A TEXT MINING BASED MAP OF ENGINEERING DESIGN: TOPICS AND THEIR TRAJECTORIES OVER TIME

Chiarello, Filippo; Melluso, Nicola; Bonaccorsi, Andrea; Fantoni, Gualtiero

Università di Pisa

ABSTRACT

The Engineering Design field is growing fast and so is growing the number of sub-fields that are bringing value to researchers that are working in this context. From psychology to neurosciences, from mathematics to machine learning, everyday scholars and practitioners produce new knowledge of potential interest for designers.
This leads to complications in the researchers' aims who want to quickly and easily find literature on a specific topic among a large number of scientific publications or want to effectively position a new research.
In the present paper, we address this problem by using state of the art text mining techniques on a large corpus of Engineering Design related documents. In particular, a topic modelling technique is applied to all the papers published in the ICED proceedings from 2003 to 2017 (3,129 documents) in order to find the main subtopics of Engineering Design. Finally, we analyzed the trends of these topics over time, to give a bird-eye view of how the Engineering Design field is evolving.
The results offer a clear and bottom-up picture of what Engineering design is and how the interest of researchers in different topics has changed over time.

Keywords: Semantic data processing, Big data, Machine learning, Topic Modelling

Contact:
Chiarello, Filippo
Università di Pisa
Italy
filippochiarello.90@gmail.com

Cite this article: Chiarello, F., Melluso, N., Bonaccorsi, A., Fantoni, G. (2019) 'A Text Mining Based Map of Engineering Design: Topics and their Trajectories Over Time', in *Proceedings of the 22nd International Conference on Engineering Design (ICED19),* Delft, The Netherlands, 5-8 August 2019. DOI:10.1017/dsi.2019.283

1 INTRODUCTION

The continuous evolution of sciences and scientific terminologies precludes rigid classifications of knowledge fields. This is particularly true for Engineering Design (ED) (Bailey, 2009), that finds contributions from many different scientific fields (Engineering, Psychology and Economics to name a few) and that has been subject to an unprecedented growth in the last 15 years.
To solve this problem in other scientific fields researchers (Trivelli *et al*., 2019; Devyatkin *et al*., 2018) uses a well-known Text Mining (TM) Methods: Topic Modelling (Rehurek and Sojka, 2010).
Topic models are useful to rationalize large collections of documents by clustering them with an unsupervised approach. The result is a list of topics that are discussed inside documents and the words (or phrases) that are the most representative of each topic. The topics are derived bottom-up from the language itself, instead of relying on top-down taxonomies usually made by experts of the field. Such an approach has clear advantages: it mitigates the biases introduced by experts (even if introduces the one made by the analysts), it is cost and time effective (it relies on computational effort instead of human effort) and its output are quantitative, thus allowing further statistical analysis.

In the present paper, we use Topic Modelling in order to map ED knowledge field and identify salient development trajectories of topics over time. To reach this goal we analyse a corpus of papers (3,129 documents) presented in the last 8 editions of the largest and most authoritative conference in Engineering Design: International Conference on Engineering Design (ICED).
The result of this process is a list of 12 distinct topics discussed in the ED field and the corresponding most common words and phrases. Furthermore, thanks to the possibility to measure how much each topic is discussed inside each document, we painted the trajectories of the extracted topics over the last 15 years in order to identify patterns of rising, falling (or mixture of the two), and persistency of the topics.

2 STATE OF THE ART

2.1 Text Mining Techniques for Engineering Design

Scholars in ED adopt a variety of methodological perspectives (from established engineering disciplines to artificial intelligence, from ethnographic to simulation and operations research) and use a large array of qualitative and quantitative techniques. In recent years TM techniques has proven to be part of these, due to the possibility to:

1- extract information that are relevant for the design process but that are hidden in massive quantity of unstructured documents (Chiarello *et al*., 2017; Chiarello *et al*., 2018a)
2- exploit publicly available sources, thus helping in resolving the problem of the non-availability of data used in ED research (Paraguez and Mayer, 2017)

In the context of ED, the main problems that TM methods are helping to facing are in the context of Knowledge Management. In particular they can be classified as:

a) *Knowledge Extraction:* Automatically distillate design-relevant knowledge from unstructured sources. These sources are typically patents, surveys, interview transcripts or requirements documents (Jin *et al*., 2016)
b) *Knowledge Representation:* Schematize information about the world in a language that is comprehensible both for the computer and for the designer in order to facilitate the design process. The result is typically a knowledge base or an ontology of design concepts (Bolloju *et al*., 2012).
c) *Knowledge Deployment:* Facilitate or asses Design Education (Autrey *et al*., 2018).
d) *Knowledge Sharing:* Improve co-operative and distributed design environments. (Goonetillake *et al*., 2002)
e) *Knowledge Expansion:* Improve innovation and evaluating the creativity in design processes (Chiu and Shu, 2005)

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

The present work, and many others found in literature (Bonaccorsi and Fantoni, 2007; Li and Tate, 2010) finds its position in solving both *Knowledge Extraction* and *Knowledge Representation* problems.

