2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8329

Evolution and scientific visualization of Machine learning field

Río-Belver, Rosa[a]; Garechana, Gaizka[a] , Bildosola, Iñaki[a] and Zarrabeitia, Enara[a]
[a] Technology, Foresight and Management Research Group, Departamento de Organización de Empresas, Universidad del Pais Vasco UPV/EHU, Spain.

Abstract

This article provides a retrospective and understanding of the development of automatic learning methods. The beginnings are visualized as a discipline within Computer Sciences in the subcategory of Artificial Intelligence, its development and the current transfer of knowledge to other areas of Engineering and its industrial applications. Based on the publications about machine learning and its application contained in the Web of Science database, records from 1986 to 2017 are downloaded. After a description of the technological profile, a new approach is introduced to the classification of a discipline based on the year of appearance of those terms that define it. Mining of technological texts and network theory has been applied to extract the terms and interpret their evolution. They are the those that define the stages of emergence, development and maturation of the discipline Machine learning. The novelty of this approach lies in the technical nature of applied research in Machine Learning, which aims to be a guide for the development of future engineering applications and to make technology transfer to industry visible.

Keywords: *Machine Learning; Tech-Mining; Scientometrics; Social Network Analysis; Visualization; Bibliometrics*

1. Introduction

Industry 4.0 is generating an unprecedented revolution in the manufacturing sector, greatly favored by the Internet of Things (IoT), the growth of Big Data and the sensorization of machines. According to the OECD, Peña Lopez (2015), the new industrial revolution is understood as the incorporation of fundamentally digital technologies conducive to achieving the smart factory.

The application of automatic learning methods defined by Alpaydin (2014) to industrial production will be one of the pillars of the new revolution. Decision-making must be decentralized and productive systems should have the ability to make basic decisions and become as autonomous as possible.

Understanding the keys to the development of the machine learning discipline makes it possible to understand the transfer process from algorithm development in the laboratory to machine programming in industry. To study a scientific discipline we have to refer to methodologies developed by Alejo (2015), Garechana et al. (2014), Gore et al. (2016), Noyons (2009), Porter et al. (2011), which shows how text-mining methods, natural language processing and network theory collaborate to produce indicators. The algorithms developed within Machine Learning are also helping the development of bibliometrics, Ranei (2017).

This article is arranged in four sections. After the introduction, the second section describes both the methodology followed and the composition of the sample to be analyzed. Next, the analyses leading to the elaboration of the technological profile, the analysis of the field topics and the visualization of the networks are carried out, finishing with the conclusions and future lines of work.

2. Method

The method followed in the preparation of this study is explained through the block diagram shown in figure 1.

The data for the analysis of the scientific field of Machine Learning and its applications were obtained from the core collection of Web of Science database. WOS is an online platform belonging to the company Clarivate Analytics that provides access to the world's leading citation databases, with multidisciplinary information from over 18,000 high impact journals, over 180,000 conference proceedings, and over 80,000 books from around the world.

The Machine learning and applications (MLAA) data set has been obtained by retrieving the articles, conference proceedings and book chapters published from 1986 to 2017. The

concepts ("Machine learning") AND (Application*) have been searched in the fields Title, Abstract, Author Keywords and Keywords plus, detecting 14805 records that were downloaded in their full record format.

