RESEARCH ARTICLE

Exploring machine learning: A bibliometric general approach using SciMAT [version 1; referees: 2 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:** 07 Aug 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)
Latest published: 07 Aug 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)

Abstract

Background: Machine learning is becoming increasingly important for companies and the scientific community. In this study, we perform a bibliometric analysis on machine learning research, in order to provide an overview of the scientific work during the period 2007-2017 in this area and to show trends that could be the basis for future developments in the field.
Methods: This study is carried out using the SciMAT tool based on results extracted from Scopus. This analysis shows the strategic diagrams of evolution and a set of thematic networks. The results provide information on broad tendencies of machine learning.
Results: The results show that SciMAT is a useful tool to carry out a science mapping analysis, and emphasizes the premise that machine learning has boundless applications and will continue to be an interesting research field in the future.
Conclusions: Some of the conclusions exposed in this study show that classification algorithms have been widely studied and represent a relevant tool for generating different machine learning applications. Nonetheless, regression algorithms are becoming increasingly important in the scientific community, allowing the generation of solutions to predict diseases, sales, and yields, for example.

Keywords

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

Open Peer Review

Referee Status: ? ?

	Invited Referees 1	2
version 1 published 07 Aug 2018	? report	? report

1 **Rajesh Kumar Tiwari**, Amity University Uttar Pradesh, India

2 **Mikhail G. Dozmorov**, Virginia Commonwealth University, USA

Discuss this article

Comments (0)

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

Author roles: Rincon-Patino J: Conceptualization, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Ramirez-Gonzalez G**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Corrales JC**: Conceptualization, 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 SciMAT [version 1; referees: 2 approved with reservations]** *F1000Research* 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)

First published: 07 Aug 2018, **7**:1210 (doi: 10.12688/f1000research.15620.1)

Introduction

The machine learning field researches different human learning processes, the theoretical analysis of possible learning algorithms and methods for several application domains[1]. Studies based on machine learning have allowed scientists and companies to predict mass mortality events[2], the quality of water[3], segment clients in private banking[4], automatically classify text[5] or for the production of crops, such as cocoa[6]. Considering the growing interest of the scientific community in machine learning research and its challenges, it is interesting and necessary to analyze the field. A good approach for that purpose is science mapping analysis because it is a different way of visualizing information that allows a new researcher to become familiar with a field. An example of science mapping analysis providing an overview of the conceptual evolution of a field in medicine is proposed in 7. In this study, we perform a science mapping analysis to explore machine learning research. The objective of the study is to allow new data analysts to know the current knowledge base about machine learning and to have an initial point to explore applications in this field.

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

Methods

Dataset for visualization analysis

We used Scopus as the bibliographic source. We looked, in the third quarter of 2017, for references of articles and conferences about machine learning, using this concept as the search keyword ('machine AND learning'), with results ranging from 2007 to 2017(Q2). The concept was searched for in the article title, abstract and keywords of articles found. These articles were sorted by date (newest first). All the articles, between 2007 and 2017, were taken in to account for performing the analysis with the aim of obtaining a general vision of the field.

We obtained 67,475 records that were saved using RIS format in different files sorted by year. Figure 1 shows a summary of the records, and shows that research in the field of machine learning has been growing steadily. It is important to observe that the results for 2017 are not in the figure since the records are only to the second half Q2 of year and database is permanently updated. However, these results were used for the analysis because they show trends, which is the primary objective of the present study.

Parameter design

We used the records from Scopus for executing the analysis with SciMAT version 1.1.04. In this tool, the unit of analysis was Words. Primarily, we did a deduplication process, grouping similar words (by plurals) and looking for synonyms or duplicates in words with the highest number of documents and repetitions, always trying to avoid bias to include the largest number of terms. After that, we divided the time interval (2007 – 2017) into six smaller periods: 2007–2009, 2010–2012, 2013–2014, 2015, 2016, 2017(Q2 – published papers up the second quarter of the year). We distributed the gaps this way in order to have a comparable number of articles in each one of them. Finally, we carried out the analysis with the following configuration: all the periods, author's words as the unit

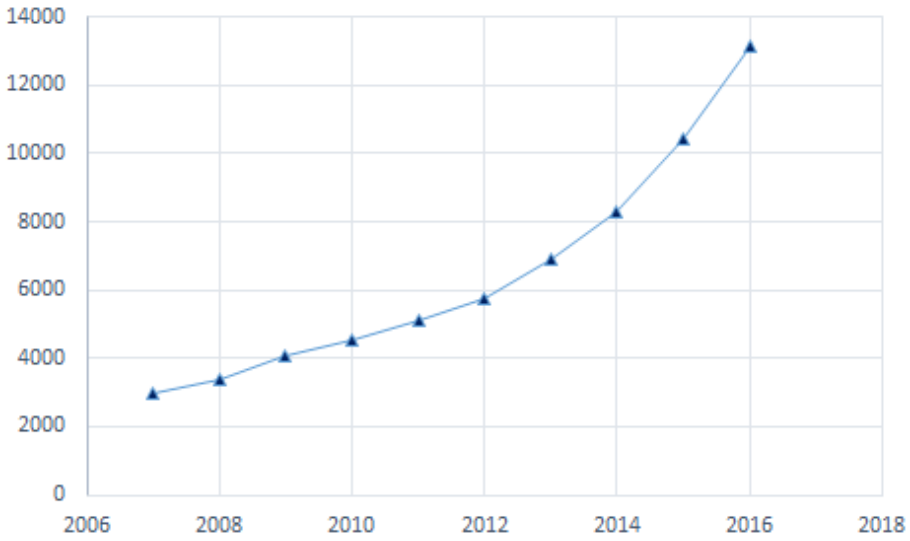

Figure 1. Number of documents for Scopus per year for machine learning 2007–2017(Q2).

of analysis, a minimum frequency for data reduction of 100 (excepting 2017 (Q2) with 50) and co-occurrence as a kind of network. Other configurations were: network reduction equal to one, association strength as a normalization measure, Simple centers algorithm with a maximum net size of seven and a minimum of five, core mapper for the document mapper, h-index and Sum citations as the quality measures and, lastly, association strength for the evolution and overlapping map.

Analysis design

SciMAT shows a spatial representation of the way disciplines, fields, specialties and documents or authors related to one another[8–10]. For this purpose, the tool implements a longitudinal framework, which takes as its base a co-word analysis and the h-index. Co-word analysis firstly provides information on the themes of a research field and, secondly, enables to analyze and to track the evolution of a study field throughout consecutive periods of time[11]. The h-index is used to measure the impact of the various identified thematic areas[8].

In this tool, we follow four steps as mentioned by 10: researching cluster detection, drawing strategic diagrams, plotting of thematic areas and carrying out a performance analysis. For the first one, the tool creates a network of keywords co-occurrence based on 12 and 13 and makes a clustering of keywords to topics, using the Simple center's algorithm. For the second step, according to 13, the cluster centrality and density rank values are relevant. The centrality measures the intensity of the interaction of a group with the others; if the cluster is boldly related to the field of research, then the link will be stronger. The density measures the intensity of internal links inside the group. According to 8 and 10, the themes of the clusters can be classified into four groups using these two measures: (1) Motor themes, which are both well developed and central to the research field; (2) basic and transversal themes, which are not sufficiently developed topics but significant for the area of investigation; (3) emerging or declining themes, which are weakly studied and marginal; (4) highly developed and isolated themes, which have well-developed internal links, but reduced external links and have only minimal importance to the field. The third step is the plotting of thematic areas, and the last one is to conduct the performance analysis. In this study, we analyzed quantitative measures, such as the number of documents and authors.

Results

To analyze the most relevant topics of investigation in different years, SciMAT uses strategic diagrams. We generated a diagram for each period of the study. The charts are divided into four quadrants: the upper-right quadrant alludes to the motor themes, the upper-left quadrant to the highly developed and isolated themes, the lower-right one to the basic and transversal themes, and the lower-left one to the emerging or declining topics[9].

Figure 2 shows the strategic diagram of the period 2007–2009, which has 27 themes. During this time span, OPTIMIZATION, CONTROL-THEORY, and IMAGE-CLASSIFICATION are some of the emerging topics. SEQUENCE-ANALYSIS-PROTEIN appears as the motor theme with the highest density (0.83) and centrality (3.21) values, followed by DATABASES-PROTEIN, with centrality equal to 2.97 and density equal to 0.43. This suggests that machine learning was widely used to study the protein molecule in that period, covering, for example, studies to predict the structure or function of that molecule.