2.2 Limitations of Expert-Based Scientific Knowledge Maps

The adoption of TM techniques, combined with domain knowledge, can be used to answer questions that just a decade ago would have been impossible to answer for field experts alone (Nuzzo *et al*., 2010; Dassisti *et al*., 2018). Furthermore, TM techniques represent a dramatic improvement for the task of Scientific Mapping with respect to the standard experts-based approach, because:

- expert based keyword definition is a very expensive activity (Tseng *et al*., 2007).
- experts rarely apply systematic rules (Noh *et al*., 2015).
- experts may be subject to a number of biases, such as the desirability bias (attributing higher probability of occurrence to preferred events) (Bonaccorsi *et al*., 2017).
- experts based methods are static, and a re-iteration in time of the process is costly (Chiarello *et al*., 2018b)

This last limitation is particularly evident for an ever-evolving field like ED. The easiness in replicability of the present also in the next years of the ICED conference or for other similar conferences (i.e. Design Computing and Cognition, CIRP Design or DESIGN) is future development of interest for our methodology.

3 METHODOLOGY

3.1 Paper Retrieval

The collection of documents published in the ICED proceedings is gathered from the Scopus API[1], one of the largest databases of peer-reviewed literature, that is well known to cover particularly well papers published in the field of engineering and technical sciences. We gathered data for all ICED conferences published in Scopus (from 2003 to 2017) and collected the abstracts and the year of each publication.

3.2 Pre-processing

The abstracts are collected as free-text. These documents have to be pre-processed before being analysed by the Topic Modelling algorithm in order to make the information structured for the analysis. In particular it is important here to remove words that could produce statistical noise (thus impeding the process of topic identification). For this reason, we applied the sequence of tasks shown in *Table 1*. The table shows the name of the task, a brief description of what the task does and an example of the transformation that is performed on the text.

3.3 Topic Modelling

The aim of Topic modelling algorithms is to discover the underlying set of topics of a given text collection. These algorithms provide quantitative measures (document-topic association and term-topic association) without the need of prior categorization, labelling, or annotation. These methods have been largely discussed in the Natural Language Processing domain (Steyvers and Griffiths, 2007). Here we describe the three main design decisions that we adopted to compute the topics of the ED knowledge field that are: the probabilistic model we used (Latent Dirichlet Allocation) and the measures that we adopted to assess the models (and select the best one).

[1] https://dev.elsevier.com/

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

#	*Task name*	*Task Description*	*Example*
1	*Tokenization*	Since documents are unstructured information, these has to be divided into linguistic units. English words are often separated from each other by white space, but the white-space rule is not always sufficient to have a suitable tokenization.	'In' 'engineering' 'design' ',' 'the' 'existence' 'of' 'a' 'product' 'is' 'subjected' 'to' 'its' 'function' '.'
2	*Lemmatization*	Determining the root of a word. The output allows to find that two words have the same root, despite their surface differences. This allows to diminish the lexical sparsity of the corpus.	Designers → designer described → describe was → be
3	*Part-of-Speech Tagging*	Assigning to each token a label of its most probable part-of-speech. This information is useful to perform task 4.	'designers' [NN] 'describes' [VB] 'hard' [ADJ]
4	*N-grams Tagging*	An n-gram is a contiguous sequence of n word. The adopted method to extract n-grams uses PoS-tagging. Once PoS-tagging representation is ready, it is possible to extract only certain sequences of part-of-speeches, the ones that whit a high level of confidence are valuable n-grams.	'new product development', 'design science', 'text mining', 'functional analysis'
5	*Sparse terms removal and stop-words removal*	Are removed the terms that: a) occurs in less than 1% of the documents b) are typical terminology of the document under analysis c) are tagged with part-of-speech that are useless for the analysis (e.g. punctuation or articles)	a) 'suffer', 'avoidable' b) 'paper', 'scientific', c) 'this', 'a', '.', '-',

Table 1 – Summary of the text pre-processing pipeline performed on the abstract.

