

A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field

Jose M. Alonso²

Ciro Castiello¹

Corrado Mencar¹

Abstract

This paper presents the results of a bibliometric study of the recent research on eXplainable Artificial Intelligence (XAI) systems. We took a global look at the contributions of scholars in XAI as well as in the subfields of AI that are mostly involved in the development of XAI systems. It is worthy to remark that we found out that about one third of contributions in XAI come from the fuzzy logic community. Accordingly, we went in depth with the actual connections of fuzzy logic contributions with AI to promote and improve XAI systems in the broad sense. Finally, we outlined new research directions aimed at strengthening the integration of different fields of AI, including fuzzy logic, toward the common objective of making AI accessible to people.

Key words: Interpretability, Understandability, Comprehensibility, Explainable AI, Interpretable Fuzzy Systems

1. Introduction

In the era of the Internet of Things and Big Data, data scientists are required to extract valuable knowledge from the given data. They first analyze, cure and pre-process data; then, they apply Artificial Intelligence (AI) techniques to automatically extract knowledge from data [20]. Getting AI into widespread real-world usage requires to think carefully of many important issues. Among them, we would like to highlight (1) Ethics, (2) Law and (3) Technology.

Recently, ACM issued a Statement on “Algorithmic Transparency and Accountability”, which establishes a set of principles, consistent with the ACM Code of Ethics, to support the benefits of algorithmic decision-making while addressing ethical and legal concerns [1]. Among such principles, *Explanation* is of relevance

Email addresses: josemaria.alonso.moral@usc.es (Jose M. Alonso), ciro.castiello@uniba.it (Ciro Castiello), corrado.mencar@uniba.it (Corrado Mencar)

¹Department of Informatics, University of Bari “Aldo Moro”, Bari, Italy

²Centro Singular de Investigacion en Tecnoloxias da Informacion (CiTIUS), Universidade de Santiago de Compostela, Campus Vida, E-15782, Santiago de Compostela, Galicia, Spain

Research Centre in Information Technologies (CiTIUS), University of Santiago de Compostela (USC) (May 2018).

Final version published in proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), 2018.

Please cite this article in press as:

J.M. Alonso, C. Castiello, C. Mencar, “A bibliometric analysis of the explainable artificial intelligence research field”, *Information Sciences*, 2018, DOI:10.1007/978-3-319-91473-2_1

Available online at http://dx.doi.org/10.1007/978-3-319-91473-2_1

for this study. According to ACM: “Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.”

In addition, a new European General Data Protection Regulation (GDPR) is expected to take effect in 2018 [11]. It takes care of the protection of natural people when personal data have to be processed and freely moved. Moreover, it emphasizes the “right to explanation” of European citizens: “[...] decision-making based on such processing, including profiling, should be allowed [...] In any case, such processing should be subject to suitable safeguards, which should include [...] the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.”

Regarding technological issues, the theme of explainability in AI is also remarked in the last challenge stated by the USA Defense Advanced Research Projects Agency (DARPA) [13]: “Even though current AI systems offer many benefits in many applications, their effectiveness is limited by a lack of explanation ability when interacting with humans.”

Accordingly, non-expert users, i.e., users without a strong background on AI, require a new generation of explainable AI (XAI) systems. Such systems are expected to naturally interact with humans by providing comprehensible explanations of decisions that are automatically made. XAI systems can be also considered as an important step forward toward Collaborative Intelligence [7] which promises a fully accepted integration of AI in our society.

In this paper, we report the results of a bibliometric study of the recent research on XAI systems. We are interested in assessing the contributions of AI scholars in XAI, as well as in the subfields of AI that are mostly involved in the development of XAI systems. More specifically, we are interested in the role of the fuzzy logic community in the progress of XAI, exploring the connections of fuzzy logic contributions with AI to promote and improve systems explainability. While moving along this way, we hope to outline new research directions aimed at strengthening the integration of different fields of AI, including fuzzy logic, toward the common objective of making AI accessible to people.

The rest of the manuscript is organized as follows. Section 2 introduces material and methods. Section 3 presents our bibliometric analysis focused on XAI. Section 4 introduces additional details while focusing on Interpretable Fuzzy Systems (IFS) only. Finally, Section 5 remarks the main points of the study and pinpoints future work.

2. Material and Methods

2.1. Bibliometric Techniques

Scientometrics is informally defined as the discipline that studies the quantitative features and characteristics of science and scientific research, technology and innovation. Within Scientometrics, Bibliometrics copes with the statistical analysis of books, articles, or other kinds of publications [6].

Usually, bibliographical data are treated by statistical mathematical methods and results are visualized in form of tables and graphs. For example, Vargas-Quesada et al. [25] proposed a methodology for creating visual representations of scientific domains. They focused on illustrating interactions among authors and papers through citations and co-citations. Later, other authors generalized the idea and developed alternative methods and tools (e.g., [5, 24]) to create maps of linked items (scientific publications, scientific journals, researchers, research organizations, countries, or keywords).

