Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 214 (2022) 1374–1381

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Information Technology and Quantitative Management

Conceptual structure of federated learning research field

A. Velez-Estevez[a,*], P. Ducange[b], I.J. Perez[c], M.J. Cobo[c]

[a]Department of Computer Science and Engineering, University of Cadiz, Cádiz, Spain
[b]Department of Information Engineering, University of Pisa, Italy
[c]Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain

Abstract

Nowadays there are a great amount of data that can be used to train artificial intelligent systems for classification, or prediction purposes. Although there are tons of publicly available data, there are also very valuable data that is private, and therefore, it can not be shared without breaking the data protections laws. For example, hospital data has great value, but it involves persons, so we must try to preserve their privacy rights. Furthermore, although it could be interesting to train a model with the data of only one entity (i.e. a hospital), it could have more value to train the model with the data of several entities. But, since the data of each entity might not be shared, it is not possible to train a global model. In that sense, Federated Learning has emerged as a research field that deals with the training of complex models, without the necessity to share data, and therefore, keeping the data private. In this contribution, we present a global conceptual analysis based on co-words networks of the Federated Learning research field. To do that, the field was delimited using an advance query in Web of Science. The corpus contain a total of 2444 documents. As the main result, it should be highlighted that the Federated Learning research field is focused on six main global areas: telecommunications, privacy and security, computer architecture and data modeling, machine learning, and applications.

Keywords: Federated Learning; bibliometric analysis; science mapping analysis; co-words; SciMAT

1. Introduction

The Federated Learning (FL) term was originally used to refer to the development and distribution of reusable learning objects [12, 13], although the concept seems not to have developed in that period. In 2016, the research area increased its popularity thanks to Google, which introduced again the term "federated learning" with a different meaning. Google coined the FL term when they tried to solve the problem of training models with data that is privacy sensitive and huge in quantity [20]. This problem arises since the data to train the models comes from different clients

* Corresponding author
E-mail address: antonio.velez@uca.es

10.1016/j.procs.2022.11.319

(e.g., mobile phones, hospitals, etc.), and cannot be stored in a central server to apply classical machine learning techniques due to privacy or communication issues. Hence, FL is a novel machine paradigm where several clients (electronic devices) train a model together, by the orchestration of a central computing node, and where the data is not shared with this server [16]. The application of this technique has been used in the health care industry, where the data is stored in different locations and may belong to different institutions. Another example, are telecommunications [24] or banks [17], where data is protected by privacy legislation.

In the last years, the popularity of Federated learning has increased, and academia and industry are paying a lot of effort to develop this field, with new models, algorithms, and applications. Therefore, the number of papers related to this topic increased exponentially. The analysis of a large number of scientific documents needs the employment of advanced techniques that can automate the process, otherwise, it would be a daunting task. In that way, the scientific knowledge could be understood as a complex system (e.g., network), where the nodes could be terms, and the edge a co-occurrence relationship among them (co-words network) [7, 8, 9]. In this network, the concept could be inferred using a community detection algorithm that looks for closely related terms that used to appear frequently together.

Thus, in this contribution, we perform a science mapping analysis to uncover the conceptual structure of the Federated Learning research field, with the aim to uncover what are the main themes covered in the field and their performance measures. To do that, the Web of Science was used as the data source.

Finally, this contribution is organized as follows: Section 2 explains the science mapping analysis methodology followed, Section 3 shows how the research field is delimited, and Section 4 shows the main results of the science mapping analysis. Finally, some conclusions are drawn in Section 5.

2. Methodology

Science mapping or bibliometric mapping is a spatial representation of how disciplines, fields, specialties, and documents or authors are related to one another [23]. It has been widely used to show and uncover the hidden key elements (documents, authors, institutions, topics, etc.) in different research fields [6, 10, 18, 21, 22].

Science mapping analysis can be performed with several software tools [8]. Particularly, SciMAT was presented in [9] as a powerful tool that integrates the majority of the advantages of the available science mapping software tools [8]. SciMAT was designed according to the science mapping analysis approach presented in [7], combining both performance analysis tools and science mapping tools to analyze a research field and detect and visualize its conceptual subdomains (particular topics/themes or general thematic areas) and its thematic evolution.