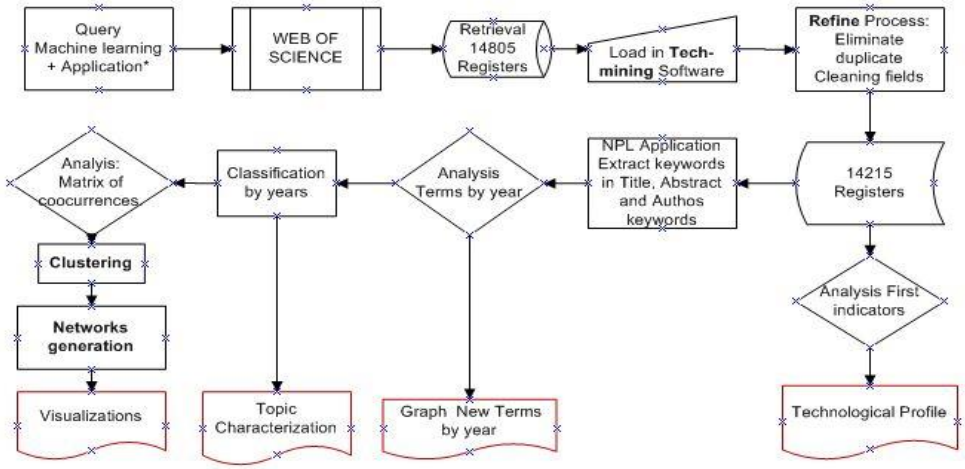

Figure 1. Method of analysis . Source: own elaboration(2018).

After the refining and cleaning processes, the records to be analyzed were reduced to 14215. To determine the life cycle of the machine learning discipline and define the temporal classification, the terms and phrases of the Title, Abstract and Author Keyword fields have been extracted using NPL. These Terms were refined and cleaned for further analysis.

A Term data set is created with the terms defined by the author themself, as they are usually ahead of the thesaurus classifications of the journals themselves. If the data extracted from the abstract and title fields is also added, it produces a good data set of terms that define the discipline.

Next, the analysis and detection of its first year of appearance is carried out, highlighting the evolution of the discipline. Once the temporality had been defined, three databases were created, one for each previously defined period.

For each sub-data set, autocorrelation maps of the Web of Science categories field were created to visualize the network structure and the strength of the connections between the nodes. The generated networks were visualized with the support of VosViewer software.

3. Analysis

3.1 Technological Profile

The countries with the highest number of publications in the field are the USA (4270), China (2002), UK (1124), Germany (914), India (820), Canada (593)...and so on. However, to analyze the technological profile focus should not be placed on the number of publications but rather on the most relevant ones in the field. To analyze the impact we have to study the number of citations of the articles.

In the case of Machine Learning and Application, a total of fifteen publications account for 49.54% of all citations. A single article in the sample receives 10390 citations, representing 12% of all citations. This is the article LIBSVM: A library for support vector machines written by Chang, C., & Lin, C. in Taiwan in 2011. Of the remaining fourteen, twelve are led and/or co-authored by the United States but we have to wait until 2007 to see highly cited publications led by other countries such as Spain, China and Taiwan.

The terms defined by the authors in the fifteen most cited articles of the analyzed discipline are as follows: NEURAL NETWORKS, TEXT CATEGORIZATION, SUPPORT VECTOR MACHINES, LANDSCAPE, MODELS, SPECIES DISTRIBUTION, GENE-EXPRESSION DATA, NEURAL-NETWORKS, COMPONENT ANALYSIS, K-MEANS ALGORITHM, PATTERN-RECOGNITION, HIDDEN MARKOV-MODELS

All highly cited articles belong to the WOS Computer Sciences Artificial Intelligence or Computer Sciences Information Systems Category.

3.2. Topic characterization

Text mining allows us to apply text classification to solve the categorization problems of a discipline. The most representative terms are extracted from the keywords defined by the author themself, to which the words and phrases identified in the title and abstract fields are added. The natural language processing application of Vantage point data mining software is used for this purpose. Once cleaned using fuzzy filters, 22181 terms are available. This number is reduced to 2612 by discarding the terms whose frequency of appearance is less than 3 in order to apply a macro that determines the first year that the term appeared.

As shown in Figure 2, most of the terms are used for the first time between 1997 and 2017, with a peak generation of 203 new terms in 2009, after which time the generation of new terms stagnated and began to decline in 2014. However, the number of records that include these terms has grown since 2000 and in 2016 the maximum of the series is published, at 2356 records.