Different types of links between pairs of items can be considered. As an example, let us briefly introduce the concept of item co-citation. Given a set of items, all potential links among pairs of items can be characterized by the standardized co-citation measure [22] as follows:

$$MCN_{ij} = \frac{Cc_{ij}}{\sqrt{c_i \cdot c_j}} \quad (1)$$

where Cc means co-citation, c stands for citation, i and j are two different items.

The link values (MCN_{ij}) define the adjacency matrix of a graph which can be analyzed and visualized with social network analysis (SNA) techniques [26]. These techniques have been already applied to multiple fields of research, such as software development (e.g., debugging multi-agent systems [23]), scientometrics (e.g., analyzing large scientific domains [18]), or fuzzy modeling (e.g., analyzing fuzzy rule-bases with fingrams [19]).

There are many metrics designed to assess the importance of a node in a bibliographical graph (e.g., centrality degree, closeness, betweenness or page rank) [25]. In addition, there are many different methods for graph visualization [4]. Among them, force-directed algorithms are the most widely used in information science [16]. Their purpose is to locate the nodes of a graph in a 2D or 3D space, so that all the edges are approximately of equal length and there are as few crossing edges as possible, trying to obtain the most aesthetically pleasing view. There are also many clustering techniques aimed at discovering communities (or bunches of highly related nodes) in accordance with the importance of each single node and how it is connected to the others [21].

2.2. Bibliographic Repositories

Bibliographic data can be read from different sources such as Web of Science (WoS) or Scopus. WoS appears not to be adequate for assessing publications and citations in Informatics. In addition, some other sources may be too large (e.g., Google Scholar) or too specialized (e.g., ACM DL, IEEEExplore, etc.). Therefore, in this work we focus on Scopus which also offers advanced search functionalities useful to select meaningful sets of items which can be considered as a ground to build our bibliometric analysis. Anyway, the selection of Scopus as a bibliographical source comes without any loss of generalization. We performed a preliminary study on data collected from WoS: the main trends and the general conclusions remained unchanged (only slight minor variations were detected).

Finally, it should be highlighted that data collected from Scopus have been cured in order to remove spurious information that would have hampered the subsequent steps of our analysis.

2.3. Bibliographic Analysis Tools

We used a couple of tools to analyze the results of search queries from Scopus:

Bibliometrix [3] - An R package for performing comprehensive quantitative research in Scientometrics and Bibliometrics. It allows importing bibliographic data from several sources (including Scopus and WoS). In addition, it evaluates co-citation as well as other kinds of measures, such as coupling, scientific collaboration and co-word analyses.

VOS viewer [24] - A software tool for constructing and visualizing bibliometric networks which can be related to citation, co-citation, bibliographic coupling, co-authorship or co-occurrence of words relations. Some clustering methods to identify related groups or communities are also provided.

3. A Global Overview on XAI

On October 20th, 2017, we ran the following query through the “Advanced Search” tool provided by Scopus:

```
Q1 = TITLE ('*interpretab*') OR TITLE ('*comprehensib*')
OR TITLE ('*understandab*') OR TITLE ('*explainab*')
OR TITLE ('*self-explanat*') OR KEY ('*interpretab*')
OR KEY ('*comprehensib*') OR KEY ('*understandab*')
OR KEY ('*explainab*') OR KEY ('*self-explanat*')
```

