Software Reengineering Bibliometrics – People, Topics, and Locations

Jochen Quante

Robert Bosch GmbH, Corporate Research
Stuttgart, Germany

The statistical analysis of publication data can provide insightful information about the state and evolution of a research field. We report about application of such an analysis on software reengineering publications. The analysis of authors, titles and abstracts results in an overview about relevant people, topics, and locations.

1 Introduction

Researchers often face the challenge of getting an overview over a research field. A good starting point is literature research. However, this easily results in thousands of documents, and it is very hard to get an overall picture. This is where statistical analyses can help. By using data mining techniques, this data can be leveraged to a level where it really gives an overview of the most relevant persons, topics, and locations. We applied such an approach to the analysis of scientific publications on software reengineering. It is based on an idea by Hassan et al.[1] who did a similar analysis ten years ago.

2 Data Collection

We start by identifying the most relevant conferences. The selection of conferences guarantees a certain degree of quality (peer reviews). It also has the advantage that people meet there, so the probability to find collaborations between people is higher than in journals. Next, a bibliographic database is queried for publications from these conferences. We use Scopus[1], which provides not only authors and title, but also a lot of additional information: Abstracts, addresses of authors, citation information, etc. This data is then exported and analyzed in a proprietary mining tool.

3 Mining Approach

The following information can be mined from this data:

- Ranked authors lists: Authors with the highest number of publications, or with the highest number of direct or indirect collaborations. This requires to match identical authors, which is a challenge of its own (e. g., typos, variants in writing the name, multiple people with same name).

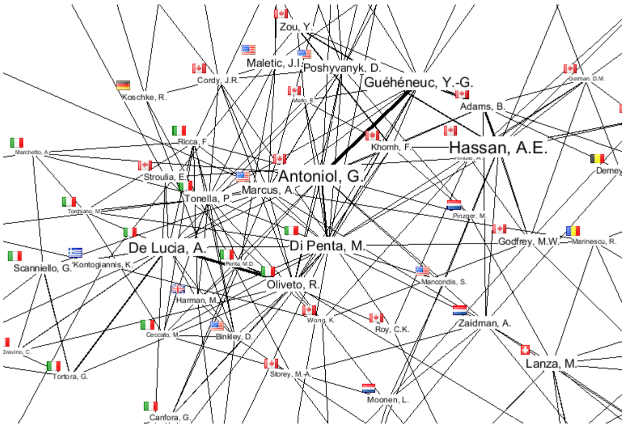

Figure 1: Collaborations between authors with at least 10 publications.

- Collaboration graph: Who has written papers together?
- Geographical authors map: Where do people come from? What are the central locations of research in this field? This information can be derived by using Scopus' affiliation information. Unfortunately, the affiliation is a free text field, which means that text mining techniques are needed to derive the location (city and country) out of this. For example, sometimes only the name of the institute is given.
- Topic map: What are the main topics of these publications? Which topics are related? Who is active in which of these topics? There are several approaches to do topic mining. We integrated word group counting and an advanced topic mining approach [2].
- Trend analysis: Which topics are becoming more frequent, which ones are fading out? When topics have been identified, the number of papers on this topic over the years can be analyzed: Is it increasing or decreasing? However, only really strong trends can be identified this way. "Weak signals" cannot be found using such an approach.

4 Results

We took as a basis all publications from CSMR, ICPC, ICSM and WCRE from 2004 until 2013, plus ICSE

[1] `http://www.scopus.com/`

Author	Country	#Pub.	#CoAu.
Hassan	Canada	58	61
Antoniol	Canada	56	70
De Lucia	Italy	46	(45)
Di Penta	Italy	45	62
Guéhéneuc	Canada	44	62
Koschke	Uni Bremen	22	24
Nierstrasz	Uni Bern	21	32
Sneed		21	16
Pinzger	Uni Klagenfurt	15	28
Knodel	IESE Kaiserslt.	15	27

Table 1: Authors with most publications and most distinct co-authors (overall and German-speaking).