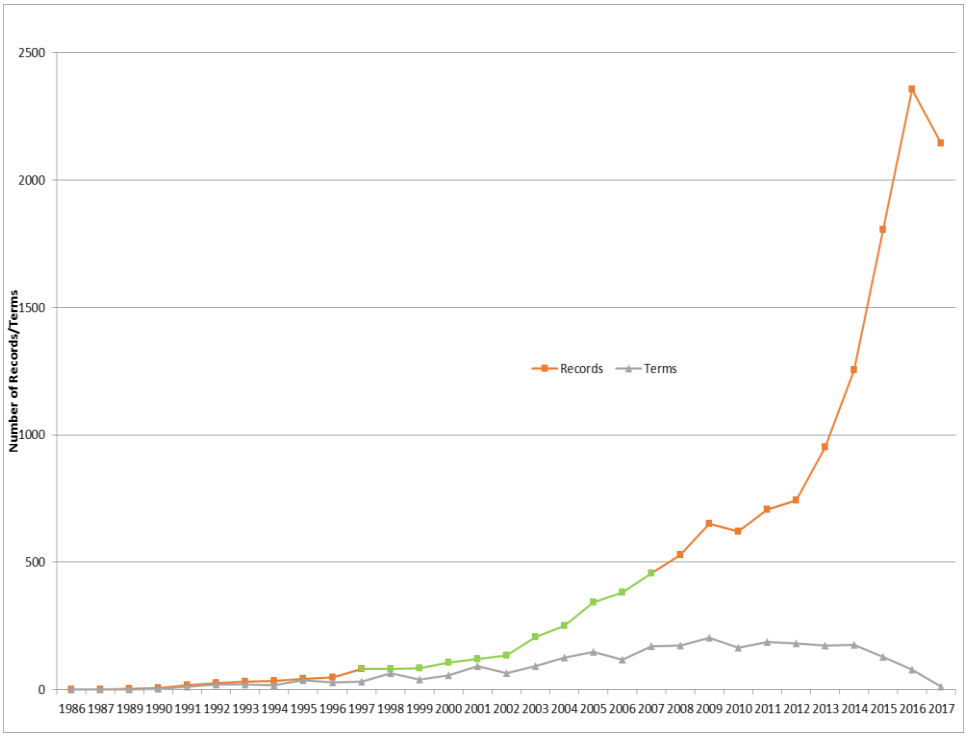

Figure 2. Number of new Author Keywords any year versus the number of records of that year. Source: own elaboration(2018).

If we carry out an analysis of the new terms that arise every year we can say that in 1990 the four terms that appear for the first time are: Machine learning (later repeated in Author keyword field 4236 times), Expert system (38), knowledge acquisition (14) and Knowledge base (14).

In 1997, 32 new terms appeared for the first time in the author's words, such as Data mining (515), Regression (72), Evolutionary algorithm (45), Rule extraction (14), Graph theory (14), fuzzy clustering (12), times series prediction (11), Fuzzy system (10) or Neural net (9). Later in 2007, 169 new terms were generated in the author's words such as: wireless sensor network (50), affective computing (37), Machine Supervised Learning (34) artificial neural network (ANN) (20), machine learning application (19). Finally, in 2017, only 11 new terms were generated, including Landslides (5), Precision medicine (5), and Age prediction (3). Based on the development of the terms, the evolution of the Machine Learning field is divided into three stages: Emergence stage from 1986 to 1996, development stage from 1997 to 2006 and maturation stage from 2007 to 2017. As the technology matures into its own field, the number of new terms each year is shrinked.

3.3 Networks and visualizations

It is considered appropriate to approach the scientific field through the use of the categories assigned by the Web of Science (WOS) to publications (papers, proceedings or book chapters), Leydersdoff (2013). All books and journals included in the Web of Science Core Collection, the leading provider of scientific and technological publications, which includes references to leading scientific publications in any discipline of knowledge since 1945, are assigned at least one of the 242 subject categories predefined by Clarivate Analytics. This makes it possible to determine the scientific classification of the document.