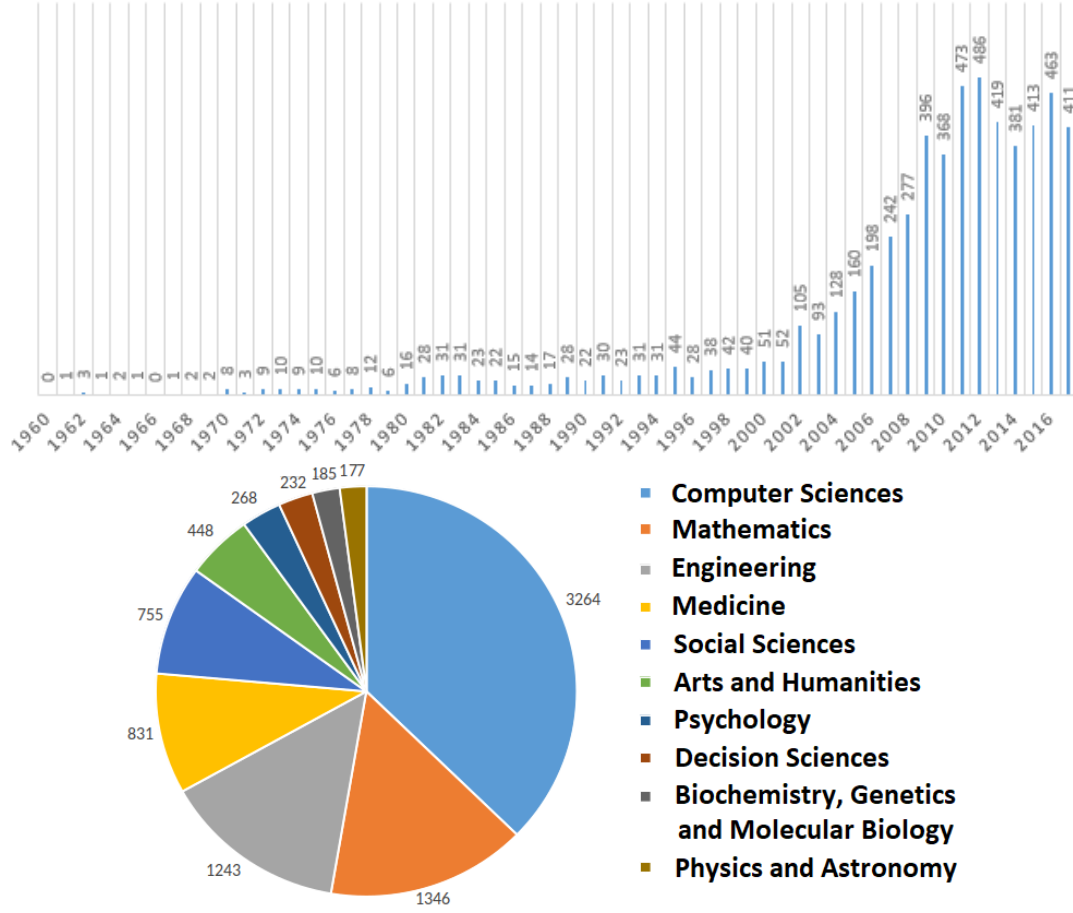


Figure 1: Histogram of XAI publications and their distribution by subject areas.

As a result, we found out 5735 documents. It is worthy to note this query is intentionally very general in order to broaden the global picture of the research field under examination. We identified only 5 general terms and their variants represented by the * symbol. We required at least one of these terms to be present in the title of the retrieved document or in the associated keywords (provided by authors or automatically indexed).

Fig. 1 depicts the number of XAI publications since 1960 (top picture) and the distribution of publications in the top-10 ranking of subject fields (bottom picture). The number of publications started to grow significantly since 2000. Accordingly, we decided to focus our analysis only on the years ranging from 2000 to 2017.

XAI represents a multidisciplinary research field, as witnessed by the variety of subject areas. Anyway, three of them (Computer Sciences, Mathematics, and Engineering) collect most of the publications. Therefore, we are going to pay attention only to publications in these research areas. In this way, the final number of publications to analyze is 3737. We downloaded from Scopus all the related bibliographical information in form of csv and bib files.

Table 1 presents the Top-5 rankings of authors (columns 2-4) and countries (columns 5 and 6) with respect to h-index, total number of citations (TC) and publications (NP). Herrera stands as the leading author in terms of h-index, TC and NP. USA is by far the leading country in terms of both TC and NP.

The leading publications are listed in Table 2 in terms of TC and average citations per year (ACY). Guillaume [12] reviewed methods for automatically designing IFS. This is the most cited publication being

Table 1: Top-5 Ranking of XAI Authors and Countries in terms of h-index, Total Citations (TC) and Number of Publications (NP).

Rank.	Authors (h-index)	Authors (TC)	Authors (NP)	Countries (TC)	Countries (NP)
1	Herrera (22)	Herrera (2287)	Herrera (45)	USA (6737)	USA (604)
2	Alcalá (13)	Alcalá (1216)	Mencar (29)	Spain (4628)	China (323)
3	Mendling (13)	Baesens (846)	Piattini (29)	Germany (2242)	Spain (318)
4	Alonso (12)	Mendling (717)	Alonso (26)	UK (1703)	Germany (270)
5	Piattini (11)	Guillaume (707)	Alcalá (25)	China (1692)	UK (175)

Table 2: Top-5 Ranking of XAI Publications in terms of Total Citations (TC) and Average Citations per Year (ACY).

