30th Annual INCOSE international symposium
Cape Town, South Africa
July 18 - 23, 2020

A Bibliometric Method for Analysis of Systems Engineering Research

Dr. Rudolph Oosthuizen
Department of Engineering and Technology Management, University of Pretoria and CSIR
South Africa
+27 82 733 6355
roosthuizen@csir.co.za

Prof. Leon Pretorius
Department of Engineering and Technology Management, University of Pretoria
South Africa
+27 83 625 1756
leon.pretorius@up.ac.za

Abstract. Since systems engineering is still a relatively young and growing discipline, it requires a periodic analysis, taking into account past research to derive the requirements for future growth. Published research provides a good indication on the progress and maturity of a scientific discipline. Bibliometric analysis is a valuable tool to assess published research. This paper establishes a method to determine the main research topics published in the Systems Engineering Journal (INCOSE) since its inception. The research and associated analysis method applies Natural Language Processing with Topic Modelling to extract the main topics from the abstracts of all the papers published in the journal. The analyzed data provides the trends in topic coverage over time.

Introduction

The International Council on Systems Engineering (INCOSE) Handbook defines Systems Engineering as "… an interdisciplinary approach and means to enable the realization of successful systems ..." (Walden et al. 2015). Systems engineering is an approach that considers the whole system as distinct from its parts. It employs processes that define customer needs and requirements before designing and validating a solution system. The solution also has to consider system cost, manufacturing, operation, training, performance, support, and disposal. The systems engineering effort integrates multiple specialty disciplines to proceed from a concept solution to production and to operation of the system (Walden et al. 2015).

Despite the general systems theory and systems thinking already existing, the field of modern systems engineering was recognized only after World War 2. An approach was required to support the development of weapon systems that continually increased in complexity. Since then, the concepts of systems design, analysis, and development evolved more formally into the discipline of "Systems Engineering". A knowledge base of systems engineering methodologies, tools, and techniques is formalized in a series of handbooks, standards, and other guides that provide a foundation for the discipline (Brill 1998).

However, systems engineering is still young and growing when compared to other engineering fields. Systems engineering became a formally defined discipline in the early 1990s (Valerdi & Davidz 2009). Systems engineering also has to continue developing processes and tools, supported by theories, to cope with the continued growth in complexity of engineering projects (Sage 1998).

The development of systems engineering depends on effective research. Research aims to create knowledge by understanding, explaining, and predicting phenomena (Lakatos 1978, Kuhn 1996). According to Popper (1972), successively rejecting falsified theories achieves progress in science. However, Kuhn (1962) proposed a softer approach of negotiation, compromise, and conceptual

growth to achieve scientific progress without a strict falsification approach. Valerdi & Davidz (2009) agree that the development of knowledge in a scientific discipline involves learning through empirical observation, formulation of theories, and experimenting to test theories.

The typical starting point for a research project in systems engineering is a need for improvement, often triggered by an industrial problem (Muller 2013). Systems engineering depends on empirical research to develop valid and proven theories. Even if the research is not conducted under clinical controlled conditions, evaluation of an artefact's success may still unravel the main causes and effects (Brown 2009, Patankar 2014, Valerdi et al 2010, Muller 2013).

Valerdi and Davidz (2009) warn that the current research tendency of intuition and revelation, instead of a rigorous scientific process may hamper the growth of systems engineering. The current research agenda seems to focus on convincing the audience of the field's potential rather than on building intellectually sound principles. Due to limits on resources and time, researchers should rather focus on long-term intellectual establishment of the field through well-defined and relevant research projects (Fitzgerald & Adam 2000, Ferris 2009, Khalid 2013, Caillaud 2016).

Antons et al. (2016) found that evaluating trends the topic landscape of a heterogeneous research field over time helps researchers to detect meaningful research opportunities. They demonstrated that analysis of the dynamic evolution of research topics of a heterogeneous over time may provide useful guidance for planning and managing future research. The rise and fall of research topic prominence may guide investment decisions and identification of new topics and concepts (Lamba & Madhusudhan 2019).

Due to the continuous growth in academic paper publication numbers, reading and sorting them into topics and to derive bibliometric data is tedious and difficult. In the past, papers were manually assigned to a predetermined topic list, based on subjective judgment of the authors and subject matter experts. However, manual approaches may miss the latent topics from large text corpus and is costly in terms of time and resources. Also, the set of predetermined categories may ignore new and emerging topics while some articles may contain two or more topics. Eker et al. (2019) propose that text-mining techniques can be applied to analyze published paper abstracts to extract the underlying topics. Implementing automated machine learning methods will improve this process (Lee & Kang 2018).

This paper therefore presents and demonstrates a methodology to examine the historic trends in systems engineering research to identify opportunities for future projects. Firstly, bibliometrics is presented as a useful and valid method to trace the development in a research field. Thereafter, data mining and Natural Language Processing (NLP) are held as modelling and analysis tools for generating data on the evolution of systems engineering research topics. The NLP techniques, with topic modelling, are then applied to demonstrate the extraction of the main research topics from abstracts of the Systems Engineering Journal. The outputs of the analysis are presented to highlight the utility of the research analysis method. Lastly, some research trends and conclusions are presented.

Bibliometrics

Scientometrics, the "science of science", provides quantitative and statistical techniques to measure the progress in the development in a research field through analysis of published literature. Scientists tend to codify their findings in publications, which are the building blocks of a scientific field. Peer reviews and expert-based judgment validate published research. A substantial increase of basic and applicable knowledge (tacit or codified) indicates scientific progress. Scientometric methods to analyze research trends are increasing in popularity (Hood & Wilson 2001, van Raan 2005).

Bibliometrics is a common systematic analysis approach to scientometrics. The bibliometric approach is suited for most behavioral, engineering, and scientific fields. Bibliometric approaches rely

on the assumption that scientific research produces knowledge, captured in published scientific literature. The mapping of bibliometric outputs visualizes patterns in scientific data to improve the knowledge of researchers on a specific domain (Jie et al. 2014, Jiang et al. 2016). Many scientific fields already apply the quantitative evaluation of publication and citation data to evaluate leading authors, growth, performance, trends, maturity, and intellectual mapping (van Raan 2005, Aria & Cuccurullo 2017, Kalantari et al. 2017).

