,89	5,34	1,16	0,00

Fig. 2 Citation landscape of papers in ESEM per year.

Identification and characterization of highly-cited papers is common and regularly reported in various disciplines, e.g. "*The top 100 papers*" in Nature magazine [11]. The top most cited ESEM papers are presented in Table 1. It is no surprise the top 20 cited papers are mostly from earlier years in ESEM, 2007-2009 and surprisingly two more recent papers from year 2012 (there are 8, 5, 5 and 2 papers for years 2007, 2008, 2009 and 2012, respectively). More recently published papers may show an increase in citation count as time goes by.

The total number of citations for the top 20 papers is 623 covering 32.8% of all citations. For these top 20 cited papers it seems noteworthy that the findings may have either been fundamentally impressive or provided exceptionally worthwhile contribution to the discipline, since unlike many publications in ESEM, these have been adopted by the researches already the year following the publication. Also, e.g. the paper with most citations, published in 2007, [12] is not only the top cited paper but also the most cited of all in 2014 amongst the top 20 cited papers in ESEM.

The citing papers tell about the visibility of ESEM in general. We observed that the self-citations by neither the authors nor papers published in the ESEM seem to play a big role in paper visibility. Self-citation by authors accounted for 16% of the citations of the top twenty papers and citations from ESEM-proceedings (papers published in ESEM) accounted for 9% of the citations. Thus, we think the data shows that ESEM papers have visibility over ESEM community boundaries.

B. *The most active authors and institutions*

The top 10 most productive authors in ESEM are: 14 papers both Seaman C. and Yang Y., 13 papers both Wang Q. and Nagappan N., 12 papers Shull F., 11 papers both Travassos G.H. and Mendes E., and 10 papers Basili V., Cruzes D.S. and Dybå T. each. Out of the top ten authors five authors have papers in the list of top 20 most cited papers in ESEM (Table 1) as follows: Nagappan (3), Mendes (2), Dybå (2), Cruzes (1) and Basili (1).

The average number of authors for the papers during the observation period was 3.19. There is a high trend of collaboration across different affiliations and countries enabling sharing of not only the publicity but also the advantages of the research work.

A Wordle (www.wordle.net) word-cloud was created utilizing the author names of the full set of papers. The visual word-cloud nicely gives a descriptive overview of the most active author names in ESEM, see Figure 3. The numbers are actually irrelevant when considering productivity in this context. More interesting perspective of productivity is provided by the proportions of the names in the cloud than just a graph with numerical figures. From the word-cloud the overview of the authors is captured by a glance.

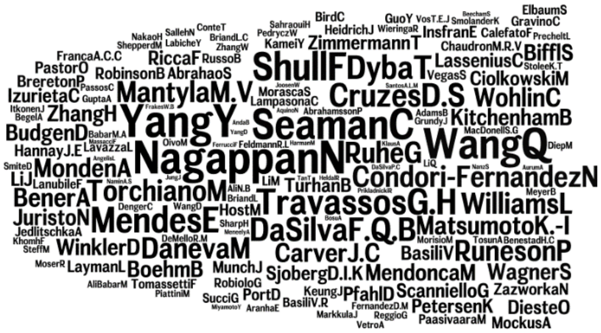


Fig. 3 Visual word cloud of the authors for all papers (2007-2015).

The top 10 most productive institutions in ESEM are (papers listed in parenthesis): Fraunhofer Institute for Experimental Software Engineering (20), Blekinge Institute of Technology (15), Microsoft Research (14), Polytechnic University of Valencia (13), University of Maryland Baltimore County (12), University of Oslo (12), Aalto University (11), Federal University of Pernambuco (11), University of Lund (11), University of Maryland (11) and University of Auckland (11).

It is claimed that in many countries many evaluating bodies, like committees determining funding, promotions or appointments, use figures like publication record or citation count in decision making [13-14]. This kind of evaluation seems fair, but at the same time the trend can be questioned whether it services the research field in an appropriate way.

TABLE 1 TOP 20 MOST CITED PUBLICATIONS (2007-2015).

