Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies

Antti Knutas, Arash Hajikhani, Juho Salminen, Jouni Ikonen, Jari Porras

Abstract: *There is an increasing number of scientific articles being published, which makes tracking the state of the art more time-consuming. There are software tools available to help with systematic mapping studies in a field of science, but most of these tools are closed source and involve several manual time-consuming steps that could be automated further. We present an open solution as a cloud-based design for bibliographic analysis that makes the research method available for a wider audience.*

Key words: *Systematic mapping study, literature review, bibliometric analysis, citation analysis, social network analysis, R, SaaS, cloud computing*

INTRODUCTION

A systematic mapping study (SMS) classifies and structures a field of interest in research by categorizing publications and analyzing their publication trends [17]. Additionally, SMS can provide a sufficient terminology to facilitate the overall analysis of studies in the field, and represents the research methods and outcomes [3]. There is an increasing number of papers being published in any given field, which makes it difficult to give as much attention to individual research papers as previously [15]. At the same time systematic mapping studies are becoming an increasingly important method of getting an overview about the state of a field of science, also in software engineering and education [7, 11]. This means that these mapping studies, which involve several manual processing and analysis steps are becoming increasingly time-consuming. A mapping study involves at a minimum a systematic search from scientific databases, archival, sorting, manual overview of articles by reading, recording selected metadata and content information, categorizing articles and writing a summary from each selected article.

For example, manually reviewing all the 990 articles in CompSysTech 2000 - 2014 conference proceedings would take a researcher three and half days if reviewing one article took only five minutes. This problem has been recognized by several research communities and several tools have been developed that assist in statistical and citation network analysis [6, 8]. However, these tools still require setup and expert knowledge in data preparation and processing, and they are closed systems. The current tools available are not extensible, meaning that each analysis problem a researcher faces has to be re-implemented from the ground up. We propose that open, extensible tools with even more automated workflows will make this bibliographic analysis available to a wider part of the community of researchers and enables more people to get statistical insight into their scientific database search results. As a way to test and demonstrate our proposal we implemented an extensible web-based interface for our literature analysis program.

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/event.cfm?id=RE248 by late 2015. Please use the official version of the paper and publication reference when citing: Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., Porras, J., 2015. Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies. CompSysTech 2015.

We present the literature analysis tool *NAILS*, which uses a series of custom statistical and network analysis functions to give the user an overview of literature datasets. The features can be divided into two primary sections: Firstly, statistical analysis, which for example gives an overview of publication frequencies, most published authors and journals. Secondly, the more novel network analysis, which gives further insight into relationship between the interlinked citations and cooperation between authors. For example, the most basic features can use citation network analysis identify the most cited authors and publication forums. Advanced features support mapping researcher cooperation and citation networks, and finding the core publications in the examined field of science. The tool's source code is freely available in Github, an open source code repository, and the web-based interface can also be accessed from the project page *(http://aknutas.github.io/nails/)*.

Our research question in this study is as follows: *How can the bibliographic analysis process for systematic mapping studies be made more straightforward and accessible for researchers?* When accomplishing this task, we also investigate as a subquestion which kind of automatic analysis results would provide additional value to researchers.

The next section of the paper reviews publications on systematic mapping studies, analysis tools and the possibilities of network analysis in bibliometric studies. The second to last section introduces the analysis software design, features and the utility of analysis results. The paper ends with the discussion and conclusion section.

RELATED WORK ON SYSTEMATIC MAPPING STUDIES

A systematic mapping study (SMS) is a secondary study that aims at classification and thematic analysis of earlier research [11, 17]. It is closely related to a wider secondary study, a systematic literature review (SLR), which aims at gathering and evaluating all the research results on a selected research topic [2, 10]. Kitchenham and Charters [11] present the best practices of both for the field of software engineering and also compare the two. The SMS is more general in search terms and aims at classifying and structuring the field of research, while the target of SLR is to summarize and evaluate the research results. Kitchenham and Charters [11] also discuss the applications and states where SMS can be especially suitable if few literature reviews have been done on the topic and there is a need to get a general overview of the field of interest. Both kinds of studies can be used to identify research gaps in the current state of research.