Keathley et al. (2015) highlight that bibliometric analysis is applicable to a specific journal as well as a set of journals to assess a research field. Performance analysis (citations) or science mapping (conceptual structure) can describe progress in a research. The main bibliometric indicators include the following (van Raan 2005, Cobo et al 2014, You et al 2014, Jiang et al. 2016, Aria & Cuccurullo 2017, Kalantari et al. 2017, Jia et al. 2018, Kotsemir 2019, Eker et al. 2019):

- The number of publications per research field is an indication of growth.
- The number of authors is an indicator of the size of a research field. Associated measures include number of authors per paper, papers per author, affiliation, and country.
- Citations provide a measure of the impact and quality of an article or author. Highly cited papers have more visibility and attract attention from other researchers. However, citation analysis may be incomplete or biased as authors tend to cite work from their own community. The number of citations may also only peak in the 3rd or 4th year after publication.
- Social network analysis studies the relationships between authors and between publications. Relationships can be derived from keywords, industries, affiliations, countries, and authors. The outputs include co-publication, co-citations, and co-occurrence of words.
- Text-mining techniques, including topic modelling, analyze paper titles, abstracts, and sometimes the full content, to identify the main topics and calculate the similarity between publications. Keyword analysis algorithms list the most frequent words in a dataset to indicate research trends. Analysis of the keywords identify hotspots and key points of research.

Topic Modelling is the focus of this paper. Keyword analysis is also applied as a comparison to the topic modeling.

Natural Language Processing

Most of the data available on the internet and other databases are unstructured free text, which is difficult for machines to process and analyze with software tools. Text mining is the process of structuring unstructured text and deriving information. Natural Language Processing (NLP) helps to convert input text into numerical values for machine learning algorithms to process. The outputs are key phrases, relationships, and patterns for evaluation and interpretation. NLP is applicable to text categorization, keyword extraction, text clustering, document classification, document summarization, language translation, speech recognition etc. (Banu & Chitra 2015, Agrawal et al. 2018).

Lately, text mining approaches, such as topic modelling, are becoming more approachable to researchers due to the processing power of hardware and accessibility of software. Topic modelling is an unsupervised text classification method that extract semantic information from text, using quantitative statistical algorithms without prior understanding. The mixture of words in a document constitutes a set of latent topics. NLP automatically analyses the occurrence and hidden relationships of these words in the documents to define each topic as a probability distribution (Cunningham & Kwakkel 2016, Jiang et al. 2016, Tong & Zhang 2016, Kunc et al. 2018, Agrawal et al. 2018).

Latent Dirichlet Allocation (LDA) provides a generative statistical model where unobserved groups explain the similarity of data. LDA is popular and proven in scientometric research of a discipline to discover semantic latent topics and structures. The model learns the distributions of topics in each document with their associated word probabilities (Jiang et al. 2016, Suominen & Toivanen 2016). The researcher has to define the number of topics to be extracted. The LDA algorithm then probabilistically forms document-topic and topic-word pairs. Each publication is then associated with a topic by the highest probability (Tong & Zhang 2016, Agrawal et al. 2018, Eker et al. 2019). Topic modeling is useful to synthesize and structure substantial bodies of literature to locate relevant publications. The outputs of topic modelling assist in the following (Antons et al. 2016):

- Assist educators in designing courses or curriculums in a field. Listing of articles associated with each topic also helps to recommend a reading list.
- Positioning of researchers in an increasingly competitive research field by identifying emerging opportunities for differentiation.
- Enable academic leaders to identifying a group of prominent researchers in a field or topic.
- Conference organizers can identify focal themes to differentiate their conference. Also, automatically clustering of submissions by topic enables developing the program.
- Editors can compare editorial priorities to the topics to assist development of a journal.
- Systems engineering practitioners may apply topic mapping to identify experts for specific innovative topics.

Method

Research Questions

The Research Questions (RQ) considered in this paper in support of developing the bibliometric method for analysis of systems engineering research are the following:

- RQ1 - What are the prominent topics published in the Systems Engineering journal since its inception?
- RQ2 - What are the temporal trends in relative prominence of the topics over this period?
- RQ3 - What is the trend of keywords over this period and does it correlate with the topics?

The answers to these research questions may contribute to providing information on the trends in systems engineering research. However, the aim is to demonstrate the utility of the bibliometric analysis method using NLP and topic modelling. This will serve as an input to further systematic research on the relevant research topics identified. Understanding how systems engineering has evolved due to research up to now, can contribute to managing and coordinating further scientific progress of the field.

Systems Engineering Journal

INCOSE aims to promote the development of systems engineering knowledge, establishing standards, and improve the professional status of systems engineering. “Systems Engineering”, INCOSE's scholarly journal, is a primary source of published peer reviewed systems engineering research. The journal publishes systems engineering research articles on various topics. The articles provide information on technologies, processes, and systems management approaches for systems engineering activities. The goals of the journal are integration and dissemination of systems engineering

knowledge, promotion of collaboration between major role-players and supporting development of appropriate professional standards (Sage 1998). The readership for the journal includes systems engineers, system programmers, and computer system developers (Sage 1998). Since 1998 it produced at least four issues per year. The current impact factor of the journal is 0,848. This journal is therefore deemed a natural starting point for bibliometric analysis of systems engineering, as identified in the research questions of the previous section.

For the purpose of this paper, the development the bibliometric method for analysis of systems engineering research will be demonstrated with an analysis of the Systems Engineering Journal publications. Although many other publications exist, including ASME, JCISE, JMD, MDPI Systems, and IEEE Systems, the Systems Engineering Journal will be used in this demonstration as it provides a controlled, smaller, and focused sample of papers. The other publications may contain topics outside the scope of systems engineering research.

Research Process

The process applied in this paper, as seen in Figure 1, entails inter alia the analysis of abstracts from all research papers in the Systems Engineering Journal with bibliometric and NLP tools. The whole research process is implemented using data science principles executed by Python based algorithms, which were adapted and developed as part of the research presented in this paper. The full details of the algorithms (software code) are however not part of the aim of this paper. An overview of the analysis method and associated tools used is provided in the following subsections.