Due to the great number of themes that appear in the strategic diagram, we decided to choose three themes, giving priority to those that may have a high impact application. Figure 3A shows the net for the theme DATABASES-PROTEIN, which has a density of 0.43, a centrality of 2.97 and a document count of 322. In this net, we can observe that an important topic is PROTEIN-STRUCTURE. Figure 3B presents the network for the theme DATASETS, which has a density of 0.06, a centrality of 1.37, a document count of 273 and strongly related topics such as CLASSIFIERS, SEMI-SUPERVISED-LEARNING and RANDOM-FOREST. Figure 3C shows the network for the theme IMAGE-CLASSIFICATION, which has a density of 0.11, a centrality of 1.3, a document count of 273 and relevant concepts, such as NEURAL-NETWORKS, SUPPORT-VECTOR-MACHINE (SVM) and IMAGE-PROCESSING.

Figure 4 shows the strategic diagram of the period 2010–2012. The diagram has 35 themes. During this time span, VIRTUAL-REALITY, ROBOTS, and METADATA are some of the emerging themes, while CLASSIFICATION-ALGORITHM is one of the basic and transversal topics. AGED and PROTEIN-ANALYSIS appear as the motor subjects with the highest density (0.72 and 0.58, respectively) and centrality (1.91 and 2.18, respectively) values. Other important themes are CHEMISTRY AND GENETICS. This is a sign that during this period, topics on biology and health began to become relevant in applied machine learning research.

From the second period, we selected three thematic networks. Figure 5A shows the net for the theme PROTEIN-ANALYSIS, which has a density of 0.58, a centrality of 2.18 and a document count of 229. This shows us that topics about proteins continue to be important in this period. Figure 5B presents the network for the theme CHEMISTRY, which is an emergent theme and has a density of 0.32, a centrality of 2.24 and a document count of 346. Figure 5C shows the network for the subject VIRTUAL-REALITY, which has a density of 0.07, a centrality of 0.98, a document count of 65 and is another emerging theme for the period 2010–2012.

The strategic diagram of the period 2013–2014 is shown on Figure 6, which has 32 themes. During this time frame, SENSORS, FACE-RECOGNITION and COMMERCE are some of the emerging topics. The IMAGE-INTERPRETATION-COMPUTER-ASSISTED topic appears as the motor theme, with the highest density (0.58) and centrality (2.61) values. This suggests that in this period there were studies on machine learning applied to different topics, such as sensor data, costs, and gesture recognition.

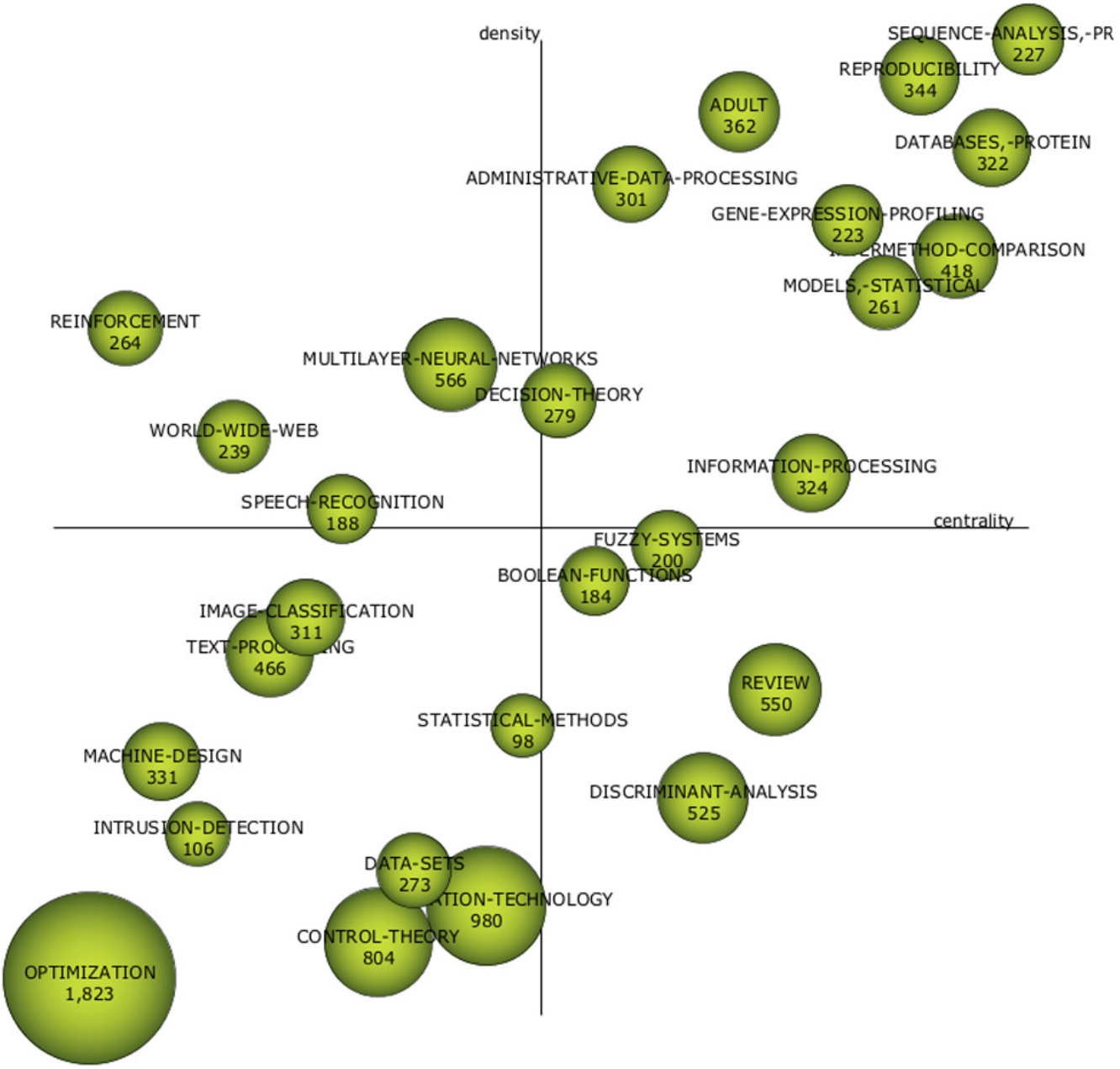

Figure 2. Strategic diagram of the period 2007–2009 for machine learning using data from Scopus.

We selected three thematic networks from the third period (2013–2014). Figure 7A shows the network for the theme AMINO-ACID-SEQUENCE, which has a density of 0.42, a centrality of 1.91 and a document count of 276. Figure 7B presents the network for the theme MOBILE-DEVICES, which has a density of 0.21, a centrality of 0.76, a document count of 159 and strongly related topics such as HUMAN-COMPUTER-INTERACTION, E-LEARNING, and UBIQUITOUS-COMPUTING. Figure 7C shows the network for the theme COMMERCE, which has a density of 0.07, a centrality of 0.88, a document count of 205 and relevant concepts, such as SOCIAL-NETWORKING and COSTS.

Figure 8 shows the strategic diagram of the period 2015. The diagram has 25 themes. During this time span, SMARTPHONES, FACE-RECOGNITION and FORESTRY

