RESEARCH ARTICLE

Exploring machine learning: A bibliometric general approach using Citespace [version 1; peer review: 1 approved, 1 approved with reservations]

Juan Rincon-Patino, Gustavo Ramirez-Gonzalez, Juan Carlos Corrales

Telematic Engineering Department, University of Cauca, Popayán, Cauca, 190001, Colombia

v1 **First published:** 10 Aug 2018, **7**:1240
https://doi.org/10.12688/f1000research.15619.1
Latest published: 10 Aug 2018, **7**:1240
https://doi.org/10.12688/f1000research.15619.1

Abstract

Background: Machine learning researches algorithms that allow a machine to learn about resolving problems in different application domains. Due to the wide number of machine learning applications, it is necessary for newcomers to the field to have alternatives to explore this field faster.
Methods: In this paper, we present a science mapping analysis on the machine learning research in the period 2007-2017. This study was develop using the CiteSpace tool based on results from Clarivate Web of Science. This analysis shows how the field has evolved, by highlighting the most notable authors, institutions, keywords, countries, categories, and journals.
Results: The results provide information on trends and possibilities in the near future, particularly in areas such as health, biology and banking, where machine learning is a valuable tool to generate solutions.
Conclusions: Machine learning is being widely studied, and several institutions in countries like the USA and China constantly generate machine learning based solutions. Diseases, such as cancer or Alzheimer's disease, studies in biology, such as the protein molecule, virtual reality, commerce, smartphones, and ubiquitous computing, are all fields where machine learning contributes to resolving problems.

Keywords

machine learning, science mapping, bibliometrics, topic analysis, citeSpace

This article is included in the Artificial Intelligence and Machine Learning gateway.

Open Peer Review

Approval Status ✔ ?

	1	2
version 1 10 Aug 2018	✔ view	? view

1. **Sally Ellingson**, University of Kentucky, Lexington, USA

2. **Chaomei Chen**, Drexel University, Philadelphia, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Research on Research, Policy & Culture gateway.

Corresponding author: Gustavo Ramirez-Gonzalez (gramirez@unicauca.edu.co)

Author roles: Rincon-Patino J: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Ramirez-Gonzalez G**: Conceptualization, Formal Analysis, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Corrales JC**: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The authors are grateful to the Telematics Engineering Group (GIT) of the University of Cauca for scientific support and Innovacción Cauca project for master's scholarship granted to J. Rincon-Patino.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Rincon-Patino J, Ramirez-Gonzalez G and Corrales JC. **Exploring machine learning: A bibliometric general approach using Citespace [version 1; peer review: 1 approved, 1 approved with reservations]** F1000Research 2018, **7**:1240 https://doi.org/10.12688/f1000research.15619.1

First published: 10 Aug 2018, **7**:1240 https://doi.org/10.12688/f1000research.15619.1

Introduction

Machine learning is a computer science field that studies the learning processes of humans and replicates themusing machines. Different algorithms allow a machine to learn and use the acquired knowledge to resolve several problems that society faces. This field is widely studied and there exists a huge number of articles that present machine learning applications. Consequently, in the present study, we seek to create a generic map about machine learning applications, which allows newcomers to know the fields that are being explored and use machine learning techniques. In this study, we carried out a science mapping analysis of the existing research on machine learning. As a starting point, we find that bibliometrics is a relevant tool to analyze academic research developed on different topics. Bibliometric analyses contribute to the progress of science in many different ways[1], for example, by allowing evaluation of progress to be made, identifying trustworthy sources of scientific publications, laying the academic foundation for assessing new developments, or identifying major scientific actors. Performance analysis and science mapping are two bibliometric approaches used to explore a research field[2]. While performance analysis is an interesting way to evaluate the impact of published papers, based on their citations, science mapping aims at exhibiting the structure of scientific research, showing its evolution and dynamical aspects[3].

The present study performs a science mapping analysis; however, this is not the only approach to discover tendencies or to give an overview of a topic. We can find existing literature reviews on specific machine learning topics such as algorithms[4], applications into visual analytics[5], and recommendation systems[6]. There are other reviews on applications for different fields, such as medical diagnosis[7], radiation oncology[8], semantic web[9], models for quality prediction[10] and methods for text categorization[11]. Also, it was possible to find a general review on machine learning[12], but without a science mapping analysis, as this study performs. In 3 we find a bibliometric analysis related to machine learning, but this work only focuses on reviewing the state of the research carried out by the journal Knowledge-Based Systems (KnoSys) from 1991 to 2014. 13 and 14 use this method in the medical field, while 15 carries out an analysis in the social work area and 16 in the intelligent transportation systems research. Furthermore, there are other approaches and important analyses for providing an overview of a topic or finding its trends, using text mining or Latent Dirichlet allocation, such as in 17 and 18, among others.

This article has the following structure: In the *Methods* section, we describe the methodology, the dataset extracted, the tool configuration, and how the analysis was performed. The *Results* section presents the results of the science mapping analysis. The conclusions are given at the end of the article.

Methods

Dataset for visualization analysis

We used Web of Science (WOS) Core Collection. This is one of the primary databases for scientific literature in the scientific world. We looked, in the third quarter of 2017, for papers and conferences about machine learning, using that concept as a keyword ('machine AND learning'), with results ranging from 2007 to 2017 Q2 (published papers up the second quarter of the year). We used the 'All databases' option to have a complete results list. Finally, the results were sorted by date. All the articles, between 2007 and 2017, were taken into account for performing the analysis with the aim of obtaining a general vision of the field.

We obtained 41,962 records from WOS Core Collection that were downloaded as plain text including the full record and cited references. The files were named as 'download' with .txt as the file extension. Figure 1 shows a summary of the records.

Parameter design

In CiteSpace version 5.1.R8 SE[19–21], we used the records from WOS database and set a time slicing from 2007–2017, using one year per slice and the default Citespace configuration in term type, links and selection criteria options. We also used the title, abstract, author keywords and keyword plus as term sources. We changed the size of the generated network to fit the graphs, so we reduced the number of documents that were part of the top cited ones on each slice. The Top N configured for the networks are presented below each figure.

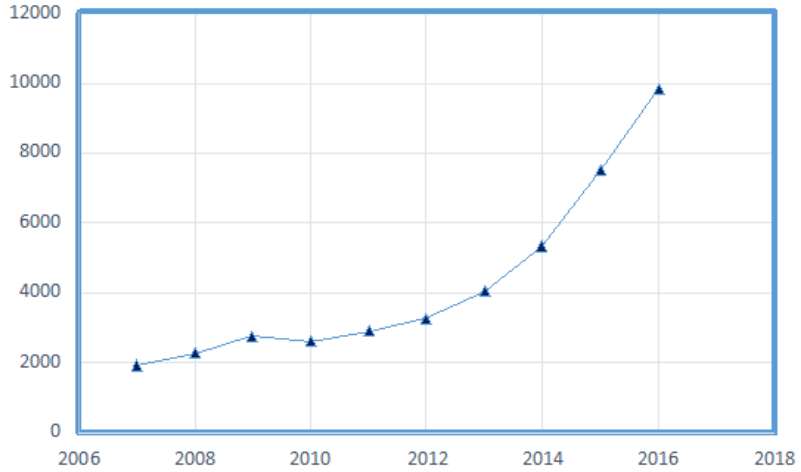

Figure 1. Number of documents from Web of Science Core Collection per year 2007–2017(Q2).

Analysis design

CiteSpace allows us to detect and visualize emerging trends and transient patterns in the scientific literature[20]; for this purpose, we applied three types of bibliometric techniques as in 22. First, co-author analysis, which investigates leading authors that are cited together[23]. It uses the authors' names, affiliation countries and institutions as units of analysis and then it shows the author, institution and country co-occurrences. Second, co-word analysis to establish links between documents[24], through keyword and category co-occurrences. Third, co-citation analysis that provides, as a result, the cited author, cited-reference and cited journal co-occurrences.