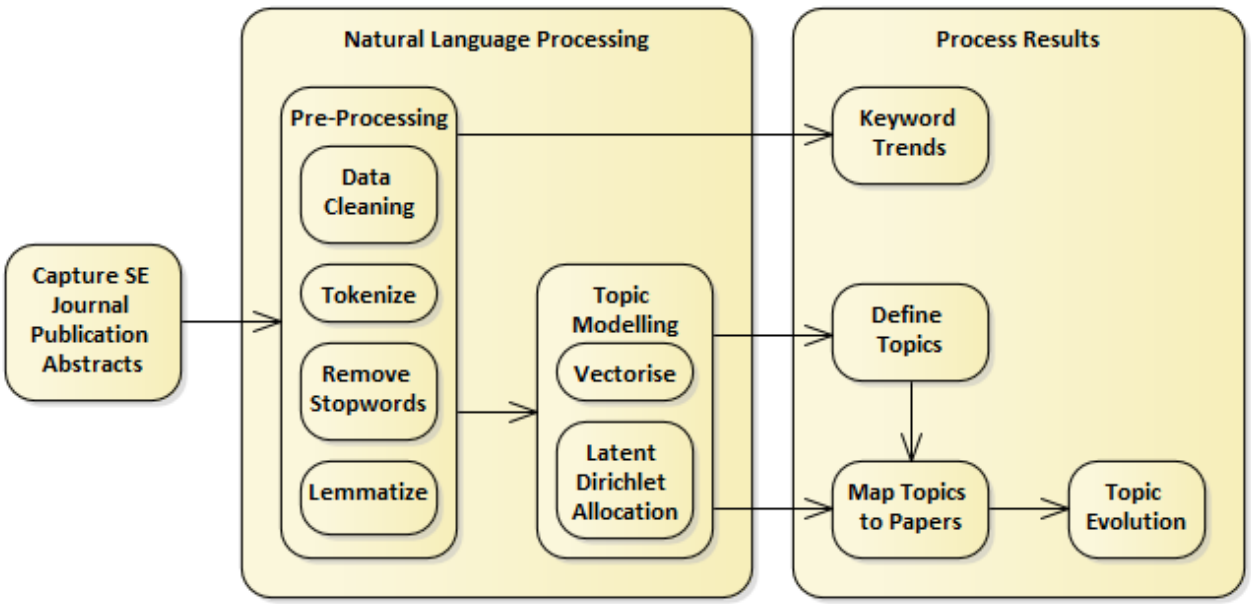

Figure 1. Bibliometric Analysis Process

Capture Systems Engineering Journal Publication Abstracts. Normally a NLP application starts with a collection of documents. The Pyblometrics library in Python was used to extract bibliometric information on the Systems Engineering Journal from Scopus. Rose & Kitchin (2019) developed Pybliometrics to ease the use of Scopus data. Pybliometrics provides a simple and consistent interface that can integrate with Python's data science ecosystem for Machine Learning and visualization tools. The application only requires the International Standard Serial Number (ISSN) of the journal to access all the required bibliometric data. The captured data for this research consists of the following fields:

- Year of Publication.
- Authors.

- Paper title.
- Abstract.
- Author Affiliations.
- Author Country.
- Author Count.
- Citations.

The data was exported to a Comma-Separated Values (CSV) file for further processing. A number of the earlier issues could not be accessed on Scopus, which necessitated the manual population of the missing bibliometric data from the journals' website. After screening the results for removing non-research papers, such as errata, correspondence, and editorial papers, 623 regular papers remained for analysis. These papers are not listed in the references, only those used to support the argumentation in this paper.

Natural Language Processing. Pre-processing transforms and structures the data (unstructured text) into a suitable format for analysis. The Spacy library in Python performs the NLP. The Scikit-learn library implements the topic modelling with a LDA function to provide the list of topics with their description classifying keywords. The detailed steps are the following:

- Data Cleaning. The abstracts needed cleaning before the actual pre-processing activities could commence. The text contained punctuations, numbers, capitalized letters, special characters, and superfluous words such as "copyright" and the journal reference. In addition, the text was converted to lower case for easier NLP.

- Tokenize. The document is treated as a string to partition the text into a list of tokens. Tokenization extracts the linguistic units (words or identifiers) that form building blocks for sentences or paragraphs, using the spaces in between. If required, the algorithm can extract phrases which have strong meaning independent of the individual words, for example "systems engineering" (Lin et al. 2016).

- Remove Stopwords. The stopwords not adding meaning to the text are removed. These words are common in a language and tend to have the maximum frequency that add noise when applying machine learning to the text. Typical stopwords are as "and", "the", "if", "a", "this", etc. There is no single universal list of stop words, as it may differ between research fields (Patel & Soni 2012). In this, case additional words removed before the topic modelling include terms such as "article" and "Paper".

- Lemmatization. Stemming linguistically normalizes text by reducing words to their root form (base or stem) by truncating the derivational affixes. This is applied to an individual word without considering the context. Lemmatization, on the other hand, extracts the word base for a lemma that is more linguistically correct. It applies a morphological analysis and a vocabulary to transform the root word. Lemmas are more lexicographically correct than a root stem. Stemmers, on the other hand, are usually easier to implement and run faster than lemmatization. Some applications may also be insensitive to the reduced "accuracy" (Patel & Soni 2012, Smith 2018, Eker et al. 2019).

- Vectorize. Python's Scikit-learn CountVectoriser transforms all the journal abstract terms into a Data Term Matrix. The two key variable to tune the CountVectoriser are the maximum and minimum document frequencies (max_df and min_df). These parameters remove terms that are too rare or too common to improve the processing. The optimum set of parameters of

the word vectorizer and LDA is required for valid topic modelling. Cross-validation using perplexity measures of the ability of a probabilistic model to predict a sample. Perplexity is a measure to indicate how well a probability model predicts a sample. Perplexity is the inverse of the geometric mean per-word likelihood. Therefore, a lower perplexity indicates the better LDA model (Jiang et al. 2016, Agrawal et al. 2018, Kunc et al. 2018). The most suited max_df and min_df would result in the lowest "perplexity", as seen from the perplexity plot in Figure 2. For this paper, max_df is set at 0.9 as it includes enough terms. A min_df of 0.2 would exclude the least amount of terms while the perplexity has a low enough value.

