Insights from Mining Eleven Years of Scholarly Paper Publications in Requirements Engineering (RE) Series of Conferences

Richa Sharma
SNU (India)
richa.sharma@snu.edu.in

Peeyush Aggarwal
BVCOE (India)
peeyushaggarwal94@gmail.com

Ashish Sureka
ABB (India)
ashish.sureka@in.abb.com

DOI: 10.1145/2894784.2894794
http://doi.acm.org/10.1145/2894784.2894794

ABSTRACT

We present insights from a bibliometric analysis and scientific paper publication mining of 551 papers in Requirements Engineering (RE) series of conference (11 years from 2005 to 2015). We study cross-disciplinary and interdisciplinary nature of RE research by analyzing the cited disciplines in the reference section of each paper. We apply topic modeling on a corpus consisting of 551 abstracts and extract topics as frequently co-occurring and connected terms. We use topic modeling to study the structure and composition of RE research and analyze popular topics in industry as well as research track. Co-authorship in papers is an indicator of collaboration and interaction between scientists as well as institutions and we analyze co-authorship data to investigate university-industry collaboration, internal and external collaborations. We present results on the distribution of the number of co-authors in each paper as well as distribution of authors across world regions. We present our analysis on the public or proprietary dataset as well as the domain of the dataset used in studies published in Requirements Engineering (RE) series of conferences.

Keywords

Bibliometric Analysis, Co-Authorship Analysis, Interdisciplinary Research, Requirements Engineering, Scientific Paper Publication Mining, Topic Modeling

1. INTRODUCTION

Requirements Engineering (RE) (a sub-field of Software Engineering) is a discipline concerning processes, tools and techniques for defining, documenting and maintaining requirements. RE is an applied and a practice-oriented field. The techniques, case-studies and research results reported by researchers and practitioners in RE conferences is an indicator of important research issues, challenges and problems encountered by practitioners. We believe that the direction in which the RE research and the scientific community moves is to a great extent driven by the need to address short-term and long-term problems encountered by RE practitioners and brining vale to industry. The impact of RE research on practice and the gap between RE research and practice is an area that has attracted several researcher's attention [1][3][4][8]. The study presented in this paper is motivated by our belief that bibliometric analysis and scientific paper publication mining can be used as a research tool to analyze various aspects of Requirements Engineering (RE) research such as industry-academia collaboration, impact of RE research on practice, cross disciplinary nature of RE, internal and external collaboration, emerging and popular topics, RE across industry verticals and domains.

Year	Research	Industry	Total
2015	28	9	37
2014	42	13	55
2013	29	13	42
2012	37	9	46
2011	34	10	44
2010	39	14	53
2009	33	15	48
2008	37	15	52
2007	43	12	55
2006	47	7	54
2005	56	9	65
Total	425	126	551

Table 1: Number of Papers (Experimental Dataset) in Requirements Engineering Conference from 2005**-**2015

The International Requirements Engineering Conference (RE)[1] is a prestigious and a long running conference (started in 1993) in the area of Requirements Engineering. RE 2016 is the 24th edition of the conference and the RE series of conferences provides an Industry track in addition to the research track intended to bring practitioners from Industry to present case-studies, experience report, problems and solutions relevant to practice. We create a database of all the research and industry track papers published in RE 2005 until RE 2015 (11 years). Table 1 shows the number of papers for each year in our dataset categorized into research and industry track. We observe that a total of 551 papers are published in RE in past 11 years out of which 425 (77.13%) are in research track and 126 (22.87%) are in industry track. Table 1 reveals that RE has a significant industry participation as more than 20% papers are from Industry track. The scope of our analysis presented in this paper is confined to only Requirements Engineering (RE) series of conference over a period of eleven years from 2005 to 2015. RE being a sub-field of SE, research papers on RE are published in several conferences on Software Engineering. However, we analyze data only from RE series of conferences as selection of SE conferences and identifying RE papers in such conferences by us can result in a selection bias. Hence, we eliminate selection bias without compromising on data quality as well as quantity by analyzing publications only from RE conference. We study past eleven years of scientific paper publications both from research as well as industry track from RE series of conferences as we believe that including data (papers in our case) which is too many years back might not be representative of current practice. We believe that our corpus of 551 papers covering entire 11 years of publications in RE conferences is representative of the current

[1] http://requirements-engineering.org/

practice in RE. A similar methodology is adopted by Tripathi et al. [7] in their study on scientific publications in the field of Mining Software Repositories (MSR). The fields extracted from raw data and their values can be downloaded as Excel which we have made publicly available[2].

