Revista Española de Documentación Científica
42(1), enero-marzo 2019, e228
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2019.1.1583

ESTUDIOS / *RESEARCH STUDIES*

Analysis of scientific production based on trending research topics. An Artificial Intelligence case study

Jesús Bobadilla*, Abraham Gutiérrez*, Miguel Ángel Patricio**, Rodolfo Xavier Bojorque***

* Universidad Politécnica de Madrid, Computer Science.
e-mail: jesus.bobadilla@upm.es | ORCID iD: https://orcid.org/0000-0003-0619-1322
e-mail: abraham@etsisi.upm.es | ORCID iD: https://orcid.org/0000-0001-6974-7514

** Universidad Carlos III, Madrid, Spain. Computer and Engineering Dept.
e-mail: mpatrici@inf.uc3m.com | ORCID iD: https://orcid.org/0000-0002-9304-826X

*** Universidad Politécnica Salesiana. Ecuador. Department of Computer Science
e-mail: rbojorque@ups.edu.ec | ORCID iD: https://orcid.org/0000-0002-6045-8692

Recibido: 23-04-2018; 2ª versión: 02-07-2018; Aceptado: 03-07-2018

Citation/Cómo citar este artículo: Bobadilla, J.; Gutiérrez, A.; Patricio, M. A.; Bojorque, R. X. (2019). Analysis of scientific production based on trending research topics. An Artificial Intelligence case study. *Revista Española de Documentación Científica*, 42 (1): e228. https://doi.org/10.3989/redc.2019.1.1583

Abstract: Scientific documentation research leads to the computation of large amounts of information from published works of the scientific community. It is necessary to explain these processes and create application frameworks. This paper provides the following: a) An *Information System* designed to extract scientific information from published papers, b) Accurate explanations of the main processing stages including data mining, natural language processing, and machine learning, and c) Categorized and explained results coming from the *Artificial Intelligence* case study. The results in this paper include the following: a) Topics and research area rankings and b) Quantity versus quality comparisons of topics and research areas.

Keywords: Research topics; Scientific production; Scientific documentation; Machine learning; Data mining; Natural language processing; artificial intelligence trends; Scopus.

Analisis de la producción científica basado en las tendencias en temas de investigación. Un estudio de caso sobre inteligencia artificial

Resumen: La investigación en el campo de la documentación científica nos lleva hacia un procesamiento automático de grandes cantidades de información proveniente de los trabajos publicados por la comunidad científica. Resulta necesario explicar estos procesos y crear sistemas que los lleven a cabo. En este artículo se proporciona: a) Un *Sistema de Información* diseñado para extraer información científica a partir del texto que proporcionan los artículos publicados, b) Explicaciones de las etapas fundamentales de procesamiento: minería de datos, procesamiento del lenguaje natural, aprendizaje automático, y c) Resultados categorizados y explicados de nuestro caso de estudio: el área *Artificial Intelligence*. Los resultados de este artículo incluyen: a) Ranking de temas y ranking de áreas de investigación, y b) Comparativa entre cantidad y calidad de los temas y de las áreas de investigación.

Palabras clave: Temas de investigación; producción científica; Documentación Científica; aprendizaje automático; recogida de datos; Scopus; procesamiento de lenguaje natural; inteligencia artificial.

1. INTRODUCTION

This section is divided into the following three subsections: 1) Related works, 2) Motivation, and 3) Machine learning introduction. The first subsection provides a reference to the most relevant publications from various knowledge areas related to this paper. In the second subsection, the objectives of the paper are summarized and the reasons that support its usefulness are described. The third subsection explains how machine learning converts data into information, supporting the results of this paper. An experiments design section is provided, followed by a discussion of the experimental results. Finally, the conclusions and future work are reviewed.

1.1. Related work

High-level topics and trends are very important in making decisions for managers and developers (Hindle et al., 2015). Usually, topics are processed using machine learning latent Dirichlet allocation (LDA) (Bleu et al., 2003). Currently, there are some machine learning alternatives to LDA, presenting some advantages over traditional methods; matrix factorization (Xue et al., 2017) is usually applied when data is sparse and unknown values require prediction, such as in collaborative filtering recommender systems (Bobadilla et al., 2013). This is not the case when processing words in a natural language processing (NLP) scenario (Sun et al., 2017) because missing words are counted as zero instead of an unknown value. Recently, a matrix factorization probabilistic method (Hernando et al., 2016) overcomes this limitation. As an alternative, clustering (Bobadilla et al., 2017) is a powerful tool to mine topics from words and group the words semantically (Wang and Koopman, 2017).