Results

A co-authorship analysis was done to explore the authors who have the greatest bibliographic production in the field of machine learning. Figure 2 shows the resulting network. The network has 301 nodes and 336 links. Each node represents an author, and its width indicates the number of author's publications proportionally. The connections between the nodes represent co-authorship of papers and their width suggests the proportion of the cooperative relationships. Finally, the different colors of the nodes and links represent the years between 2007 and 2017(Q2). From Figure 2, following a precise analysis supported in CiteSpace and without an additional analysis of duplicates, it can be highlighted that Wang Y, Zhang Y, Liu Y and Zhang L are the authors that have published the highest number of papers on machine learning.

After the previous co-authorship analysis, it was relevant to study the authors' institutions and countries. Figure 3 shows a network with the leading countries in which machine learning is an important subject of study, and the relationships between them. The network has 23 nodes and 85 links. From Figure 3, we can observe that the United States of America (USA) is the most productive country, followed by the People's Republic of China, Germany, and England. Regarding the distribution, 24,761 papers correspond to the USA, 10,808 to China, 4,479 to Germany, 4,365 to England, 3,866 to India, 3,407 to Spain and 3,045 to Canada. The nodes with the highest centrality, as indicated by purple rings, suggest that the USA plays a major role in machine learning research with authors from other countries, followed by Canada, England, Brazil and Australia. The centrality

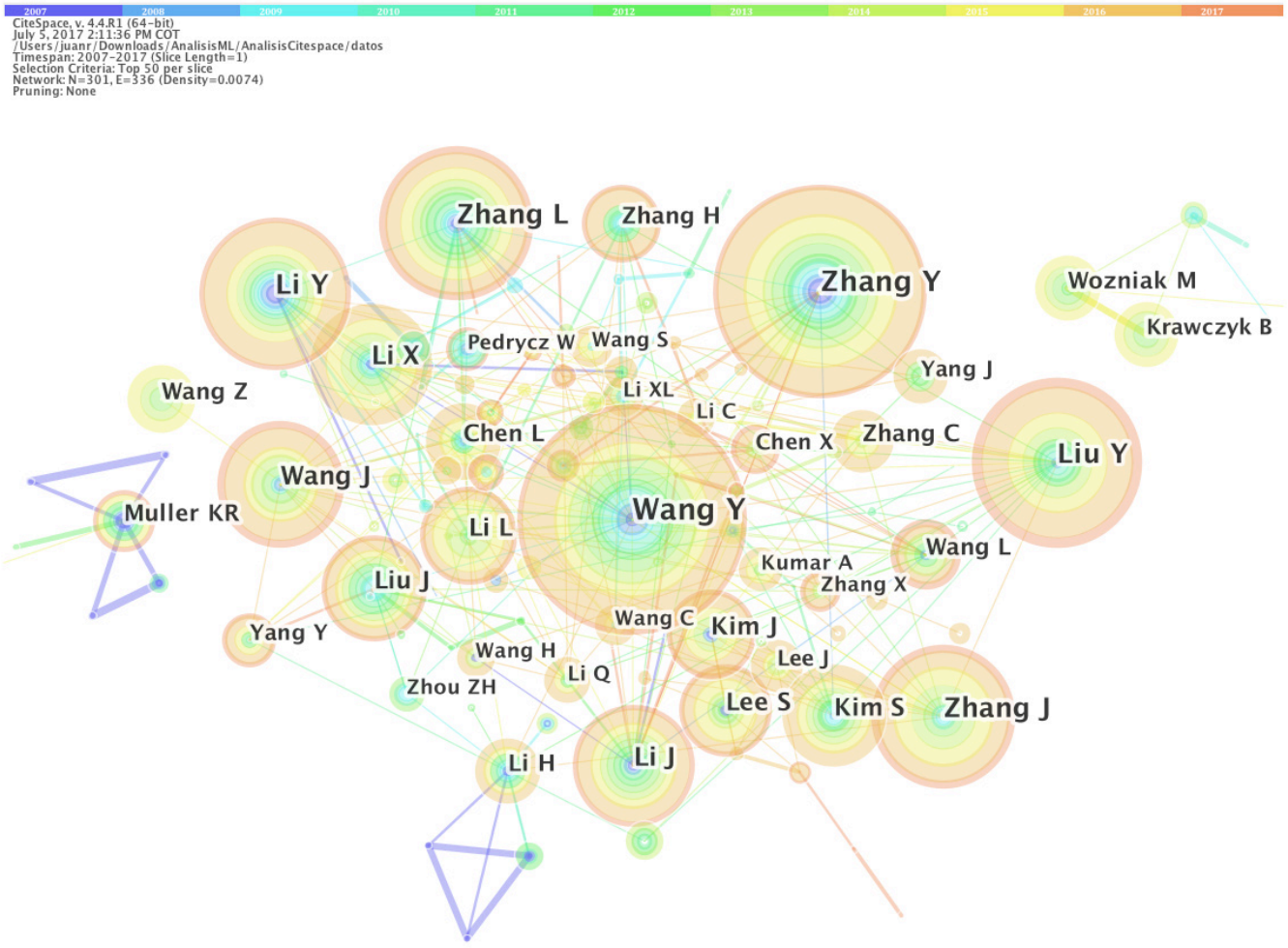

Figure 2. Co-authorship network for machine learning 2007–2017(Q2) from Web of Science Core Collection.

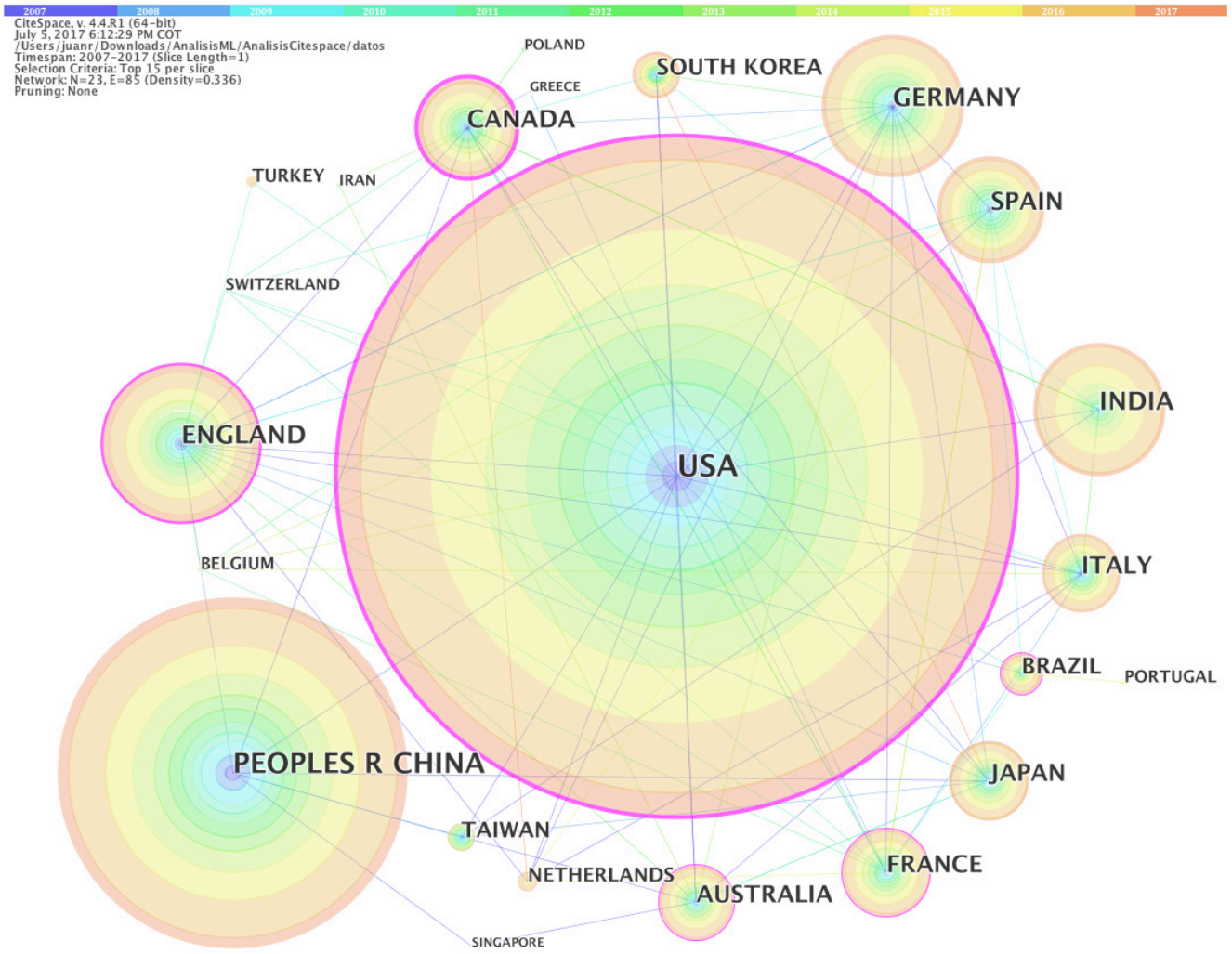

Figure 3. Countries network for machine learning 2007–2017(Q2) from Web of Science Core Collection.