Document	´08	´09	´10	´11	´12	´13	´14	´15	Total	WSC[a]	WEC[b]
Dyba T., Dingsoyr T., Hanssen G.K., Applying systematic reviews to diverse study types: An experience report, 2007	4	5	9	13	6	13	20	2	72	67	66
Cataldo M., Herbsleb J.D., Carley K.M., Socio-technical congruence: A framework for assessing the impact of technical and work dependencies on software development productivity, 2008	0	3	5	10	9	14	13	1	55	50	55
Nagappan N., Ball T., Using software dependencies and churn metrics to predict field failures: An empirical case study, 2007	2	7	6	15	5	11	6	1	53	46	48
Petersen K., Wohlin C., Context in industrial software engineering research, 2009	0	0	2	4	12	8	15	4	45	24	40
Begel A., Nagappan N., Usage and perceptions of Agile software development in an industrial context: An exploratory study, 2007	2	1	8	5	6	7	6	4	39	36	34
Riaz M., Mendes E., Tempero E., A systematic review of software maintainability prediction and metrics, 2009	0	0	2	6	9	9	7	1	34	28	31
Dyba T., Dingsoyr T., Strength of Evidence in Systematic Reviews in Software Engineering, 2008	0	2	3	9	2	8	6	0	30	26	25
Condori-Fernandez N., Daneva M., Sikkel K., Wieringa R., Dieste O. Pastor O., A systematic mapping study on empirical evaluation of software requirements specifications techniques, 2009	0	0	4	9	2	9	3	0	27	25	24
Svahnberg M., Aurum A., Wohlin C., Using students as subjects - An empirical evaluation, 2008	0	0	3	3	7	6	6	2	27	27	25
Kamei Y., Monden A., Matsumoto S., Kakimoto T., Matsumoto K.-I., The effects of over and under sampling on fault-prone module detection, 2007	4	2	3	3	3	7	3	0	25	22	23
Olbrich S., Cruzes D.S., Basili V., Zazworka N., The evolution and impact of code smells: A case study of two open source systems, 2009	0	0	5	2	2	9	4	2	24	19	23
Begel A., Nagappan N., Pair programming: What's in it for me?, 2008	0	1	3	4	6	6	4	0	24	23	20
Jalali S., Wohlin C., Systematic literature studies: Database searches vs. backward snowballing, 2012	0	0	0	0	0	7	12	4	23	19	21
Gupta A., Jalote P., An experimental evaluation of the effectiveness and efficiency of the test driven development, 2007	0	1	5	4	5	3	5	0	23	23	23
Cinneide M.O., Tratt L., Harman M., Counsell S., Moghadam I.H., Experimental assessment of software metrics using automated refactoring, 2012	0	0	0	0	0	11	7	4	22	17	20
Juristo N., Vegas S., Using differences among replications of software engineering experiments to gain knowledge, 2009	0	0	3	5	3	6	5	0	22	17	16
Yoon K.-A., Kwon O.-S., Bae D.-H., An approach to outlier detection of software measurement data using the k-means clustering method, 2007	0	0	5	5	3	4	2	2	21	20	21
Nugroho A., Chaudron M.R.V., A survey into the rigor of UML use and its perceived impact on quality and productivity, 2008	0	2	1	3	3	4	6	0	19	11	17
Mendes E., A comparison of techniques for web effort estimation, 2007	3	1	1	3	7	4	0	0	19	6	16
Host M., Runeson P., Checklists for software engineering case study research, 2007	3	2	2	2	5	3	1	1	19	16	18
	18	27	70	105	95	149	131	28	623	522	566

[a]Without self-citations [b]Without ESEM-proceedings citations

Figure 4 shows the contribution of papers at the country level. The country contributing the most papers to ESEM is United States, Italy and Germany following far behind but close to one another.

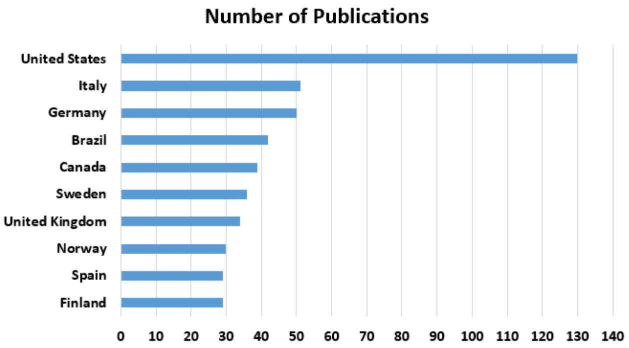


Fig. 4 Contribution of papers at the country level (2007-2015).