Word embedding methods are being used for topic detection (Naili et al., 2017). Particularly, *word2vec* (Altszyler et al., 2016) and *topic2vec* perform well when relating topics. Co-word analysis (Ravikumar et al., 2015) can be used to improve the trend results of topics. Computational overhead for topic model training may be reduced by selectively removing terms from the vocabulary of text corpora. Document space density can be reduced by the removal of frequently occurring terms, but it increases with greater numbers of topics (Lu et al., 2017).

Altmetrics (Haustein et al., 2014) and webometrics (Kaya et al., 2010) focus on the bibliometric and informetric study of different online information systems. This study will include altmetrics and webometrics quality measures; in this phase, information is merged (Karlsson et al., 2014; Yu, 2015) from various sources in a big data mining process. In the short term, the information source will be based on the *Scopus* data, and the quality measures will come from the classic Bibliometrics (Manolopoulus and Katsaros, 2017) field. In particular, the information types used in Lis-Gutierrez et al., (2017), production volume, document typology, number of citations, institutional affiliation of researchers, sources, main journals, and researcher country of origin, will be used. From the country information of researchers, scientific and institutional collaborations will be extracted (Ortoll et al., 2014). The quantity and quality of the scientific output of the topmost 50 countries in four basic sciences have been previously studied (NejatiSeyyed and Jenab, 2010). Disaggregated by topics, h-index improvements as a quality measure can be applied to the above results (Nedra et al., 2015).

One of the priority objectives of this study is to make a comparison between the quality of production and its quantity. A primary indicator of quality is taken as the number of *Scopus* citations (Mingers and Leydesdorff, 2015; Yazdani et al., 2015), which is used in many studies in the scientometrics area. The number of citations in *Google Scholar* has not been used due to the controversy in this academic search engine (Orduna-Malea et al., 2017). In any case, *Google Scholar* offers an original and different vision of the most influential academic documents (measured from the perspective of their citation count) (Martín-Martín et al., 2016). Quantity versus quality bibliometric studies show correlations between the number of papers and citations (Hayati, 2009). This research is typical in the scientific documentation field. This type of study opens the door to the achievements of related results, such as the correlation between the most productive authors and those most cited (Abramo et al., 2014).

Most research based on scientific production datasets (Aksnes, 2003; Aksnes and Sivertsen, 2004) focuses on the bibliometric analysis of journals; they reveal their main scientometric factors (Aguillo et al., 2010; Bornmann and Mutz, 2011). This paper also analyses selected scientific journals. However, the focus is on a different aspect, which is the determination of research topics and their evolution. In this way, the final objective of this work moves away from the quality of journals and focuses on discovering scientific topics in publications. Therefore, in this paper, due to length limitations, the computer science technological area is examined to include the growing subfield of artificial intelligence. The intention is to extend this study to all *Journal Citation Reports* (*JCR*) areas.

Information and communication technology (ICT) contributes to the economic growth of most countries. This production area is especially challenging due to its enormous variability over time. To be able to conduct precise planning and identify technological issues with the greatest potential, it is necessary to perform studies of research topic evolution in ICT. Preparing ICT graduates is a strategic task for governments, which is addressed in Anicic et al. (2016) according to the following parameters: curriculum design and delivery, knowledge and skills of future ICT professionals, teaching methods, collaboration between academia and industry, and future employment and career development of ICT professionals in the labour market. It is possible to use network approaches based on large datasets such as Web of Science (Khan and Niazi, 2017). The Web of Science was studied from 2014 to 2017, and it shows co-citation patterns of documents, co-occurrence patterns of terms, and the most influential articles, among others. An alternative work from (Mustafee et al., 2014) uses co-citation analysis as a knowledge method.