3.3.1 Latent Dirichlet allocation

Blei *et al*. (2002) developed the Latent Dirichlet Allocation (LDA), one of the most well-established topic modelling algorithms. The main assumption on which LDA is based is that the documents of a collection (a corpus) are written by authors as collection of multiple topics and that each document is composed by a particular topic distribution. In the context of the collection of all ICED articles published between 2003 and 2017, this assumption has the reasonable implication that the collection discusses a set of topics in the field of ED; the individual articles are heterogeneous with regards to the specific subset of topics they address.

In LDA the topic distribution is assumed to have a sparse Dirichlet prior distribution (Ng et el., 2011), designed following the intuition that documents are likely to cover only a small set of topics and that topics use only a small set of words. In practice, this implies a better disambiguation of words and a more precise assignment of documents to topics (especially in relatively short texts as scientific abstracts).

Considering the goal of the present paper , given a set of D documents represented by T different tokens (see section 3.2 for a definition of token) and chosen a number of K topics (see section 3.3.2 for the selection of the best K) the main output of interest of the LDA are two variables defined as follow:

- α: maps the topics on the documents. Given a document, α indicates for each of the K topics which is the probability that the document taken in to consideration belongs to each topic. Thus, a K dimensional vector is computed for each document.
- β: maps the words on the topics. Given a topic, β indicates for each of the T tokens which is the probability that the topic taken in to consideration contains each token. Thus, a T dimensional vector is computed for each topic.

These two measures will be important to understand the results of section 4.3 and 4.4.

3.3.2 Model assessment

In many clustering algorithms there is the need to decide the number of clusters a-priori. Unfortunately, LDA is a member of this family. There is in fact the need to set K, the number of topics that best represent the set of

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

documents in analysis. To do that, since one of the aims of the present paper is to give an unbiased and objective map of the ED knowledge field, we adopted a bottom-up/top-down approach for number of topics selection. First (as shown in this section) we evaluated the quality of LDA models for a large number of topics; then (see section 3.4) we used the opinion of independent experts to make the final decision.

State of the art approaches for the computation of K, follows the idea of computing distances (or similarities) between pair of topics varying K. We used four different measures to evaluate the output of a topic model:

A. *Caojuan2009* (Cao *et al*., 2009): Minimize the average cosine distance between every pair of topics. The best topic number K has a minimal final distance between topics in Latent Dirichlet Allocation.
B. *Arun2010* (Arun *et al*., 2010): Minimize the symmetric KL-Divergence of the salient distribution that are derived from the matrices of factors. These matrices are the re-projections of the documents on the topics and of the topics on the tokens. The divergence values are higher for non-optimal K values.
C. *Griffiths2004* (Griffiths and Steyvers 2004): Maximize the likely of the data given the model built considering K topics. This is a problem of model selection using Bayesian statistics.
D. *Deveaud2014* (Deveaud *et al*., 2014): Maximize the information divergence between all pairs of topics. The optimal value k is the value for which LDA modelled the most scattered topics.

3.4 Experts cross-validation

As stated in section 3.3.1 the most important output of LDA are (for the present study) α and β. These are nothing but high dimensional vectors, hardly interpretable by humans. As such, it remains the responsibility of the expert who read topic models output to go further and make sense of the results. This process can be divided in two sub-phases:

1- Make the final decision on the best number of topics K, and thus choosing the best model.
2- Assign an intelligible label to each topic.

3.4.1 Final Decision on K

The output of the model assessment phase (described in section 3.3.2) are the measures used to evaluate the topic models varying the number of topics K (for an example of output see figure 1). Looking at the measures (which has to be minimized or maximized by the optimal K) it is possible to identify a neighbour of optimal values for K, made of a set of *n* different K values (*n* can vary dependently to the evolution of the curve).

To make the final decision among the 10 models trough consensus building, experts needs to be visually assisted so that it is possible for them to get a bird eye view description of the topics. For this reason, the graph described in section 3.5 (and for which figure 2 is an example) has been produced 10 times, for each of the values of K candidate to be the optimal one. This set of result is analysed by a panel of 4 experts that discuss and deliberate on the optimal output. The final decision is taken considering that the topics has to give a complete view of the ICED research activity and that each topic has to be consistent (i.e. it has to be easy to assign a label to it).

3.4.2 Topic Labelling

Various works are conducted to find the best method for labelling topics, both manually and automatically. This task entails a significant cognitive load in interpretation, prone to subjectivity (Lau *et al*., 2011). For the present study, for ease of interpretation in research publications (Wang and McCallum, 2006), the experts manually assign the label by choosing the most representative tokens for each topic. In particular, 4 independent experts assign a label to each topic, choosing from the top-5 tokens in terms of β (the tokens that with the highest probability are contained in the topic under analysis). The experts take their decision independently, then for each topic, the label with the maximum number of votes is chosen. In case of tie the experts has to discuss on the labels and vote again (which has not been the case of the present work).

