


Augmenting research cooperation in production engineering with data analytics

Thomas Thiele¹  · André Calero Valdez² · Sebastian Stiehm¹ · Anja Richert¹ · Martina Ziefle² · Sabina Jeschke¹

Received: 16 September 2016 / Accepted: 19 January 2017
© German Academic Society for Production Engineering (WGP) 2017

Abstract Understanding how members of a research team cooperate and identifying possible synergies may be crucial for organizational success. Using data-driven approaches, recommender systems may be able to find promising collaborations from publication data. Yet, the outcome of scientific endeavors (i.e. publications) are only produced sparingly in comparison to other forms of data, such as online purchases. In order to facilitate this data in augmenting research cooperation, we suggest to combine data-driven approaches such as text-mining, topic modeling and machine learning with interactive system components in an interactive visual recommendation system. The system leads to an augmented perspective on research cooperation in a network: Interactive visualization analyzes, which cooperation could be intensified due to topical overlap. This allows to reap the benefit of both worlds. First, utilizing the computational power to analyze large bodies of text and, second, utilizing the creative capacity of users to identify suitable collaborations, where machine-learning algorithms may fall short.

Keywords Text-mining · Topic-modeling · Recommender systems · Human-computer interaction · Interactive machine learning · Deep neural networks

1 Introduction

Future perspectives on Industry 4.0 aim to enhance production by integrating virtual and physical systems into cyber-physical production systems. Yet, this process mainly tackles the manufacturing part of engineering. The underlying drivers of this process, like developments in data mining on the one hand and machine learning and visual analytics on the other hand, can be used to allow decision makers to manage relationships between research and development projects. Here, the aim is to connect technologies or methods in an innovative way. Following this idea, Industry 4.0 and merging of technologies not only imply future solutions for challenges of scale and scope in production, but also support the generation of cross-sectional innovations by revealing synergies between research topics. To utilize these recommendations, we focus on graph-based visualization of synergies and complexity in this contribution.

But which data allow predicting innovation potentials? And how can we transform data into meaningful incubators for joint research? We address these aspects by exemplifying, how a data analytics process can be designed in order to detect synergies between research topics based on publication data. As processing of real data often requires the time-consuming annotation of data to generate meaningful results, the process will provide concepts to minimize this issue. In addition, deep neural networks are used to automatize the discovery of relationships between research topics.

Funded by Deutsche Forschungsgemeinschaft under: DFG EXC-128.

✉ Thomas Thiele
thomas.thiele@ima-zlw-ifu.rwth-aachen.de
André Calero Valdez
calero-valdez@comm.rwth-aachen.de

¹ IMA/ZLW & IfU-RWTH Aachen University, Aachen, Germany

² Human-Computer Interaction Center, RWTH Aachen University, Aachen, Germany

By visualizing the outcome of data analytics, the human is included as an actuator and controlling element. Data analytics and interactive visualization serve as enabler; the human has to transform the presented recommendation into further actions. Hence, the idea adopts the Industry 4.0 perspective on the merge towards cyber-physical systems: By creating an effective interface to understand data analytics results, the human perspective on future cooperation is augmented. The resulting system serves as an information provider for cooperation processes. It determines, which thematic overlaps exist as well as the content, on which these overlaps are based.

2 Related work

To achieve a prediction towards new potential cooperation, we give insights into relevant works that are connected to that idea. In order to enable the human to access these potentials, the derived information have to be presented in a manner that is suitable for the cognitive model of the human. Hence, relevant research in the field of data visualization is depicted.

2.1 Data analytics for synergy detection

Within this analysis, scientific texts serve as underlying data for the analysis. Due to the—from a statistical point of view—unstructured composition of texts [18], these have to be transformed into a machine readable format. Text Mining offers a wide variety of methods for this challenge, ranging from the parsing of texts to filtering to unsupervised and supervised learning processes [35]. Whereas the first mentioned method deals with the preparation of the unstructured data towards further analysis, unsupervised and supervised learning processes are operations that allow to deduce patterns in the data.