Our literature survey leverages studies on mining scientific knowledge from scholarly publications and bibliometric analysis in the field of Software Engineering. Freitas et al. analysis 677 papers in Search Based Software Engineering (SBSE) [2], Raulamo-Jurvanen et al conduct a citation and topic analysis of 513 papers in Empirical Software Engineering and Measurement (ESEM) [6], Tripathi et al. conduct a study of 187 papers in Mining Software Repositories (MSR) [7]. However, to the best of our knowledge, the work presented in this paper is the first study on scholarly publication analysis in Requirements Engineering. In context to existing work, the study presented in this paper makes the following novel contributions:

Research Contributions: This paper presents the first in-depth and focused study on scholarly publication analysis of 551 papers across 11 years of RE series of conference (from year 2005 until 2015) to identify latent topics using topic modeling techniques, extract popular topics and topic compositions, analyze cross-disciplinary and interdisciplinary nature of RE research, study extent of internal as well as external collaboration and industry-academia collaboration, examine the type and domain of dataset used and conduct an analysis on number of authors.

2. EMPIRICAL ANALYSIS AND RESULTS

In this Section, we present our research methodology and results of our empirical analysis. We present the results of our analysis on Interdisciplinarity, Topic Modeling, Authorship Numbers and Regions, Collaboration and Public and Proprietary Dataset.

2.1 Interdisciplinarity

We conduct empirical analysis to measure interdisciplinarity of RE. Measuring multi-disciplinarity and relation of RE with other disciplines is useful to understand the inter-relatedness, integration and diffusion of ideas between fields and a body of research. Also, such metrics and insights are useful for policy makers, research community and funding agencies as high interdisciplinarity of a body of research is associated with solving complex problems and promotion of scientific development and innovation [5]. We look at the reference section of each paper in our dataset and extract the journal and conference name for all the citied papers and articles. We believe that citing or referring a discipline is a reasonable indicator of measuring cross-disciplinary research. If the conference or journal name is on Software Engineering (such International Conference on Software Engineering) or any sub-field within Software Engineering (such as International Symposium on Software Reliability Engineering) then we classify it as a neighboring field and ignore it. Our interest is to measure interaction of RE with distant fields such as law, medicine and sociology.

Table 2 displays the list of all the disciplines (in alphabetical order) cited by papers in our dataset. Table 2 reveals that RE is very integrative or multidisciplinary field which builds relies on several concepts, theories, tools, techniques and data from various distant fields. Table 2 shows that the knowledge sources from which the RE research draws are diverse. We observe that several papers are more integrative in comparison to other papers. We found some papers which do not cite any distant field whereas some papers cite two and three distant fields. For example, we have one paper on requirements engineering, medical application and game playing. Another example is a paper on requirements engineering, legal compliance and health care systems. We calculate the number of unique distant fields referred in a paper and plot a pie chart showing the distribution of the number of distant fields per paper in the experimental dataset. We observe that the maximum number of unique distant field per paper is 3. The pie chart in Figure 1 reveals that 32% of the papers do not refer to any field outside the broad area of Software Engineering. The pie chart shows that 68% of the RE papers are multidisciplinary and 7% studies consists of interaction between 3 fields.

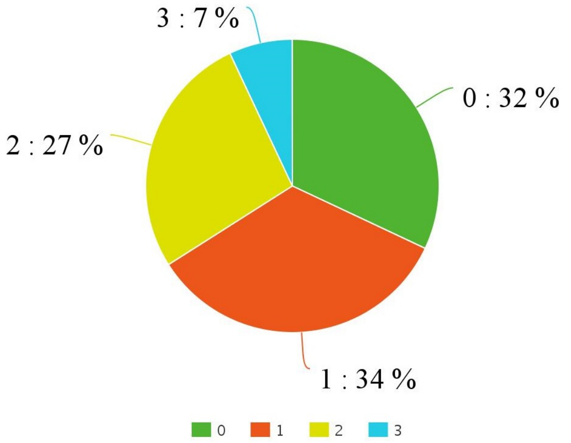

Figure 1: Pie Chart showing the Distribution of Number of Distant Fields per Paper