of these nodes is 0.44 for the USA, 0.42 for Canada, 0.23 for England, 0.18 for Brazil and 0.16 for Australia.

Figure 4 shows the institutions' network, which presents the organizations with the highest production of articles on machine learning. The network has 54 nodes and 159 links. The Chinese Academy of Sciences, Carnegie Mellon University, Stanford University, Massachusetts Institute of Technology, Nanyang Technological University, University of California and Harvard University are part of the institutions that have published the largest number of articles. Additionally, Harvard University (0.17), Stanford University (0.12), Massachusetts Institute of Technology (0.12) and Columbia University (0.11) have the highest centrality, which means that they occupy key positions on the relevant paths in machine learning research.

To find the main subjects of the publications and, due to the fact that during the last decade the topics in machine learning research may have changed, a co-category analysis was performed. We did a preliminary analysis, using the categories generated by WOS, as shown in Table 1.

COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE (12,594, 30.013%) and ENGINEERING ELECTRICAL-ELECTRONIC (10,715, 25.535%) are the two categories that have the highest number of publications, followed by COMPUTER SCIENCE THEORY METHODS, COMPUTER SCIENCE INFORMATION SYSTEMS and COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS. Out of all these categories, we conclude that COMPUTER SCIENCE (and its sub-categories) is the leading one. Apart from this category, other relevant fields for research in machine learning may be biology, telecommunications and automation control systems.

To perform a deeper analysis, we built a network of co-occurring subject categories, as shown in Figure 5. The resulting network has 27 nodes and 80 links. COMPUTER SCIENCE - INTERDISCIPLINARY APPLICATIONS (0.47), COMPUTER SCIENCE

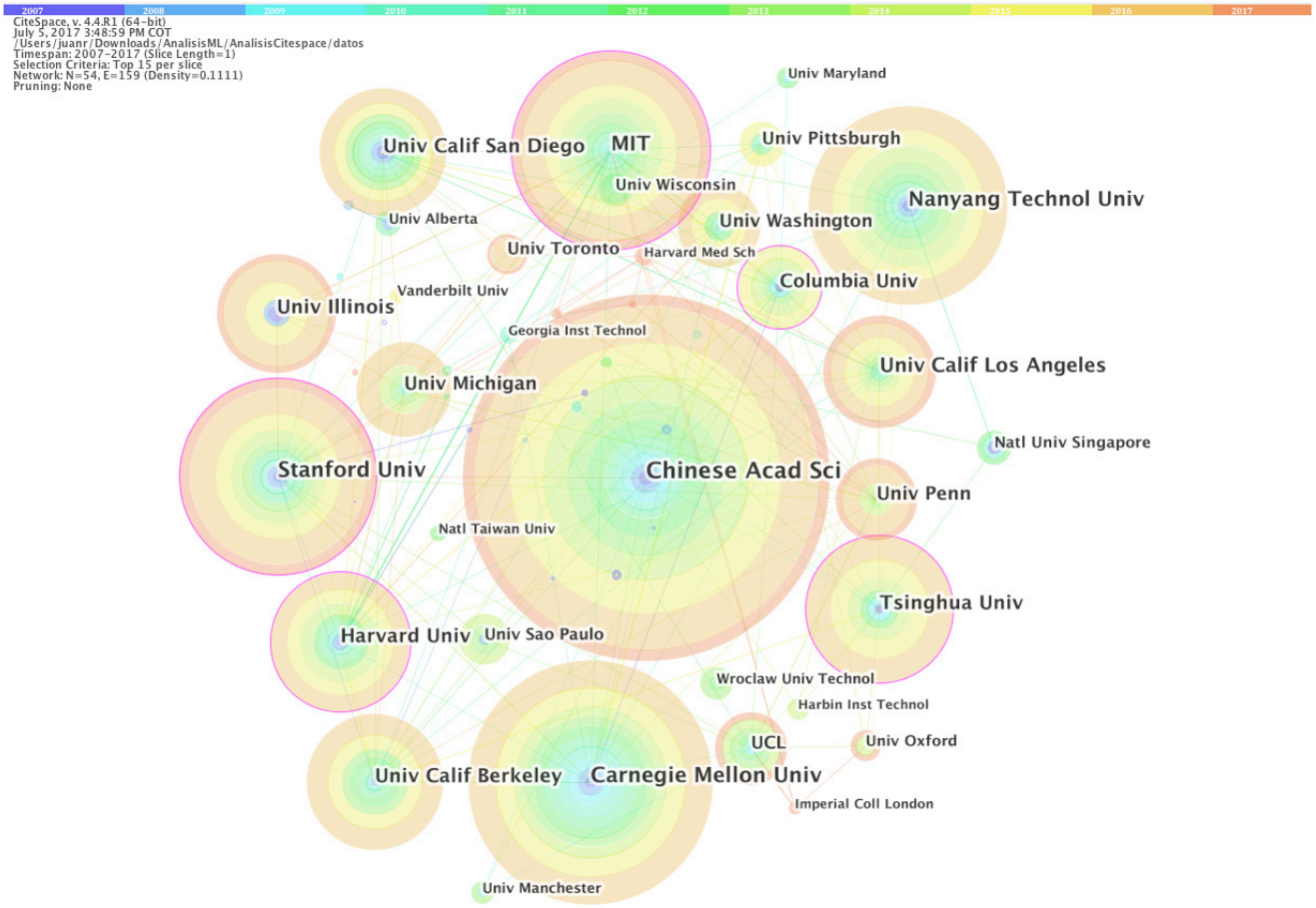

Figure 4. Institutions network for machine learning 2007–2017(Q2) from Web of Science Core Collection.

Table 1. Top 10 research fields from major publications in Web of Science Core Collection 2007–2017(Q2).