Following Hastie et al. [24], *supervised learning* can be formally described as a density estimation problem, with the aim to determine the conditional densities of given input observations based on the assumption of a given model of training variables. Within this work, neural networks are used as supervised classifiers (see Sect. 3). Deep neural networks reassemble the concept of the human brain: multiple hidden layers of neurons are connected to an input and an output layer [43]. *Unsupervised learning* tries to achieve this goal without the trained model. Based on a similarity measure, overlaps of mostly high-dimensional models are derived [35]. This is achieved “by doing”: The algorithm sorts the data and measures the outcome against a predefined quality function. Our work especially refers to Topic Modeling based on Latent Dirichlet Allocation (see Sect. 3).

The detection of synergies often refers to a combination of unsupervised and supervised learning methods. From a data-driven perspective, a synergy can be defined as a match between patterns of data, e.g. in this case between the word combinations “temperature, model, effects, function” and “cooling, variables, parameters, machine-engineering”. The degree of this match is mostly defined by a certain metric as a result of an algorithmic analysis. In the early 2000s, the mining of association rules have become quite popular. Xiong et al. [54] used association rules to skewed word distributions in order to derive clusters within this data. Although the paper exemplifies patterns of words forming a common topic, this has only been applied to around 900 item sets. Today’s big data (and our example) exceeds this by far. Other examples show the application of topic modeling and especially Latent Dirichlet Allocation for the recommendation of scientific publications [53]. By extending Latent Dirichlet Allocation with an collaborative filtering to create a so-called collaborative topic regression.

Although data analysis can be characterized as the core element for the prediction of innovation potentials, calculated results have to be made usable for the human. Therefore, approaches of data visualization are depicted in the following that allow a recommendation of results suitable for the end user.

2.2 Data visualization as the enabler for the human

Recommender systems have been developed since the early 90ies [21, 42] and have come a long way. They are typically used in commercial scenarios, where users are informed about other options to take along in their virtual shopping cart [1, 46], or in tourism scenarios [4, 19], learning [17, 33] or even in cooking [15]. Modern, so called hybrid-recommender systems are not only based on content data (from text-mining for example), but also on user- [6], user-generated [20, 55], or user-network data [22, 50]. Park et al. provide a good overview of different recommender systems, algorithm, recommendation methods, and data models in their review from 2012 [38]. While there has been a lot of research going into improvement of algorithms and procedures, the interface to the user and its impact on the decision has been rather neglected [12, 32, 34]. Therefore, it is important to investigate the effect of the user interface on decision making and also examine how the interface to the recommendation reveals dimension of the algorithms.

Most algorithms output high-dimensional data from \mathbb{R}^n , which is hard to interpret by the human, whose visual perception is mostly limited to \mathbb{R}^2 (\mathbb{R}^3 or 3D is created in the mind, and not directly perceived). Looking at collaborations between researchers, non-spatial mental models could also apply. The mental model of the user influences heavily

how trustworthy the data seems and how the interaction with the interface unfolds.

And as at the end of all data analysis, a user must make a decision, it pays to look at how data is visualized. However, not only the visualization, but also the interaction [26, 39] with the visualization must be controlled. Adding additional complexity to a system can be avoided if usability-evaluations are done effectively and efficiently [9] and the visualization is chosen with an appropriate method (e.g. design study methodology [45]).

In a review of interactive recommender systems, He et al. [25] compare different approaches of how interactive recommender systems reveal themselves to the user. They are not only a post-hoc visualization to the “core” data analysis, but an integral part of the data-driven approach to new recommendations.

Visualizing the data helps in shaping the mental model [30, 40] and has been shown to be helpful in recommendations as well [36]. An appropriate visualization increases controllability, transparency and supports exploration [51] of recommendations.

2.2.1 Collaboration recommender systems

The performance of visual and interactive recommender systems depends highly on the usage context [3, 13]. Therefore, giving bibliometric recommendations is particularly hard as data (i.e. publications) is sparsely distributed. Because over-indulging in clever algorithms is less important than using more data [16], good bibliometric recommendations can hardly be created without using visualization.

Simply visualizing the data may also not be helpful when the usage motives of its potential users are unclear. Integrating such a system in a continuous feedback mechanism, such as a scientific cooperation portal, requires the analysis of the motivation and leverages of researchers to use the portal in general [7, 11, 23]. Otherwise, recommendations will never be retrieved.