- Latent Dirichlet Allocation. The first step in the current NLP analysis is to determine the optimum number of topics for extraction. According to the perplexity mapping, nine topics with the selected max_df and min_df values would also provide the best model for this current research, as seen in Figure 3.

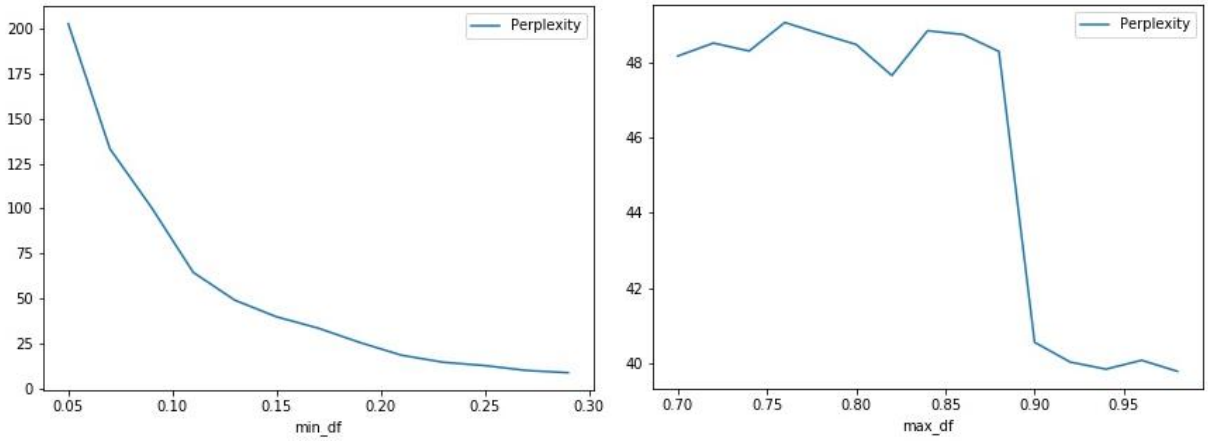

Figure 2. Perplexity Scores for Max_df and Min_df

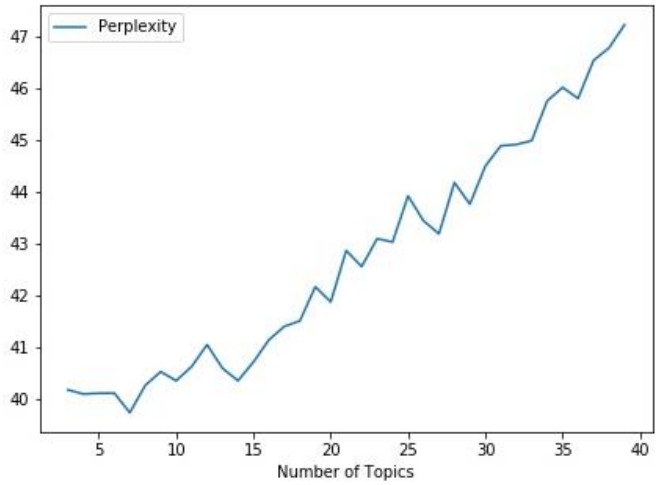

Figure 3. Perplexity Scores for Topic Numbers

Process Results. The output of the LDA is the nine topics as described by the most prominent keywords per topic. These results are processed further in Python and Excel to produce the required results:

- Map Topics to Papers. The main topic identified per paper is appended as column to the output table of the initial journal scraping using Python. This table is the source of information from the results processing.

- Define Topics. The LDA algorithm can only cluster documents by their topics, without an indication of what they are. Manual analysis still has to interpret the NLP results and assign a dominant topic name to each cluster. Using expert domain knowledge and systems engineering insight, the list of words describing each topic is manually converted into a descriptive topic title. This is a subjective process. This is tested by manually comparing (spot-check) a sample of paper topics to the titles of the papers.

- Topic Evolution. An important aspect of reviewing published literature is to analyze the topic evolution patterns in the research field over time. The patterns can uncover research trends to guide future studies. The number of papers per topic per year are calculated in Python. As the total number of papers published per year were not constant, this was normalized by dividing the results by the total papers published per year. The time trends of the research topics were calculated using a fourth order binomial regression using the Scikit-learn library in Python (Antons et al. 2016, Lee & Kang 2018). The outputs are plotted over the publication years to investigate possible trends.

- Keyword Trends. The trends and priority of keywords in the abstracts over time without considering obvious words such as "Systems Engineering", "Paper", etc. is calculated in Python. The keyword trend should follow the topic trend, therefore this will also serve as a validation of the unsupervised machine learning based LDA.

Results

The results of applying the method described above on the Systems Engineering Journal are discussed in this section. The result for RQ1 is a list of topics with their descriptive keywords listed in Table 1. For context the number of papers per topic was added. It is clear that up to now, most of the research have covered aspects on the “Systems Engineering Process”.

Table 1: Topic Identification

Topic	Name	Description	Number of Papers
1	Systems Engineering Process	system, engineering, system engineering, engineer, process, development, identify, use, new, need, apply, provide	151
2	Systems Management	management, cost, development, new, approach, method, application, identify, develop, case, performance, base	35
3	Decision Support	decision, method, support, application, provide, develop, approach, case, apply, base, new, system	29
4	Requirements Engineering	requirement, product, development, need, method, develop, process, use, system, case, support, engineering	59
5	Product Development	process, model, development, provide, product, describe, support, system, approach, framework, develop, identify	31
6	Complex Systems Engineering	system, complex, approach, performance, engineer, base, support, need, develop, provide, model, apply	134
7	Systems Architecting	architecture, system, framework, provide, base, development, case, describe, performance, develop, need, complex	65

8	Systems Modelling	model, base, use, approach, system, case, provide, describe, cost, apply, develop, method	59
9	Systems Design	design, system, approach, performance, base, provide, requirement, method, use, new, engineering, case	59

The most prominent keywords were analyzed to allocate topic names. In some of the cases, the values associated to the top 2 or 3 keywords are overwhelming, negating the value of the other words. In many of the cases, this was obvious, but for a few cases the identification was more challenging (e.g. Systems Engineering Applications).