TC Rank.	Publication (Authors, Year, Source, TC) [Ref]
1	S. Guillaume, 2001, IEEE T Fuzzy Syst, 435 [12]
2	V. Alevn and K. Koedinger, 2002, Cognitive Sci, 418 [2]
3	Y. Jin, 2000, IEEE T Fuzzy Syst, 333 [15]
4	S. García et al., 2009, Soft Comput, 308 [10]
5	H. Ishibuchi and Y. Nojima, 2007, Int J Approx Reason, 266 [14]
ACY Rank.	Publication (Authors, Year, Source, ACY) [Ref]
1	L. Martínez and F. Herrera, 2012, Inform Sciences, 39.8 [17]
2	S. García et al., 2009, Soft Comput, 38.5 [10]
3	M.J. Gacto et al., 2011, Inform Sciences, 33.3 [9]
4	V. Alevn and K. Koedinger, 2002, Cognitive Sci, 27.9 [2]
5	S. Guillaume, 2001, IEEE T Fuzzy Syst, 27.2 [12]

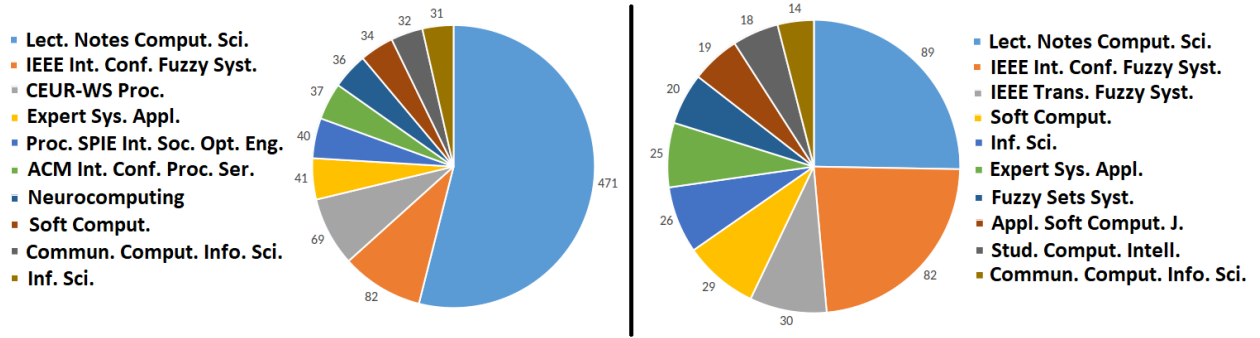


Figure 2: Pie charts with the leading sources in XAI (left) and IFS (right).

also the fifth as for ACY. Alevn and Koedinger [2] described how to improve students' learning with a computer-based approach endowed with self-explanation. This is the second most cited publication, ranked fourth in terms of ACY. Jin [15] authored the third most cited publication, presenting a fuzzy modeling approach designed to improve the interpretability of high-dimensional systems. This publication is out of the Top-5 in terms of ACY. Garcia et al. [10] reviewed statistical techniques to get a good interpretability-accuracy trade-off in genetics-based machine learning. This is the fourth most cited paper and the second one in terms of ACY. The fifth publication in terms of TC (out of the ACY Top-5) comes from Ishibuchi and Nojima [14] who applied a multi-objective genetics-based machine learning approach to build fuzzy systems with a good interpretability-accuracy trade-off. The scenario is completed by Martínez and Herrera [17] (first paper in terms of ACY), who proposed a linguistic model for solving decision-making problems, and Gacto et al. [9] (third paper in terms of ACY), who reviewed interpretability indexes for assessing IFS. Notice that Herrera co-authored 3 of the Top-5 publications in terms of ACY: this emphasizes his leading role in the XAI research field (see Table 1).

The leading sources in XAI are depicted in the pie chart on the left of Fig. 2. Most papers are published in conference proceedings. Nevertheless, the Top-5 papers (see Table 2) appear in well-recognized journals.

Fig. 3 shows a graph with the most popular author keywords in the publications under study. Each node is associated to a keyword and its size is proportional to the number of documents where the keyword

appears. *Interpretability* is the main keyword since it is associated to the larger node. *Understandability* and *classification* are the second and the third main keywords. Links between nodes relate keywords which usually appear together in the same documents.

This graph gives a global overview about the main topics of interest in the XAI research field, with groups of closely related nodes painted in the same color. On the one hand, *interpretability* is closer to topics usually addressed in the fuzzy logic community (e.g., *fuzzy modeling* or *rule selection*). On the other hand, *understandability* is surrounded by keywords related to software engineering. The gap between the main keywords is filled by other relevant nodes such as *comprehensibility* or *self-explanation*. Moreover, *interpretability* and *comprehensibility* are related to a group of keywords including popular topics in AI (e.g., *classification*, *data mining* or *knowledge discovery*). A community of keywords is partially disconnected from the rest of the graph (e.g., *semantic web*, *ontology*, and so on), and some single nodes lie away from others (e.g., *interpretability logic* or *image interpretability*). That is due to their relatedness to some specific research lines. Notice that NIIRS stands for National Imagery Interpretability Rating Scale which is a subjective scale for rating the quality of images.