Various interfaces in research collaborations [5, 10, 14, 29, 37, 52, 56] have been suggested, of which some have already been reviewed by He et al. [25]. Graph-based visualizations seem to reflect the consensus on how to visualize publication data.

But even when the correct “type” of visualization is used, it is necessary to incorporate the user-diversity as it may influence how users perceive the recommender system [2, 31, 39]. Experience and domain knowledge may influence what knowledge a user can deduct from a given visualization. It is thus necessary to evaluate the quality of a recommender system by the end-user using an evaluation framework such as ResQue introduced by Pu et al [41].

3 Own approach and methods

The presented approach combines the aforementioned methods to a combination of unsupervised topic modeling and classification. As the results of this combination have to be made available for the human as recommendation and decision support, this work especially addresses the visual representation and cognition of our data visualization. The used processes from data to the interaction of the human with derived information is depicted in Fig. 1.

3.1 Data analytics process chain

For the purpose of synergy detection in scientific cooperation, we use a tailor-made framework that covers the necessary processes [47–49]. The data analytics part of this framework relies on five subsequent steps:

1. Text Mining of scientific publication
2. Generation of a probabilistic topic model
3. Classification of topics
4. Graph-based data visualization
5. Human interaction

Whereas the first three steps are covered in the following, the visualization part is described in Sect. 3.2. First, we rely on Text Mining to make our underlying data, consisting of scientific publications, machine readable. The publications that have to be analyzed are grouped per entity. Within this step, we define the semantic level of our analysis: An entity can be characterized as a research project, for example. Another possibility for an entity can be seen in different universities or research networks. For this contribution, we focus on research cooperation between different projects in a research network for technology. The grouping of publications represents the only

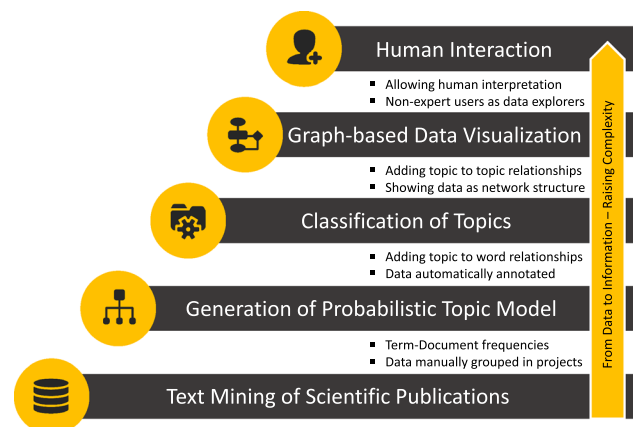


Fig. 1 Process model and analytics chain

manual annotation step within our process. All other processes rely on that information and add other labels based on machine learning models.

One of these models is generated in the second step, where we include Topic Modeling via Latent Dirichlet Allocation: The frequency of words per paper is analyzed regarding hidden patterns. The result is a probabilistic model, in which different groups of words form a topic. Thinking within our data structure, the words are labeled with this relationship. In addition, the words forming a topic are ranked regarding their importance for the topic.

The third step is based on the supervised learning process of classification. The training data for this process is the topic of one project. Hence, the classifier uses labeled topic data to generate predictions between input data and the trained topics. The result can be characterized as a matchmaking between topics. In order to define a topic for the classification, not only the words are used, but also the importance of each word for the topic as well as their relative frequency are included. Regarding the classifier, we are currently working on deep learning as this branch of machine learning has proofed its ability to outperform the human in certain areas [44]. The neural network adds another label to our data set: A relationship for each topic towards the topics in other research projects is derived. Hence, a list of potentially affiliated projects is generated and in the next presented to the user via data visualization.

3.2 Visual recommender systems

The visualization front-end is based on concepts derived from the Tigrs-system [5] using an interactive multimodal [8] graph-based visualization to represent both *topics*, *words* and *project-teams*. These are nodes in the graph, while the interconnections are represented as edges (see e.g. 3). The visualization framework allows hovering over nodes, interactively highlighting all connected nodes and giving additional information for the node when available. A search field allows to filter data for a given task.