2.2 Topic Modeling

We use Topic Modeling (Latent Dirichlet Allocation) techniques to automatically identify the topics that characterize the papers in our dataset. Our objective is to infer the latent topics (wherein a topic is a distribution of terms) from the paper abstracts (or dataset of 551 abstracts is publicly available and can be downloaded[3]) and then answer questions like which are the popular topics, which topics are emerging and what is the general composition of the papers across various topics. We pre-process the abstract by first removing all the stop-terms (non-content bearing and common terms), converting the text into lower-case and then we apply term stemming for reducing inflected words to their root form. We experiment with various number of topics and set the number of topics as 10, 15 and 20. The number of topics needed to represent the latent topics contained in the corpus is based on heuristics of total number of papers and size of the abstract. LDA is applied for different number of iterations. We make the result of topic modeling publicly available (can be downloaded[4]). Table 3 shows the list of Topics (manually labeled by us) and the key terms associated to the topic. The topics and terms in Table 3 reveals the structure of RE research. We used a graphical user interface based Latent Dirichlet Allocation based Topic Modeling tool hosted on Google Code[5].

The topic modeling output gave us best results for 15 topics out of which one is miscellaneous and remaining 14 are presented in Table 3. As shown in Table 3, a topic (labeled by us manually)

[2] http://bit.ly/1MQvSM4

[3] http://bit.ly/1NYoqR2

[4] http://bit.ly/1kxjlqi

[5] https://code.google.com/p/topic-modeling-tool/

Table 2: List of All the Disciplines (in alphabetical order) cited by papers in our dataset

Aerospace Engineering	Computer Networks and Communication	Game Design and Playing	Nuclear Safety
Agriculture and Food	Computer Simulation	Generative Programming	Performance Engineering
Ambient Intelligent Systems	Computer Supported Collaborative Work	Graph Theory	Political Science
Artificial Intelligence	Control Systems	Healthcare and Medicine	Predictive Modeling
Assistive Technology	Creativity and Cognition	Human Computer Interaction	Privacy and Trust
Automation Engineering	Cross Culture Management	Information Retrieval	Probabilistic and Asymptotic Method
Automotive Systems	Crowdsourcing	Intelligent Mechatronics	Production and Operations
Automotive Systems	Data Mining	Knowledge Engineering and Ontology	Risk and Uncertainty
Banking and Finance	Database Management Systems	Law	Robotics
Behavioral Science	Decision Support Systems	Library Science	Security Engineering
Bioinformatics and Bioengineering	Distributed Systems	Machine Learning	Social Network Analysis
Biometrics	E-Commerce	Management Science	Sociology
Business Process Management	Economics	Manufacturing Industry	Spacecraft Design
Case-Based Reasoning	Education	Marketing	Strategic Management
Chemical and Environmental Engineering	E-Learning and Social Learning	Mobile and Handheld Computing	Structural and Multidisciplinary Optimization
Cloud Computing	Electronic Voting System	Molecular Biotechnology	Supply Chain Management
Cognitive Modeling	Emergency Management	Multi-Agent Systems	Telemedicine and Telecare
Cognitive Psychology	Engineering Management	Multimedia Information Processing	Text Mining
Commonality Analysis	Epistemology	Multi-Valued Logic	Transportation Engineering
Computational Linguistics	Ethnography	Nanotechnology	Usability Engineering
Computational Swarm Intelligence	Evolutionary Computing	Natural Language Processing	Visualization
Computer Aided Design	Expert Systems	Neuropsychology	Water Science and Technology