Therefore, in this contribution, SciMAT was employed to develop a co-words [5, 2] science mapping analysis [3, 8]. Thus, according to [7], the analysis is performed following a three stages methodology:

1. *Detection of the research themes*. The research themes are detected by applying a co-word analysis [5] to raw data for all the published documents in the research field, followed by a clustering of keywords to topics/themes [11], which locates keyword networks that are strongly linked to each other and that correspond to centres of interest or to research problems that are the subject of significant interest among researchers. The similarity between the keywords is assessed using the equivalence index [4].
2. *Visualizing research themes and thematic network*. In this phase, the detected themes are visualized by means of two different visualization instruments: strategic diagram [14] and thematic network [7]. Each theme can be characterized by two measures [4]: *centrality* and *density*. Centrality measures the degree of interaction of a network with other networks. On the other hand, density measures the internal strength of the network. Given both measures, a research field can be visualized as a set of research themes, mapped in a two-dimensional strategic diagram (Figure 1) and classified into four groups:

   (a) Themes in the upper-right quadrant are both well developed and important for the structure of the research field. They are known as the *motor-themes* of the specialty, given that they present strong centrality and high density.
   (b) Themes in the upper-left quadrant have well-developed internal ties but unimportant external ties and so, they are of only marginal importance for the field. These themes are very *specialized and peripheral*.

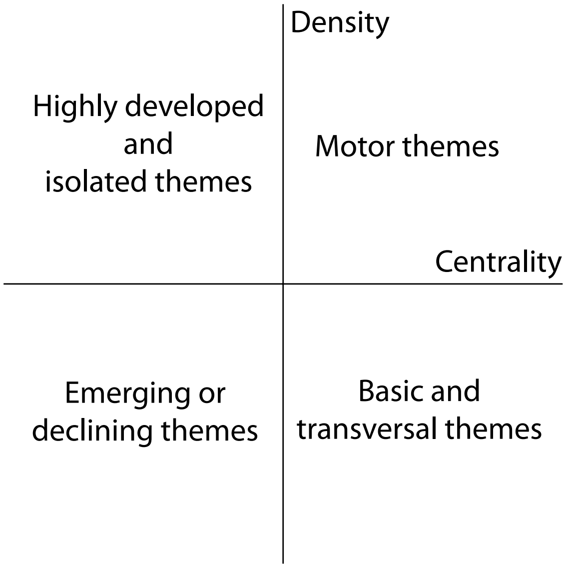


Fig. 1: The strategic diagram.

(c) Themes in the lower-left quadrant are both weakly developed and marginal. The themes in this quadrant have low density and low centrality and mainly represent either *emerging or disappearing* themes.
(d) Themes in the lower-right quadrant are important for a research field but are not developed. This quadrant contains *transversal and general*, basic themes.

3. *Performance analysis*. In this phase, the relative contribution of the research themes to the whole research field is measured (quantitatively and qualitatively) and used to establish the most prominent, most productive and highest-impact subfields. Some of the bibliometric indicators to use are: the number of published documents, number of citations, and different types of h-index [1, 15, 19]. For each theme, the performance measures are computed considering the documents associated with it. Thus, for instance, the h-index is computed using the citations of the theme's documents.

3. Dataset

In order to carry out the performance and science mapping analysis, the research documents related to the research field of Federated Learning must be collected and also, preprocessed. Since Web of Science is one of the most important bibliographic databases, we used it to delimit the Federated Learning research field, and retrieve the documents through the following advance query:

TS=("federated learning" OR "federated machine learning" OR "federated ML" OR "federated artificial intelligence" OR "federated AI" OR "federated intelligence" OR "federated training")

Thus, the query retrieved a total of 2444 documents from 2007 to 2022. Citations of these documents are also used in this study, thus, they were counted up to July 2022.

In Figure 2, the number of papers per year is shown. We can observe how from 2018 the research field started to grow exponentially, with 2021 accounting for twice times the articles of the whole previous period. Finally, since the dataset was collected in July 2022, we can see a small number of documents in this year, but with the exponential growth trend, probably at the end of 2022 will be more papers than those published in 2021.