The visualization uses a force-based layout algorithm to maximize distances of unrelated parts of the graph. Elements that are interconnected group more closely together. This allows to naturally detect clusters in graphs, and thus similarity of topics or projects. In conjunction with the meta-data, this allows the user to acquire a deeper knowledge of the visualized data and derive recommendations for future collaboration.

No list-based recommendation is given, the visualization is used to identify recommendations. This enhances controllability, transparency and fosters the search for novel options to collaborate.

4 Results and evaluation

Our results and the evaluation are based on two levels. First, the results from the data analysis: Does the data provide enough information to allow any recommendation or decisions for further research cooperation? Second, we consider the human factor: In how far is the human capable in understanding the derived information?

4.1 Data analytics results

Figure 2 shows the initial data visualization. The small image depicts the whole graph of projects, topics and words centered on the user's project. As this representation is a quite confusing representation of the data set, we included highlighting of thematic paths. The highlighting of an element includes the accentuation of all connected elements up to two edges away. All other elements are depicted with a much lower opacity. To provide an easy to use description of each topic, only the top ranked words within each topic are visualized in the graph. This allows an individual navigation through the semantic spaces of each relationship in the data.

Figure 3 shows a zoomed area in the graph, in which a relationship between the two projects is highlighted. The topic nine (at the bottom of Fig. 3) belongs to the project A2 and deals with the modeling of time effects in heat models. The connection to topic ten of project D2 (in the middle of Fig. 3) is of special interest, as words like "parameters", "cooling", "manufacturing" and "machine" indicate that this topics deals with cooling parameters of manufacturing machines.

4.2 User evaluation

In a first small user study ($N = 20$), we evaluated how users interact with a graph-based visual recommender system, namely Tigrs. We measured accuracy of recommendations, transparency, controllability and responsiveness of the visualization (Scale: 1–6, low–high).

4.2.1 The sample

As participants we recruited members of a research organization. The average age of the participants was $M = 29.8$ years ($SD = 3.59$, range = 24–35) and 25% of them were female. 11 had finished their Masters (or similar), while 2 already had a Ph.D. In total, we had eight communication scientists, nine computer scientists, four mechanical engineers, four psychologists, one sociologist, one language and communication scientist, one