While less deep in analysis than a full systematic literature review, a systematic mapping study includes the following steps [17]: 1) Systematically determining the search terms and databases. 2) Performing test searches to validate the search terms. 3) Running a full search and storing the results. 4) Deduplication, sorting and application of inclusion and exclusion criteria. 5) Fully reviewing the papers according to established criteria.

The challenge of analyzing author and citation interactions can be approached with social network analysis. Social network analysis (SNA) is an interdisciplinary technique for the analysis of social networks [14], where social relationships are viewed in the terms of network theory. In social network analysis communication between individual or social units are mapped into a communication matrix and then visualized in graphs. In graph theory there are different mathematical tools available, which can be used to for example estimate the relative influence of nodes in the graph or analyze the graph by the nodes' connection patterns [1, 12]. In scientific bibliometric analysis the communication patterns, or citations, of different authors can be analyzed by modeling publications or authors as nodes and mapping citations or co-publications as node edges. The method of social network analysis for bibliometric studies has been applied for citation network visualization [9], and for detecting and analyzing co-authorship networks [16].

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/event.cfm?id=RE248 by late 2015. Please use the official version of the paper and publication reference when citing: Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., Porras, J., 2015. Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies. CompSysTech 2015.

AUTOMATING SYSTEMATIC MAPPING STUDIES WITH NAILS

The literature analysis system presented in this paper consists of two major parts: the series of analysis scripts, implemented in R, and the web-based batch processing interface. The web interface is responsible for queuing up jobs and serving results, and the R analysis component performs the statistical and network analysis. The server uses an extensible plug-in architecture and the open analysis components can be modified to have new features or new plugins can be added to the system.

The two-part design of the analysis system and the separation of components are intended to make software design for different environments easier. This way the open source R functions can for example be included in other desktop analysis software packages and similarly more analysis components can be added to server without having to recompile the server software itself.

The entire process can be deployed to a group of dynamically scaling cloud instances. The analysis process is initiated by a user from the front-end web server, which keeps track of input files and queues up job requests into a relational database. The queue of job requests is then batch processed by a separate analysis program, which can be on a single server in low traffic situations, or on several different server instances when necessary. After the analysis is complete, the analysis program uploads the results to storage and updates the database with completed job status. The request process between different server components is illustrated in the Figure 1.