Squires et al. (2012) proposed a set of topics for the Systems Engineering Body of Knowledge (SEBoK) that comprise the discipline of systems engineering. The SEBoK focuses on the system engineering elements that has solid roots in theory and application as well as emerging topics from research and advances in technology and tools. These topics at that stage were emerging and developing in systems engineering to highlight some gaps in the SEBoK (Squires et al. 2012). With some minor adjustments in the wording of the topics, these can be mapped to the topics from Table 1 :

- SE Efficiency and Responsiveness (Systems Engineering Process).
- Architecture Reuse and Efficiency (Systems Architecting).
- Model-based SE (Systems Modelling).
- Complex systems and System-of-Systems Analysis (Complex Systems Engineering).
- System Affordability (Systems Management).
- SE Measurement Evolution (Decision Support).
- Total system solution perspective (Product Development and Systems Design).
- Advanced Methods of Understanding Stakeholder Needs (Requirements Engineering).

Although many more systems engineering topics (about 37) or sections in the INCOSE Handbook exist, these are the topics extracted from the journal. One reasons for this discrepancy may be that not all the systems engineering topics are being formally researched and published in the journal. The topic modelling may be performed for more topics with fine grain differences, but this will decrease the accuracy of the model. Also trying to identify and describe each topic will also be more difficult.

Approaches to counter these problems may include performing the topic modelling on the whole paper as well as comparing the outputs to a similar analysis on the INCOSE Handbook. However, these fall outside the scope of this paper and will be addressed in future work. Validity of the proposed method outputs can enhanced through careful preprocessing of the text, utilizing various experts in the field for topic labeling as well as post hoc and sensitivity analyses to check for robustness (Antons et al. 2016). Similar approaches have been applied to develop technology roadmaps (Letaba 2018).

The next step determines the evolution of the topics over journal's publication history for RQ2. Figure 4 shows the number of papers for the topic "Systems Engineering Process" published per year to show the difference between the normalized publications of papers per topic per year. It is clear that the trend line provides a better indication on the temporal popularity of the topic.

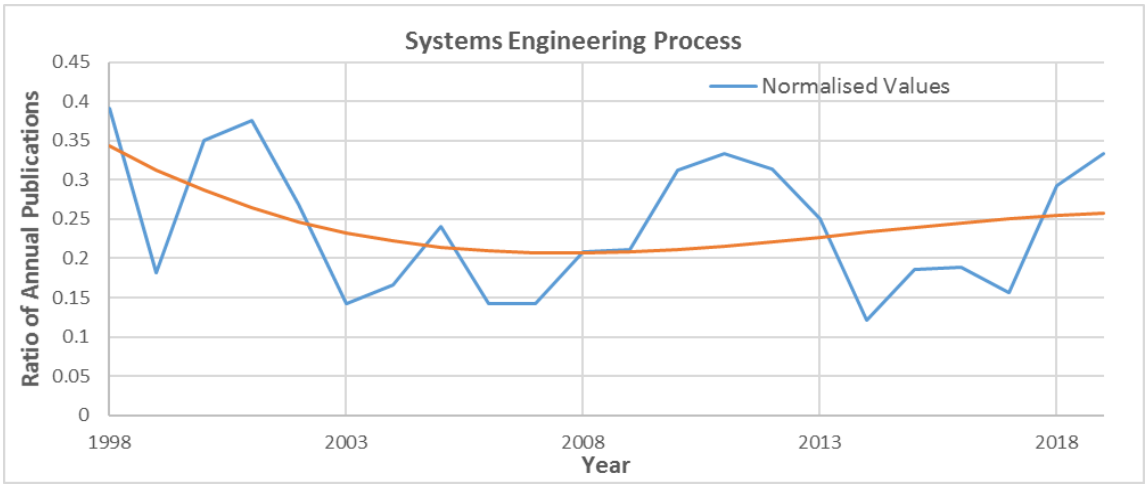

Figure 4. Evolution of Systems Engineering Process Topic

The topics with a strong rising trend is shown in Figure 5, while the remaining topics are in Figure 6 to enhance readability. However, the topics of Systems Modelling and Requirements have turned the corner and is on the way up. Noteworthy is the drastic fall of Complex Systems Engineering relative to the other topics since being the most popular topic at the beginning of the decade.