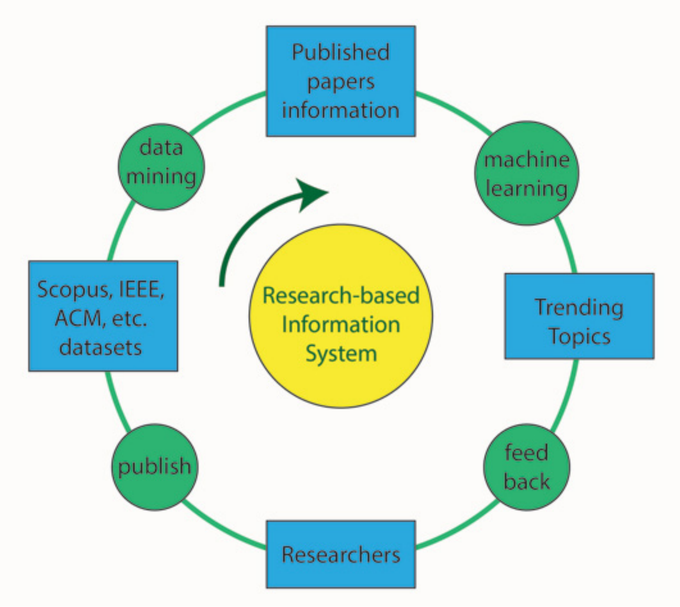
1.2. Motivation

This paper shows the necessary tasks to extract the most popular topics from relevant scientific publications. The *Information System* developed in this work to carry out the experiments has been designed in such a way that operations can be repeated by taking different research areas as a starting point. By way of an example, the research area of *Computer Sciences, Artificial Intelligence* has been chosen. In the same way, the results can be obtained from different potential research areas. To do it, we can data mine any of the Journal Citations Report areas and then apply to the data base, the same natural language processing methods and the same machine learning algorithms that we explain in this paper. Since it is possible to make a data mining of any existing research area, the concepts explained in this paper can be fully extended.

Knowing trending topics from a research area has considerable advantages for society, including 1) allowing a focus on teachings and curricula towards academic subjects with more future potential, 2) providing information to government institutions about their strategic lines of action, 3) significantly improving the impact of grants, subsidies and institutional investments, 4) guiding students towards subjects with better perspectives, and 5) directing companies towards business models with greater possibilities of succeeding in the medium term.

Figure 1 shows the feedback process that can be used to improve research production tasks. The scientific community is the starting point that carries out research activities, as shown in the bottom of Figure 1. Starting from scientific publications, research paper repositories are maintained by various publishers, as shown on the left side of Figure 1. The data mining process from the publisher datasets is completed, obtaining a database with the

Figure 1. Research-based information system architecture

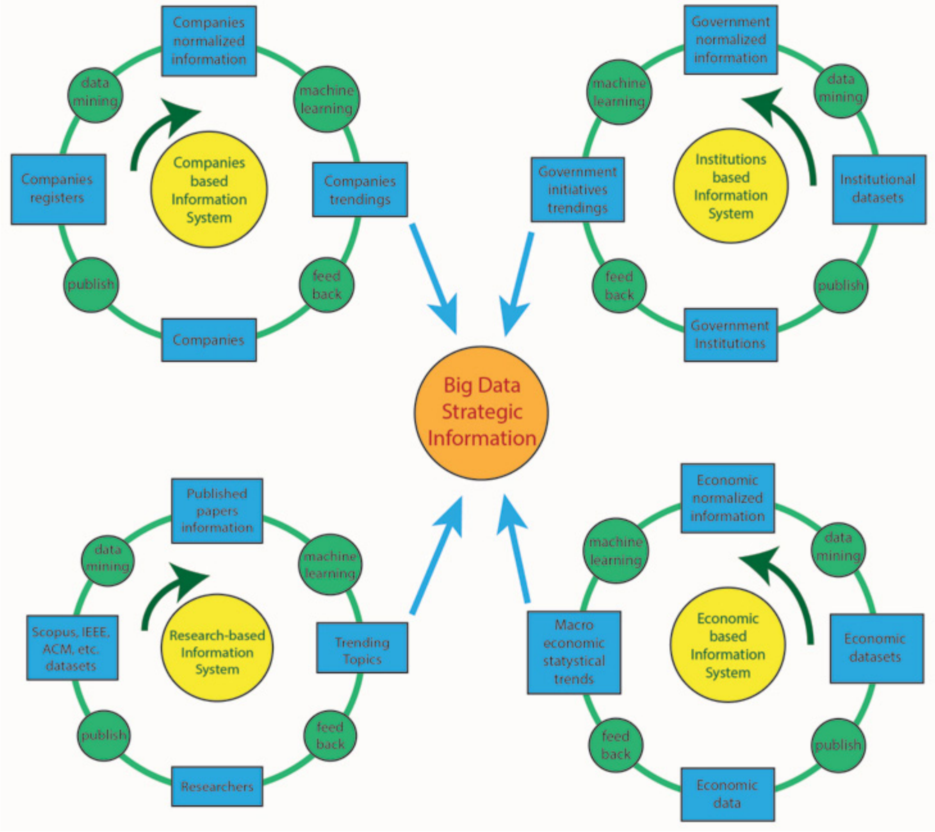

Published papers information
machine learning
Trending Topics
feed back
Researchers
publish
Scopus, IEEE, ACM, etc. datasets
data mining
Research-based Information System