consists of a cluster of frequently co-occurring terms. We observe that terms trace, artefact and links co-occur and are connected to the field of requirement traceability. Similarly, we notice that terms terms evolution, change and environment co-occur and are frequent used within the context of requirement evolution. We believe that there are important topics in our corpus which the topic modeling algorithm has not been able to find. This is because our corpus size is relatively small (551 documents) and the size of the document is also small (only abstracts). The topic modeling output gives each document as a mixture of various topics. We applied topic modelling only on the abstract and not the paper and thus map each of the 551 abstract to only one of the 14 topics in Table 3 or Miscellaneous (incase if there is no dominating topic). Figures 2 and 3 shows topic distribution for research track and industry track papers. The full-form for the two letter acronym denoting the topics in Figure 2 and 3 are mentioned in Table 3. Figure 2 reveals that more than 8% of the papers are classified into requirements specification, evolution, elicitation, goal-based analysis and domain description. Figure 2 shows a different distribution for industry track papers in comparison to the research track papers. We observe that papers on requirements engineering in business process management and industry practices are much higher in industry track in comparison to the research track. On the other hand, papers on goal-based requirement analysis, requirement tractability and law-compliance requirements are relatively higher in research track in comparison to the industry track. We observe that topics like security-critical requirements, requirements specification, requirement evolution, modeling and quality requirements are equally distributed and popular in both the tracks.

2.3 Authorship Numbers and Regions

We conduct a bibliometric analysis of co-authored scientific articles in our dataset to study research collaboration. Co-authorship in papers is an indicator of collaboration and interaction between scientists as well as institutions. Our objective is to study university-industry collaboration, internal and external collaboration and examine to what extent university and industry are producers of knowledge. Studying collaboration between researchers through statistical indicators is important from science policy perspective. We extract the author names and their respective affiliation from each article in our dataset. Table 4 displays the distribution of number of co-authors for 551 articles in our experimental dataset.

Table 4 reveals that 7.07% of the articles are solo-authored. We observe that 27.04%, 30.30% and 19.41% articles are co-authored by 2, 3 and 4 authors respectively. Table 4 shows that 16.2% of the papers have more than 5 co-authors. Experimental analysis shows that more than 90% of the publications involve collaboration and the number of single authored articles are less than 10%. It is interesting to note that 77% of the articles are co-authored by 2 to 4 authors. We observe a similar trend over 11 years (except minor difference) and there are no significant changes over the years in terms of the distribution of number of authors.

Table 3: Topic Model Output (List of Topics and Terms associated to each Topic)

	Topic Label	Associated Terms		Topic Label	Associated Terms
1	Requirements Traceability (TR)	trace, link, source, automated, artifacts	8	Requirements Specification (SP)	use-case, specification, language, natural language, documents
2	Requirements Modeling (MD)	formal approaches, models, modeling	9	Goal-Based Requirement Analysis (GB)	design goals, business goals
3	Security Requirements (SR)	security, risk, impact	10	Safety-Critical Requirements (SC)	safety critical, compliance, legal, control, risk, uncertainty
4	Requirements Evolution (EV)	evolution, change, environment	11	Quality Requirements (QL)	quality, maintenance, cost
5	Legal and Law-Compliant Requirements (LG)	legal, compliance, regulation	12	Requirements Elicitation (EL)	elicitation, communication, domain knowledge, service
6	Domain Requirements Description (DM)	domain requirements, product features, feature models	13	Requirements Change Management (CR)	change, complexity, scenario
7	Business Process Modeling in Software Requirements (BP)	business process, organization, project, management	14	Requirements Engineering Industry Practices (IP)	industry practices, management, empirical, industrial, case study, experience, practitioner