On the other hand, technology maps have the potential to become fundamental tools in science policy planning and therefore in the innovative development of a country. However, their interpretation is difficult because they are complex structures whose representation is difficult to interpret. In figure 3, on the right side, we can see the tentacles of the category called Computer science Artificial Intelligence, the scientific category father of machine learning for the period 1986-1996. In the upper left corner we can see a network composed of nodes, WOS categories, which are connected by means of edges. The strength of the line represents the number of records in the line, the stronger the representation the more common records between the nodes. In the period 1986-1996, there were 205 publications categorized into 66 items and 132 edges. This is a relatively low number and, as can be seen, the main collaborations are carried out between the same science; Computer Sciences Artificial, C.S. information, C. S. interdisciplinary, CS Cybernetics, although there are tenuous connections with Information Sciences, Automation Control systems and Electrical Engineering. This is an EMERGENCE STAGE, where science is developing and focusing on itself.

In the subsequent period 1997-2006, the central part of Figure 3, is seen as the network grows and doubles the number of nodes, WOS Categories , reaching 133. The records from this period date back to 1788, so that more connections and edges are generated (563). Computer Sciences Artificial Intelligence maintains its central position in the network, however its relationships are extended to Medical, Bioinformatics, Biotechnology, Imaging Science photographic, Business Finance, Neuroscience, Biochemical Research methods,… We can define it as a DEVELOPMENT STAGE. Keywords at this time include terms such as: supervised learning; Bayesian decision theory; parametric, semi-parametric, and nonparametric methods; multivariate analysis; hidden Markov models; reinforcement learning; kernel machines; graphical models; Bayesian estimation; and statistical testing.

Finally, in the last ten years, 2007-2017, the network has become too extensive. This is a MATURATION STAGE where most of the 12222 records are generated and the nodes almost double again, 216 nodes and 1295 edges. The area expands to almost all WOS categories (216/242). The graph on the right shows how its position has changed, Computer Sciences Artificial Intelligence has lost its central position in the network, no longer

domineering the game but remaining as a base. Areas of major applicability such as Industrial Engineering, Electrical Engineering, Robotics, Material Sciences, Nanosciences, Biomedical Engineering, Optics, Instrumentation... show their clear progress in the life cycle of science. The following figures can be see better in https://doi.org/10.6084/m9.figshare.6302039.

4. Conclusion and future work

The application of text mining techniques combined with visualizations allows us to understand and interpret the evolution of a scientific discipline. Machine learning was born in the heart of Computer Sciences as a subdiscipline of Artificial Intelligence and has few links with other areas. From 1997 to 2006 it began to grow and branch out, connecting other areas of CS. From 2007 to 2017 we can see how the CS category branches out, goes beyond its own scope and expands into areas of applied techniques. In the future, it will become cross-cutting knowledge, as using examples from past experiences has been the basis for problem solving. The change is driven by the generation of large amounts of data on past experience and the high capacity of processors to process them and therefore the generation of solutions and automatic outputs that move the system forward.

For the generation of new terms the number of publications increased until 2009, when it stagnated and began to decrease, however, this is the period of greatest number of publications due to the strong appearance of Industrial Engineering and related areas. As a future extension of this study, WOS data will be combined with patent databases and the flows generated through the non-patent literature collected in the industrial property registers will be analyzed. The aim is to highlight, from various points of view, the current incorporation of Machine Learning and its collaboration in the Industrial Organization 4.0.