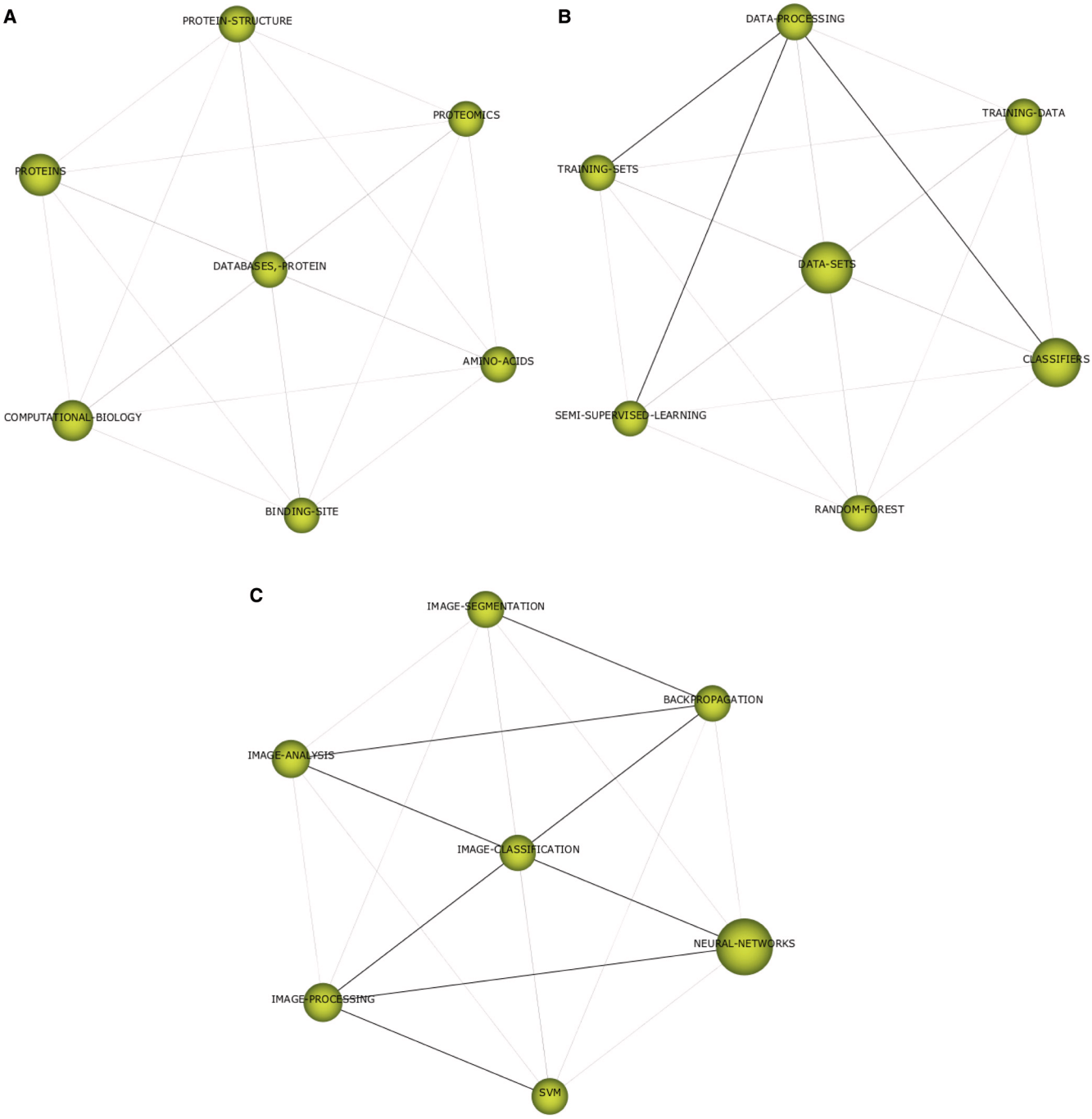

Figure 3. Selected thematic networks for period 2007–2009 for machine learning using data from Scopus. (**A**) DATABASES-PROTEIN; (**B**) DATASETS; (**C**) IMAGE-CLASSIFICATION.

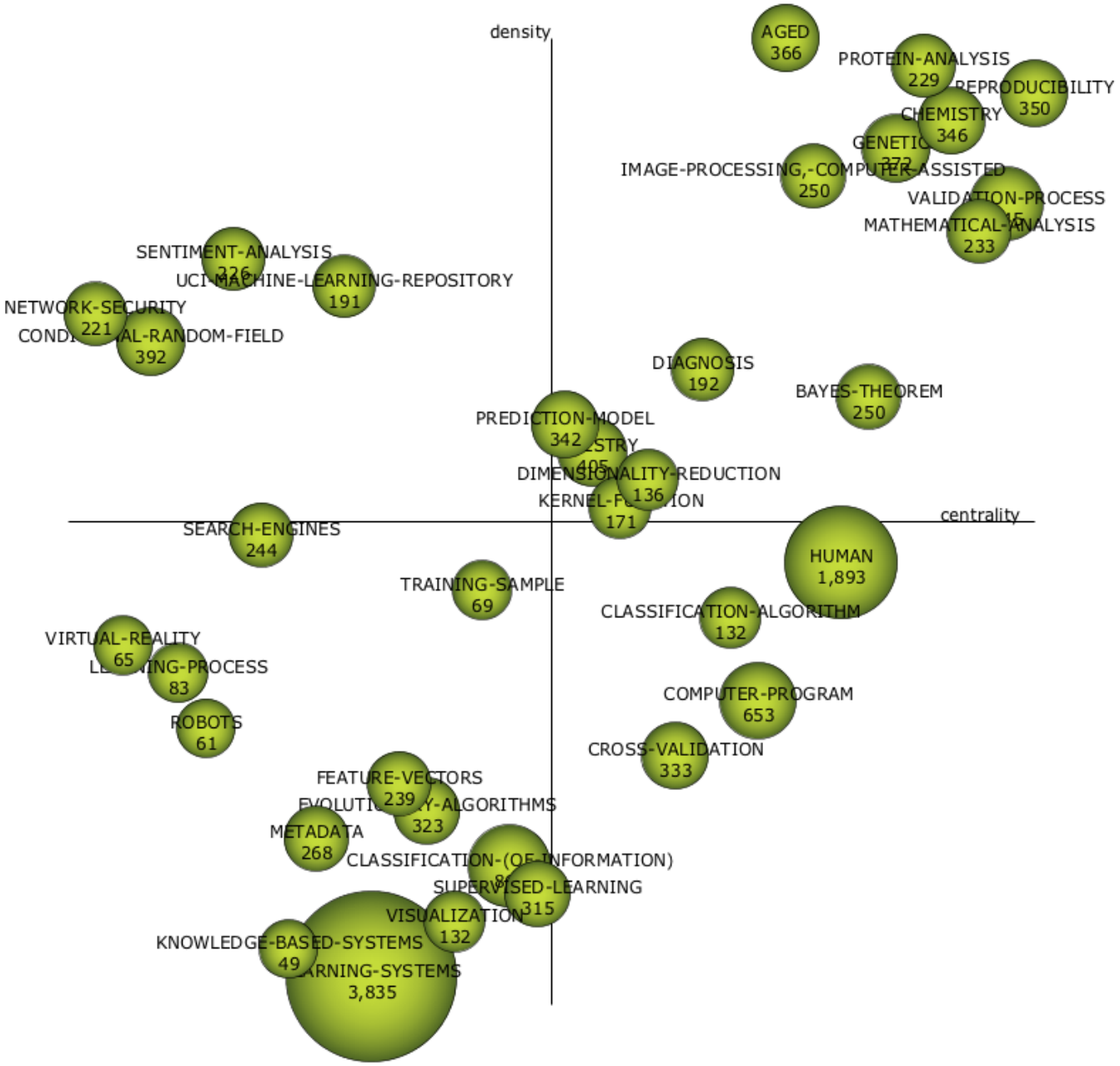

Figure 4. Strategic diagram of the period 2010–2012 for machine learning using data from Scopus.

(label generated for algorithms such as Random-Forest or Decision-Trees) are some of the emerging themes. NUCLEAR-MAGNETIC-RESONANCE-IMAGING appears as the motor subject with the highest density (0.57) and centrality (2.81) values. Other important themes are COMPUTATIONAL-BIOLOGY and MEDICAL-IMAGING. We found that during this period biology and health were once again relevant topics in machine learning research.

From the fourth time frame, we selected three thematic networks. Figure 9A shows the network for the theme NUCLEAR-MAGNETIC-RESONANCE-IMAGING, which has a density of