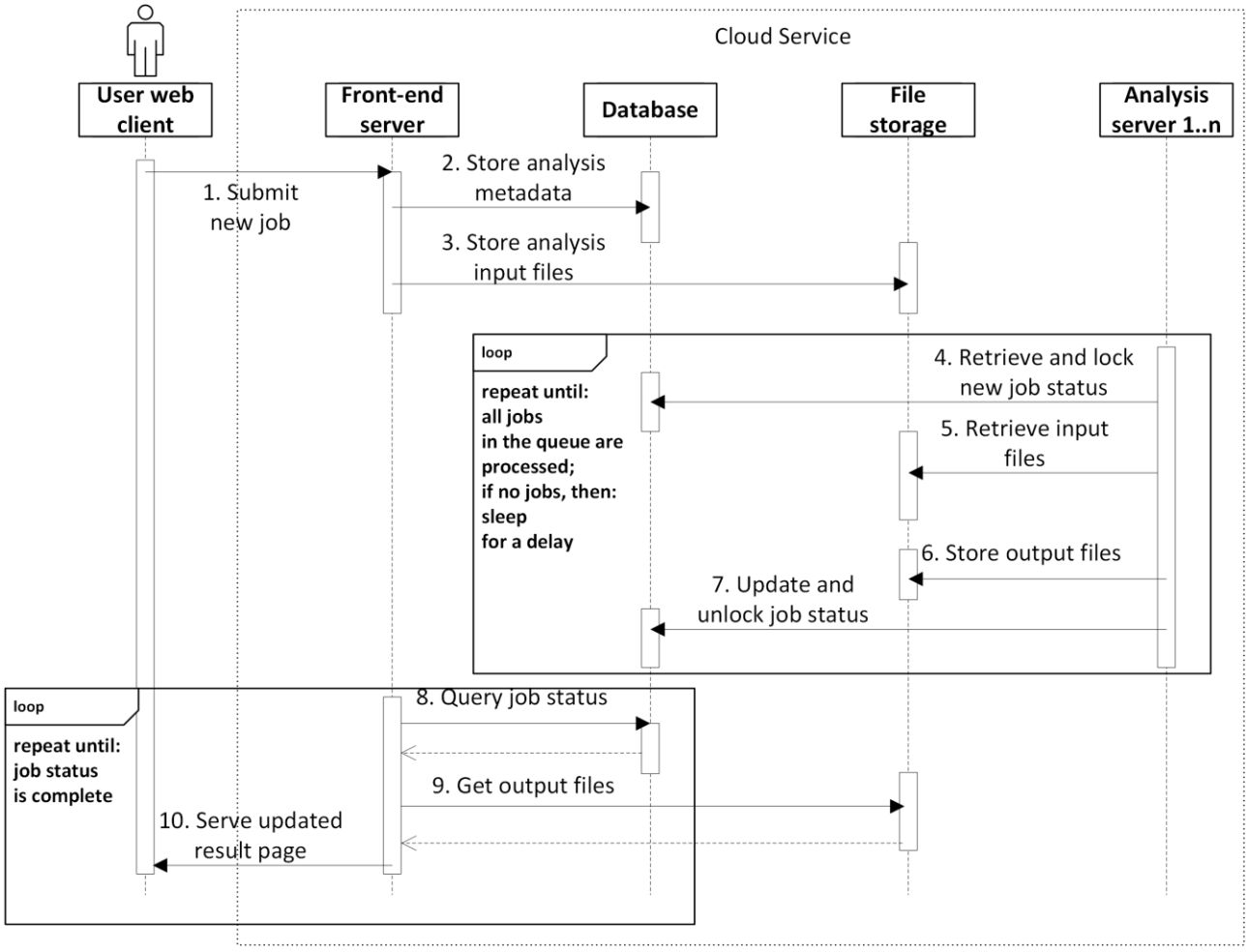

Figure 1. Traversal of a job request through the system components

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/event.cfm?id=RE248 by late 2015. Please use the official version of the paper and publication reference when citing: Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., Porras, J., 2015. Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies. CompSysTech 2015.

The separation of components might seem excessive at a first glance, but the software components can be run on a single machine at low traffic situations. Conversely the components can be deployed on separate servers with multiple simultaneous program instances if necessary and cloud storage service with content distribution services be used for file storage.

Functionality and Services

The system works on publication records available for download from Thomson Reuters Web of Science Core Collection. The system analyses seven essential variables from each publication, which include the authors, keywords, publication forum, article type and cited articles. The user downloads the literature data from Web of Science and uploads it to the analysis system via a web interface. The system then removes duplicate records and performs an exploratory data analysis on provided literature data. The analysis identifies for instance the most cited articles and authors, most common keywords, and journals with most publications. These statistics are accompanied with visualizations for a quick data overview. Additionally the system extracts the citation network data from the literature. Having an access to a citation network enables calculating how many times each reference has been cited by a paper inside the analyzed dataset. This feature is useful for identifying influential sources within a field of science, especially because it finds often-cited papers and books not listed in the primary literature dataset.