When we turn to consider author co-citation, we look for pairs of authors being cited by the same publications. Fig. 4 shows the co-citation map obtained by the VOS viewer (the minimum number of total citations by author is set to 50). Size of nodes is proportional to the number of citations, while link weights

The graph was generated by the VOS viewer employing the suggested default parameters for layout visualization and clustering of nodes. Other clustering approaches may be applied, but choosing the best approach is out of the scope of this paper.

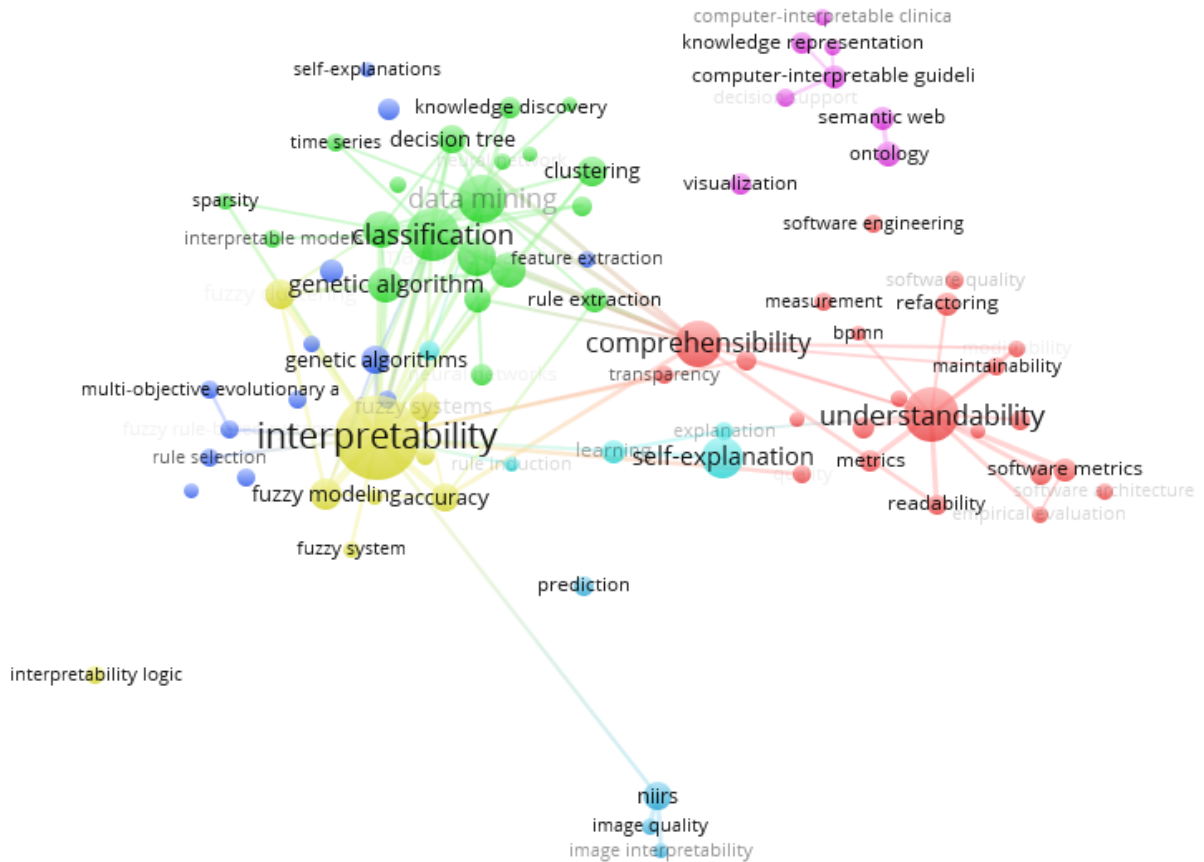


Figure 3: Map of author keywords in XAI.

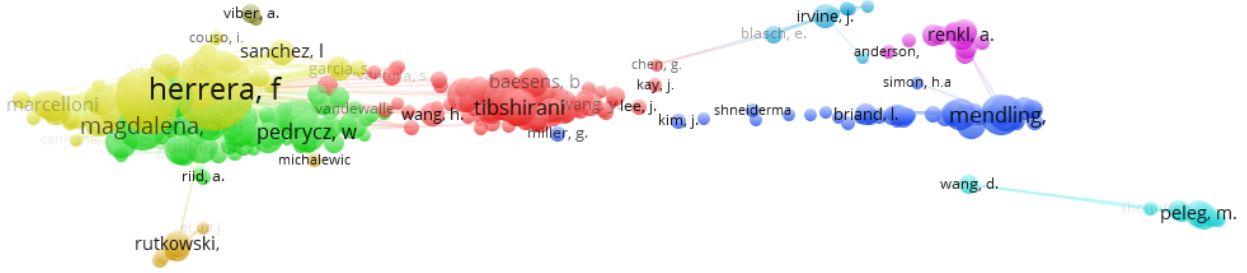


Figure 4: Map of author co-citation in XAI.