C. Hot research topics

When searching for relevant papers of importance researchers tend to either search for a specific journals or topics. The importance of the paper title is obvious – firstly, to capture the attention of the reader and secondly, to depict the specific statement in a compact and comprehensive style. The paper titles of the full data set were visualized using Wordle as shown in Figure 5. In this visualization, common English words (e.g. “in”, “of”) and also the common words in this context (e.g. “software”, “study”, “empirical”, “engineering” and “development”) have been removed from the titles.

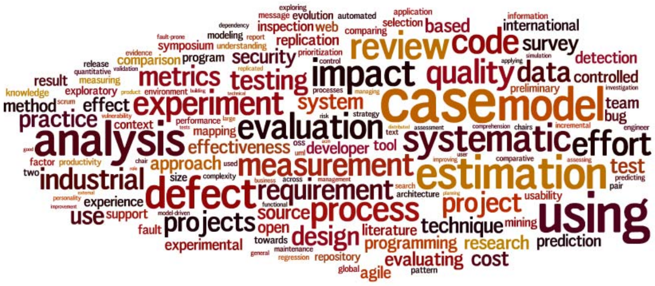


Fig. 5 Word cloud of publication titles (2007-2015)

As discussed in Section 1, for topic analysis, we conducted text-mining of paper titles and abstracts using the LDA topic modeling algorithm. Topic modeling creates a statistical model from a set of documents. This model presents topics which are in fact collection of words most probable for each topic. Each document can then be categorized into a single topic that most accurately represents the contents of the document. The optimal number of topics, measured with log likelihood as explained in [10], was 54. The most popular topics measured by the number of papers are listed in Table 2. The topic names in the top row of the Table 2 are authors’ interpretations and not produced by the topic modeling algorithm.

D. Threats to validity

The data retrieved from Scopus was studied as such. The data set is limited by the selected database, venue and rather short observation period. There were no extra precautions to verify the validity of the data set. Scopus (like other publication databases) has been claimed to include duplicate and/or inconsistent data, regarding e.g. citations, titles or authors (e.g. missing citations, inconsistent style of handling author names or incorrect titles). The study covered all papers from ESEM to the point of conducting the study, thus the set of publications is valid in the context of ESEM.

Not all data was available in Scopus as expected. As in the case for querying the entire data set for number of citations excluding the self-citations, that data was not available as such. In Scopus, there is an option for excluding the self-citations from the data. When querying such data a notification about the data set being too large to handle was generated. The response was rather unexpected for such characteristic functionality of publication database, in particular considering the nature of Scopus, claimed to be the largest abstract and citation database of peer-reviewed literature, and rather small data set used in the study.

The statistical topic modelling of research topics should capture similarity in the semantic content of papers based on statistical word distribution. However, the relatively small number of papers affected the results so that some incoherent topics were created, e.g. papers on literature search strategies, GQM+strategies, and offshoring strategies were put under a same topic. On a larger corpus one expects to get less incoherent topics [10].

IV. Conclusions

The study has given an overview of latent information in the data and related statistics in the context of ESEM papers. We found a large number of un-cited papers but that effect seemed to wear off, e.g. out of papers published prior 2010 81% were cited. This suggest that ESEM papers have a long shelf life and counting citation only of the past two years (as done by the impact factors computed from ISI web of science) might not be suitable. We also found that the top cited papers of ESEM have good visibility outside the ESEM community.

The most popular research topics in ESEM have been systematic reviews, testing, defects, cost estimation and team work. The identification of those topics can help both established and new researchers to spot the active and more impactful topics – to proceed with further incremental research on those areas.

TABLE 2 10 MOST PROBABLE STEMMED TERMS FOR TOP FIVE TOPICS.