In addition to providing an exploratory analysis report, the system extracts and exports data about citation and author cooperation networks that can be visualized e.g. with the Gephi [4] open graph visualization platform. This dataset of citation connections can be used to calculate the relative influence of publications in the network, for example using the eigenvector centrality analysis. Eigenvector is a measure of node centrality, which can be applied to identify nodes that play central roles in the network structure. It can be seen as a weighted sum of not only direct connections, but indirect connections of every length [5]. Compared to simpler geometrical measures like degree centrality (i.e. total number of citations), eigenvector centrality also considers the influence of the connected nodes and takes into account the entire pattern of the graph. Where degree centrality gives a simple count of number of connections a node has, eigenvector centrality assigns higher values to connections to higher-ranking nodes [13]. For example, with this calculation method a node with few high-ranking connections might outrank a node with a larger number of low-ranking connections.

Analysis Case Study

For the purpose of demonstration a sample dataset was retrieved from the Thomson Reuters Web of Science with the search term of "computer supported collaborative learning" and year limit of 1990-2015. 1806 records were stored and processed with the analysis system. The processing and output rendering phase took 32 seconds on a dual core 2GHz Xeon test server. The entire analysis process from database search, analysis server upload and result download took three and half minutes.

The keyword summary section from the exploratory data analysis is displayed in the Figure 2 as a sample feature. It allows one to get an overview what are other common research topics in the dataset. In the sample dataset it can be seen that distance education and higher learning are current research topics in computer-supported collaborative learning.

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/event.cfm?id=RE248 by late 2015. Please use the official version of the paper and publication reference when citing: Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., Porras, J., 2015. Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies. CompSysTech 2015.