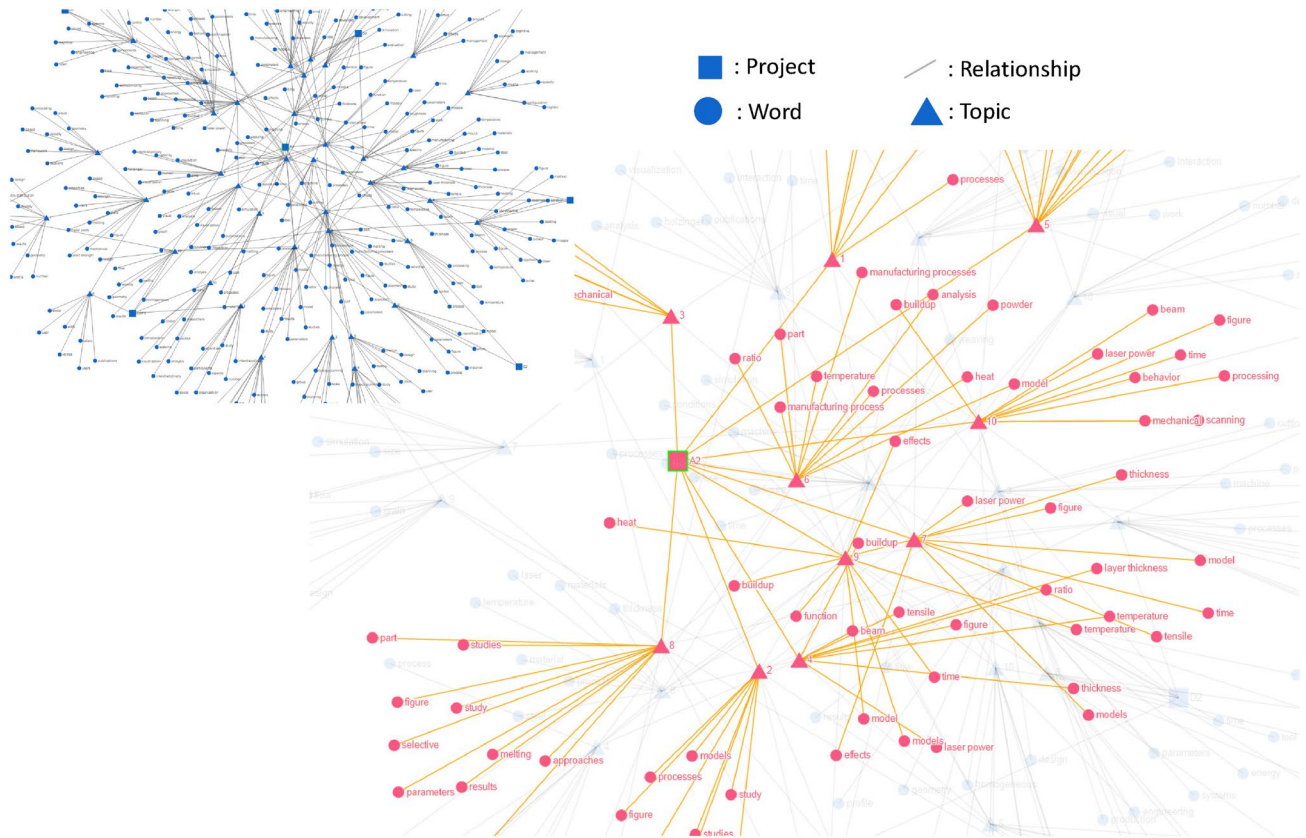


Fig. 2 Graph-based data visualization: (1) complete graph (*small picture*), (2) highlighted project graph

electrical engineer and one architect in our sample (multiple selections allowed).

4.2.2 Results

The overall accuracy of the system was very high ($M = 4.81$, $SD = 0.575$). Its sub metrics, accuracy ($M = 4.87$, $SD = 0.663$) and relative accuracy ($M = 4.76$, $SD = 0.581$), had similar descriptive results. This means that the system provides good recommendations and is a viable option to seek novel recommendation. The trust in the given recommendations was relatively high ($M = 4.4$, $SD = 0.64$). Participants got the impression that the given recommendations were actually sensible. Participants perceived the system’s interaction and visualization very positively. Its transparency ($M = 4.87$, $SD = 0.663$), control ($M = 4.87$, $SD = 0.663$) and particularly responsiveness ($M = 4.87$, $SD = 0.663$) were very high. Participants understood its concept and felt in control while interacting with the system.

Furthermore, we asked users to rate, why a graph-based recommendation was superior (or inferior) over a list-based recommendation and which one they would prefer in a research cooperation based scenario. The strongest reason

for graph-based recommendation was, that it supports the group understanding in the cooperation. This was followed by an overview of the work, exploration of new content and the visualization being more informative. Only when a specific recommendation was sought, graph-based visualizations seemed to be less preferred.

5 Outlook

We have demonstrated, that the use of both text-mining for topic modeling as well as multi-modal graph-based recommender systems are valuable tools in identifying possible synergies in research cooperation. Both tools can be used to augment the research process individually and jointly. By combining data analytics and an effective interface, we position the user in a space, where they control the data and not the other way around. By implementing the user in the loop, even relatively small-sample data (such as bibliometric data) can be effectively used in interactive recommender systems, incorporating the interactive knowledge discovery method [27, 28]. This approach can also be used in steering scenarios, when the organization wants to ensure that