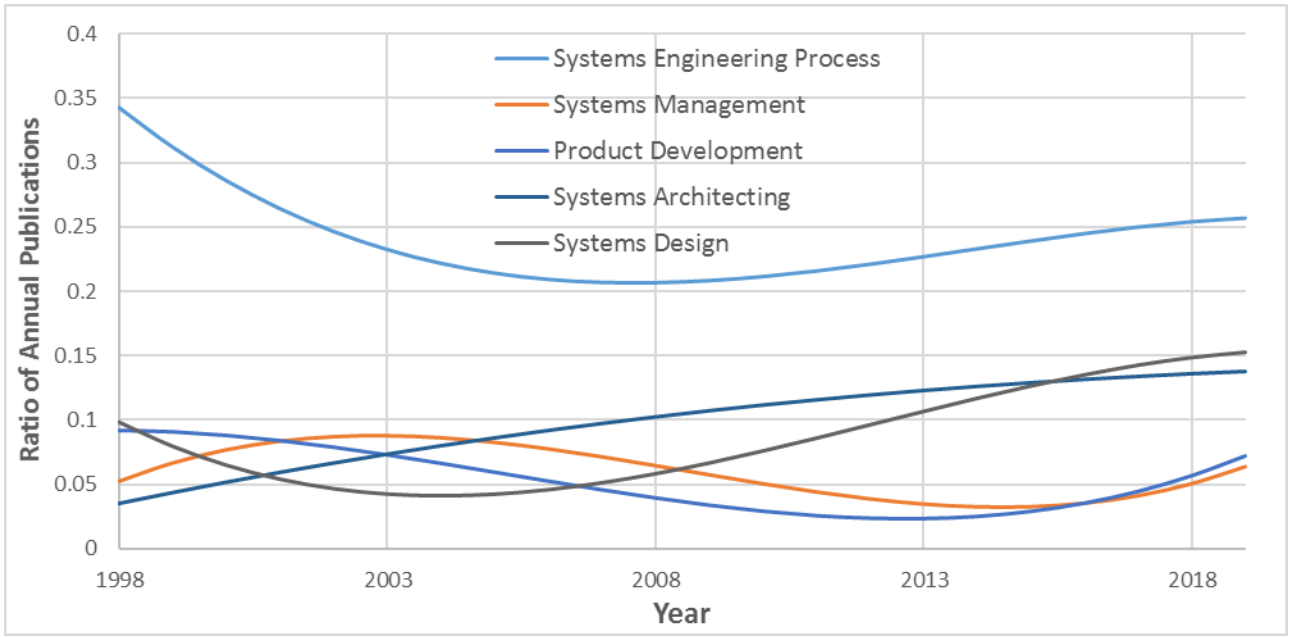

Figure 5. Systems Engineering Journal Topic Trends

The trends from Figures 5 and 6 can inform the planning of future research projects. Past growth can indicate the future potential of a research topic to reduce the risk for investment. A negative linear time trend identifies "cold" topics while positive trends indicate "hot". "Evergreen" topics have no significant time trends (relatively stable over time) above median number of publications. "Wall-flower" topics have no significant time trends below median number of publications (Antons et al. 2016).

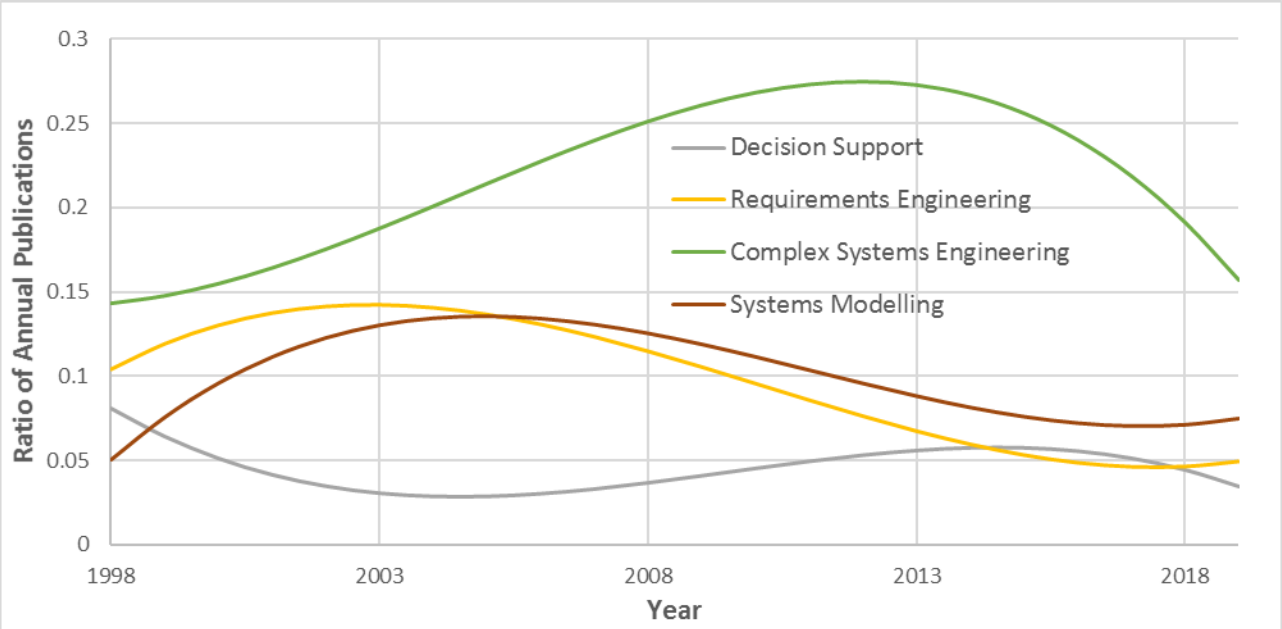

Figure 5. Systems Engineering Journal Topic Trends

Despite micro level inaccuracies possible due to unsupervised machine learning, the method still serves as a useful starting point and guide that helps identify topic dynamics and bridge opportunities by linking previously distinct topics (Antons et al. 2016):

- Hot topics tend to attract research interest. Researchers could strengthen their position in already strong topics by filling the blanks through introducing multiple levels of analysis.
- Hot topics with relatively small amounts of published articles is an indication of growing interest.
- Cold topics can be revived through integration with topics on a better development trajectory.
- Wall-flower topics can be energized by changing the perspective.

A by-product of NLP is the word frequency over the corpus of the text analyzed. Although topics provide richer results than keyword, they also present information and context of the research trends in the discipline to the. The LDA extracts a number of terms from the text corpus to define the topics. In this case 30 were extracted. Figure 6 shows the word frequency of these words in the abstracts of all the published Systems Engineering Journal papers to address RQ3. It is notable that the top keywords can be correlated with the identified topics. In this case, system and engineering are seen as common words and are removed. The word "System" has a count of 3010 and is not shown on the graph.

Method Threats and Shortcomings

The proposed method is not without its inherent limitations. The list of topics presented in this paper is not complete or exhaustive as is limited to topics with sufficient discriminant validity. It should be remembered, that topic modelling through LDA is unsupervised machine learning that may not produce perfect results in all of the cases. The assumption is that "micro"-level inaccuracies will be average out over a large sample corpus. It may not be possible to accurately allocate identify a topic for each paper. Therefore, it will be difficult to identify specific papers for detailed analysis in terms of their citations or authors.