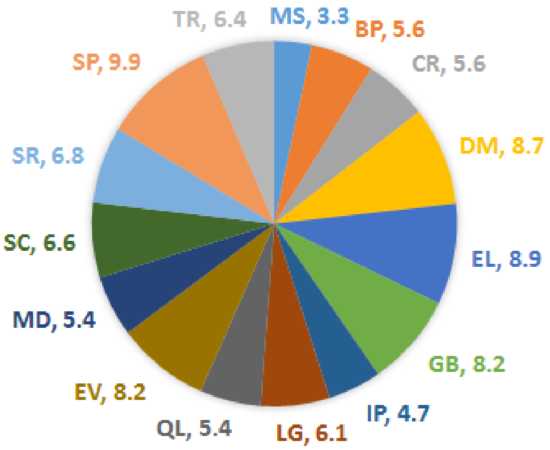

Figure 2: Topic Distribution for Research Track Papers

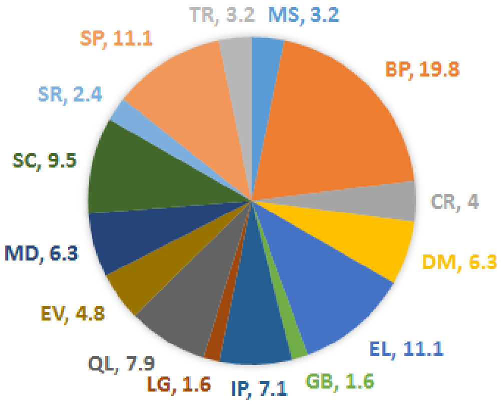

Figure 3: Topic Distribution for Industry Track Papers

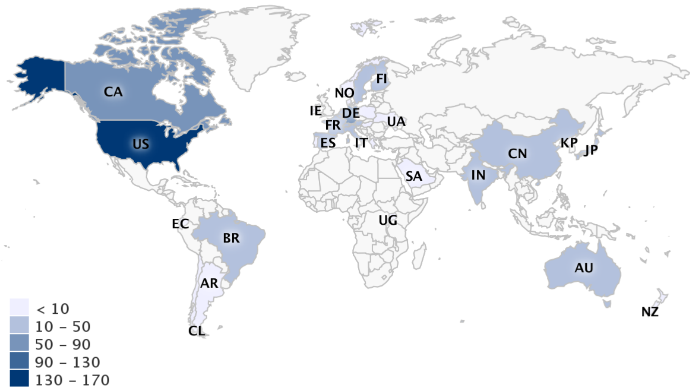

Figure 4: Map Displaying Scholarly Output of Various Countries in the World

We study the regional demographics of authorship for the 551 papers in our dataset. We extract the country of every author in the dataset and observe that the papers are from 36 different countries. We extract the unique list of countries from each paper and notice that cross country collaborations also. Figure 4 shows a map revealing the scholarly output of 36 countries. Our analysis reveals that USA (143 papers), UK (79 papers), Canada (75 papers) and Germany (74 papers) have the highest scholarly output. Next in the list are Italy (33), Netherland (26), Spain (24), China (21) and Switzerland (21). We observe that out of 36 countries, 18 countries have less than 10 papers.

2.4 Collaboration

We study the nature and scale of collaboration from the perspective of internal or external collaboration. We compute statistical indicators for collaboration between institutions. We define internal collaboration as one form of collaboration in which all the co-authors (single or multiple-authors) are from one Institution. We define external collaboration as a form of collaboration which involves participation of two or more institutions (irrespective of industry or university) in the production of the study and scientific output. Figure 5 shows a bar chart displaying the percentage of papers having external collaboration across 11 years. Figure 5 reveals that 58% of the articles involved co-authors from more than one Institution. We observe that external collabora-

Year	1	2	3	4	5	6	7	8	9	10	11
2015	3	14	7	6	6	1	0	0	0	0	0
2014	1	17	11	16	4	4	2	0	0	0	0
2013	2	13	16	5	3	3	0	0	0	0	0
2012	3	11	14	8	5	2	1	1	1	0	0
2011	2	14	10	12	3	2	1	0	0	0	0
2010	5	11	18	11	6	2	0	0	0	0	0
2009	2	9	17	9	6	2	1	1	0	0	1
2008	4	15	12	14	3	1	3	0	0	0	0
2007	6	13	20	6	9	0	0	1	0	0	0
2006	4	18	16	10	4	1	0	0	0	1	0
2005	7	14	26	10	6	1	1	0	0	0	0
Total	39	149	167	107	55	19	9	3	1	1	1

Table 4: Distribution of Number of Co-Authors in Each Paper in the Experimental Dataset

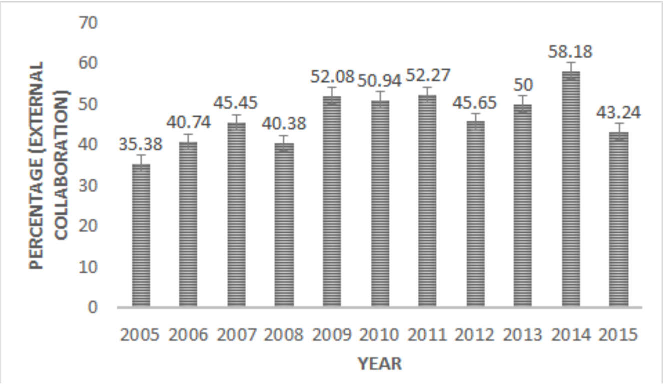

Figure 5: Bar Chart displaying Percentage of Papers having External Collaboration across 11 years

tion ranges from 35% to 58% which is an indicator of a good propensity towards interaction of scientist between organizations.