most relevant information of the published papers, as shown in the top of Figure 1. The large amount of information in published papers that is stored in the database is processed using modern machine learning methods and algorithms, and the expected research trending topics are obtained, as shown in the right side of Figure 1. Finally, researchers benefit from the feedback provided by the system, directing their research towards topic areas that promise greater projection.

The methods, algorithms, designs and technologies presented in this article could be considered as one of the constituent parts of a broader architecture that pursues the following ambitious objective: to provide the information processes of a technological *think tank*. The *think tank*, in this paper example, offers strategic information related to the Information and Communication Technology field. Such an Information System is composed of a variety of abstraction levels, including data mining from various data sources, machine learning, automatic decision making and expert decision making. Figure 2 schematizes the complete architecture, where different information sources are provided including research, government institutions, companies and macroeconomic data. The central circle in Figure 2 represents the results of the big data process. The information generated can be used to determine strategic actions. These strategic actions will improve company, research, economic and government results. In this way, the process is fed back, and continuous improvement of the productive tasks is obtained.

Additionally, all the processed information can be used to provide useful tools to the research community: Scientific Documentation Recommender Systems. Although traditional recommender systems are based on explicit ratings (e.g.: users voting movies), there is a growing field where recommender systems are based on implicit ratings (e.g: songs listened by each user). These concepts can be extended through recommender systems where we assign a value to each existing topic of each data mined paper. As in real recommender systems, the resulting dataset will be very sparse, since each paper only is related with a reduced number of topics; in the same way that each user only votes a reduced

Figure 2. Think tank Information System architecture


Companies normalized information
machine learning
Companies trendings
feed back
Companies
publish
Companies registers
data mining
Companies based Information System
Government normalized information
data mining
Institutional datasets
publish
Government Institutions
feed back
Government initiatives trendings
machine learning
Institutions based Information System
Big Data Strategic Information
Published papers information
machine learning
Trending Topics
feed back
Researchers
publish
Scopus, IEEE, ACM, etc. datasets
data mining
Research-based Information System
Economic normalized information
data mining
Economic datasets
publish
Economic data
feed back
Macro economic statystical trends
machine learning
Economic based Information System

number of films or she only hears a reduced number of songs. Such scientific documentation recommender systems will be used to recommend papers or topics to researchers, as well as related researchers to share their work.

1.3. Machine learning introduction

Machine learning is a necessary process to obtain useful information from research publications, economic data, institutional reports, etc. (Figure 2). The machine learning methods allow tasks to be completed, such as the following: a) To extract topics from texts, b) To relate topics and publications to each other, c) To detect complex correlations in the data, d) To establish clustering results, e) To provide predictions and recommendations, and f) To find temporal and/or spatial patterns and trends.

This paper focuses on machine learning processes that are able to complete the following: a) Extract topics from published research information, b) Establish a ranking of extracted topics, and c) Find relationships and clustering of existing topics.