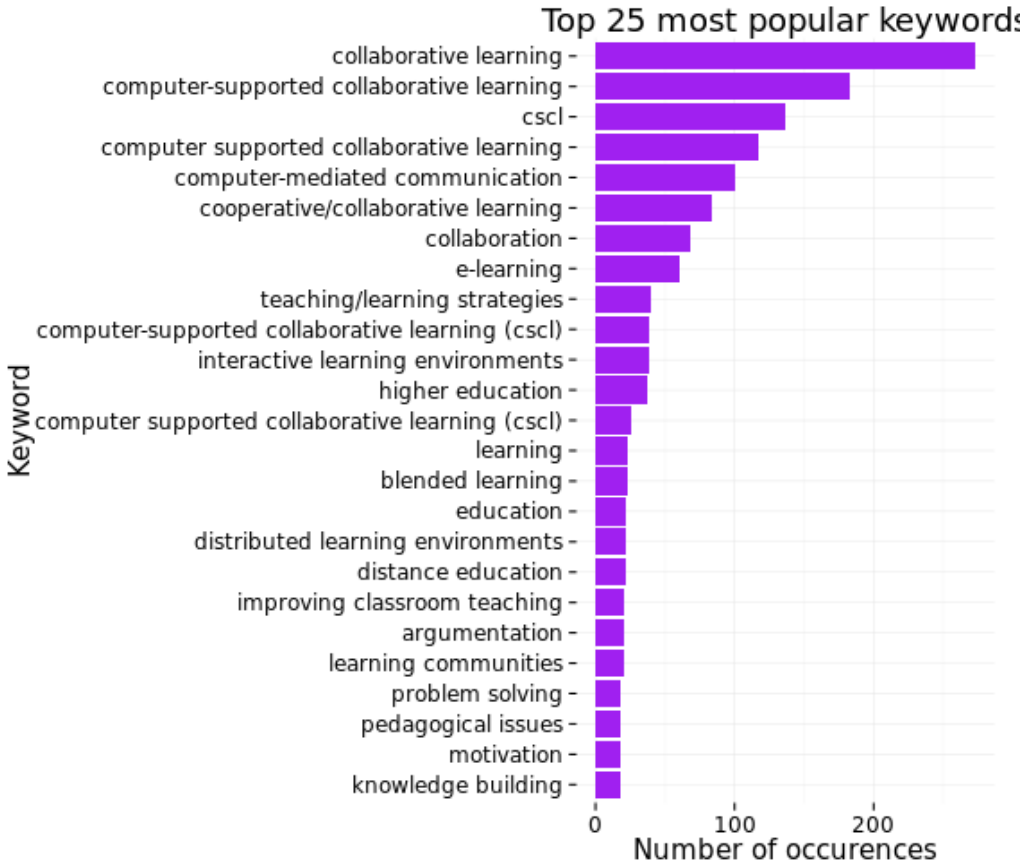

Figure 2. Most commonly occurring keywords in the example dataset

Another result from exploratory data analysis is the publication citation counts. An example result for citation counts from the sample dataset is displayed in the Table 1. Another measure of relevance is centrality values, displayed in the rightmost column (eigenvector), which discussed further in the next paragraph.

Table 1. Four most cited articles in the example dataset

Author	Year	Journal	Identifier	Citations	Eigenvector
KREIJNS, K.	2003	COMPUT. HUM. BEHAV.	10.1016/S0747-5632(02)00057-2	274	0,622
ALAVI, M.	1994	MIS. QUART.	10.2307/249763	238	0,005
STAHL, G.	2006	CAMB. HANDB. PSYCHOL.	V, P409	220	0,243
DE WEVER, WB.	2006	COMPUT. & EDUC.	10.1016/J.COMPEDU.2005.04.005	165	0,284

Additionally, we applied an influence analysis to the network using eigenvector centrality measure. In the Figure 3 we present a visualization of the network analysis results using the Gephi [4] visualization software. In the graph the size and shade denote node, or article, influence, with darkest and largest nodes being the most influential. Because of size limitations we display only the 250 most influential nodes according to the eigenvector centrality analysis results. Additionally, we marked the most referred article in

black (node D) and the three most central articles with white and light gray (nodes A, B, C). In the figure the benefits of centrality analysis become apparent. The literature review article by Kreijns (node D) has most citations, but is not as central to the field of science as for example the three other nodes, which discuss fundamental issues of CSCL and are commonly cited by other influential articles in the dataset.

The value of centrality analysis is highlighted by the results from the sample dataset. Basic citation count would not have highlighted the fundamental articles, because the citations are more diffused among several valued papers, but literature review articles are more rare and often cited, despite not bibliographically interacting as much with the field of science.