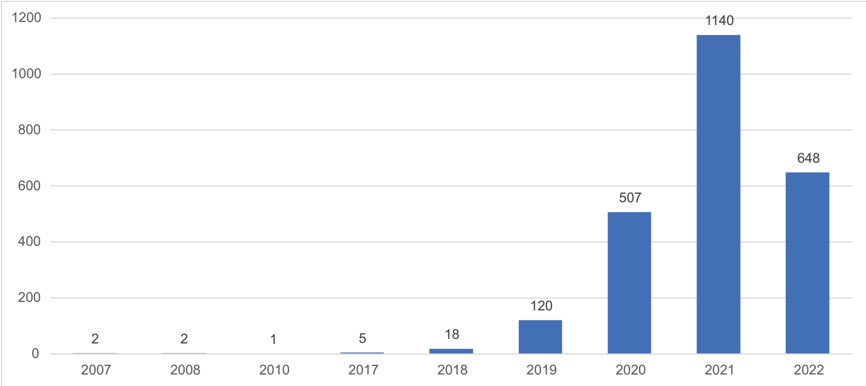


Fig. 2: Number of documents per year.

4. Conceptual Analysis

Federated Learning is a topic that has gained a lot of attention in recent years. It covers a great variety of techniques that come from different research areas of computer science and other related fields. In this section, we present the global conceptual structure of this field, showing the main topics in which the field is structured.

Thus, to show the most important themes of Federated Learning, a strategic diagram for the whole period (2007-2022) was made following the step explained in the methodology. We should remark that the volume of the spheres is proportional to the number of documents associated with each theme.

Table 1: Performance measures of the themes

Theme	#documents	#citations	h-index
DATA-MODELS	747	7,704	41
DEEP-LEARNING	714	7,915	41
BLOCKCHAIN	506	5,787	39
DIFFERENTIAL-PRIVACY	349	3,054	26
RESOURCE-ALLOCATION	340	4,947	35
REINFORCEMENT-LEARNING	204	3,085	27
PERFORMANCE-EVALUATION	181	1,632	22
DISTRIBUTED-LEARNING	175	5,546	18
INDUSTRIAL-INTERNET-OF-THINGS	163	1,767	22
COMPUTER-ARCHITECTURE	155	1,616	21
UNMANNED-AERIAL-VEHICLES	104	1,109	19
ATTACKS	98	1,329	15
OVER-THE-AIR-COMPUTATION	93	1,120	14
TRANSFER-LEARNING	74	4,387	14
WIRELESS-SENSOR-NETWORKS	63	480	11
MATRIX-FACTORIZATION	56	229	7
HOMOMORPHIC-ENCRYPTION	54	346	9
AUCTION	45	471	8
DEVICE-TO-DEVICE-COMMUNICATION	42	464	11
COVID-19	40	404	11
LONG-SHORT-TERM-MEMORY	34	103	6
INTELLIGENT-TRANSPORTATION-SYSTEMS	27	144	5
CONVERGENCE-RATE	22	147	7

According to the strategic diagram shown in Figure 3, the research on Federated learning is structured in twenty-three themes. In the strategic diagram, the upper-right quadrant contains the motor themes, that is, themes where the scientific community is paying a lot of effort, and also they act as support in the remaining themes. Moreover, the bottom-right quadrant contains other important themes that are the basis for the development of the field. Therefore, the Federating Learning research field is paying special attention the the following twelve themes (motor and basic themes): *data-models*, *resource-allocation*, *blockchain*, *over-the-air-computation*, *distributed-learning*, *deep-learning*, *unmanned-aerial-vehicles*, *wireless-sensor-networks*, *performance-evaluation*, *reinforcement-learning*, *differential-privacy*, *computer-architecture*.