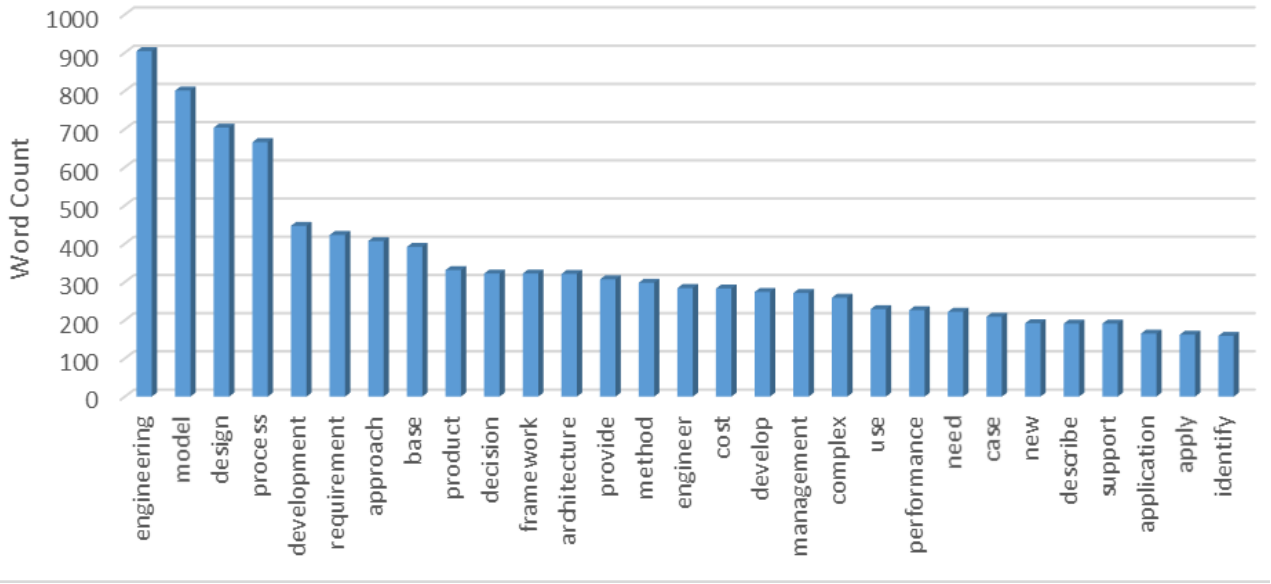

Figure 6. Systems Engineering Word Count

Antons et al. (2016) expressed that this is affected by the selected TF-IDF cutoff values. Some terms may not be sufficiently discriminatory, because they appear with similar frequency in many documents throughout the text corpus. As a result, the derived topics exclude the most topics common to most articles. Therefore, the analysis of the Systems Engineering Journal will exclude "Systems Engineering" as a specific topic. However, this is not a problem as we are interested in the lower level underlying topics. Also, this analysis remains descriptive and cannot isolate the specific factors that shed light on the temporal behavior of published research topics.

Syed and Spruit (2017) found that document frequency, document length, and vocabulary size have mixed effects on the extracted topic quality. Building LDA models from abstracts of a single journal with a lower number of paper tend have a higher frequency of noise terms. This requires careful cleaning and preparation of the data. Processing full-text data of these "smaller" document sets are less affected by these words. The topic distribution of a small dataset of abstracts provide a broad range of topics while full-text topics cause fine-grained results. However, utilizing large collections of document result in extracted models that are less affected by incorrect or noise terms from processing only abstracts. Here processing the full document has less of an effect.

Since this paper only focuses on the Systems Engineering Journal, the whole field may have a wider range of topics. Other journals publishing systems engineering research may have different topics or different behavior of similar topics. Future research will therefore identify other publication platforms for an inclusive analysis.

Conclusions and Future Work

This paper establishes and demonstrates a methodology for analysis of published systems engineering research to guide the planning of future research projects. NLP and machine learning were applied on a small and controlled sample of abstracts published over the history of the Systems Engineering Journal. The NLP and topic modelling extracted latent themes from the unstructured text. This provides bibliometric information to investigate the development of the discipline through published research. The mapping of research and publishing trends over time highlighted the prominence of modelling in systems engineering.

As most of the process is automated with established Python algorithms, it can be repeated further research into other systems engineering publications or related fields (e.g. different systems engineering specialities, engineering management and technology management). The size of the text

corpus can also be drastically increased by including othe publication platforms. Therefore, the method applied in this paper provides the data for an indepth investigation into published systems engineering research. The bibliometric data can be used to assess the growth and maturity of the field. This will assist in identifying key research opportunities to establish a research framework and roadmap. Despite the possible shortcommings of the method, it will be useful to detect patterns and trends accros a large corpus that will be too difficult or time consuming to process manually.

For future research on the method, as well as its application for analysis of systems engineering and related research, the corpus of publications can be expanded. Firstly, the focus will be on INCOSE sponsored publications such as the annual symposium proceedings, SEBoK and the INCOSE Handbook. The distribution and trends in topics needs to be established to further confirm the utility of the method. The titles and other information from the journals could also be added to increase the text corpus.

The next phase will be to incorporate the fringe systems engineering publications such as IEEE Systems, ASME JCISE, ASME, JMD, MDPI Systems etc. The method will be useful to find systems engineering research papers in these platforms that publish other fields as well. Topic modelling provides a richer search capability than only using keywords. Lastly, the extracted topics can be labelled to support analysis of other bibliometric aspects, such as citations, leading authors, and key papers.

References

Adam, F. and Fitzgerald, B., 2000. The status of the information systems field: historical perspective and practical orientation.

Agrawal, A., Fu, W. and Menzies, T., 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering. Information and Software Technology, 98, pp.74-88.

Antons, D., Kleer, R., & Salge, T. O., 2016. Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories. Journal of Product Innovation Management, 33(6), 726-749.

Aria, M. and Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), pp.959-975.

Banu, G.R. and Chitra, V.K., 2015. A Survey of Text Mining Concepts. International journal of innovations in engineering and technology, ISSN-2319-1058, 5(2).

Brill, J. H. 1998. "Systems Engineering—A Retrospective View." Systems Engineering 1 (4): 258-266.

Brown, S.F., 2009, April. Naivety in Systems Engineering Research: are we putting the methodological cart before the philosophical horse. In 7th Annual Conference on Systems Engineering Research (CSER 2009).