Topic	Systematic reviews (n=30)	Testing (n=27)	Defects (n=21)	Cost estimation (n=17)	Team work (n=16)
1	review	test	defect	estim	team
2	systemat	coverag	defects	cost	meet
3	synthesi	exploratori	remov	cocomo	coordin
4	slrs	manual	evolutionari	error	teams
5	primari	tester	fals	scope	belief
6	reviews	alloc	exposur	homogen	daili
7	slr	cases	reduct	modelbas	performance
8	tertiari	algorithm	postreleas	paramet	pilot
9	themat	costeffect	name	durat	obstacl
10	synthes	application	quantifi	evaluation	schedul

In the set of top-cited papers, we noticed that there are papers on all technical, methodological and review types present, see #2, #4, and #6 ranked papers in Table 1, respectively. The identification and classification of the top-cited papers provide various benefits for researchers and practitioners, e.g. the results would (1) help new researchers to see the types of approaches & research methods used and contributions presented in highly-cited papers, and (2) help practitioners spot the highest quality work in specific areas of SE and aim at utilizing techniques, tools or findings reported in those studies in their real-world SE challenges. However, we also recognize that bibliometrics and citation statistics may easily lead to not so obvious problems, as they can e.g. drive people to work only in fashionable sub-fields as suggested in a recent column in the Nature magazine [14].

Our future work directions include replicating this analysis for other publication venues. This allows comparison between research venues and provides more depth to our analysis.

References

[1] C. Wohlin, “An analysis of the most cited articles in software engineering journals - 1999”, Information and Software Technology, 47, 15, pp. 957–964, 2005.

[2] C. Wohlin, “An analysis of the most cited articles in software engineering journals - 2000”, Information and Software Technology, 49, 1, pp. 2–11, 2007.

[3] W. E. Wong, T. H. Tse, R. L. Glass, V. R. Basili, and T. Y. Chen, "An Assessment of Systems and Software Engineering Scholars and Institutions (2001-2005)," *Journal of Systems and Software,* vol. 81, 6, pp. 1059-1062, 2008.

[4] W. E. Wong, T. H. Tse, R. L. Glass, V. R. Basili, and T. Y. Chen, "An Assessment of Systems and Software Engineering Scholars and Institutions (2002-2006)," *Journal of Systems and Software,* vol. 82, 8, pp. 1370-1373, 2009.

[5] V. Garousi and T. Varma, "A Bibliometric Assessment of Canadian Software Engineering Scholars and Institutions (1996-2006)," *Canadian Journal on Computer and Information Science,* vol. 3, 2, pp. 19-29, 2010.

[6] R. Farhoodi, V. Garousi, D. Pfahl, and J. P. Sillito, "Development of Scientific Software: A Systematic Mapping, Bibliometrics Study and a Paper Repository," *Int'l Journal of Software Engineering and Knowledge Engineering, In Press,* vol. 23, 04, pp. 463-506, 2013.

[7] R. L. Glass, I. Vessey, and V. Ramesh, "Research in software engineering: an analysis of the literature," Information and Software Technology, vol. 44, 8, pp. 491-506, 2002.

[8] N. Coulter, I. Monarch, and S. Konda, "Software engineering as seen through its research literature: A study in co-word analysis," Journal of the American Society for Information Science, vol. 49, 13, pp. 1206-1223, 1998.

[9] A. Hoonlor, B. K. Szymanski, and M. J. Zaki, "Trends in computer science research," *Commun. ACM,* vol. 56, 10, pp. 74-83, 2013.

[10] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences,* vol. 101, Suppl 1, pp. 5228-5235, 2004.

[11] R. V. Noorden, B. Maher, and R. Nuzzo, "The top 100 papers," *Nature,* vol. 514, 7524, pp. 550-553, 2014.

[12] T. Dybå, T. Dingsoyr and G.K. Hanssen, “Applying systematic reviews to diverse study types: An experience report”, IEEE, pp. 225-234, 2007.

[13] D.Adam, “Citation Analysis: The counting house”, Nature, vol. 415, 6873, pp. 726–729, 2002.

[14] R. Werner, "The focus on bibliometrics makes papers less useful," Nature, vol. 517, 7534, pp. 245, 2015.