Table 3: Top-5 Ranking of IFS Authors and Countries.

Rank.	Authors (h-index)	Authors (TC)	Authors (NP)	Countries (TC)	Countries (NP)
1	Herrera (20)	Herrera (1900)	Herrera (42)	Spain (3255)	Spain (177)
2	Alcalá (13)	Alcalá (1216)	Mencar (29)	Italy (850)	China (116)
3	Alonso (12)	Guillaume (707)	Alonso (26)	France (766)	Italy (81)
4	Magdalena (11)	Ishibuchi (646)	Alcalá (25)	UK (749)	UK (68)
5	Mencar (10)	Nojima (568)	Fanelli (24)	Poland (552)	India (42)

come from the co-citation index defined by eq. (1). Most nodes are concentrated in the left-hand side of the map. Again, Herrera stands out as the main node.

4. Detailed Analysis on Interpretable Fuzzy Systems

We replicated the previous analysis with a modified query:

Q2 = Q1 AND ‘‘fuzzz*’’

By adding ‘‘fuzzz*’’ to Q1 we focus our search on publications in the XAI field that are related to fuzzy sets and systems. Hereafter, we refer to this field of research as IFS. In addition, we filtered the collected results by adopting the same constraints imposed in the previous section: (1) years range [2000-2017] and (2) subject areas [Computer Sciences, Mathematics and Engineering]. As a result, we got 1054 documents, consisting in about 28% of the whole set of documents previously analyzed.

The Top-5 rankings of authors and countries is detailed in Table 3 concerning IFS. Most authors in Table 3 are present also in Table 1, thus certifying the relevance of the fuzzy community in the context of XAI. However, USA (the leading country in Table 1) is now out of the Top-5. Moreover, European countries take up the Top-5 in terms of TC. China and India appear only when looking at NP. These data reflect the relevance of European scholars in the fuzzy community and their outstanding leadership in IFS.

Table 4: Top-5 Ranking of IFS Publications.

TC Ranking	Publication (Authors, Year, Source, TC) [Ref]
1	S. Guillaume, 2001, IEEE T Fuzzy Syst, 435 [12]
2	Y. Jin, 2000, IEEE T Fuzzy Syst, 333 [15]
3	H. Ishibuchi and Y. Nojima, 2007, Int J Approx Reason, 266 [14]
4	M.J. Gacto et al., 2011, Inform Sciences, 200 [9]
5	L. Martínez and F. Herrera, 2012, Inform Sciences, 199 [17]
ACY Ranking	Publication (Authors, Year, Source, ACY) [Ref]
1	L. Martínez and F. Herrera, 2012, Inform Sciences, 39.8 [17]
2	M. Fazzolari et al., 2013, IEEE T Fuzzy Syst, 34 [8]
3	M.J. Gacto et al., 2011, Inform Sciences, 33.3 [9]
4	S. Guillaume, 2001, IEEE T Fuzzy Syst, 27.2 [12]
5	H. Ishibuchi and Y. Nojima, 2007, Int J Approx Reason, 26.6 [14]