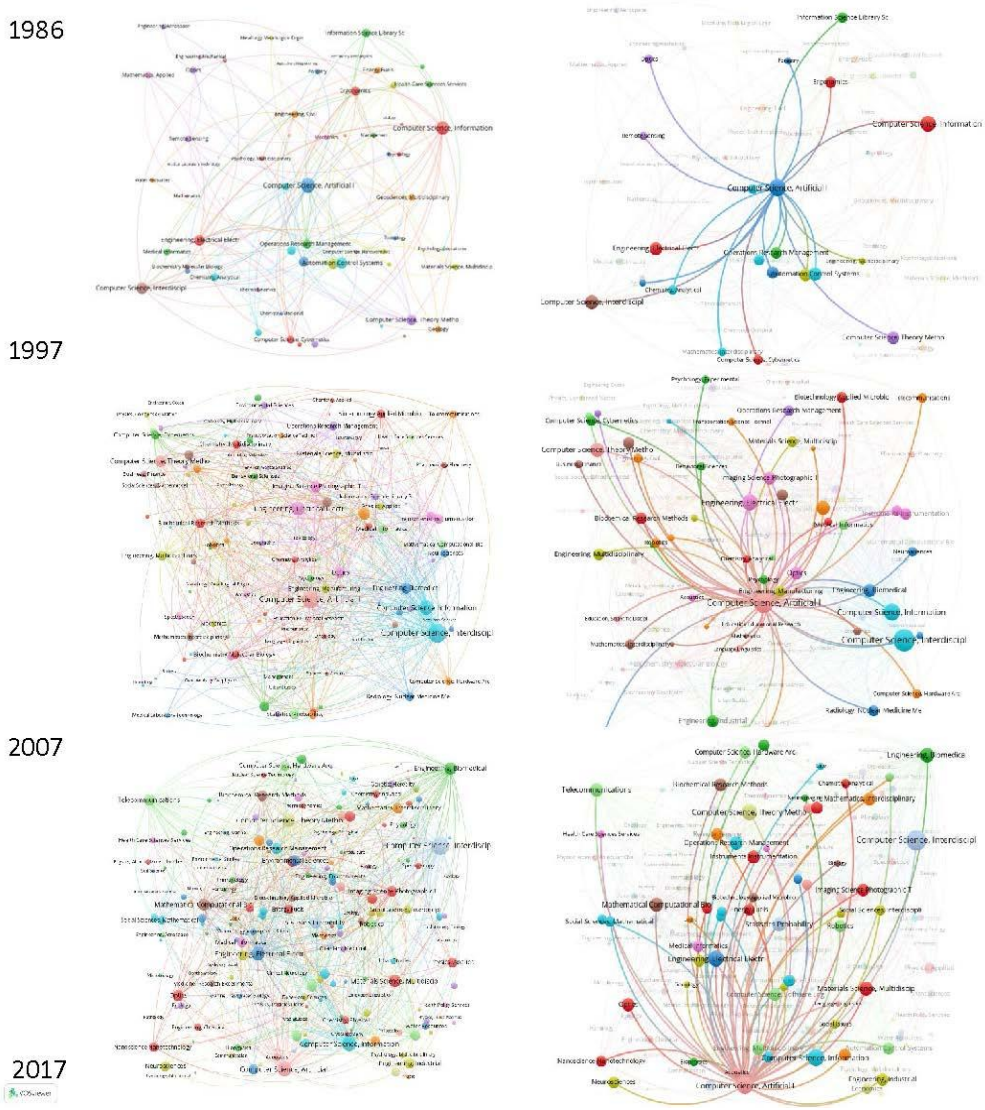

Figure 3. Machine learning evolution in the Web Science Categories. Source: own elaboration(2018).

References

Alejo-Machado, O. J., Manuel Fernandez-Luna, J., & Huete, J. F. (2015). Bibliometric study of the scientific research on "learning to rank" between 2000 and 2013. Scientometrics, 102(2), 1669-1686.

Alpaydin, E. (2014). Introduction to machine learning. Cambridge: The MIT Press.

Garechana, G., Rio-Belver, R., Cilleruelo, E., & Larruscain J. (2014). Clusterization and mapping of waste recycling science. evolution of research from 2002 to 2012. Journal of the Association for Information Science and Technology, 66, 1431-1446.

Gore, R., Diallo, S., & Padilla, J. (2016). Classifying modeling and simulation as a scientific discipline. Scientometrics, 109(2), 615-628.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new web-of-science categories. Scientometrics, 94(2), 589-593.

Noyons, E. C. M., & Calero-Medina, C. (2009). Applying bibliometric mapping in a high level science policy context. Scientometrics, 79(2), 261-275.

Peña-López, I. (2015). OECD digital economy outlook 2015. On line.

Porter, A. L., Guo, Y., & Chiavatta, D. (2011). Tech mining: Text mining and visualization tools, as applied to nanoenhanced solar cells. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 1(2), 172-181.

Ranaei, S., & Suominen, A. (2017). Using machine learning approaches to identify emergence: Case of vehicle related patent data. 2017 Portland International Conference on Management of Engineering and Technology (Picmet),1-8.