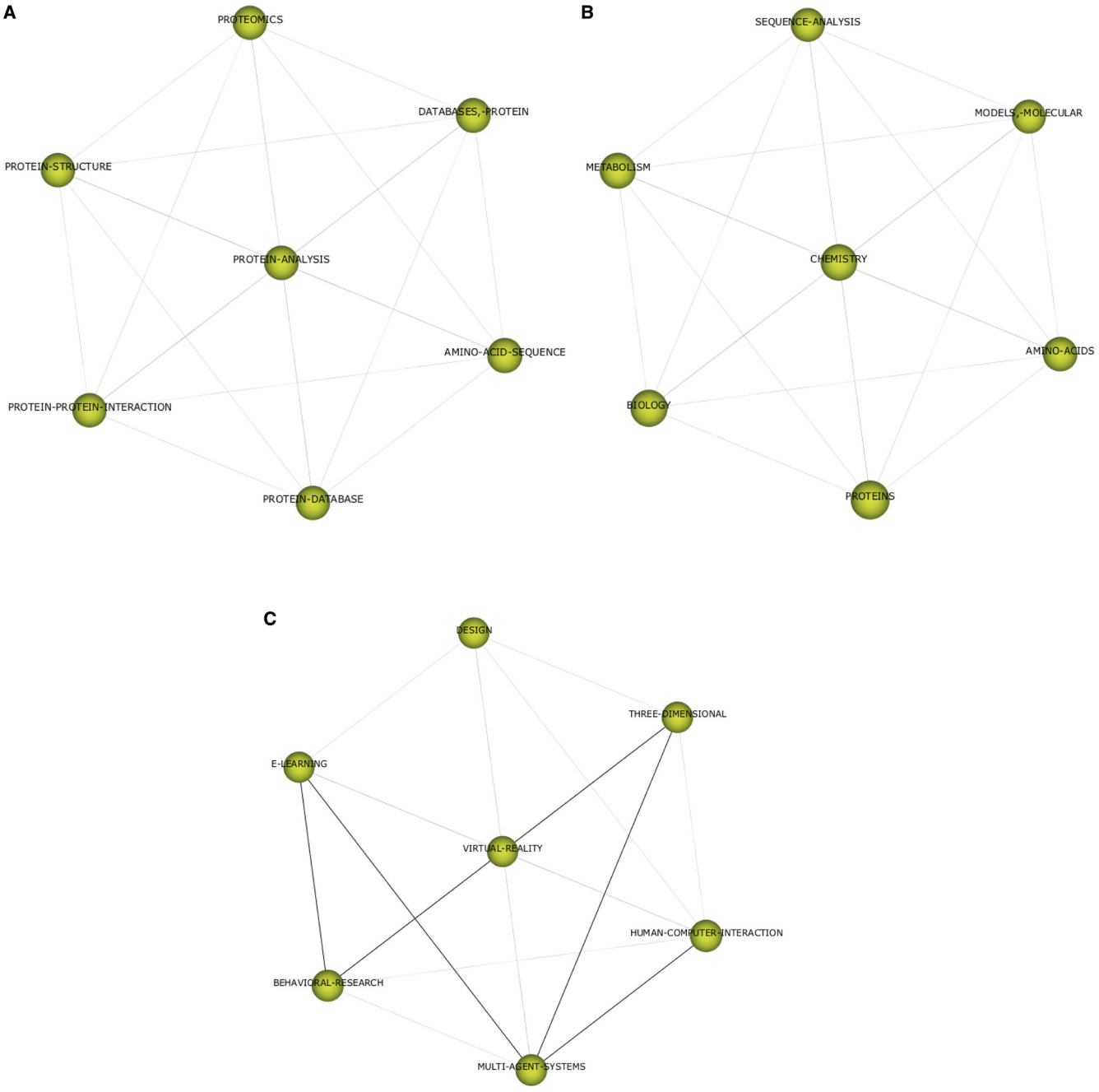

Figure 5. Selected thematic networks for period 2010–2012 for machine learning using data from Scopus. (**A**) PROTEIN-ANALYSIS; (**B**) CHEMISTRY; (**C**) VIRTUAL-REALITY.

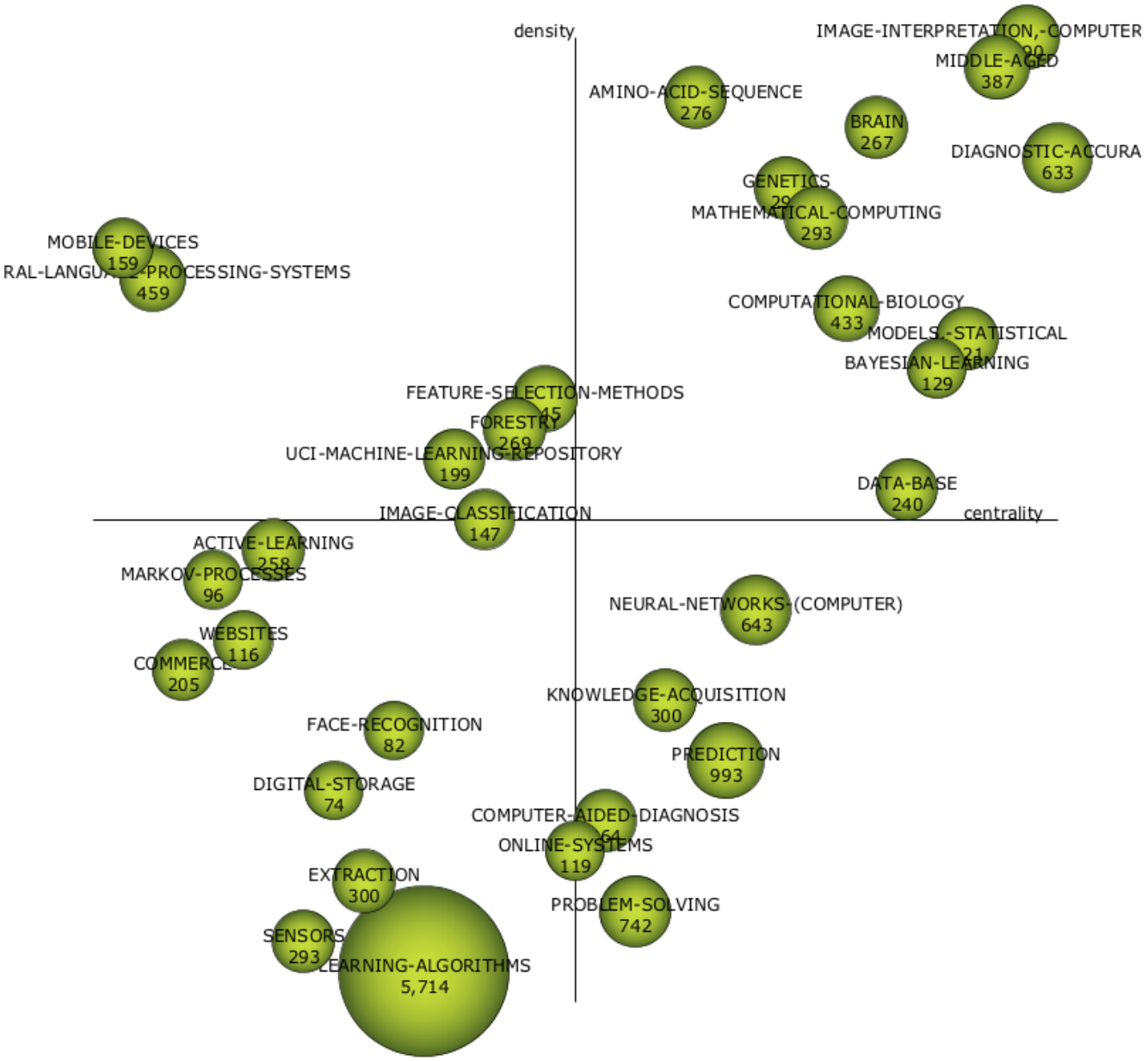

Figure 6. Strategic diagram of the period 2013–2014 for machine learning using data from Scopus.

0.57, a centrality of 2.81 and a document count of 251. Figure 9B presents the network for the theme COMPUTATIONAL-BIOLOGY, which is an emergent theme and has a density of 0.44, a centrality of 1.8 and a document count of 320. Figure 9C shows the network for the topic SMARTPHONES, which has a density of 0.07, a centrality of 0.13, a document count of 144 and is another emerging theme for the 2015 period.

The strategic diagram of the period 2016 is shown in Figure 10, which has 24 themes. During this time span, UBIQUITOUS-COMPUTING and COMMERCE are some of the emerging themes. The MIDDLE-AGED theme –which refers to applications developed for middle-aged people– appears as the motor topic with the highest density (0.76) and centrality (2.04) values.

Figure 7. Selected thematic networks for period 2013–2014 for machine learning using data from Scopus. (**A**) AMINO-ACID-SEQUENCE; (**B**) MOBILE-DEVICES; (**C**) COMMERCE.