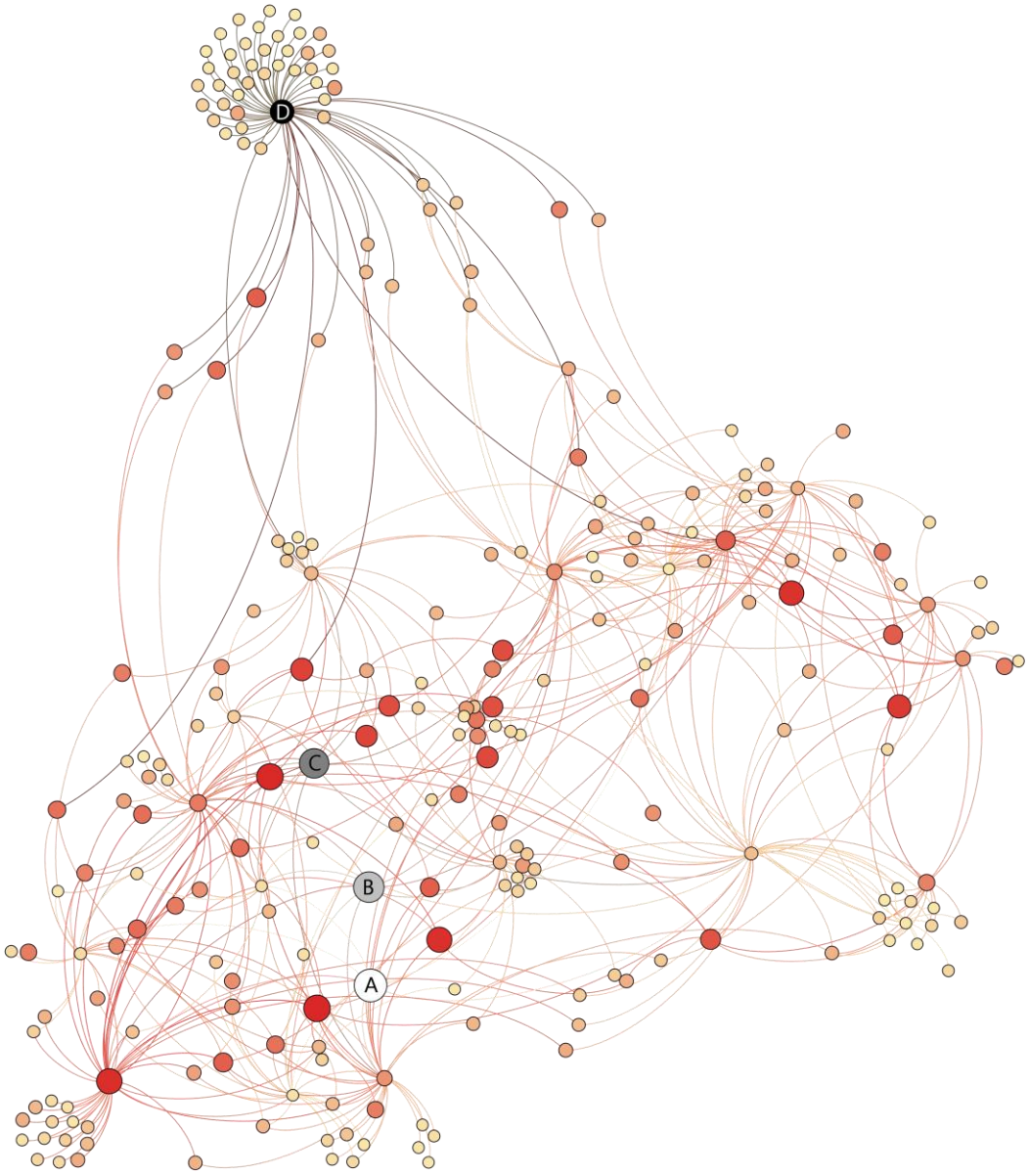

Figure 3. Visualization of a social network analysis results from the example dataset

DISCUSSION AND CONCLUSION

The exploratory analysis figures and network analysis results can help a researcher to get a quick visual overview and then deeper insight into the investigated field of science. By having the data automatically analyzed and visualized, the researcher can identify the

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/event.cfm?id=RE248 by late 2015. Please use the official version of the paper and publication reference when citing: Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., Porras, J., 2015. Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies. CompSysTech 2015.

core publications, publication trends, common research themes and the direction of latest research. With centrality analysis the core publications can be identified more reliably than by measuring the total number of citations, because it takes into account citation weighing and considers the centrality of citing articles.

In multidisciplinary fields network analysis enables identifying the interplay of citations and the contributions from other disciplines. This enables the researcher to see how the multidisciplinary nature of the field has formed and from which papers. Another opportunity that emerges by adding the dimension of time to the network, which visualizes the evolution of the publication network. This feature illustrates how the literature in a given field of science came to exist over time and allows one to identify publication forums and influential publications at different periods.

The research question of this paper was how to make systematic mapping studies more straightforward and accessible for researchers. We presented an open, extensible cloud-based literature analysis architecture as a solution and an implementation of that architecture. The presented tool, *NAILS*, allows the user to get a statistical and network overview of bibliographical datasets by uploading it to the cloud-based analysis service. The service uses an open source, extensible plugin architecture, which can serve as a platform for researchers who implement additional analysis features.

The sample implementation is a basic version and could benefit from additional features. The largest limitation is that data import works now only with Web of Science input data and thus cannot process articles not included in that database. The second limitation is that while the system analyses both statistical and citation network data, at the moment it visualizes only statistical data, requiring an user-installed software package for network visualization and the display of centrality values. Future work will include adding these features using the open plugin architecture and including additional major data sources, like Scopus. Additionally, having automatic import tools as browser plugins for different scientific datasets would make initiating analysis jobs even easier for researchers, but automatic downloads would also involve complicated copyright issues.