Category	Total	%
COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE	12,594	30.013
ENGINEERING ELECTRICAL-ELECTRONIC	10,715	25.535
COMPUTER SCIENCE THEORY METHODS	8,202	19.546
COMPUTER SCIENCE INFORMATION SYSTEMS	6,925	16.503
COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS	4,791	11.417
COMPUTER SCIENCE SOFTWARE ENGINEERING	2,433	5.798
MATHEMATICAL COMPUTATIONAL BIOLOGY	2,216	5.281
TELECOMMUNICATIONS	2,215	5.279
COMPUTER SCIENCE HARDWARE ARCHITECTURE	1,933	4.607
AUTOMATION CONTROL SYSTEMS	1,649	3.930

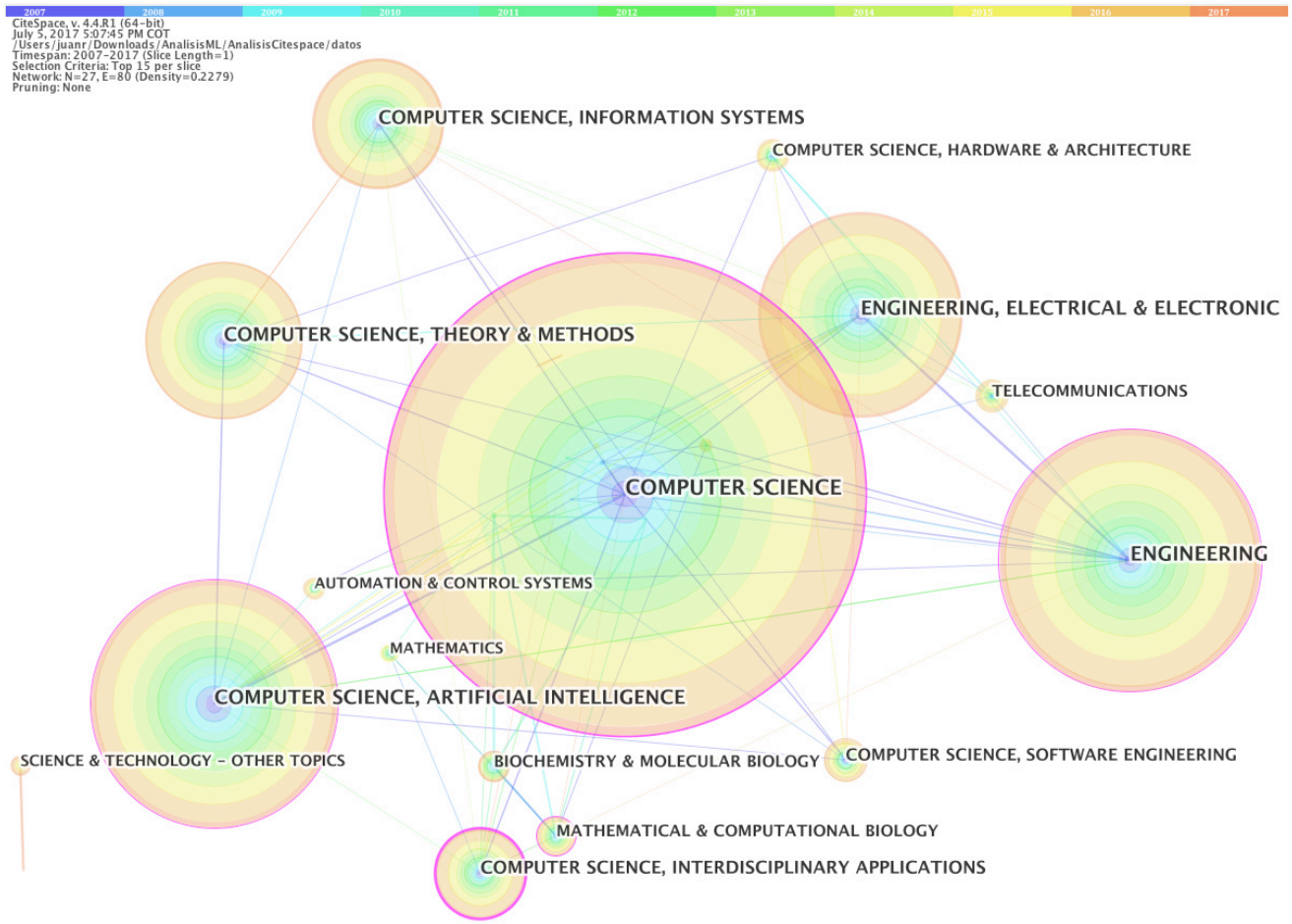

Figure 5. Network of co-occurring subject categories for machine learning 2007–2017(Q2) from Web of Science Core Collection.

(0.37), ENGINEERING (0.20) and MATHEMATICAL & COMPUTATIONAL BIOLOGY (0.18) are the nodes with the highest centrality, suggesting that they are the main topics that link machine learning studies carried out on different periods. We could find that COMPUTER SCIENCE - INTERDISCIPLINARY APPLICATIONS, due to its centrality value, is a relevant category between the other concepts. This means it can be the basis of future works.

A keyword analysis allows us to observe emerging trends, since it provides information on the content of articles published on the subject. For this purpose, we constructed several networks of co-occurring keywords. First, we built a network with N=15, where N is the size of the top cited or occurred items from each slice (one year in this case). Figure 6 presents the resulting network, and has 23 nodes and 88 links. It is important to remember that each node in the network has several rings around it, and their colors refer to the years in which that keyword appears.

The most important keywords appearing in Figure 6, as ordered by their citation counts, are classification (5,546), support vector machine (3,347), algorithm (2,681) and neural network (2,450), followed by model (2,253), system (1,898), prediction (1,893), feature selection (1,559), data mining (1,282) and network (1,196). By their centrality, the main keywords are classification (0.56), support vector machine (0.18), pattern recognition (0.17) and neural network (0.10). From these keywords, we can observe that the classification algorithms, such as support vector machine, have been widely studied and represent an important intellectual turning point, acting as bridges that link concepts over different periods. We can find all the concepts connected to this main node. Other relevant algorithms are the ones used for regression purposes, such as neural networks, and the ones used for grouping purposes, such as k-nearest neighbors.

Second, a network of co-occurring keywords with N=50 was constructed, the resulting net being shown in Figure 7, with 95 nodes and 420 links. The keyword with the highest citation count appearing in the network is classification, with 5,546 citations, followed by support vector machine (3,347), algorithm (2,681), neural network (2,450), model (2,253), system (1,898), prediction (1,893), feature selection (1,559), data mining (1,335), network (1,304), recognition (1,283), regression (1,110), artificial neural network (1,048), random forest (971), identification (966), selection (935), optimization (853), classifier (818), genetic algorithm (743) and decision tree (675). This network highlights

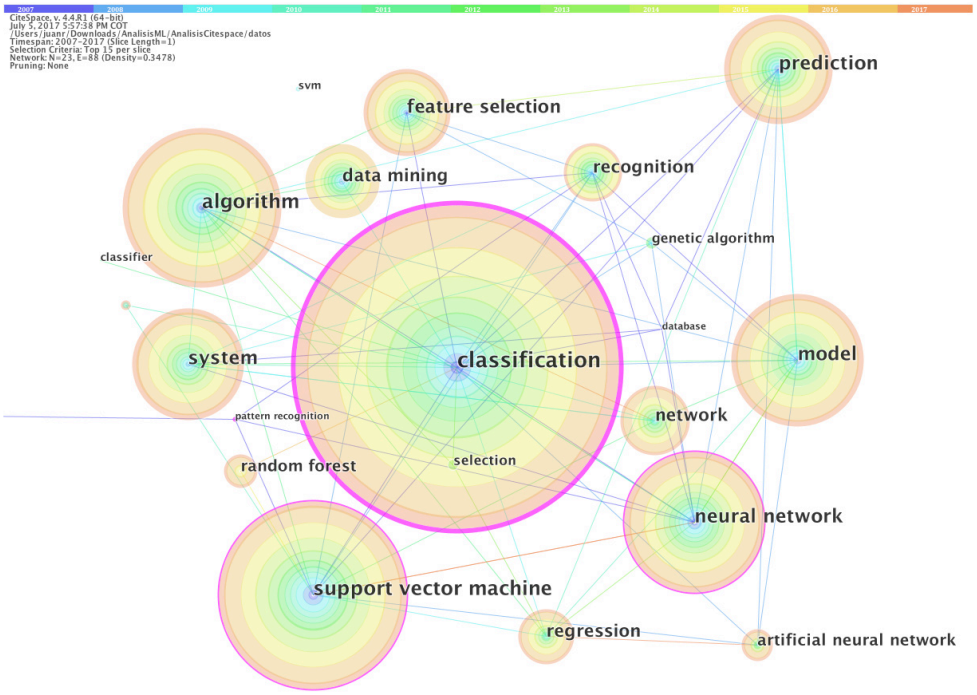

Figure 6. Network of co-occurring keywords with N=15, for machine learning 2007–2017(Q2) from Web of Science Core Collection.