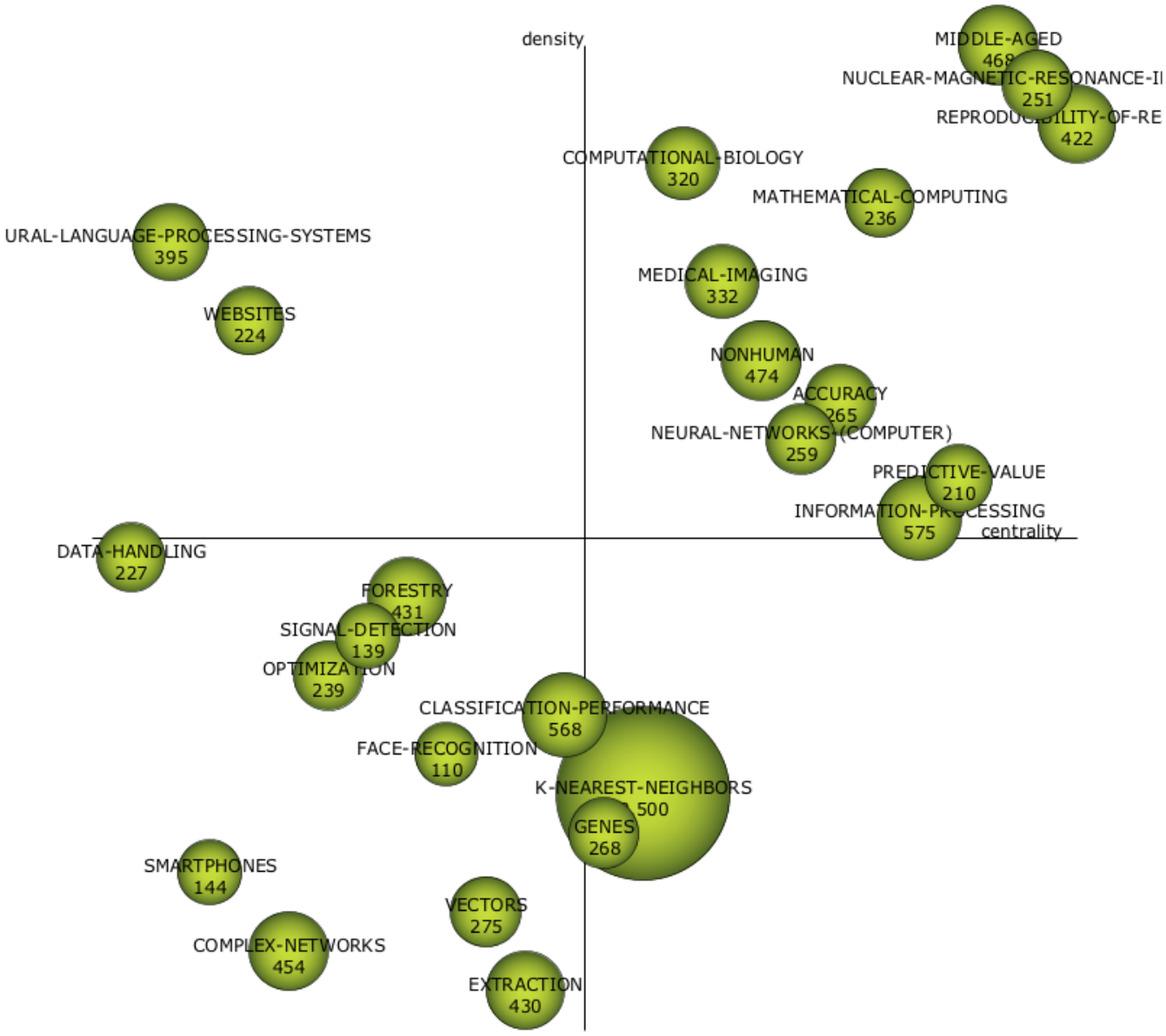

Figure 8. Strategic diagram of the period 2015 for machine learning using data from Scopus.

We selected three thematic networks from the fifth period (2016). Figure 11A shows the network for the theme MEDICAL-IMAGING, which has a density of 0.26, a centrality of 1.45 and a document count of 312. Figure 11B presents the network for the theme INTRUSION-DETECTION, which has a density of 0.3, a centrality of 0.89 and a document count of 289. Figure 11C shows the network for the theme UBIQUITOUS-COMPUTING, which has a density of 0.07, a centrality of 1.07, a document count of 132 and relevant concepts, such as AUTOMATION, SMARTPHONES and the INTERNET.

Figure 12 shows the strategic diagram of the period 2017(Q2). The diagram has 16 themes. During this time span, HUMAN is the only emerging theme, while FORESTRY is one of the

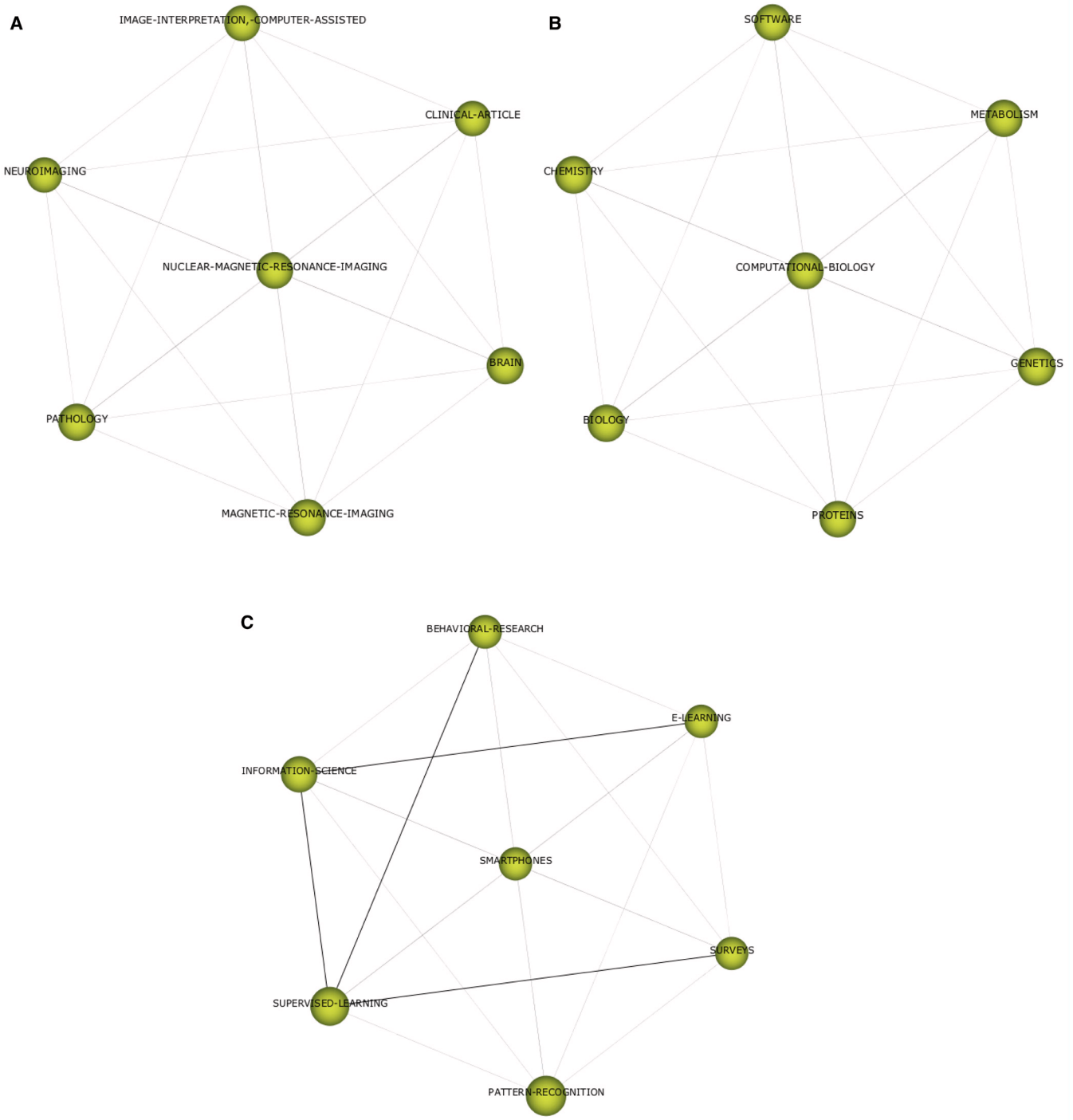

Figure 9. Selected thematic networks for period 2015 for machine learning using data from Scopus. (**A**) NUCLEAR-MAGNETIC-RESONANCE-IMAGING; (**B**) COMPUTATIONAL-BIOLOGY; (**C**) SMARTPHONES.