REFERENCES

[1] Abraham, A. and Hassanien, A.E. 2010. Computational Social Network Analysis: Trends, Tools and Research Advances. Springer.

[2] De Almeida Biolchini, J.C. et al. 2007. Scientific research ontology to support systematic review in software engineering. *Advanced Engineering Informatics*. 21, 2 (Apr. 2007), 133–151.

[3] Bailey, J. et al. 2007. Evidence relating to Object-Oriented software design: A survey. *ESEM* (2007), 482–484.

[4] Bastian, M. et al. 2009. Gephi: An open source software for exploring and manipulating networks. *International AAAI conference on weblogs and social media* (2009).

[5] Bonacich, P. 2007. Some unique properties of eigenvector centrality. *Social Networks*. 29, 4 (Oct. 2007), 555–564.

[6] Börner, K. 2011. Science of Science Studies: Sci2 Tool. *Communications of the ACM*. 54, 3 (2011), 60–69.

[7] Borrego, M. et al. 2014. Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields. *Journal of Engineering Education*. 103, 1 (Jan. 2014), 45–76.

[8] Van Eck, N.J. and Waltman, L. 2014. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*. 8, 4 (Oct. 2014), 802–823.

This is a pre-print version of an article. The actual version will be published in ACM DL at http://dl.acm.org/event.cfm?id=RE248 by late 2015. Please use the official version of the paper and publication reference when citing: Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., Porras, J., 2015. Cloud-Based Bibliometric Analysis Service for Systematic Mapping Studies. CompSysTech 2015.

[9] Van Eck, N.J. and Waltman, L. 2014. Visualizing bibliometric networks. *Measuring Scholarly Impact*. Springer. 285–320.

[10] Kitchenham, B. et al. 2009. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*. 51, 1 (Jan. 2009), 7–15.

[11] Kitchenham, B.A. and Charters, S. 2007. *Guidelines for performing systematic literature reviews in software engineering*. Technical Report #EBSE-2007-01. Department of Computer Science, University of Durham.

[12] Knoke, D. et al. 2008. *Social network analysis*. Sage Publications Los Angeles, CA.

[13] Newman, M.E. 2008. The mathematics of networks. *The new palgrave encyclopedia of economics*. 2, (2008), 1–12.

[14] Otte, E. and Rousseau, R. 2002. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*. 28, 6 (Dec. 2002), 441–453.

[15] Parolo, B. et al. 2015. Attention decay in science. *Available at SSRN 2575225*. (2015).

[16] Perianes-Rodríguez, A. et al. 2010. Detecting, identifying and visualizing research groups in co-authorship networks. *Scientometrics*. 82, 2 (2010), 307–319.

[17] Petersen, K. et al. 2008. Systematic mapping studies in software engineering. *12th International Conference on Evaluation and Assessment in Software Engineering* (2008), 1.

ABOUT THE AUTHORS

Antti Knutas, M.Sc., Lappeenranta University of Technology, Phone: +358-0294-462-111, E-mail: antti.knutas@lut.fi.

Arash Hajikhani, M.Sc., Lappeenranta University of Technology, Phone: +358-0294-462-111, E-mail: arash.hajikhani@lut.fi.

Juho Salminen, M.Sc., Lappeenranta University of Technology, Phone: +358-0294-462-111, E-mail: juho.salminen@lut.fi.

Associate Professor Jouni Ikonen, D.Sc., Lappeenranta University of Technology, Phone: +358-0294-462-111, E-mail: jouni.ikonen@lut.fi.

Professor Jari Porras, D.Sc., Lappeenranta University of Technology, Phone: +358-0294-462-111, E-mail: jari.porras@lut.fi.