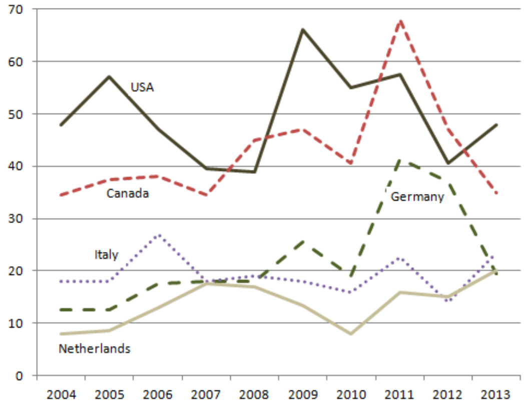

Figure 2: Where most publications came from (year vs. number of publications).

publications with keyword “maintenance”. This corpus contains 2,547 publications by 3,433 different authors from 63 different countries.

Looking at authors shows that the most active ones come from Canada. The ones with the highest number of publications also have the highest number of co-authors (see Table 1). Table 1 also shows the “Top 5” German-speaking authors. Collaborations can best be shown in a graph. Figure 1 shows collaborations between authors who have published more than 10 papers. The strongest collaborations can naturally be found between Professors and their (former) Ph. D. students. For other collaborations, it can be noted that they are more common between people who reside close to each other.

Another question is where the center of research in this area is. Figure 2 shows where most papers came from during the last 10 years. USA and Canada are on top, but Germany has also been quite active for some time. Such a visualization can be useful to see if other players (e. g., from China) are coming up.

Another aspect of publications concerns content: You want to know what the main topics of research are. When looking at word group frequencies in titles, terms like “reverse engineering”, “source code” or “software maintenance” are identified. Looking at the whole abstracts results in more specific terms like “web applications”, “execution traces” or “aspect oriented programming”. A more advanced technique is *topic mining* [2]. Applying this technique to all abstracts from ICPC results in topics as shown in Figure 3. Such a *topic map* also shows how topics are related: Each small circle is a publication, and the topic areas cover all papers that deal with this topic. This technique can be used to create a “landscape” of a research topic.

City	Country	#Auth.	#Pub.
Montreal	Canada	83	201
Delft	Netherlands	27	90
Kingston	Canada	36	76
Lugano	Switzerland	24	61
Salerno	Italy	22	58
Bern	Switzerland	17	49
Vienna	Austria	35	41
Bremen	Germany	21	35
Kaiserslautern	Germany	24	29
Stuttgart	Germany	14	24

Table 2: Where most publications and authors come from (overall and Germany only).

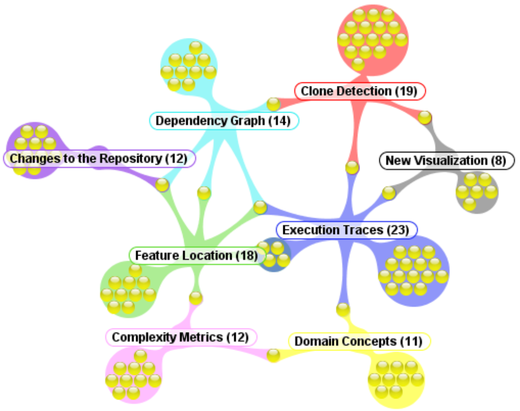

Figure 3: Topic map for ICPC publications (excerpt).

5 Summary

We have shown how bibliometrics and text mining can give insights into a given research topic. Such techniques can be helpful for getting a first overview for further analysis of a research field. It can also be used to continuously monitor a field in order not to miss relevant trends and changes.

References

[1] A. E. Hassan and R. C. Holt. The small world of software reverse engineering. In *Proc. of 11th Working Conference on Reverse Engineering (WCRE 2004)*, pages 278–283, 2004.

[2] S. Osinski and D. Weiss. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54, 2005.