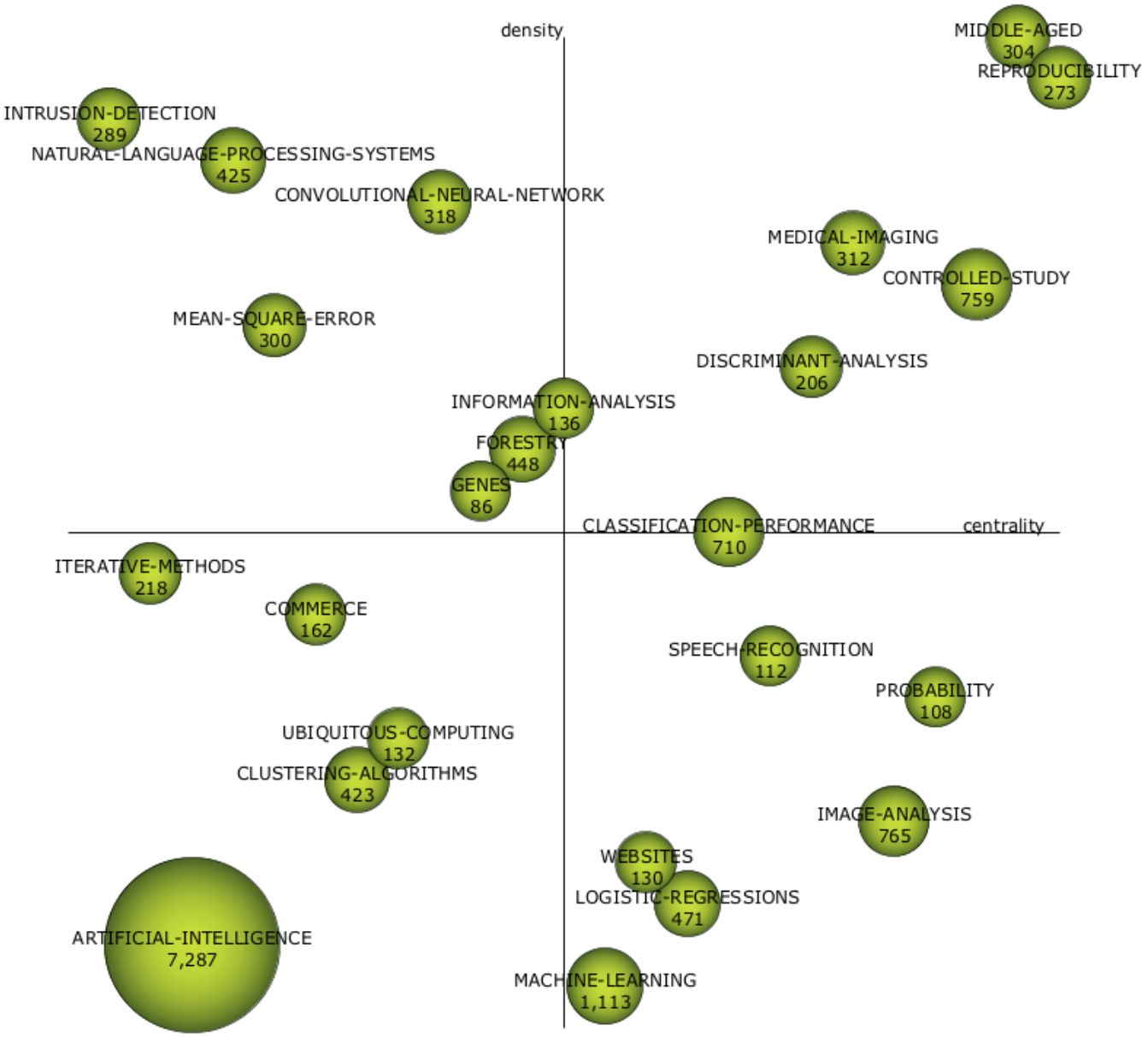

Figure 10. Strategic diagram of the period 2016 for machine learning using data from Scopus.

basic and transversal topics. This shows us the importance of algorithms such as Random-Forest or Decision-Trees during the last decade in the research on machine learning. MEDICAL-IMAGING appears as the motor theme with the highest density (0.51) and centrality (2.05) values. Once again, topics on health are relevant in machine learning research.

From 2017(Q2), we selected three thematic networks. Figure 13A shows the net for the theme MEDICAL-IMAGING, which has a document count of 137. Figure 13B presents the network for the topic FORESTRY (label generated for algorithms such as Random-Forest or Decision-Trees), which has a density of 0.17, a centrality of 1.84 and a document count of 220.

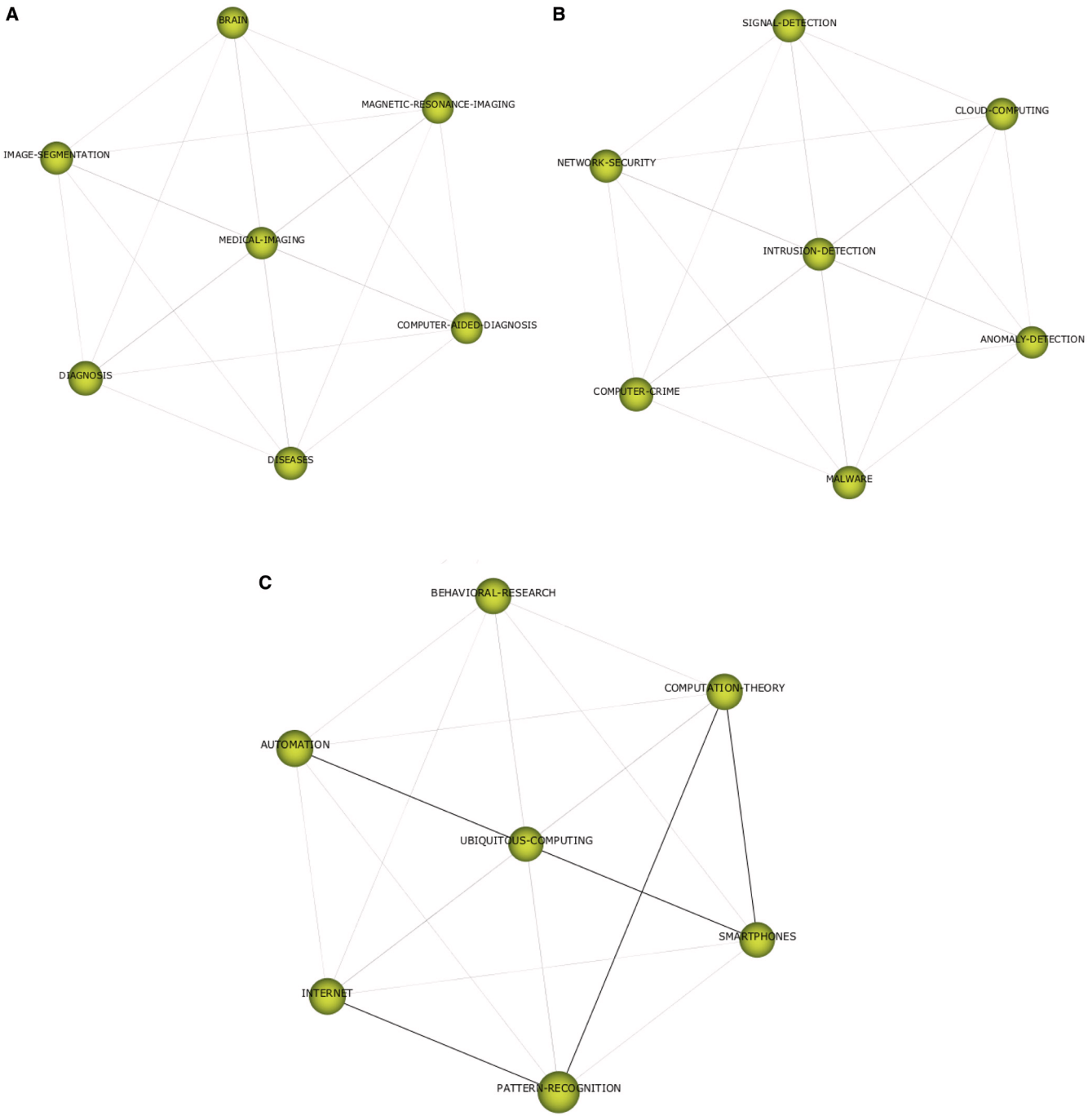

Figure 11. Selected thematic networks for period 2016 for machine learning using data from Scopus. (**A**) MEDICAL-IMAGING; (**B**) INTRUSION-DETECTION; (**C**) UBIQUITOUS-COMPUTING.