3.5 Visualization

The visual outputs of the process are two charts:

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

a) the distributions of the first five most probable words in terms of β for all topics in orders to shows the topics and their content. Working with *ggplot2* (Wickham 2016) we visualized a term-topic matrix inspired by the *Termite* tool (Chuang *et al*., 2012).
b) the evolution of the extracted topics over time in terms of α in order to identify patterns of rising, falling, and persistency of the topics. Also here working with *ggplot2* (Wickham 2016) we fitted the data points with a cubic time-trend regression.

4 RESULTS

4.1 Paper retrieval

We collected 3,129 documents, downloading all the papers in Scopus published in the proceedings of the ICED conference from 2003 to 2017. Along with the abstracts of the documents, we collected the year of publication.

4.2 Model Assessment

We use four different metrics to evaluate the models trained on the ,129 documents. The goal is to identify the best range of number of topics K (see section 3.4.1 for more details). The range of potential optimal values of k has been decided to be 10 a-priori, since no deterministic rules exists to select it. This decision has two main advantages:

1. the number is big enough to cover a wide range of the space search
2. the number it is small enough so that the experts can analyse the results of each model (see section 3.4.1 for more details)

In figure 1 we show the evolution of the curves of the metrics that has been computed for different value of K and in red the neighbour of optimal values for K. Referring to figure 1, we have on the x-axis the different value of K (from 2 to 30) and on the y the normalised value of the metrics (computed subtracting the minimum for each metrics and dividing by the maximum). A discussion and a description of the single metrics has been presented in section 3.3.2.
As it is evident from figure 1:

- B, C and D have a monotonic behaviour.
- A grows in a non-monotonic way (this metric showed the same behaviour in many other cases like in the work by Dassisti (2018))
- C reaches a constant behaviour around 12 topics and intersects *D* for k=6
- *A* has local minima at 7, 11 and 17 topics.

Taking in to consideration that, the experts positioned the neighbour of optimal values for K between 7 and 17, as shown by the red dashed lines in figure 1.

4.3 Final Decision on K

Four experts working in the ED field (2 Professors, a Researcher and a Practitioner) conducted the analysis of the results for the 10 values of K independently (looking at the material described in section 3.4.1). After a discussion on the results, the experts agreed on the common position that the best number of topics to map the ICED conference papers is 12. This number has been chosen because it *is a suitable number to represent at the right level of grain the ED knowledge field. In other words, the number of topics is neither too big so that it would be hard for the algorithm to identify differences between the topics, neither too small so that many topics would be embedded inside other ones.* Furthermore, it is the point of saturation of the *Griffiths2004 (A) metrics.*