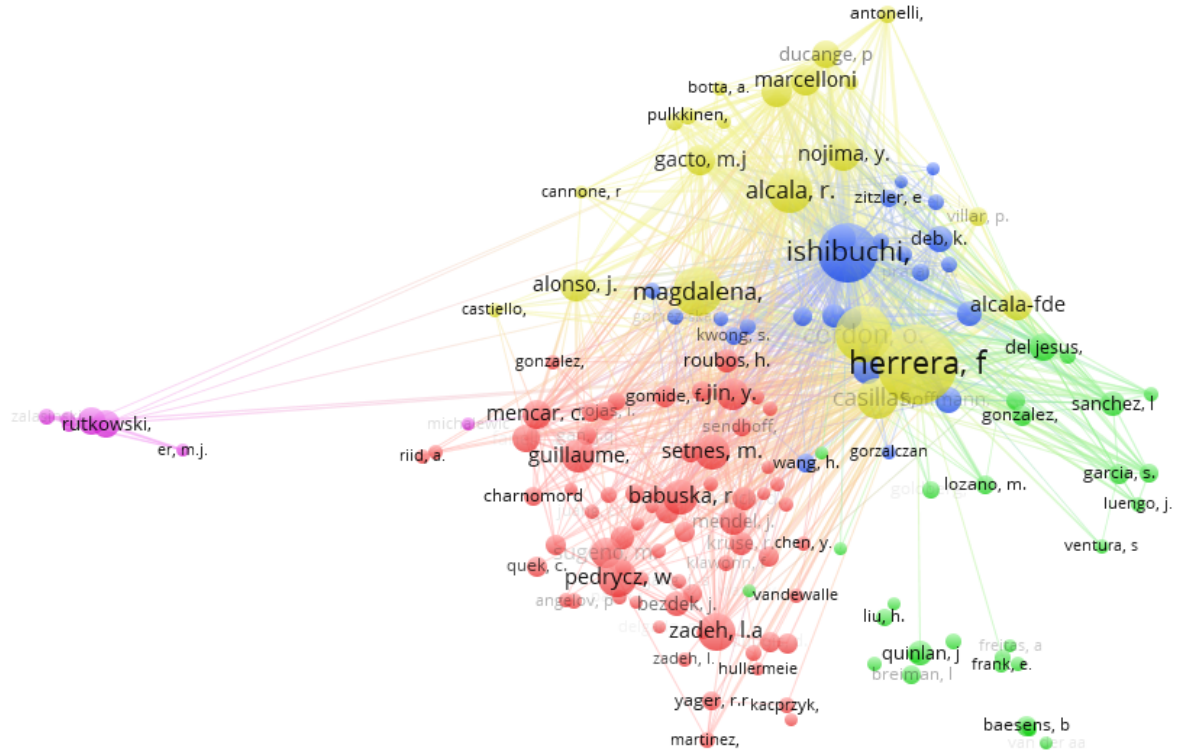


Figure 6: Map of author co-citation in IFS.

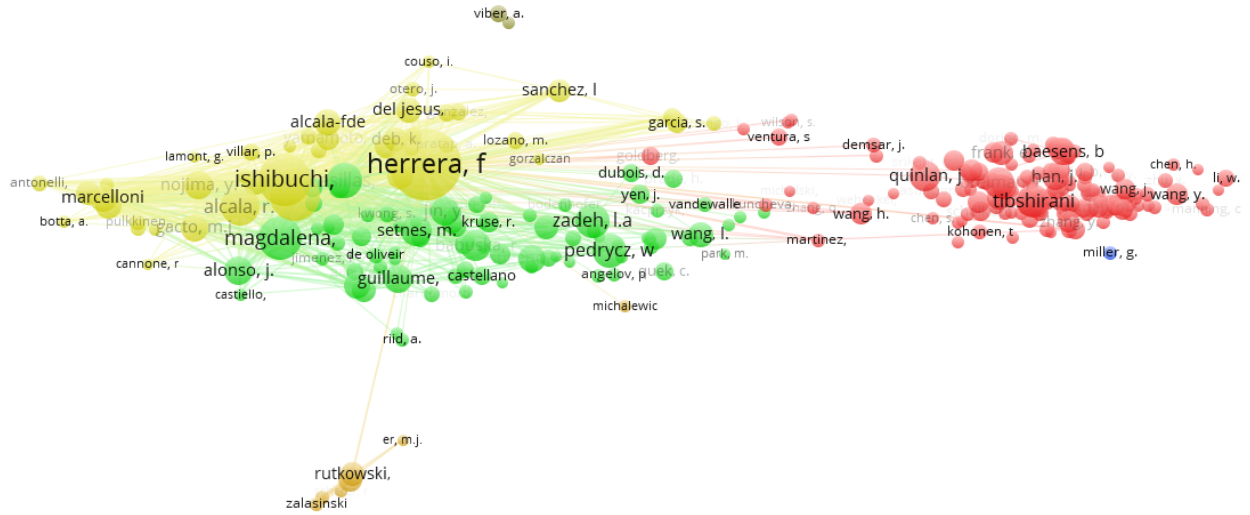


Figure 7: Map of author co-citation in XAI (zoom in the left hand side of Fig. 4).

where a zoom of the left-hand side of Fig. 4 is provided: authors related to fuzzy logic appear to be aggregated in two compact clusters (the yellow and the green ones in Fig. 7) demonstrating also a tight interconnection of the related research activities. On the other hand, authors in XAI not related to fuzzy logic appear to be loosely distributed, as a sign of a more scattered collaboration.

This analysis suggests at least two lines of development. Firstly, there is a need to clarify and distinguish

the notions of *interpretability*, *comprehensibility*, *understandability* and *explainability* to provide a common terminological ground inside the varied XAI context. This could also shed light on refined conceptualizations where fuzzy logic could significantly contribute. Moreover, an opportunity emerges to tighten the connections of studies between fuzzy and non-fuzzy worlds of XAI, which now appear unnecessarily separated. We strongly believe that cross-fertilization between these communities is needed to successfully face the challenges posed by XAI.