We study and gather evidences on collaboration and knowledge flow between industry and academia by measuring the degree of joint authorship between scientists in industry and university. One of the aims of our study presented in this paper is to assess the degree of collaboration between University and Industry in the area of Requirements Engineering (RE) by mining author affiliation data from scientific paper publications. We compute the number and percentage of papers published in our experimental dataset corpus having co-authors from both University and Industry. The number and percentage of papers involving authors from both University and Industry is an indicator of the extent of University-Industry collaboration. Figure 6 displays a bar chart showing the extent of university-industry collaboration as a percentage of external collaborations (which can between universities, between industries and between universities and industries) and as a percentage of total number of studies in the dataset. The bar chart in Figure 6 reveals that for the year 2007, 27% of the articles involved co-authors from both industry and academia and 60% of the external collaborations belonged to the university-industry collaboration category. We observe that the percentage of articles having university-industry collaboration ranges from 13% to 27%. University-Industry collaboration as a percentage of external collaboration ranges from 33% to 62%. We do not observe any noticeable downward or upward trend and or analysis reveals that the level of university-industry collaboration is generally within the 15% to 25% band.

2.5 Public and Proprietary Dataset

Table 5: Percentage of Studies involving Dataset Analysis. Percentage of Papers (With Respect to Papers having Dataset) having Public, Proprietary and Both Public and Proprietary Dataset Analysis (CD: Percentage of Studies using Dataset)

Year	SD	Proprietary	Public	Both
2005	60.0	79.5	7.7	12.8
2006	53.7	82.8	3.4	13.8
2007	49.1	96.3	0.0	3.7
2008	55.8	86.2	3.4	10.3
2009	62.5	83.3	0.0	16.7
2010	62.3	81.8	3.0	15.2
2011	72.7	84.4	3.1	12.5
2012	58.7	77.8	0.0	22.2
2013	66.7	78.6	10.7	10.7
2014	76.4	78.6	7.1	14.3
2015	73.0	77.8	7.4	14.8
Average	62.8	82.5	4.2	13.4

Generalization of findings and results is an important aspect in Requirements Engineering (RE) research. The characteristics of publicly available and open source data is not always the same as closed or proprietary data. Our objective is to investigate the extent of usage of publicly available and proprietary data in requirements engineering research. Table 5 shows percentage of studies involving analysis on a dataset and a distribution of those studies across proprietary, public and both (public and proprietary) dataset analysis. Our analysis reveals that a large number of studies use dataset which are requirements specification documents. Requirements specification documents are proprietary to the organization and are generally not made available in public domain. We observe that for a small number of studies, research work is based on user reviews, policy or regulation documents that are available in public documents. Therefore, we observe that a significant percentage (around 60%) of datasets used in RE research are proprietary in nature as summarized in Table 5. Another interesting observation that we found is only 63% (averaged over a period of 11 years) of the research work is based on datasets. Following are few examples of the proprietary dataset used in studies in our corpus:

1. Contingency Requirements for an autonomous rotorcraft project
2. Requirements for Siemens Telecommunication System
3. Instrument Cluster Specification at Daimler Chrysler
4. Scenarios from Air Traffic Management System
5. Specifications for NASA's science instruments
6. Requirements specific to Teradyne instruments

Following are few examples of the publicly available dataset used in studies in our corpus:

1. London Ambulance Case-Study
2. iTrust project requirements
3. App Reviews (data collected: review text, title, app name, category, store, submission date, username, and star rating)
4. Four goal models from papers on goal-oriented RE
5. Documents of Three IT laws: HIPAA documents and amendments to HITECH Act, and the India 2011 IT Rules

The need for benchmark datasets has been realized in RE domain over the last few years. As a result, collaborative efforts have led