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

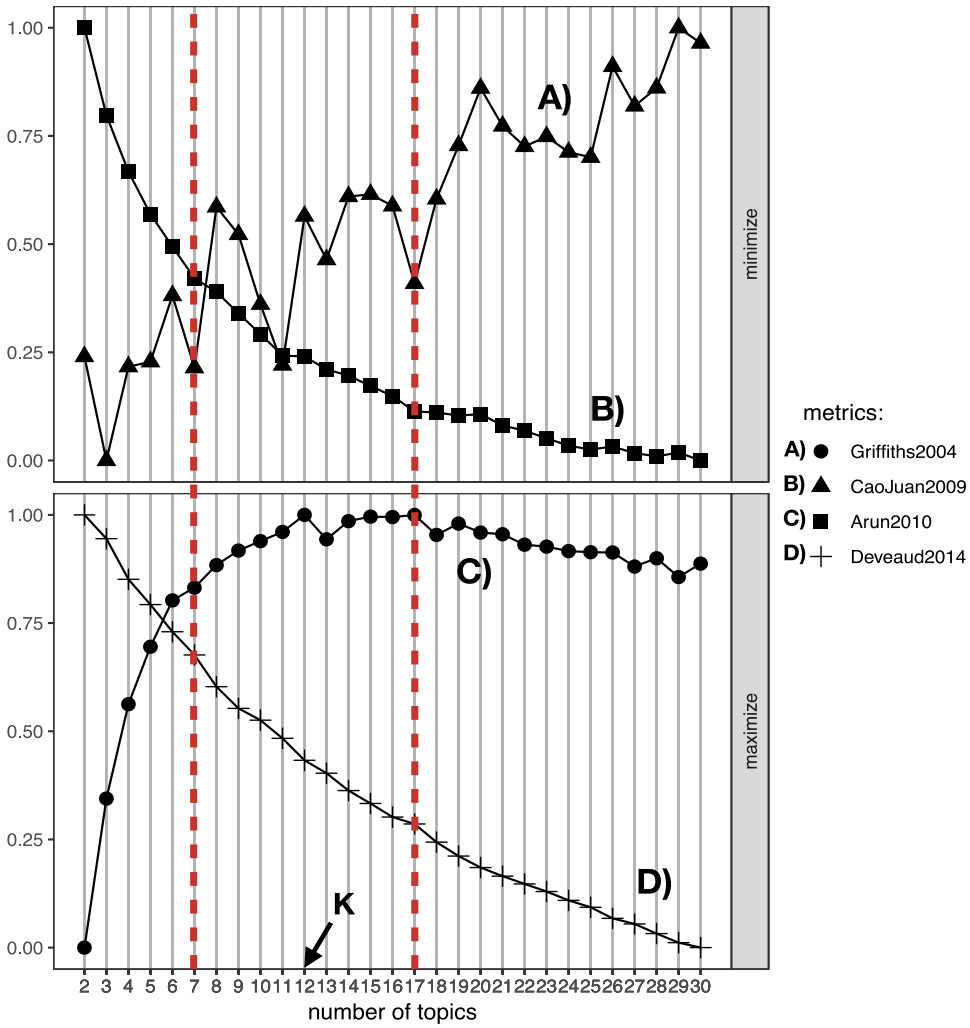

Figure 1 – Measures of the quality of the LDA models when varying the number of topics K

4.4 Labels assignment

The 12 topics are labelled as follows: *Creativity*, *Customer Satisfaction*, *Design Education*, *Manufacture*, *Problem Solving*, *Process Modelling*, *Product Families*, *Product Modelling*, *Requirements*, *Sustainable Development*, *Uncertainty* and *User Centred Design*. See section 3.4.2 for a description of the labelling approach.

4.5 Topic Model Interpretation

To support effective evaluation of term distributions associated with LDA we used a matrix view inspired by *Termite* (Chuang *et al*. 2012) where the y-axes correspond to terms and x-axes to topics. This term-topic structure (Figure 2) shows term distributions of the first five most probable words in terms of β for all latent topics. This matrix supports comparison across both topics and terms unlike the standard practice of using lists of per-topic words. Furthermore, we used the dimension of the points to encode β. The higher is the diameter of the circle the higher is the probability to find the word in the corresponding topic. As the topics are represented by its term dimension, an overlapping of topics, or of a part of it, means that they share some words but with different probabilities. So that, the topic-term matrix view gives a clear overview of the possible links among the topics by seeing those words been represented by more than one points (e.g. the term *new product development* is contained in 6 different topics).

The topic-term matrix represents in a synthetic view the main topics discussed in the last 15 years. Therefore, some topics as for example *creativity*, *manufacture*, *problem solving*, *sustainable development* and *user centred design* are expected to appear since they are superimposable with

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

conference tracks or main topics for some ICED conferences. (e.g. sustainable development in ICED 2015 in Milan and 2017 in Vancouver).
No surprise to see well represented the *manufacture* topic since it is the immediate internal customer of design. In the last year it showed a renewed interest owing to the hype around additive manufacturing (3D printing).