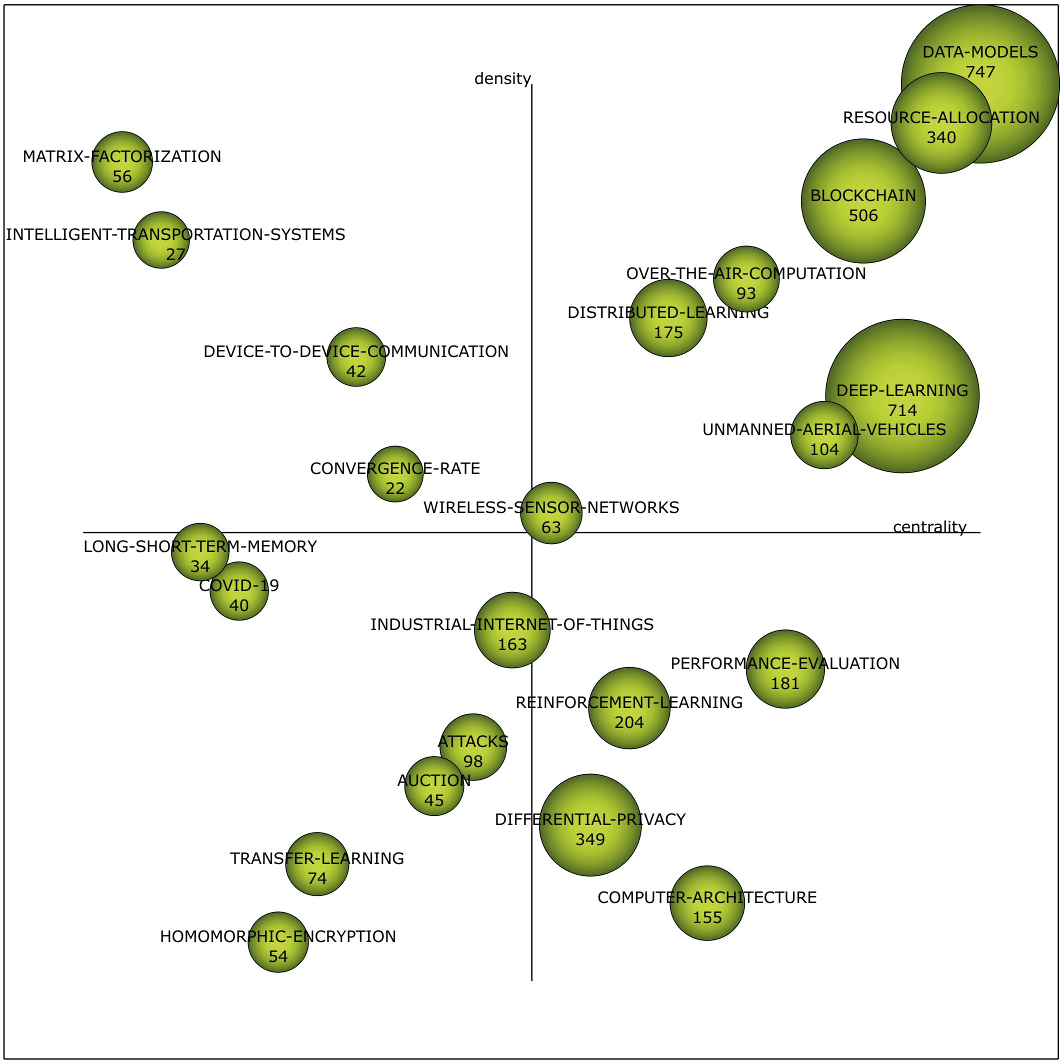
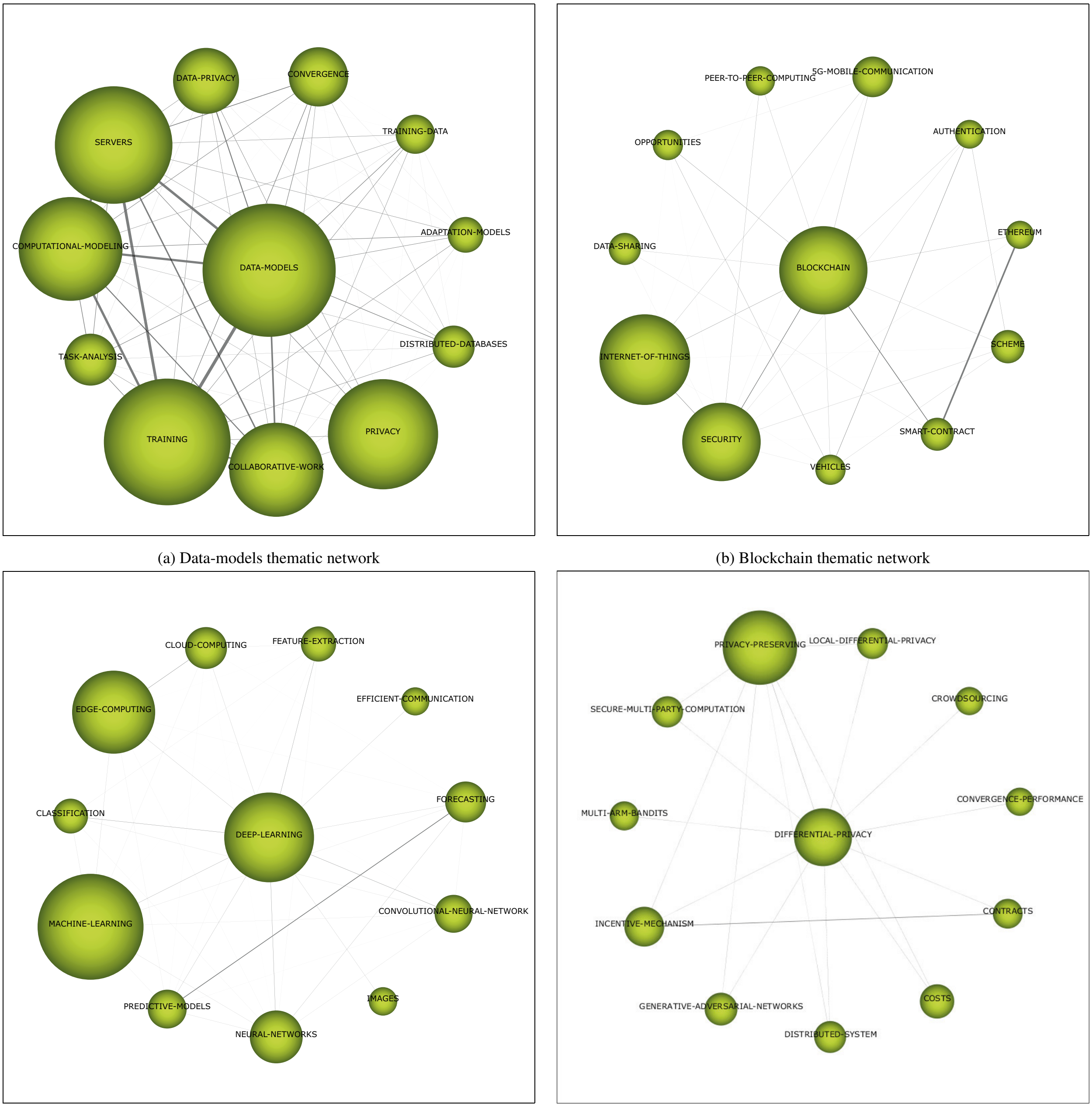