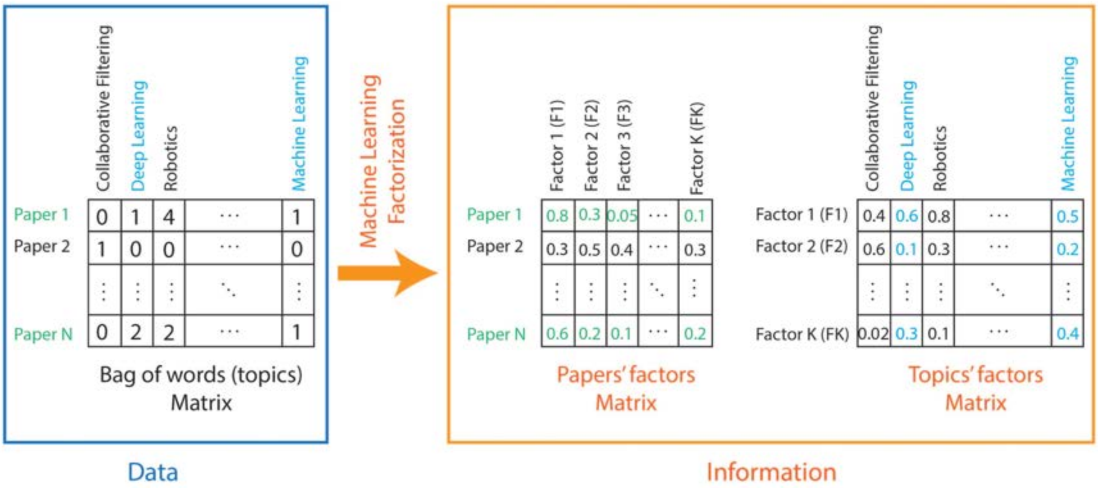
The starting point to obtain the desired results is the information coming from the research papers. Each paper provides a broad set of information, and this study is especially interested in keywords, index-keywords, titles and abstracts. From this information, a table can be made with the format shown in the left part of Figure 3: "Bag of words (topics) Matrix". This matrix contains the number of times each topic appears in each publication.

Matrix factorization is a classic machine learning process. Normally, we start with a bag of words matrix (natural language processing), a ratings matrix (recommender systems), speech features (speech processing), pixels (image compression), etc. Classic factorization processes provide two matrices as a result, the matrix of paper factors and the matrix of topic factors (Figure 3 example). Factors in both matrices have the same unknown meaning (hidden factors). To understand this concept, each factor indicates one or several characteristics from the papers and topics, e.g., F4 may indicate that the paper or topic belongs to the computer vision area, while F7 may be linked to the areas of speech recognition and speech synthesis.

It is important to note that the number of factors *K* is much lower than the number of topics (usually a few thousand) and much lower than the number of papers (usually tens of thousands or hundreds of thousands). In this way, the bag of words matrix is much larger than the sum of the sizes of the factor matrices (papers and topics). Although the factorization process requires intensive calculations, the subsequent information processing is simple and effective.

From the bag of words matrix (left side of Figure 3) we can obtain the importance of each topic and we can obtain a ranking of topics. From the matrices of factors (right side of Figure 3), the following can be determined: a) relationships between papers, b) relationships between topics, d) groups (clusters) of papers, and e) groups (clusters) of topics. The topics ranking, in a simple approximation, can be established by counting the number of times each topic appears in the set of papers. This mechanism can be improved by setting several weights such as the impact factor of the journal, places where each topic appears in each paper, etc.

Figure 3. Machine learning applied to scientific production



	Collaborative Filtering	Deep Learning	Robotics		Machine Learning
Paper 1	0	1	4	···	1
Paper 2	1	0	0	···	0
	⋮	⋮	⋮	⋱	⋮
Paper N	0	2	2	···	1

	Factor 1 (F1)	Factor 2 (F2)	Factor 3 (F3)		Factor K (FK)
Paper 1	0.8	0.3	0.05	···	0.1
Paper 2	0.3	0.5	0.4	···	0.3
	⋮	⋮	⋮	⋱	⋮
Paper N	0.6	0.2	0.1	···	0.2

	Collaborative Filtering	Deep Learning	Robotics		Machine Learning
Factor 1 (F1)	0.4	0.6	0.8	···	0.5
Factor 2 (F2)	0.6	0.1	0.3	···	0.2
	⋮	⋮	⋮	⋱	⋮
Factor K (FK)	0.02	0.3	0.1	···	0.4

Figure 3 is used to illustrate each relationship, and clusters can be made from the information contained in the factors matrices as follows: each row of the papers factors matrix contains factor values that define each paper. Each column of the topics factors matrix contains the factor values that define each topic. As an example, paper *1* is represented heavily by factor *1* (*F1*), as well as the topic "Robotics". It is likely that paper *1* addresses the topic "Robotics", as shown in the data matrix.

Figure 3 shows how paper *1* and paper *N* (both in green) are related. This relationship is difficult to see in the original matrix, where there are thousands of topics. However, if we go to the papers' factors matrix, it is very simple and precise to compare the factors of both papers and to establish a similarity between them. Keep in mind that the *K* number of factors is usually established in a few tens. Using the same reasoning, it is easy to determine the similarity between each pair of topics. In the example, the topics factors matrix in blue shows that "Deep learning" and "Machine learning" are related.