- Engineering 2013/2014, Springer International Publishing, pp 737–749
9. Calero Valdez A, Brauner P, Schaar AK, Holzinger A, Ziefle M (2015) Reducing complexity with simplicity-usability methods for industry 4.0. In: Proceedings 19th triennial congress of the IEA, vol 9, p 14
 10. Calero Valdez A, Bruns S, Greven C, Schroeder U, Ziefle M (2015) What do my colleagues know? dealing with cognitive complexity in organizations through visualizations. In: Learning and collaboration technologies, Springer, pp 449–459
 11. Calero Valdez A, Schaar AK, Bender J, Aghassi S, Schuh G, Ziefle M (2016) Social media applications for knowledge exchange in organizations. Innovations in knowledge management. Springer, Berlin, pp 147–176
 12. Calero Valdez A, Ziefle M, Verbert K (2016) HCI for recommender systems: the past, the present and the future. In: International Conference on Recommender Systems, RecSys'16 Boston, USA, ACM
 13. Calero Valdez A, Ziefle M, Verbert K, Felfernig A, Holzinger A (2016) Recommender systems for health informatics: State-of-the-art and future perspectives. In: Holzinger, A (ed) Machine Learning for Health Informatics, Lecture Notes in Computer Science LNCS 9605, Springer, pp 391–414
 14. Conforti R, de Leoni M, La Rosa M, van der Aalst WM, ter Hofstede AH (2015) A recommendation system for predicting risks across multiple business process instances. Decis Support Syst 69:1–19
 15. De Clercq M, Stock M, De Baets B, Waegeman W (2016) Data-driven recipe completion using machine learning methods. Trends Food Sci Technol 49:1–13
 16. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10):78–87
 17. Drachsler H, Verbert K, Santos OC, Manouselis N (2015) Panorama of recommender systems to support learning. In: Recommender systems handbook, Springer, pp 421–451
 18. Feldman R, Sanger J (2007) The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge
 19. Gavalas D, Konstantopoulos C, Mastakas K, Pantziou G (2014) Mobile recommender systems in tourism. J Netw Comput Appl 39:319–333
 20. Godoy D, Corbellini A (2015) Folksonomy-based recommender systems: a state-of-the-art review. Int J Intell Syst 31(4):314–346
 21. Goldberg D, Nichols D, Oki BM, Terry D (1992) Using collaborative filtering to weave an information tapestry. Commun ACM 35(12):61–70
 22. Gretarsson B, O'Donovan J, Bostandjiev S, Hall C, Höllerer T (2010) Smallworlds: visualizing social recommendations. In: Computer Graphics Forum, Wiley Online Library, vol 29, pp 833–842
 23. Hamann T, Schaar AK, Calero Valdez A, Ziefle M (2016) Strategic knowledge management for interdisciplinary teams-overcoming barriers of interdisciplinary work via an online portal approach. In: International conference on human interface and the management of information. Springer International Publishing, pp 402–413
 24. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
 25. He C, Parra D, Verbert K (2016) Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. Expert Syst Appl 56:9–27
 26. Hijikata Y, Kai Y, Nishida S (2012) The relation between user intervention and user satisfaction for information recommendation. In: Proceedings of the 27th annual acm symposium on applied computing. ACM, New York, NY, USA, SAC '12, pp 2002–2007. doi:10.1145/2245276.2232109
 27. Holzinger A (2014) Trends in interactive knowledge discovery for personalized medicine: cognitive science meets machine learning. IEEE Intell Inform Bull 15(1):6–14
 28. Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf 3(2):119–131. doi:10.1007/s40708-016-0042-6
 29. Karni Z, Shapira L (2013) Visualization and exploration for recommender systems in enterprise organization. In: IS&T/ SPIE electronic imaging. International Society for Optics and Photonics, p 86640E
 30. Klerx J, Verbert K, Duval E (2014) Enhancing learning with visualization techniques. In: Handbook of research on educational communications and technology, Springer, pp 791–807
 31. Knijnenburg BP, Reijmer NJ, Willemsen MC (2011) Each to his own: how different users call for different interaction methods in recommender systems. In: Proc. of the Fifth ACM Conf. on recommender systems, ACM, New York, NY, USA, RecSys '11, pp 141–148. doi:10.1145/2043932.2043960
 32. Konstan JA, Riedl J (2012) Recommender systems: from algorithms to user experience. User Model User Adapt Interact 22(1–2):101–123
 33. Manouselis N, Drachsler H, Verbert K, Santos OC (2014) Recommender systems for technology enhanced learning: research trends and applications. Springer Science & Business Media
 34. McNeel SM, Riedl J, Konstan JA (2006) Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: CHI'06 extended abstracts on Human factors in computing systems, ACM, pp 1097–1101
 35. Miner G (2012) Practical text mining and statistical analysis for non-structured text data applications. Academic Press, Amsterdam
 36. Mutlu B, Veas E, Trattner C, Sabol V (2015) Vizrec: a two-stage recommender system for personalized visualizations. In: Proceedings of the 20th international conference on intelligent user interfaces companion, ACM, pp 49–52
 37. O'Donovan J, Smyth B, Gretarsson B, Bostandjiev S, Höllerer T (2008) Peerchooser: visual interactive recommendation. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 1085–1088
 38. Park DH, Kim HK, Choi IY, Kim JK (2012) A literature review and classification of recommender systems research. Expert Syst Appl 39(11):10059–10072
 39. Parra D, Brusilovsky P (2015) User-controllable personalization: a case study with setfusion. Int J Hum Comput Stud 78:43–67
 40. Picard RW, Papert S, Bender W, Blumberg B, Breazeal C, Cavallo D, Machover T, Resnick M, Roy D, Strohecker C (2004) Affective learning—a manifesto. BT Technol J 22(4):253–269
 41. Pu P, Chen L, Hu R (2012) Evaluating recommender systems from the user's perspective: survey of the state of the art. User Model User Adapt Interact 22(4–5):317–355
 42. Resnick P, Varian HR (1997) Recommender systems. Commun ACM 40(3):56–58
 43. Russell SJ, Norvig P, Canny JF, Malik JM, Edwards DD (2003) Artificial intelligence: a modern approach, vol 2. Upper Saddle River, Prentice hall
 44. Rutkin A (2016) Anything you can do. New Sci 229(3065):20–21
 45. Sedlmair M, Meyer M, Munzner T (2012) Design study methodology: reflections from the trenches and the stacks. IEEE Trans Vis Comput Gr 18(12):2431–2440
 46. Stavrianou A, Brun C (2015) Expert recommendations based on opinion mining of user-generated product reviews. Comput Intell 31(1):165–183