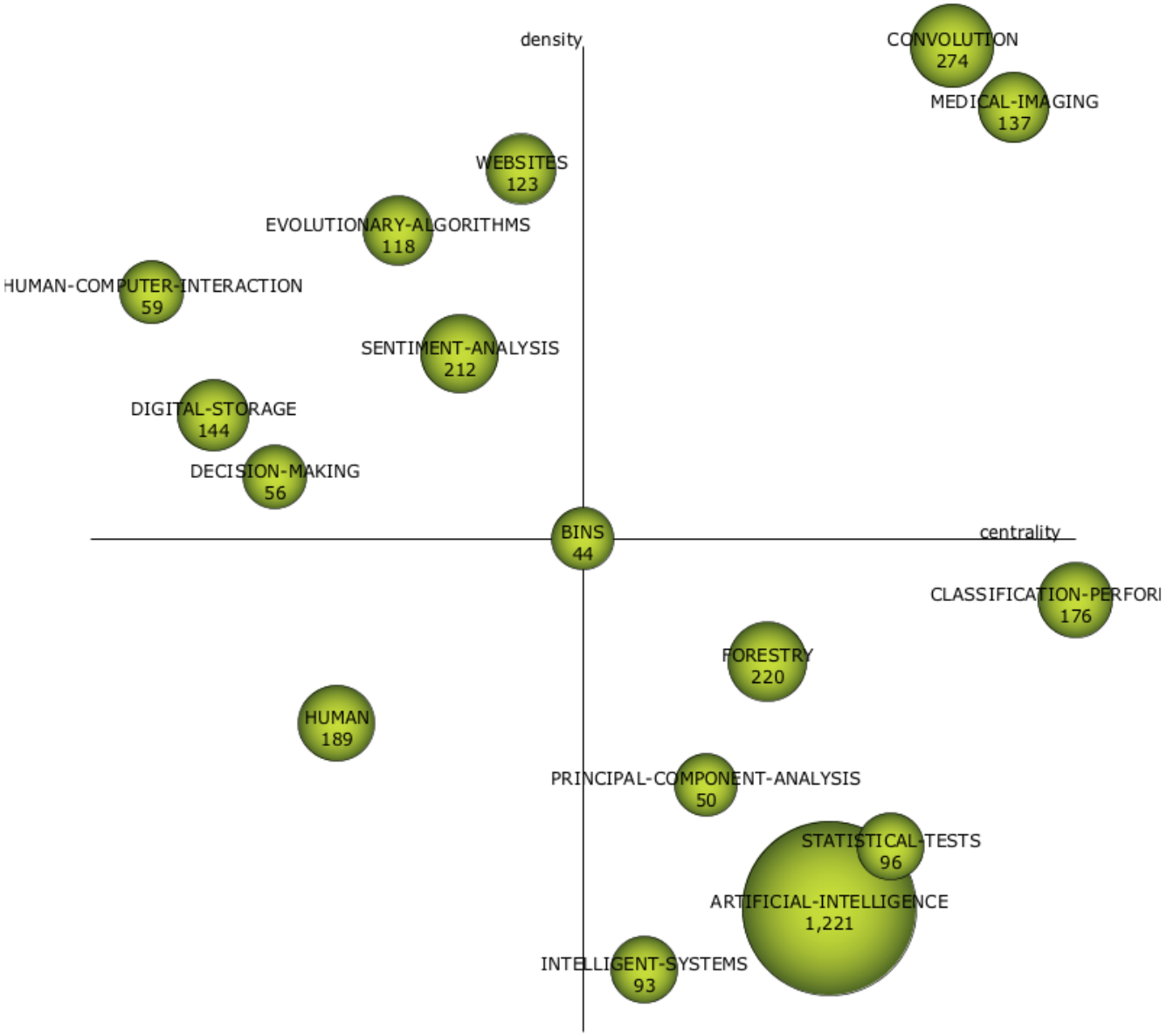

Figure 12. Strategic diagram of the period 2017(Q2) for machine learning using data from Scopus.

Dataset 1. Data obtained from Scopus and SciMat project file, to be opened in SciMat

http://dx.doi.org/10.5256/f1000research.15620.d212425

Conclusions

Exposing emerging trends in the field of machine learning allows researchers to increase their understanding of the changes and the evolution over time of this research field. One of the primary objectives of a science mapping analysis is to highlight trends and possible relationships between the relevant topics of a research field. SciMAT is a useful tool to carry out a study based on this approach, which offers fundamental themes, based on a cluster generation. The results of the present study show that machine learning is an important and widely studied scientific area. The tendencies indicate that machine learning applications will still be of interest to the scientific community.

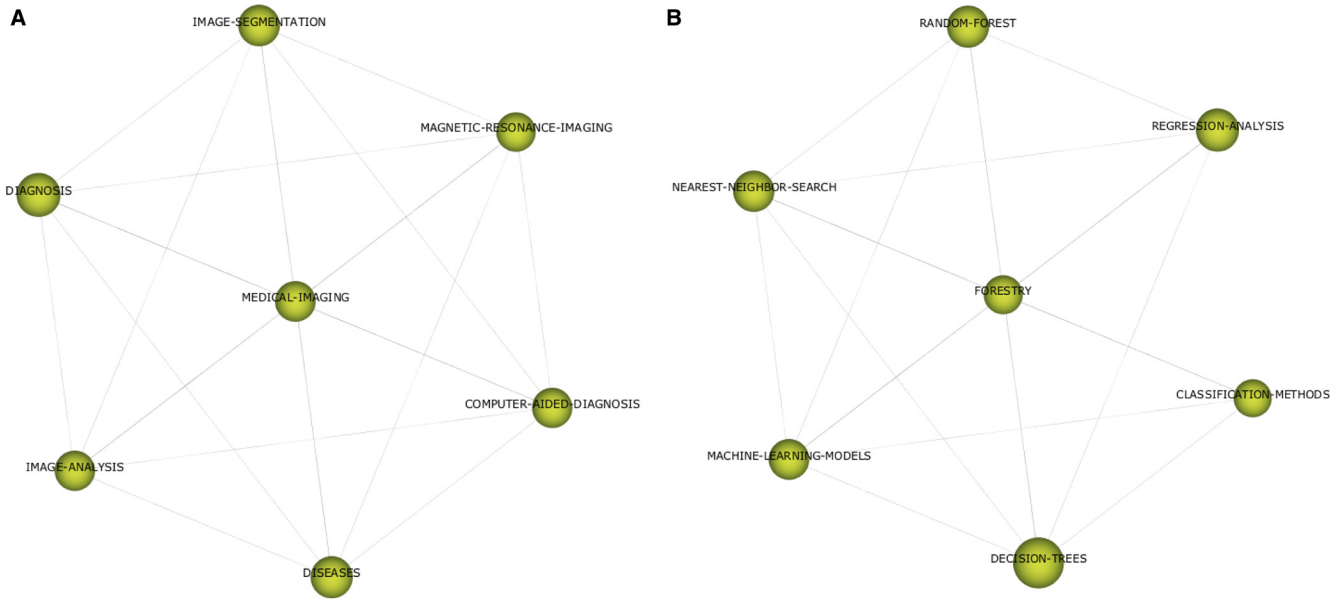

Figure 13. Selected thematic networks for period 2017(Q2) for machine learning using data from Scopus. (**A**) MEDICAL-IMAGING; (**B**) FORESTRY.

The use of machine learning to predict diseases such as cancer[14] or Alzheimer's disease[15], and in fields such as biology[16], rehabilitation system[17], commerce[18], smartphones[19] and ubiquitous computing[20], will be a trend in the near future.

Data availability

Dataset 1: Data obtained from Scopus and SciMat project file, to be opened in SciMat. DOI, 10.5256/f1000research.15620.d212425[21]

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. Michalski RS, Carbonell JG, Mitchell TM: **Machine Learning: An Artificial Intelligence Approach.** Springer Berlin Heidelberg, 2013.
Publisher Full Text
2. Crisci C, Ghattas B, Perera G: **A review of supervised machine learning algorithms and their applications to ecological data.** *Ecol Modell.* 2012; **240**: 113–122.
Publisher Full Text
3. López ID, Figueroa A, Corrales JC: **Adaptive Prediction of Water Quality Using Computational Intelligence Techniques**. In *Computational Science and Its Applications -- ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part II,* Cham: Springer International Publishing, 2017; 45–59.
Publisher Full Text
4. Smeureanu I, Ruxanda G, Badea LM: **Customer segmentation in private banking sector using machine learning techniques.** *J Bus Econ Manag.* 2013; **14**(5): 923–939.
Publisher Full Text
5. Sebastiani F: **Machine Learning in Automated Text Categorization.** *ACM Comput Surv.* 2002; **34**(1): 1–47.
Publisher Full Text
6. Plazas JE, López ID, Corrales JC: **A Tool for Classification of Cacao Production in Colombia Based on Multiple Classifier Systems.** In *Computational Science and Its Applications -- ICCSA 2017: 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part II,* Cham: Springer International Publishing, 2017; 60–69.
Publisher Full Text
7. 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
8. 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
9. Cobo MJ, López-Herrera AG, Herrera-Viedma E, *et al.*: **SciMAT: A new science mapping analysis software tool.** *J Am Soc Inf Sci Technol.* 2012; **63**(8): 1609–1630.
Publisher Full Text
10. Cobo MJ, López-Herrera AG, Herrera-Viedma E, *et al.*: **An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field.** *J Informetr.* 2011; **5**(1): 146–166.
Publisher Full Text
11. Garfield E: **Scientography: Mapping the tracks of science.** *Curr Contents Soc*