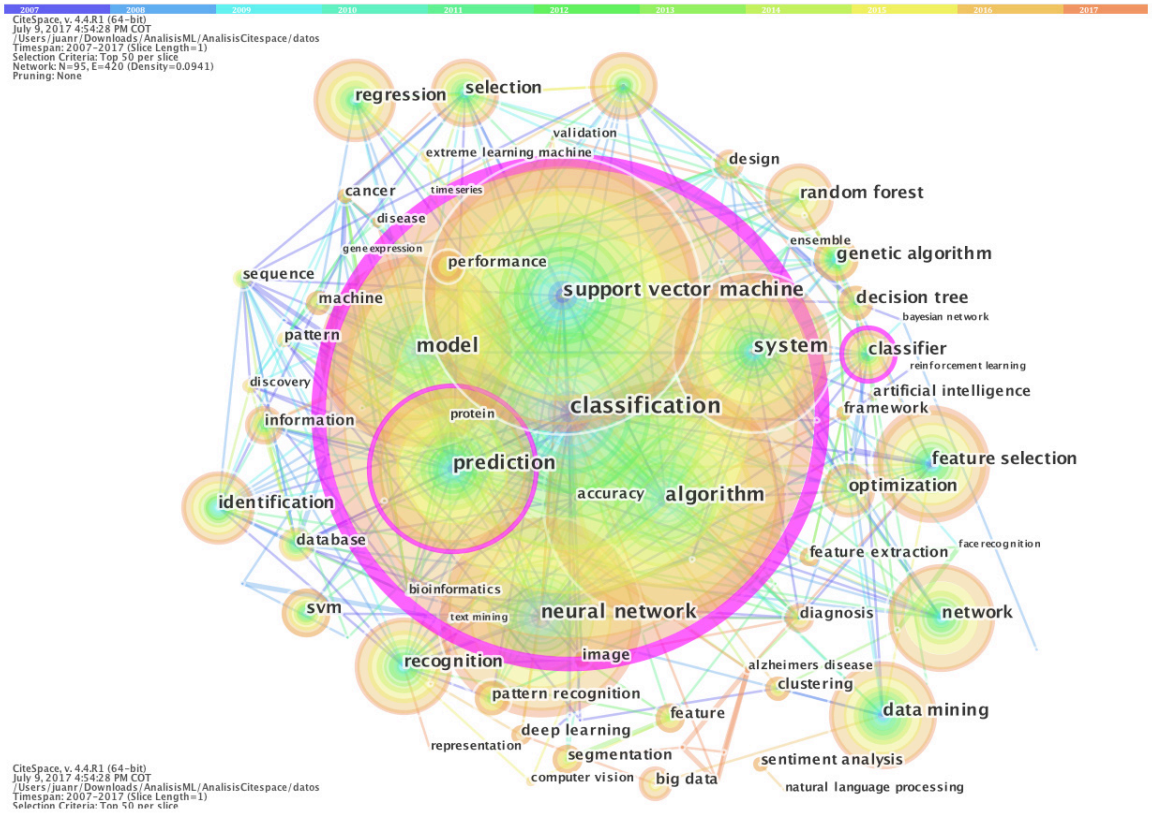

Figure 7. Network of co-occurring keywords with N=50, for machine learning 2007–2017(Q2) from Web of Science Core Collection.

once again classification (centrality = 0.42) as a widely studied subject, being an important turning point between the other concepts and having a great potential for future works. The prediction keyword, with a centrality equal to 0.13, is another turning point in this network.

Lastly, using the net of co-occurring keywords presented in Figure 7, we applied a filter, eliminating subjects that are transversal (such as data or information) and elements that belong to the proper development of any work with machine learning (such as classification or random forest). Figure 8 shows the resulting network. The most important keyword appearing on the net, by its citation counts, is data mining (1,335), followed by pattern recognition (652), database (624), diagnosis (599), cancer (449) and big data (420). Other relevant keywords are Image (414), sentiment analysis (325), disease (240), bioinformatics (209), Alzheimer's disease (188), protein (170) and computer vision (131). In the network, we can observe that data mining is an important concept in the published works, and that machine learning is becoming relevant in the health field, for the diagnosis of diseases such as cancer or Alzheimer's, by using databases collected from different sources, such as EEG signals or multiple sensors.

A co-citation analysis is an interesting way to measure the relationship between documents. It allows us to represent the proximity between the publications of the data set and the relevant cited articles in external sources. In this case, we did a journal co-citation analysis, which addresses the journals of the items analyzed. It is important to observe that, in this study, when we mention journals, we also include conference proceedings. Table 2 presents the top 10 source journals for machine learning research, based on the statistics from the WOS. LECTURE NOTES IN COMPUTER SCIENCE is the journal with the highest number of publications, having published 2,107 articles on machine learning research and being published by Springer, followed by LECTURE NOTES IN ARTIFICIAL INTELLIGENCE (1,132) and PROCEEDINGS OF SPIE (646). From Table 2, we can notice that no journal widely collects the publications made on the subject of machine learning. This dispersion in the journals confirms the multiple applications of machine learning.

In order to find the most important cited journals and to evaluate the influences and co-citation patterns of the studies in machine learning, we did a journal co-citation analysis, which resulted in the network shown in Figure 9. The network has 23 nodes and

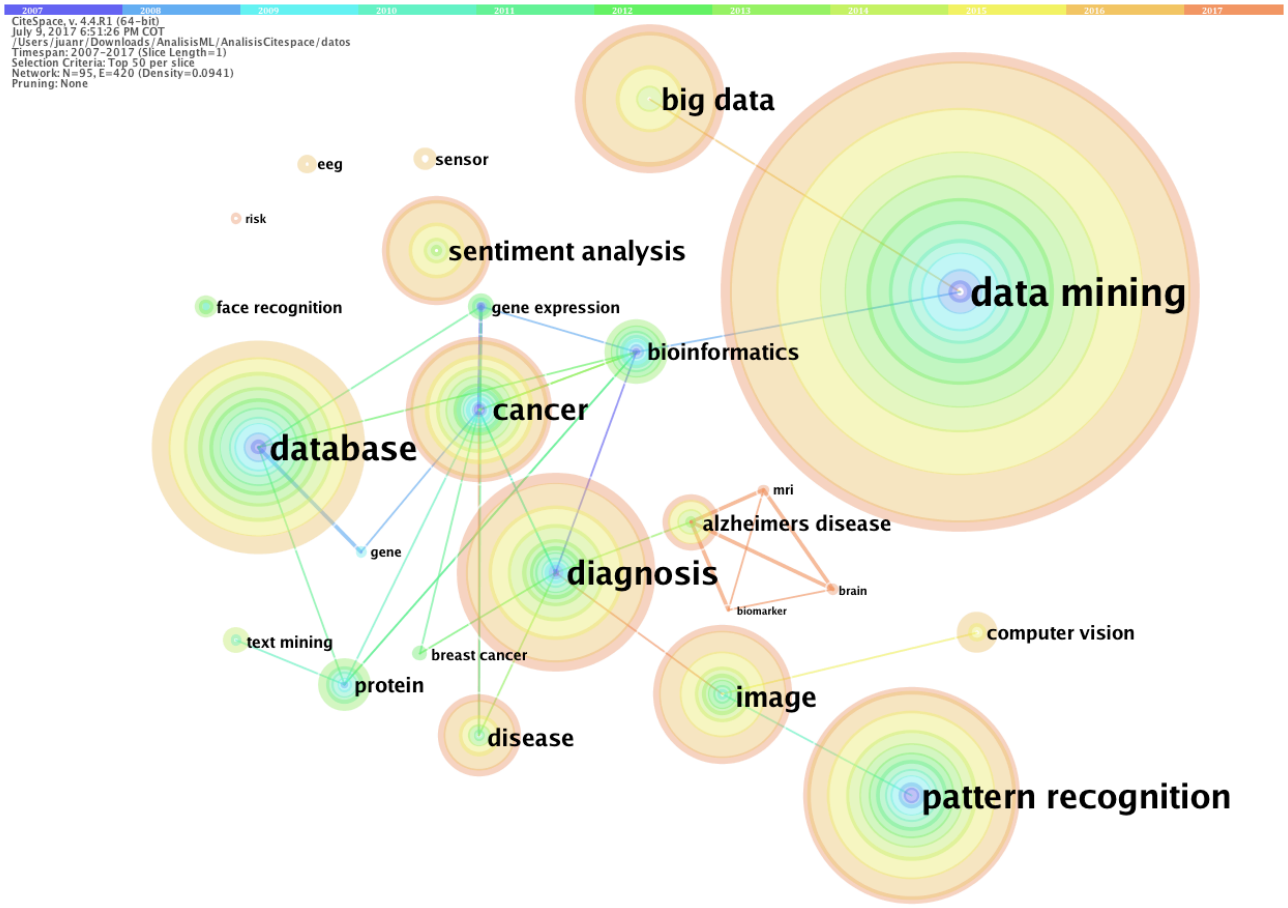

Figure 8. Network of co-occurring keywords with N=50 for filtered topics of machine learning 2007–2017(Q2) from Web of Science Core Collection.