Fig. 3: Strategic diagram.

Furthermore, taking into account the performance measures shown in Table 1, we should highlight the themes *data-models*, *deep-learning* and *blockchain* since they are the ones with more production. Moreover, if we pay attention to the citation count, it could be observed that most of the themes are over 1k citations. That is, in general terms, the topics related to Federated Learning are demonstrating great attention by the global research community. Specifically, the themes *data-models* and *deep-learning* stand out with almost 8k citations. The themes *blockchain* and *distributed learning* also stand out reaching more than 5k citations.



(a) Data-models thematic network

(b) Blockchain thematic network

(c) Deep Learning thematic network

(d) Differential privacy thematic network.

Fig. 4: Thematic networks of some themes.

According to the structural measure (density and centrality), the theme *data-model* is the most important one, since it is laid out in the upper right corner. That is, the research community is making a lot of effort on that topic and is also used in the development of the others themes. Mainly, it delved in two aspects related to how to build and organize the data, and especially the training data, in servers, and distributed databases, meanwhile keeping the privacy of the data. The thematic network of this theme is shown in Figure 4a. We should remark that *data-models* is the second theme with the most citation count. Another important theme due to its structural measure is *resource-allocation* which is focused on how to manage the energy consumption to train the models in an efficient way, manage the resources and power needed and the wireless communication.