Behav Sci. 1994; **7**(45): 5–10.
Reference Source

12. Callon M, Courtial JP, Turner WA, *et al.*: **From translations to problematic networks: An introduction to co-word analysis.** *Soc Sc Inform.* 1983; **22**(2): 191–235.
Publisher Full Text

13. Callon M, Courtial J, Laville F: **Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry.** *Scientometrics.* 1991; **22**(1): 155–205.
Publisher Full Text

14. Wang Y, Tetko IV, Hall MA, *et al.*: **Gene selection from microarray data for cancer classification—a machine learning approach.** *Comput Biol Chem.* 2005; **29**(1): 37–46.
PubMed Abstract | Publisher Full Text

15. Moradi E, Pepe A, Gaser C, *et al.*: **Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects.** *Neuroimage.* 2015; **104**: 398–412.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Swan AL, Mobasheri A, Allaway D, *et al.*: **Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology.** *OMICS.* 2013; **17**(12): 595–610.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Yeh SC, Huang MC, Wang PC, *et al.*: **Machine learning-based assessment tool for imbalance and vestibular dysfunction with virtual reality rehabilitation system.** *Comput Methods Programs Biomed.* 2014; **116**(3): 311–318.
PubMed Abstract | Publisher Full Text

18. Yan J, Zhang C, Zha H, *et al.*: **On Machine Learning Towards Predictive Sales Pipeline Analytics**. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015; 1945–1951.
Reference Source

19. Faragher RM, Harle RK: **SmartSLAM - An Efficient Smartphone Indoor Positioning System Exploiting Machine Learning and Opportunistic Sensing**. In *Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2013).* 2013; 1006–1019.
Reference Source

20. Ventura D, Casado-Mansilla D, López-de-Armentia J, *et al.*: **ARIIMA: A Real IoT Implementation of a Machine-Learning Architecture for Reducing Energy Consumption**. In *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services: 8th International Conference, UCAmI 2014, Belfast, UK*, Cham: Springer International Publishing, 2014; 444–451.
Publisher Full Text

21. Rincon-Patino J, Ramirez-Gonzalez G, Corrales JC: **Dataset 1 in: Exploring machine learning: A bibliometric general approach using SciMAT.** *F1000Research.* 2018.
http://www.doi.org/10.5256/f1000research.15620.d212425

Open Peer Review

Current Referee Status: ? ?

Version 1

Referee Report 15 August 2018

doi:10.5256/f1000research.17040.r37068

? **Mikhail G. Dozmorov**
Department of Biostatistics , Virginia Commonwealth University, Richmond, VA, USA

The paper “Exploring machine learning: A bibliometric general approach using SciMAT” by Rincon-Patino *et al.* presents an overview of areas of machine learning applications. The study is conducted using science mapping analysis with the SciMAT tool. A bibliography containing “machine” and “learning” keywords over the 10-year period (2007-2017) was extracted from Scopus. This period was split into six smaller periods, and the state of machine learning in each period was illustrated by two figures, strategic diagram, and selected thematic network. The paper is well written. The following recommendations are intended to improve the readability and message of the paper.

- Strategic diagrams may not be familiar for users who aren’t familiar with the SciMAT tool. Thus, the description of strategic diagrams given at the beginning of the Results section is better to be placed to the legend of the first figure. Also, clarify the meaning of “the basic and transversal themes.” Furthermore, the description of the lower-left quadrant, “the emerging or declining topics,” is confusing, it should be one or the other.

- Although the study claims to present a longitudinal analysis of the development of machine learning areas, it is the analysis and the description of individual time periods, presented on individual figures. I am missing the compare-and-contrast view of the analysis, which would have been possible if the figures would be presented side-by-side. Given the good readability of the strategic diagrams, it is perfectly possible to combine them into one figure. The same can be done for the selected thematic networks, by making edges shorter, thicker, and fonts - larger.

- It is unclear whether the manuscript is about the overview of machine learning fields, or the tool SciMAT and the science mapping analysis. Besides simply describing what the results are at each time period, it would be good to have some conclusions about the longitudinal trends of the field.

- Although the rationale for choosing the non-equal smaller periods in understandable, the periods themselves should be better defined. At the very least, state that the presented periods are inclusive, that is, 2007-2009 spans the 3-year period. Better, adopt (MM/DD/YYYY-MM/DD/YYYY] notation.

- Figure 1 shows the growing number of documents containing “machine” and “learning” keywords. However, it is known that the number of publications per year grows by itself. It would be more informative to show the proportion of published documents that contain machine" and “learning” keywords. The usefulness of representing the proportion of publications can be illustrated with the following R code:

```
library(MDmisc) # devtools::install_github('mdozmorov/MDmisc')
library(ggplot2)
p <- get_pubmed_graph("machine learning", yearstart = 2007, yearend = 2017, normalize = TRUE, xlab = "Year", ylab = "Proportion of all publications")
p
ggsave("pubmed_machine_learning.png", p, device = "png", height = 4)
```

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?
Yes

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

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I 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.

Referee Report 09 August 2018

doi:10.5256/f1000research.17040.r36927

Rajesh Kumar Tiwari
Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow, India

The article highlights about the bibliometric analysis of scientific works in the area of machine learning, published in Scopus indexed Journals during the period of 2007-2017. The manuscript is well written. However, I have following suggestions to make:

1. Correction in title of Manuscript: Exploring machine learning: A bibliometric general approach using SciMAT **Tool**

2. Definition of Machine learning should be incorporated in the introduction Part. Authors can refer and cite following articles: PMID: 23740390[1] and PMID: 28302041[2]

3. Apart from mass mortality events, the quality of water, segment clients in private banking, automatically classify text, production of crops, the machine learning applications are also widely used in the Drug designing in pharmaceutical industries, academia and research. Authors should mention the same in the introduction part of the article and can refer and cite the following articles with PMID: 22346230[3], PMID: 26526829[4], PMID: 28382857[5], PMID: 29256344[6]

4. Methods: As the data set is collected from Scopus, the same should be cited in the manuscript.

5. The original source of SciMAT Tool should be cited in the manuscript.

6. The authors must elaborate more about the basis of division of the data set. It is mentioned that the time interval divided into six smaller periods: 2007–2009, 2010–2012, 2013–2014, 2015, 2016, 2017. As the division is not uniform. Was the gap made such a way so as to have comparable number of articles in each one of them only?

7. The machine learning algorithms are widely used in robotics and pharmaceutical properties prediction. They can also be included in conclusion part.

8. Apart from SciMAT, are some other similar tools available? Can the comparative analysis of results using the same data set from different tools be done to know that how conclusive the reported results are? This can be included as future prospect of the current research.

References

1. Kumar R, Sharma A, Tiwari RK: Can we predict blood brain barrier permeability of ligands using computational approaches?. *Interdiscip Sci*. 2013; **5** (2): 95-101 PubMed Abstract | Publisher Full Text
2. Kumar R, Sharma A, Siddiqui MH, Tiwari RK: Promises of Machine Learning Approaches in Prediction of Absorption of Compounds.*Mini Rev Med Chem*. 2018; **18** (3): 196-207 PubMed Abstract | Publisher Full Text
3. Kumar R, Sharma A, Varadwaj PK: A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine.*J Nat Sci Biol Med*. 2011; **2** (2): 168-73 PubMed Abstract | Publisher Full Text
4. Kumar R, Sharma A, Siddiqui MH, Tiwari RK: Prediction of Metabolism of Drugs using Artificial Intelligence: How far have we reached?. *Curr Drug Metab*. 2016; **17** (2): 129-41 PubMed Abstract
5. Kumar R, Sharma A, Siddiqui MH, Tiwari RK: Prediction of Human Intestinal Absorption of Compounds Using Artificial Intelligence Techniques.*Curr Drug Discov Technol*. 2017; **14** (4): 244-254 PubMed Abstract | Publisher Full Text
6. Kumar R, Sharma A, Siddiqui MH, Tiwari RK: Prediction of Drug-Plasma Protein Binding Using Artificial Intelligence Based Algorithms.*Comb Chem High Throughput Screen*. 2018; **21** (1): 57-64 PubMed Abstract | Publisher Full Text

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?
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.

Referee Expertise: Machine learning classification, prediction models using Artificial Intelligence, Statistical analysis, Computational Biology

I have read this submission. I 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.