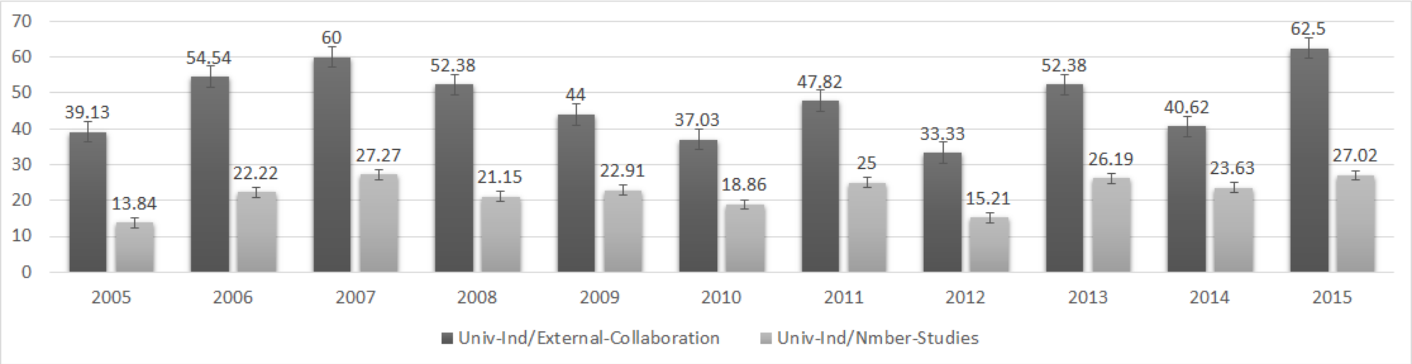

Figure 6: Bar Chart displaying University-Industry Collaboration as a Percentage of External Collaboration as well as a Percentage of Total Number of Studies in the Dataset

to compilation of requirements specification documents in few universities. iTrust (Refer to agile.csc.ncsu.edu iTrust wiki website) project is one such example that aims at collecting project artifacts for comparative analysis and serve as a benchmark project for both RE and SE research.

3. CONCLUSION

Our analysis shows that that RE is very integrative or multidisciplinary field. 68% of the RE papers in our dataset are multidisciplinary (cites papers belonging to at-least one distant field) and 7% studies consists of interaction between 3 fields. Our analysis reveals that the percentage of single authored articles are rare (only 7%). Similarly, number of articles with more than four authors are around 15%. After applying topic modeling, we observe a different distribution of topics for industry track papers in comparison to the research track papers. Results shows that countries like USA, UK, Canada and Germany have the highest scholarly output. We observe that out of 36 countries, 18 countries have less than 10 papers. We find that 58% of the articles involved co-authors from more than one Institution. We observe that the percentage of articles having university-industry collaboration ranges from 13% to 27% across 11 years. Our findings shows that a large percentage of studies using requirement specific documents for data analysis are propriety and not publicly available. We do not observe any noticeable downward or upward trend and or analysis reveals that the level of university-industry collaboration is generally within the 15% to 25% band.

4. REFERENCES

[1] Davis, A., and Hickey, A. Requirements researchers: Do we practice what we preach? *Requirements Engineering 7*, 2 (2002), 107–111.

[2] de Freitas, F. G., and de Souza, J. T. Ten years of search based software engineering: A bibliometric analysis. *SBSE 6956* (2011), 18–32.

[3] Kaindl, H., and et al., B. Requirements engineering and technology transfer: Obstacles, incentives and improvement agenda. *Requirements Engineering 7*, 3 (2002), 113–123.

[4] Pais, S., Talbot, A., and Connor, A. Bridging the research-practice gap in requirements engineering. *Bulletin of Applied Computing and Information Technology, 7(1)* (2009).

[5] Porter, A. L., Roessner, D. J., and Heberger, A. E. How interdisciplinary is a given body of research? *Research Evaluation 17*, 4 (2008), 273–282.

[6] Raulamo-Jurvanen, P., Mantyla, M., and Garousi, V. Citation and topic analysis of the esem papers. *ESEM* (2015), 1–4.

[7] Tripathi, A., Dabral, S., and Sureka, A. University-industry collaboration and open source software (oss) dataset in mining software repositories (msr) research. *SWAN* (2015), 39–40.

[8] Wieringa, R. Requirements researchers: are we really doing research? *Requirements Engineering 10*, 4 (2005), 304–306.