The motor theme *blockchain* focus on new models of authentication, and data sharing by means of smart contract over 5G or P2P communications models. It is mainly employed in the internet of things and vehicles. Regarding its performance measures, it is the third theme with the highest productivity and also the third one with the highest impact. Its thematic network is shown in Figure 4b.

Deep-learning is the theme with more academic impact, in fact, it is an important topic in the artificial intelligence and machine learning research field. Also, this theme is the second one in the number of published documents. Considering its structural measures, it is considered a motor theme. In the field of Federated Learning, deep learning is studied in order to develop new models that could be implemented in a federated way, to perform classifications or predictive models. Moreover, it delved into how to apply this type of machine learning model into cloud or edge computing. Its thematic network is shown in Figure 4c.

The basic theme *differential-privacy* (Figure 4d) deals with an important topic in the federated learning paradigms, that is, the privacy. Thus, the theme focuses on privacy-preserving, the security of multiparty computation, and local privacy. This theme reached more than 3k citations, and it is ranked fourth in documents production.

Regarding the application of Federated learning, we could remark the themes *unmanned-aerial-vehicles*, *industrial-internet-of-thing*, and *intelligent-transportation-systems*. The former is mainly related to drones and autonomous aerial vehicles, their communication with 5G and the necessity of low latency. Industrial IoT delved into cyber-physical systems, smart manufacturing, digital twins, and how to defend these devices from cyber-attacks. Moreover, this theme is also related to the theme *attack*. Lastly, the *intelligent-transportation-systems* aims to apply federated learning to identify the travel mode based on GPS trajectories, flow prediction, or license plate recognition, among others.

5. Conclusions

In this contribution, a science mapping analysis based on co-words of the Federated Learning research field has been performed. To do that, an advanced query was used to delimit the field in Web of Science, and therefore, retrieve all the published documents, including articles, reviews and also proceeding papers. Although in bibliometric analyses is not common to use proceeding papers, this research field is recent, and is moving on to the top conferences, therefore, in this contribution we consider also the knowledge published in the conferences. Once the corpus was cleaned (e.g. removing broad terms, and joining those terms that represent the same concept), a co-work network was modeled based on authors' keywords and ISI Keywords Plus. Finally, the whole network was splitted by means of a clustering algorithm, and the results were laid out in a strategic diagram.

Thus, according to the science mapping analysis carried out, we can conclude that the Federated Learning research field is conceptually structured in twenty-three themes. Among them, *data-models*, *resource-allocation*, *blockchain*, *over-the-air-computation*, *distributed-learning*, *deep-learning*, *unmanned-aerial-vehicles*, *wireless-sensor-networks*, *performance-evaluation*, *reinforcement-learning*, *differential-privacy*, *computer-architecture* stand out due to their structural measures (centrality and density). On the other hand, *data-models* and *deep-learning* have an outstanding citation impact, and the *data-models*, *deep-learning* and *blockchain* stand out due to their productivity.

Globally, the twenty-three themes could be grouped into six main areas: telecommunications (i.e. performance-evaluation, wireless-sensor-networks, and device-to-device-communications), privacy and security (i.e. differential-privacy, and homomorphic-encryption), computer architecture (i.e industrial-internet-of-things, over-the-air-computation, distributed-learning, etc.) and data modeling (i.e. data models), machine learning (i.e. deep-learning, long-short-term-memory, reinforcement-learning, etc.), and applications (i.e. unmanned-aerial-vehicles, industrial-internet-of-things, COVID-19, etc.).

Acknowledgements

This work has been supported by the project PID2019-105381GA-I00 (iScience) funded by MCIN / AEI / 10.13039/501100011033 and by the program of promotion and development of research activity of the University of Cádiz (2022-2023).

References

[1] Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F., 2009. h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics 3, 273–289. doi:10.1016/j.joi.2009.04.001.

[2] Batagelj, V., Cerinšek, M., 2013. On bibliographic networks. Scientometrics 96, 845–864.

[3] Börner, K., Chen, C., Boyack, K.W., 2003. Visualizing knowledge domains. Annual Review of Information Science and Technology 37, 179–255. doi:10.1002/aris.1440370106.

[4] Callon, M., Courtial, J.P., Laville, F., 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. Scientometrics 22, 155–205. doi:10.1007/BF02019280.