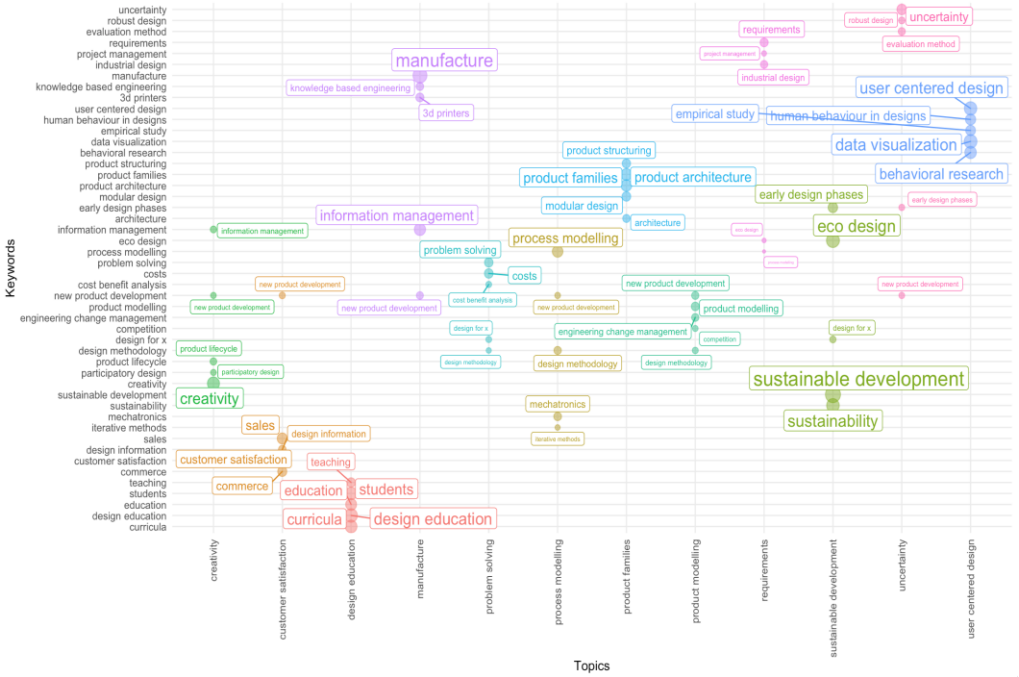

Figure 2 – Visualisation of the main topics of ICED conference

User centred design and *Education* are homogeneous topics with a constant distribution among their subtopics/Keywords. Also, the above mentioned two clusters are reasonable since they explore the design process with the user at the centre and one of the main targets of design: young designers during their education.
The topic *product families* collect many of the research about complexity, variability, postponement, modularisation and in some way anticipates the concepts behind Industry 4.0 (Chiarello *et al.*, 2018b) where the goal is mass customisation: a production system able to mass-produce (almost) infinite different batches of a single product.
Product modelling and *process modelling* are kept separated by the algorithm, but they appear too small if compared with the other topics described above. Moreover, a single picture where different ages of engineering design are joined together provides a too coarse grain with respect to more dynamic views. This problem has been addressed in section 4.6.

4.6 Evolution of Topics over time

As stated in section 3.3.1, for each document is computed α that is the probability for that document to belong to a topic. To get a hint of the popularity of the 12 topics over the last 15 years of the ICED conference, in figure 3 we show for each year and for each topic the sum of α. This method is conceptually better then assigning each document to a topic (for example the topic that has the higher α for that topic), since the same document can contain information about different topics. *The main evidences from figure 3 are:*

- *Creativity*, *Problem solving*, and D*esign Education* are almost stable. This are clearly the main underlying topic for the ICED conference.
- *Customer Satisfaction* and *Requirements* have a decreasing behaviour, while *User Centred Design* is growing and is probably capturing the interest of scholars and practitioners. It is a more modern design philosophy to deal with users and it has embedded both *Customer satisfaction* and *Requirements*.