Table 2. Top 10 source journals for Machine Learning retrieved from the Web of Science Core Collection 2007–2017(Q2).

Journal	Total	%
LECTURE NOTES IN COMPUTER SCIENCE	2,107	5.021
LECTURE NOTES IN ARTIFICIAL INTELLIGENCE	1,132	2.698
PROCEEDINGS OF SPIE	646	1,539
IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS IJCNN	355	0.846
COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE	331	0.789
ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING	279	0.665
PROCEDIA COMPUTER SCIENCE	277	0.660
IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY CONFERENCE PROCEEDINGS	233	0.555
INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP	219	0.522
FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS	173	0.412

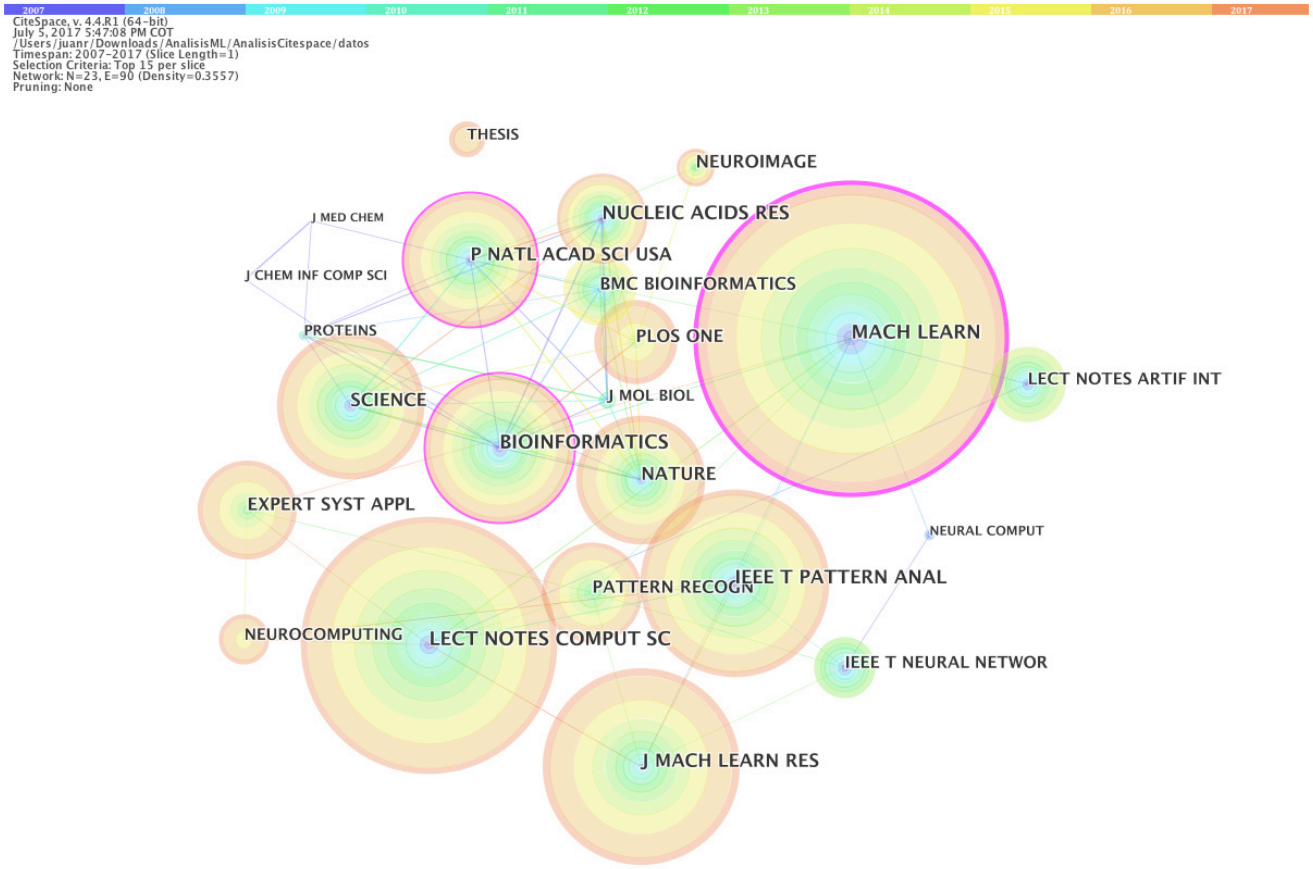

Figure 9. Journal co-citation network for machine learning 2007–2017(Q2) from Web of Science Core Collection.

90 links. Concerning co-citation frequency, the most influential journals are MACHINE LEARNING (15,767) and LECTURE NOTES IN COMPUTER SCIENCE (14,684), followed by BIOINFORMATICS (14,067), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (11,586) and NUCLEIC ACIDS RESEARCH (10,949).

To identify and to analyze the relationships between authors who have works cited in other publications and the evolution of research communities, we performed an author co-citation analysis. Figure 10 shows the resulting author co-citation network, which has 29 nodes and 131 links. Leo Breiman, a statistician at the University of California, is the author with the highest number of citations (5,270), followed by John Ross Quinlan (2,442), Bernhard Scholkopf (2,125), Vladimir N. Vapnik (2,043), Corinna Cortes (1,948) and Mark Hall (1,897).

A reference co-citation analysis allows us to observe which one is the most cited reference in the articles that belong to the dataset used. Figure 11 shows the resulting network of the reference co-citation analysis. The network has 56 nodes and 235 links. Of these references, HALL M (2009), WITTEN IH (2005) and CHIH-CHUNG CHANG (2011) occupy the top three positions (with citations counts equal to 1089, 1039 and 928, respectively) followed by PEDREGOSA F (2011) and HASTIE TREVOR (2009). The nodes with the highest centrality are BISHOP CM (2006, 0.27), DEMSAR J (2016, 0.26), HASTIE TREVOR (2009, 0.24) and WITTEN IH (2005, 0.22), showing their publication year and centrality. This suggests they are important turning points between the other nodes and interesting references for future publications.