[5] Callon, M., Courtial, J.P., Turner, W.A., Bauin, S., 1983. From translations to problematic networks: An introduction to co-word analysis. Social Science Information 22, 191–235. doi:10.1177/053901883022002003.

[6] Cobo, M.J., López-Herrera, A.G., Herrera, F., Herrera-Viedma, E., 2012a. A Note on the ITS topic evolution in the period 2000-2009 at T-ITS. IEEE Transactions on Intelligent Transportation Systems 13, 413–420. doi:10.1109/TITS.2011.2167968.

[7] Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2011a. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. Journal of Informetrics 5, 146–166. doi:10.1016/j.joi.2010.10.002.

[8] Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2011b. Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for Information Science and Technology 62, 1382–1402. doi:10.1002/asi.21525.

[9] Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2012b. SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology 63, 1609–1630. doi:10.1002/asi.22688.

[10] Cobo, M.J., Martínez, M.A., Gutiérrez-Salcedo, M., Fujita, H., Herrera-Viedma, E., 2015. 25 years at Knowledge-Based Systems: A bibliometric analysis. Knowledge-Based Systems 80, 3–13. doi:10.1016/j.knosys.2014.12.035.

[11] Coulter, N., Monarch, I., Konda, S., 1998. Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science 49, 1206–1223. doi:10.1002/(SICI)1097-4571(1998)49:13<1206::AID-ASI7>3.3.CO;2-6.

[12] Encheva, S., Tumin, S., 2007. Cooperative learning objects in a federated learning system, in: Iskander, M. (Ed.), Innovations in e-learning, Instruction Technology, Assessment, and Engineering Education. doi:10.1007/978-1-4020-6262-9_63.

[13] Encheva, S., Tumin, S., 2008. On improving quality of the decision making process in a federated learning system, in: Luo, Y. (Ed.), Cooperative design, Visualization, and Engineering, Proceedings.

[14] He, Q., 1999. Knowledge Discovery Through Co-Word Analysis. Library Trends 48, 133–159.

[15] Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences 102, 16569–16572. doi:10.1073/pnas.0507655102.

[16] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., Eichner, H., Rouayheb, S.E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., Zhao, S., 2021. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning 14, 1–210. doi:10.1561/2200000083.

[17] Kawa, D., Punyani, S., Nayak, P., Karker1, A., Jyotinagar, V., 2019. Credit risk assessment from combined bank records using federated learning. International Research Journal of Engineering and Technology 6, 1355 – 1358.

[18] Martínez, M.A., Cobo, M.M.J., Herrera, M., Herrera-Viedma, E., 2015. Analyzing the Scientific Evolution of Social Work Using Science Mapping. Research on Social Work Practice 5, 257–277. doi:10.1177/1049731514522101.

[19] Martínez, M.A., Herrera, M., López-Gijón, J., Herrera-Viedma, E., 2014. H-Classics: Characterizing the concept of citation classics through H-index. Scientometrics 98, 1971–1983. doi:10.1007/s11192-013-1155-9.

[20] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y., 2016. Communication-efficient learning of deep networks from decentralized data doi:10.48550/ARXIV.1602.05629.

[21] Murgado-Armenteros, E.M., Gutiérrez-Salcedo, M., Torres-Ruiz, F.J., Cobo, M.J., 2015. Analysing the conceptual evolution of qualitative marketing research through science mapping analysis. Scientometrics 102, 519–557. doi:10.1007/s11192-014-1443-z.

[22] Rodriguez-Ledesma, A., Cobo, M.J., Lopez-Pujalte, C., Herrera-Viedma, E., 2015. An overview of animal science research 1945-2011 through science mapping analysis. Journal of Animal Breeding and Genetics 132, 475–497. doi:10.1111/jbg.12124.

[23] Small, H., 1999. Visualizing science by citation mapping. Journal of the American Society for Information Science 50, 799–813. doi:10.1002/(SICI)1097-4571(1999)50:9<799::AID-ASI9>3.0.CO;2-G.

[24] Tran, N.H., Bao, W., Zomaya, A., Nguyen, M.N.H., Hong, C.S., 2019. Federated learning over wireless networks: Optimization model design and analysis, in: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, pp. 1387–1395. doi:10.1109/INFOCOM.2019.8737464.