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

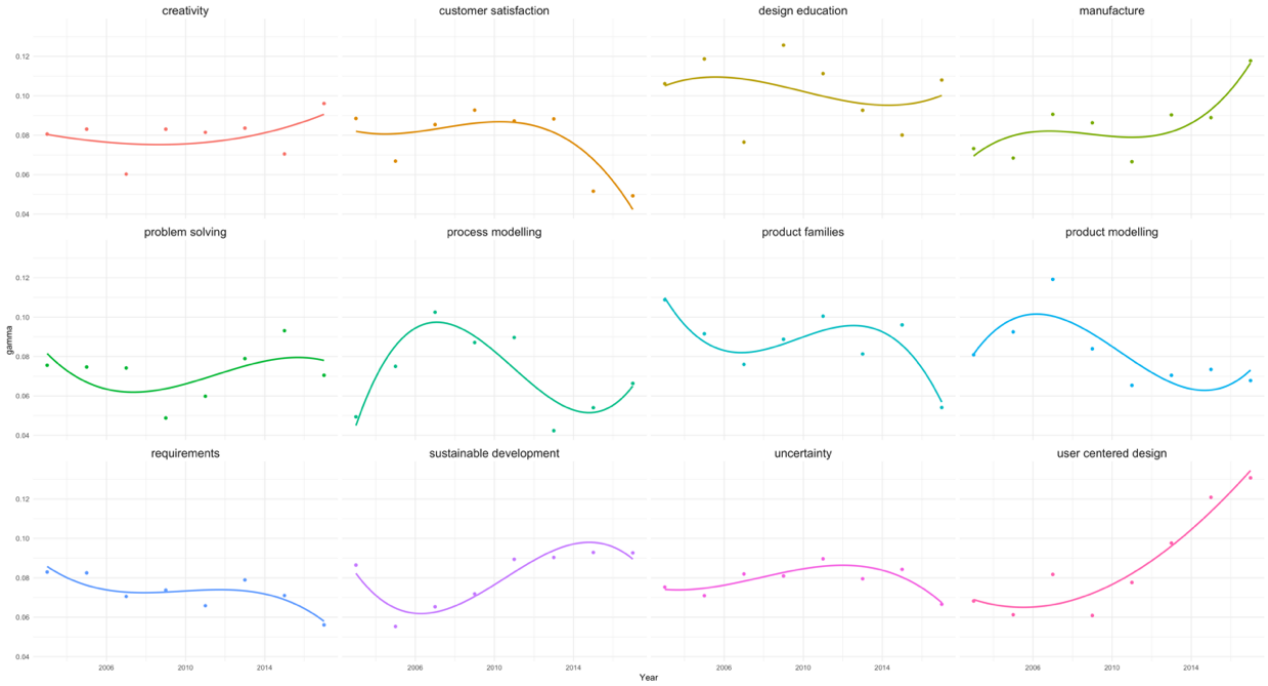

Figure 3 – Trend of the popularity of the topics over time

- *Manufacture is growing, maybe* under the pressure of Industry 4.0.
- As said before, *Product modelling* and *process modelling* are kept separated by the algorithm, but in this dynamic view of the topics it appears that they have almost the same behaviour over time, demonstrating the correlation that exists between the two topics. Furthermore Industry 4.0 will probably push also these two topics.
- *Sustainable Development* and *Product Families* has a related behaviour over time, maybe as an evidence of the current support from ICED for handling the sustainability of modular product families (Bahns *et al*., 2015)

5 CONCLUSION

This paper is a first attempt to represent the domain of Engineering Design in an integrated way. For each of the topics it could be possible to retrieve the specific terms or keywords and examine in a closer way the relevant documents.

We believe that it might be used as a support for an effort of self-reflexivity of the Engineering Design community. Scientific maturity is associated to better understanding of the research agenda, its change over time, and the direction of promising and important research. We hope to contribute to this effort.

REFERENCES

Bonaccorsi, A., Apreda, R. and Fantoni, G. (2017), "Cognitive and Motivational Biases in Technology Foresight", Technological Forecasting and Social Change (Under review).

Arun, R., Suresh, V., Madhavan, C. V. and Murthy, M. N. (2010), "On finding the natural number of topics with latent dirichlet allocation: Some observations", *In Pacific-Asia conference on knowledge discovery and data mining*, (pp. 391–402), Springer, Berlin, Heidelberg.

Autrey, J. L., Sieber, J., Siddique, Z. and Mistree, F. (2018), "Leveraging Self-Assessment to Encourage Learning Through Reflection on Doing", *International Journal of Engineering Education*, Vol. 34 No. 2, pp. 708–722.

Bahns, T., Beckmann, G., Gebhardt, N. and Krause, D. (2015), "Sustainability of modular product families", In *20th International Conference on Engineering Design, ICED15*, Milan, Italy, July pp. 27–30.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), "Latent dirichlet allocation", *Journal of machine Learning research*, Vol. 3 No. Jan, pp. 993-1022.

Bolloju, N., Schneider, C. and Sugumaran, V. (2012), "A knowledge-based system for improving the consistency between object models and use case narratives", *Expert Systems with Applications*, Vol. 39 No. 10, pp. 9398–9410.

Bonaccorsi, A., and Fantoni, G. (2007, August), "Expanding the functional ontology in conceptual design", In *International Conference on Engineering Design*.