Dataset 1. Data obtained from Web of Science and Citespace project file, to be opened in Citespace

http://dx.doi.org/10.5256/f1000research.15619.d212426

Conclusions

Understanding the dynamics of the machine learning field has practical and significant implications for researchers from different disciplines. In this study, we developed a science mapping analysis of machine learning. From this integrative approach, we identified the trends, state, and evolution in the field. From the results obtained, we can conclude that the USA is the most

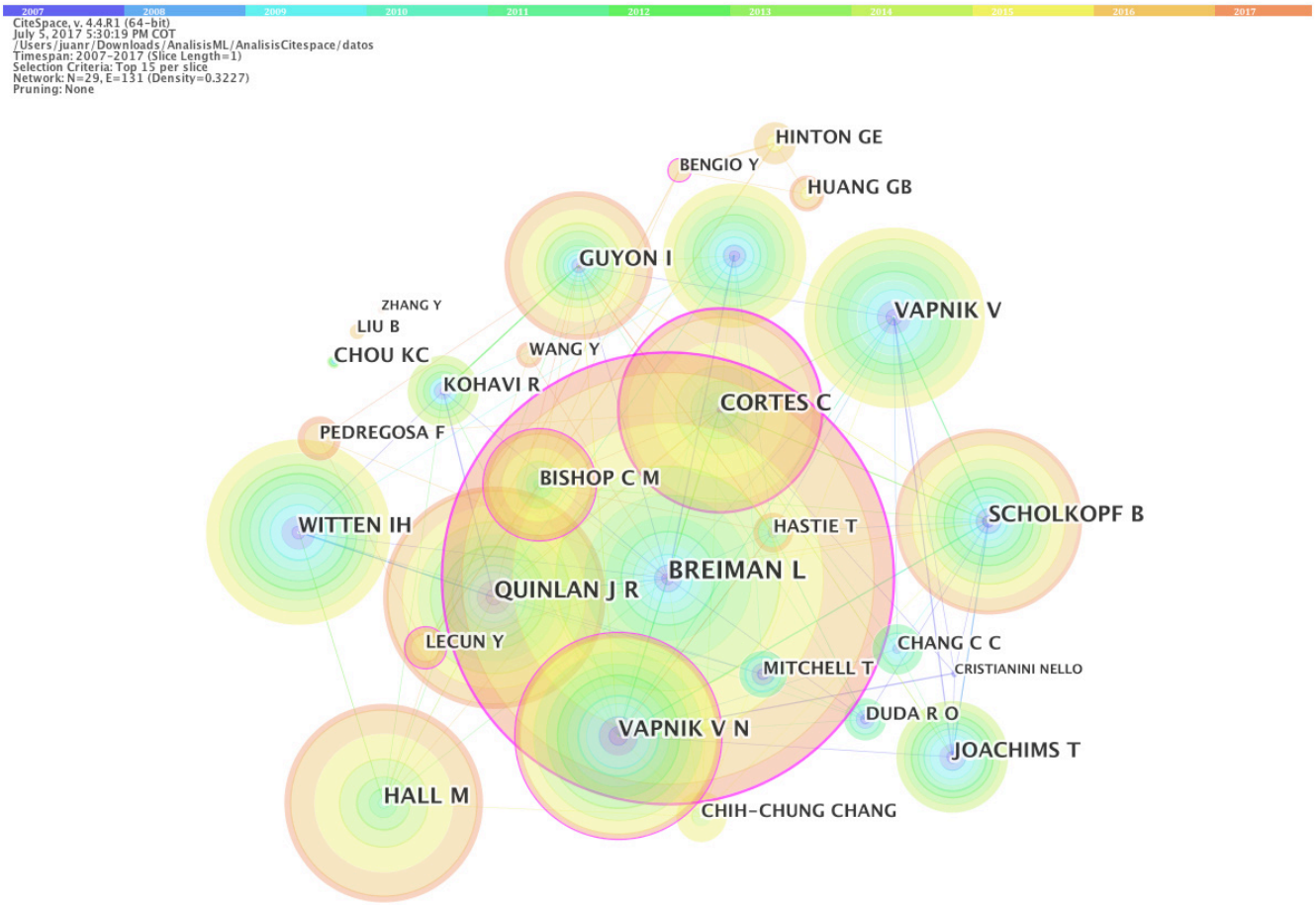

Figure 10. Author co-citation network for machine learning 2007–2017(Q2) from Web of Science Core Collection.

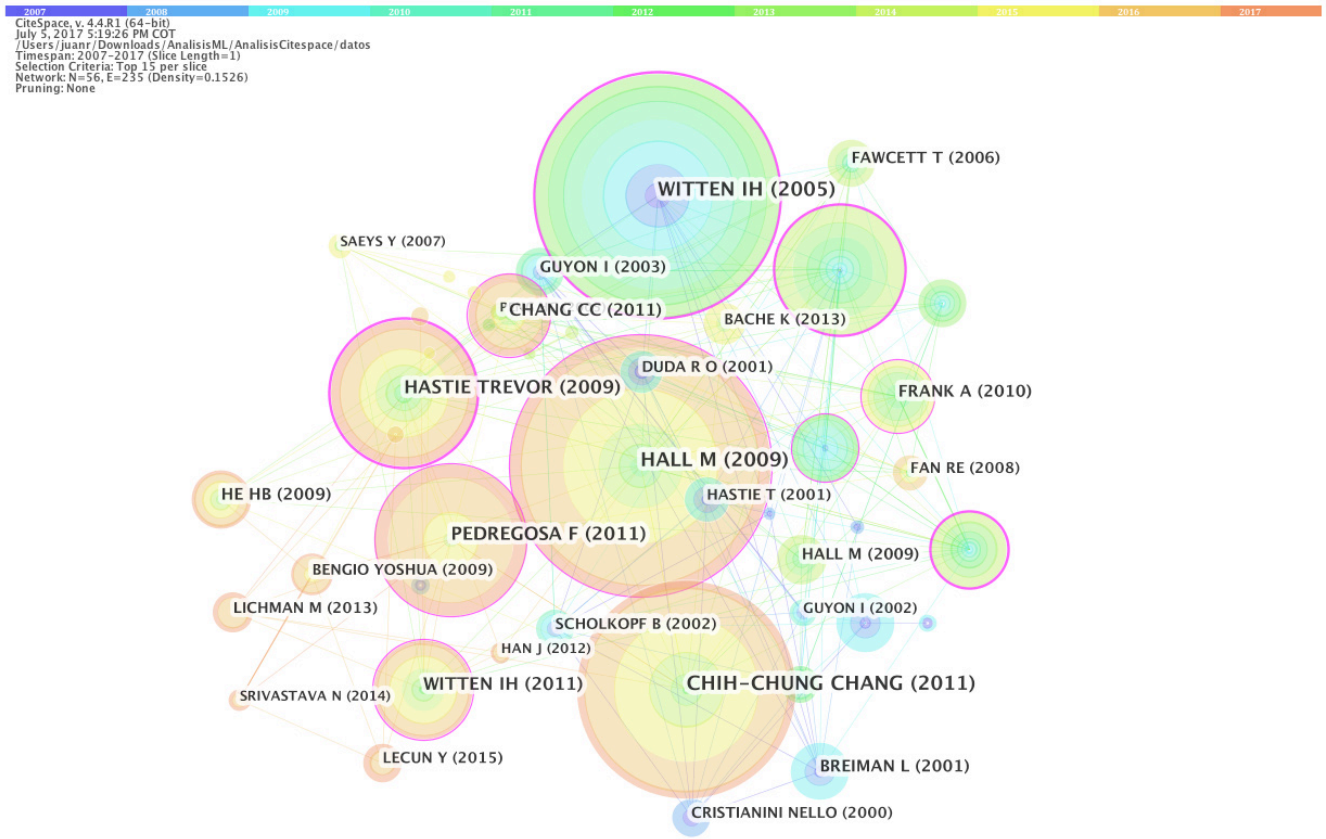

Figure 11. Reference co-citation network for machine learning 2007–2017(Q2) from Web of Science Core Collection.

productive country in the field of machine learning, with double the publications of the People's Republic of China. The Chinese Academy of Sciences, Carnegie Mellon University, Stanford University, Massachusetts Institute of Technology, Nanyang Technological University, University of California, and Harvard University are part of the institutions that have published the largest number of articles. It is useful to mention that *Machine Learning*, *Lecture Notes in Computer Science* and *Bioinformatics* are the journals with most frequently cited documents. However, no journal widely collects publications written on the subject. There are a wide number of topics that have attracted the interest of scientists and could continue to be important in the future: diseases, such as cancer or Alzheimer's disease, studies in biology, such as the protein molecule, virtual reality, commerce, smartphones and ubiquitous computing, are all important themes related to the applications of machine learning as shown by this study. This shows that machine learning can improve a large number of applications in society.

Data availability

Dataset 1: Data obtained from Web of Science and Citespace project file, to be opened in Citespace. DOI, 10.5256/f1000research.15619.d212426[25]

Competing interests

No competing interests were disclosed.