Acknowledgements

This work was supported by RYC-2016-19802 (Ramón y Cajal contract), and two MINECO projects TIN2017-84796-C2-1-R (BIGBISC) and TIN2014-56633-C3-3-R (ABS4SOW). All of them funded by the Spanish “Ministerio de Economía y Competitividad”. Financial support from the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2016-2019) and the European Union (European Regional Development Fund - ERDF), is gratefully acknowledged.

References

- [1] ACM US Public Policy Council, A., 2017. Statement on Algorithmic Transparency and Accountability. URL: http://www.acm.org/binaries/content/assets/public-policy/2017{_}usacm{_}statement{_}algorithms.pdf.
- [2] Aleven, V., Koedinger, K., 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Sci* 26, 147–179. doi:10.1016/S0364-0213(02)00061-7.
- [3] Aria, M., Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *J Informetr* 11, 959–975. doi:10.1016/j.joi.2017.08.007.
- [4] di Battista, G., Eades, P., Tamassia, R., Tollis, I., 1998. Graph drawing: algorithms for the visualization of graphs. Prentice Hall, Upper Saddle River, N.J.
- [5] Cobo, M., López-Herrera, A., Herrera-Viedma, E., Herrera, F., 2011. Science mapping software tools: Review, analysis, and cooperative study among tools. *J Assoc Inf Sci Tech* 62, 1382–1402.
- [6] De Bellis, N., 2009. Bibliometrics and citation analysis: from the science citation index to cybermetrics. Scarecrow Press.
- [7] Epstein, S.L., 2015. Wanted: Collaborative intelligence. *Artif Intell* 221, 36–45. doi:10.1016/j.artint.2014.12.006.
- [8] Fazzolari, M., Alcalá, R., Nojima, Y., Ishibuchi, H., Herrera, F., 2013. A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions. *IEEE T Fuzzy Syst* 21, 45–65. doi:10.1109/TFUZZ.2012.2201338.
- [9] Gacto, M., Alcalá, R., Herrera, F., 2011. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Inform Sciences* 181, 4340–4360. doi:10.1016/j.ins.2011.02.021.
- [10] García, S., Fernandez, A., Luengo, J., Herrera, F., 2009. A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Comput* 13, 959–977. doi:10.1007/s00500-008-0392-y.
- [11] Goodman, B., Flaxman, S., 2016. European union regulations on algorithmic decision-making and a “right to explanation”, in: ICML Workshop on Human Interpretability in Machine Learning (WHI), New York, NY. pp. 1–9.
- [12] Guillaume, S., 2001. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE T Fuzzy Syst* 9, 426–443. doi:10.1109/91.928739.
- [13] Gunning, D., 2016. Explainable Artificial Intelligence (XAI). Technical Report. Defense Advanced Research Projects Agency. Arlington, USA. DARPA-BAA-16-53.
- [14] Ishibuchi, H., Nojima, Y., 2007. Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *Int J Approx Reason* 44, 4–31. doi:10.1016/j.ijar.2006.01.004.
- [15] Jin, Y., 2000. Fuzzy modeling of high-dimensional systems: Complexity reduction and interpretability improvement. *IEEE T Fuzzy Syst* 8, 212–221. doi:10.1109/91.842154.
- [16] Kobourov, S.G., 2012. Force-directed drawing algorithms, in: Tamassia, R. (Ed.), Handbook of Graph Drawing and Visualization. CRC Press. chapter 12.
- [17] Martínez, L., Herrera, F., 2012. An overview on the 2-tuple linguistic model for computing with words in decision making: Extensions, applications and challenges. *Inform Sciences* 207, 1–18. doi:10.1016/j.ins.2012.04.025.
- [18] Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Muñoz-Fernández, F.J., 2004. A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics* 61, 129–145.
- [19] Pancho, D., Alonso, J., Cordon, O., Quirin, A., Magdalena, L., 2013. Fingrams: Visual representations of fuzzy rule-based inference for expert analysis of comprehensibility. *IEEE T Fuzzy Syst* 21, 1133–1149. doi:10.1109/TFUZZ.2013.2245130.
- [20] Philip Chen, C., Zhang, C., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inform Sciences* 275, 314–347.
- [21] Porter, M., Onnela, J., Mucha, P., 2009. Communities in networks. *Not Am Math Soc* 56, 1082–1166.
- [22] Salton, G., Bergmark, D., 1979. A citation study of computer science literature. *IEEE T Prof Commun* 22, 146–158.

- [23] Serrano, E., Quirin, A., Botia, J., Cerdón, O., 2010. Debugging complex software systems by means of pathfinder networks. *Inform Sciences* 180, 561–583.
- [24] Van Eck, N., Waltman, L., 2010. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics* 84, 523–538.
- [25] Vargas-Quesada, B., Moya-Anegón, F., 2007. *Visualizing the Structure of Science*. Springer-Verlag.
- [26] Wasserman, S., Faust, K., 1994. *Social Network Analysis: Methods And Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.