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press

Cao, J., Xia, T., Li, J., Zhang, Y. and Tang, S. (2009), "A density-based method for adaptive LDA model selection", *Neurocomputing*, Vol. 72 No. 7–9, pp. 1775–1781.
Chiarello, F., Cimino, A., Fantoni, G. and Dell'Orletta, F. (2018a), "Automatic users extraction from patents", *World Patent Information*, Vol. 54, pp. 28–38.
Chiarello, F., Fantoni, G. and Bonaccorsi, A. (2017), "Product description in terms of advantages and drawbacks: Exploiting patent information in novel ways", In *DS 87-6 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 6: Design Information and Knowledge*, Vancouver, Canada, 21-25.08. 2017 pp. 101–110.
Chiarello, F., Trivelli, L., Bonaccorsi, A. and Fantoni, G. (2018b), "Extracting and mapping industry 4.0 technologies using Wikipedia", *Computers in Industry*, Vol. 100, pp. 244–257.
Chiu, I. and Shu, L. H. (2005, January), "Bridging cross-domain terminology for biomimetic design", In *ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, (pp. 93–101). American Society of Mechanical Engineers.
Chuang, J., Manning, C. D. and Heer, J. (2012, May), "Termite: Visualization techniques for assessing textual topic models", In *Proceedings of the international working conference on advanced visual interfaces*, (pp. 74–77). ACM.
Dassisti, M., Chiarello, F., Fantoni, G., Priarone, P. C., Ingarao, G., Campana, G. and Forcelles, A. (2018), "Benchmarking the sustainable manufacturing paradigm via automatic analysis and clustering of scientific literature: A perspective from Italian technologists", In *16th Global Conference on Sustainable Manufacturing*, (pp. 1–7).
Deveaud, R., SanJuan, E. and Bellot, P. (2014), "Accurate and effective latent concept modeling for ad hoc information retrieval", *Document numérique*, Vol. 17 No. 1, pp. 61–84.
Devyatkin, D., Nechaeva, E., Suvorov, R. and Tikhomirov, I. (2018), "Mapping the Research Landscape of Agricultural Sciences", *Форсайт*, Vol. 12 No. 1. (eng).
Goonetillake, J. S., Carnduff, T. W. and Gray, W. A. (2002), "An integrity constraint management framework in engineering design", *Computers in Industry*, Vol. 48 No. 1, pp. 29–44.
Griffiths, T. L. and Steyvers, M. (2004), "Finding scientific topics", *Proceedings of the National academy of Sciences*, Vol. 101 No. 1, pp. 5228–5235.
Jin, J., Liu, Y., Ji, P. and Liu, H. (2016), "Understanding big consumer opinion data for market-driven product design", *International Journal of Production Research*, Vol. 54 No. 10, pp. 3019–3041.
Lau, J. H., Grieser, K., Newman, D. and Baldwin, T. (2011, June), "Automatic labelling of topic models", In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*-Vol. 1, pp. 1536–1545. Association for Computational Linguistics.
Trivelli, L., Apicella, A., Chiarello, F., Rana, R., Fantoni, G. and Tarabella, A. (2019), "From Precision Agriculture to Industry 4.0: unveiling technological connections in the agrifood sector", *British Food Journal*, http://doi.org/10.1108/BFJ-11-2018-0747
Li, Z. and Tate, D. (2010), "Patent Analysis for Systematic Innovation: Automatic Function In-terpretation and Automatic Classification of Level of Invention using Natural Language Processing and Artificial Neural Networks", *International Journal of Systematic Innovation*, Vol. 1 No. 2.
Ng, K. W., Tian, G. L. and Tang, M. L. (2011), *Dirichlet and related distributions: Theory, methods and applications*, Vol. 888. John Wiley & Sons.
Noh, H., Jo, Y. and Lee, S. (2015), "Keyword selection and processing strategy for applying text mining to patent analysis", *Expert Systems with Applications*, Vol. 42 No. 9, 4348–4360.
Nuzzo, A., Mulas, F., Gabetta, M., Arbustini, E., Zupan, B., Larizza, C. and Bellazzi, R. (2010), "Text Mining approaches for automated literature knowledge extraction and representation", *In MedInfo* pp. 954–958.
Parraguez, P. and Maier, A. (2017), "Data-driven engineering design research: opportunities using open data", In *DS 87-7 Proceedings of the 21st International Conference on Engineering Design (ICED 17)*, Vol. 7, pp. 21–25.
Ponweiser, M. (2012), "Latent Dirichlet allocation in R".
Bailey, R. (2009), "Educating engineers for multiscale systems design in a global economy: The Technology Leaders program, ASEE Annual Conference and Exposition", *Conference Proceedings*, pp. 2153–5965
Rehurek, R. and Sojka, P. (2010), "Software framework for topic modelling with large corpora", In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.
Steyvers, M. and Griffiths, T. (2007), "Probabilistic topic models", *Handbook of latent semantic analysis*, Vol. 427 No. 7, pp. 424–440.
Tseng, Y. H., Lin, C. J. and Lin, Y. I. (2007), "Text mining techniques for patent analysis", *Information Processing & Management*, Vol. 43 No. 5, pp. 1216–1247.
Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009, June), "Evaluation methods for topic models", In *Proceedings of the 26th annual international conference on machine learning*, (pp. 1105–1112). ACM.
Wang, X. and McCallum, A. (2006, August). "Topics over time: a non-Markov continuous-time model of topical trends", In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 424–433). ACM.
Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

https://doi.org/10.1017/dsi.2019.283 Published online by Cambridge University Press