Grant information

The authors are grateful to the Telematics Engineering Group (GIT) of the University of Cauca for scientific support and Innovacción Cauca project for master's scholarship granted to J. Rincon-Patino.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Martínez MA, Cobo MJ, Herrera M, *et al.*: **Analyzing the Scientific Evolution of Social Work Using Science Mapping.** *Res Soc Work Pract.* 2015; **25**(2): 257–277. Publisher Full Text

2. Noyons ECM, Moed HF, Luwel M: **Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study.** *J Am Soc Inf Sci.* 1999; **50**(2): 115–131. Publisher Full Text

3. Cobo MJ, Martínez MA, Gutiérrez-Salcedo M, *et al.*: **25 years at Knowledge-Based**

Systems: A bibliometric analysis. *Knowl Based Syst.* 2015; **80**: 3–13.
Publisher Full Text

4. Muhamedyev RI: **Machine learning methods: An overview.** *Comput Model NEW Technol.* 2015; **19**(6): 14–29.
Reference Source

5. Endert A, Ribarsky W, Turkay C, *et al.*: **The State of the Art in Integrating Machine Learning into Visual Analytics.** *Comput Graph Forum.* 2017; **36**(8): 458–486.
Publisher Full Text

6. Kim MC, Chen C: **A scientometric review of emerging trends and new developments in recommendation systems.** *Scientometrics.* 2015; **104**(1): 239–263.
Publisher Full Text

7. Kononenko I: **Machine learning for medical diagnosis: history, state of the art and perspective.** *Artif Intell Med.* 2001; **23**(1): 89–109.
PubMed Abstract | Publisher Full Text

8. Bibault JE, Giraud P, Burgun A: **Big Data and machine learning in radiation oncology: State of the art and future prospects.** *Cancer Lett.* 2016; **382**(1): 110–117.
PubMed Abstract | Publisher Full Text

9. Price S: **A review of the state of the art in Machine Learning on the Semantic Web.** *Proc 2003 UK Work Comput Intell.* 2004; 292–299.
Reference Source

10. Al-Jamimi HA, Ahmed M: **Machine Learning-Based Software Quality Prediction Models: State of the Art**. In *2013 International Conference on Information Science and Applications (ICISA).* 2013; 1–4.
Publisher Full Text

11. Dasari DB, Venu Gopala Rao K: **Text Categorization and Machine Learning Methods: Current State Of The Art.** *Glob J Comput Sci Technol.* 2012.
Reference Source

12. Flach PA: **On the state of the art in machine learning: A personal review.** *Artif Intell.* 2001; **131**(1–2): 199–222.
Publisher Full Text

13. Moral-Muñoz JA, Cobo MJ, Peis E, *et al.*: **Analyzing the research in *Integrative & Complementary Medicine* by means of science mapping.** *Complement Ther Med.* 2014; **22**(2): 409–418.
PubMed Abstract | Publisher Full Text

14. Chen C, Hu Z, Liu S, *et al.*: **Emerging trends in regenerative medicine: a scientometric analysis in *CiteSpace*.** *Expert Opin Biol Ther.* 2012; **12**(5): 593–608.
PubMed Abstract | Publisher Full Text

15. Martínez MA, Cobo MJ, Herrera M, *et al.*: **Analyzing the Scientific Evolution of Social Work Using Science Mapping.** *Res Soc Work Pract.* 2015; **25**(2): 257–277.
Publisher Full Text

16. Cobo MJ, Chiclana F, Collop A, *et al.*: **A Bibliometric Analysis of the Intelligent Transportation Systems Research Based on Science Mapping.** *IEEE Trans Intell Transp Syst.* 2014; **15**(2): 901–908.
Publisher Full Text

17. Zhang Y, Chen H, Lu J, *et al.*: **Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016.** *Knowl Based Syst.* 2017; **133**(Supplement C): 255–268.
Publisher Full Text

18. Zhang Y, Zhang G, Chen H, *et al.*: **Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research.** *Technol Forecast Soc Change.* 2016; **105**: 179–191.
Publisher Full Text

19. Chen C: **Information Visualization: Beyond the Horizon.** Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
Publisher Full Text

20. Chen C: **CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature.** *J Am Soc Inf Sci Technol.* 2006; **57**(3): 359–377.
Publisher Full Text

21. Chen C: **Searching for intellectual turning points: Progressive knowledge domain visualization.** *Proc Natl Acad Sci.* 2004; **101 Suppl 1**: 5303–5310.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Song J, Zhang H, Dong W: **A review of emerging trends in global PPP research: analysis and visualization.** *Scientometrics.* 2016; **107**(3): 1111–1147.
Publisher Full Text

23. McCain KW: **Cocited author mapping as a valid representation of intellectual structure.** *JASIS.* 1986; **37**(3): 111–122.
Publisher Full Text

24. Rip A, Courtial JP: **Co-word maps of biotechnology: An example of cognitive scientometrics.** *Scientometrics.* 1984; **6**(6): 381–400.
Publisher Full Text

25. Rincon-Patino J, Ramirez-Gonzalez G, Corrales JC: **Dataset 1 in: Exploring machine learning: A bibliometric general approach using Citespace.** *F1000Research.* 2018.
http://www.doi.org/10.5256/f1000research.15619.d212426

Open Peer Review

Current Peer Review Status: ✔ ?

Version 1

Reviewer Report 24 September 2018

https://doi.org/10.5256/f1000research.17039.r37594

?

Chaomei Chen
College of Computing and Informatics, Drexel University, Philadelphia, PA, USA

The description of the process is clear. The interpretation of the results is accurate.

I recommend the authors to consider the following options to strengthen the study:

1. Search query

   The data collection used "machine AND learning". A more robust search query should take into account additional keywords that may be important to ensure an adequate coverage, for example, AI or deep learning.

2. Versions of CiteSpace

   It is mentioned in text that CiteSpace version 5.1.R8 SE was used. However, several figures show the signature of version 4.4.R1.

3. Coauthorship network

   More recent versions of CiteSpace support the use of fullnames of authors as opposed to using initials and the lastname. Using fullnames is preferable in such cases.

4. Burst detection

   Burst detection may be a good addition to the study. For example, it will provide more specific information on which institutions are particularly active in recent years.

5. Dual-Map Overlay

   Another potentially useful function is the dual-map overlay feature. It allows researchers to identify where relevant studies are published and which areas are highly influential in terms of how they are cited.

6. Co-citation networks

CiteSpace has several more specific functions to analyze co-citation network, for example, generating clusters and automatically selected appropriate cluster labels. These functions are highly recommended for this type of study.

Using the dataset shared by the authors, I created a visualization to illustrate how one may take advantage of these functions for such studies:

http://cluster.ischool.drexel.edu/~cchen/citespace/images/f1000/f1000.png

In summary, the current study is clearly reported and should be reproducible. On the other hand, there are several functions that are readily available in CiteSpace but they are not utilized in the current study. I hope the authors may consider updating their studies with the features I recommended here.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am the designer of CiteSpace, the tool used in this study.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 18 September 2018

https://doi.org/10.5256/f1000research.17039.r37591

Sally Ellingson
Division of Biomedical Informatics, University of Kentucky, Lexington, KY, USA

This paper uses a mapping analysis to summarize machine learning literature from 2007-2017. They create a dataset by extracting papers and conference material from the Web of Science collection using the keywords 'machine AND learning' which resulted in 41,962 records. They used CiteSpace to visualize the data in several different ways: publications per year, co-authorship network, country network, institution network, co-occurring subjects and keywords, journal co-citation network, etc. The presented graphics include a wealth of information using various node sizes and colorings by year. The work is clearly and accurately presented with some current literature cited. The methods are clearly defined and their dataset and project files to recreate the research are given in a link. I think the paper gives an interesting overview of the directions of machine learning and important researchers, research hubs, and domain topics. It also presents a study that can be followed for looking at any research area. I would suggest doing another proofread, but find the article to be technically sound and interesting.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.