Caillaud, E., Rose, B. and Goepp, V., 2016. Research methodology for systems engineering: some recommendations. IFAC-PapersOnLine, 49(12), pp.1567-1572.

Cobo, M.J., Chiclana, F., Collop, A., de Ona, J. and Herrera-Viedma, E., 2014. A bibliometric analysis of the Intelligent transportation systems based on science mapping.

Cunningham, S.W., Kwakkel, J.H., 2016. Analytics and Tech Mining for Engineering Managers. Momentum Press.

Eker, S., Rovenskaya, E., Langan, S. and Obersteiner, M., 2019. Model validation: A bibliometric analysis of the literature. Environmental Modelling & Software, 117, pp.43-54.

Ferris, T.L., 2009. On Methods of Research for Systems Engineering (Doctoral dissertation, Loughborough Uni).

Hall, D.M., 2016. Applications of text analytics in the intelligence community. Naval Postgraduate School Monterey United States.

Hood, W. and Wilson, C., 2001. The literature of bibliometrics, scientometrics, and informetrics. Scientometrics, 52(2), pp.291-314.

Jia, Y., Wang, W., Liang, J., Liu, L., Chen, Z., Zhang, J., Chen, T. and Lei, J., 2018. Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences. Computer methods and programs in biomedicine, 166, pp.19-32.

Jiang, H., Qiang, M. and Lin, P., 2016. A topic modeling based bibliometric exploration of hydropower research. Renewable and Sustainable Energy Reviews, 57, pp.226-237.

Jie, L., Xiaohong, G., Shifei, S. and Jovanovic, A., 2014. Bibliometric Mapping of "International Symposium on Safety Science and Technology (1998-2012)". Procedia Engineering, 84, pp.70-79.

Kalantari, A., Kamsin, A., Kamaruddin, H.S., Ebrahim, N.A., Gani, A., Ebrahimi, A. and Shamshirband, S., 2017. A bibliometric approach to tracking big data research trends. Journal of Big Data, 4(1), p.30.

Keathley, H., Bean, A., Chen, T., Vila, K., Ye, K. and Gonzalez-Aleu, F., 2015. Bibliometric analysis of author collaboration in engineering management research. In Proceedings of the International Annual Conference of the American Society for Engineering Management. (p. 1). American Society for Engineering Management (ASEM).

Khalid, A., 2013. Systems Engineering Graduate Research as Part of Curriculum–Summary of Research. Procedia Computer Science, 16, pp.967-975.

Kuhn, T .S, 1962. The Structure of Scientific Revolutions. The University of Chicago Press, United States of America.

Kunc, M., Mortenson, M.J. and Vidgen, R., 2018. A computational literature review of the field of System Dynamics from 1974 to 2017. Journal of Simulation, 12(2), pp.115-127.

Lakatos, I., 1978. The Methodology of Scientific Research Programmes. Eds. Worral, J, and Currie, G, Cambridge, Cambridge University Press.

Lamba, M., & Madhusudhan, M., 2019. Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study. Scientometrics, 1-29.

Lee, H., & Kang, P., 2018. Identifying core topics in technology and innovation management studies: A topic model approach. The Journal of Technology Transfer, 43(5), 1291-1317.

Letaba, P., 2018. Complex Technology Roadmap Development in the Context of Developing Countries (Doctoral dissertation, University of Pretoria).

Muller, G., 2013. Systems engineering research methods. Procedia Computer Science, 16, pp.1092-1101.

Patankar, R.G., Parmar, M.P. and Patel, N.D., 2014. An Approach to choose Systems Engineering using Research Methodology. International Journal of Modern Trends in Engineering and Research, 2(3), pp.90-95.

Popper, K.R. 1972. The Logic of Scientific Discovery. Hutchinson, London.

Sage, A. P. 1998. "Systems engineering: Purpose, function, and structure". Systems Engineering, 1(1): 1-3.

Squires, A., Olwell, D., Roedler, G. and Ekstrom, J.J., 2012, July. Gaps in the body of knowledge of systems engineering. In INCOSE International Symposium (Vol. 22, No. 1, pp. 1967-1976).

Suominen, A. and Toivanen, H., 2016. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. Journal of the Association for Information Science and Technology, 67(10), pp.2464-2476.

Syed, S. and Spruit, M., 2017, October. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In 2017 IEEE International conference on data science and advanced analytics (DSAA) (pp. 165-174). IEEE.

Tong, Z. and Zhang, H., 2016. A Text Mining Research Based on LDA Topic Modelling. In International Conference on Computer Science, Engineering and Information Technology (pp. 201-210).

Valerdi, R., & Davidz, H. L. 2009. Empirical Research in Systems Engineering: Challenges and Opportunities of a New Frontier. Systems Engineering, vol. 12, no. 2, pp. 169-181.

Valerdi, R., Brown, S., & Muller, G. 2010. Towards a framework of research methodology choices in Systems Engineering. In 8th Annual Conference on Systems Engineering Research, Hoboken, NJ.

Van Raan, A.F., 2003. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. Technology Assessment-Theory and Practice, 1(12), pp.20-29.

Walden, D.D., Roedler, G.J., Forsberg, K.J., Hamelin, R.D., Shortell, T.M., 2015. INCOSE Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities. John Wiley & Sons. 4th edition.

You, G.R., Sun, X., Sun, M., Wang, J.M. and Chen, Y.W., 2014, June. Bibliometric and social network analysis of the sos field. In 2014 9th International Conference on System of Systems Engineering (SOSE) (pp. 13-18). IEEE.

Lin, J.R., Hu, Z.Z., Zhang, J.P. and Yu, F.Q., 2016. A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM. Computer-Aided Civil and Infrastructure Engineering, 31(1), pp.18-33.

Patel, F.N. and Soni, N.R., 2012. Text mining: A Brief survey. International Journal of Advanced Computer Research, 2(4), p.243.

Smith, J.R., 2018. The Application of Text Mining and Data Visualization Techniques to Textual Corpus Exploration (No. AFIT-ENS-MS-18-M-163). Air Force Institute of Technology Wright-Patterson AFB Wright-Patterson AFB United States.

Rose, M.E. and Kitchin, J.R., 2019. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. SoftwareX, 10, p.100263.