47. Thiele T, Jooß C, Richert A, Jeschke S (2015) Terminology based visualization of interfaces in interdisciplinary research networks. In: 19th Triennial Congress of the IEA
48. Thiele T, Sommer T, Schröder S, Richert A, Jeschke S (2016) Human-in-the-loop processes as enabler for data analytics in digitalized organizations. In: Mensch und Computer 2016—Workshopbeiträge, MCI Digital Library / Gesellschaft für Informatik e.V
49. Thiele T, Sommer T, Stiehm S, Richert A, Jeschke S (2016) Exploring research networks with data science: a data-driven microservice architecture for synergy detection. In: Proceedings of the 4th international conference on future internet of things and cloud workshops, Vienna, Austria, 22-24 August 2016, pp 246–251
50. Tinghuai M, Jinjuan Z, Meili T, Yuan T, Abdullah AD, Mznah AR, Sungyoung L (2015) Social network and tag sources based augmenting collaborative recommender system. *IEICE Trans Inf Syst* 98(4):902–910
51. Verbert K, Parra D, Brusilovsky P, Duval E (2013) Visualizing recommendations to support exploration, transparency and controllability. In: Proceedings of the 2013 international conference on intelligent user interfaces, ACM, pp 351–362
52. Verbert K, Parra D, Brusilovsky P (2016) Agents vs. users: visual recommendation of research talks with multiple dimension of relevance. *ACM Trans Interact Intell Syst* 6(2):11
53. Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '11, pp 448–456. doi:[10.1145/2020408.2020480](https://doi.org/10.1145/2020408.2020480)
54. Xiong H, Tan PN, Kumar V (2003) Mining strong affinity association patterns in data sets with skewed support distribution. In: Data mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE, pp 387–394
55. Yang X, Guo Y, Liu Y, Steck H (2014) A survey of collaborative filtering based social recommender systems. *Comput Commun* 41:1–10
56. Yazdi MA, Calero Valdez A, Lichtschlag L, Ziefle M, Borchers J (2016) Visualizing opportunities of collaboration in large research organizations. In: International conference on HCI in Business, Government and Organizations. Springer International Publishing, pp 350–361