Each factor matrix can be used to make groups (clusters) of papers and topics. A simple method consists of grouping according to the highest value of each factor, e.g., papers *1* and *N* will belong to the same group (say: group 1), since their highest factors (0.8 and 0.6) both belong to F1. Paper 2 will belong to group 2 because the highest value (0.5) belongs to F2. Topics can be grouped in the same way: "Collaborative Filtering" in group 2, etc.

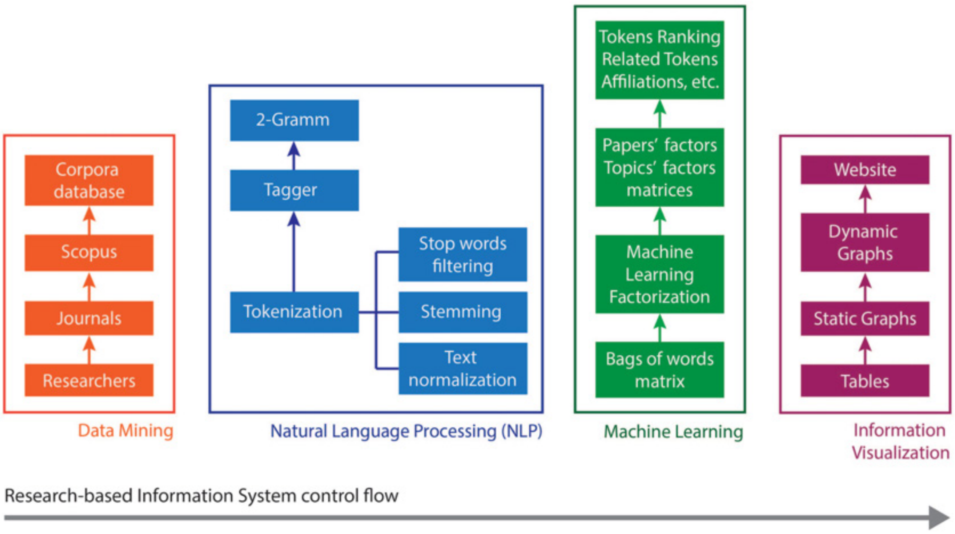
2. EXPERIMENTAL DESIGN

The objective of this work is to find the research trending topics from selected knowledge areas and to expose useful information related to them: topics ranking, groups of topics, balance between the quality and quantity of research in each topic, the results broken down by affiliation countries, journals comparatives, etc. The achievement of the above objectives entails the creation of a complex information system, ranging from the data mining to the presentation of results. Figure 4 shows a block diagram in which the most representative tasks of our research-based information system are included.

Four large blocks can be differentiated in the information system that supports experiments: Data mining, Natural Language Processing, Machine Learning, and Information Visualization. Below the functions are detailed for each of these blocks, helping with the content shown in Figure 4.

1. *Data Mining*: This block is responsible for collecting data from scientific papers written by the researchers and sent to different specialized journals. There are several publishers that offer individual search tools for published papers. *Scopus*, the Elsevier's citation database, is used; it covers more than 34,000 journals from more than 11,000 publishers. All of the provided information is collected from each paper (except its body) and we incorporate it into a database. From the database, queries can be performed by filtering the desired information, such as papers from a specific journal, authors from a set of

Figure 4. Research-based information system control flow


Corpora database
Scopus
Journals
Researchers
Data Mining
2-Gramm
Tagger
Tokenization
Stop words filtering
Stemming
Text normalization
Natural Language Processing (NLP)
Tokens Ranking Related Tokens Affiliations, etc.
Papers' factors Topics' factors matrices
Machine Learning Factorization
Bags of words matrix
Machine Learning
Website
Dynamic Graphs
Static Graphs
Tables
Information Visualization
Research-based Information System control flow

countries, etc. The data mined information has been introduced in a database. Researchers can access it through: rs.etsisi.upm.es

2. *Natural Language Processing (NLP)*: This block allows us to obtain tokens (one or two words in size) that can be selected as topics. The starting point is each of the texts that represent contents of a paper: "keywords + index-keywords + title + abstract". The *Tokenization* task is divided into the following sections: a) *Text Normalization*: stripping accents and special characters, b) *Stemming*: removing several word endings (-s, -es, -ing, etc.), and c) *Stop words filtering*: removing non-representative words (a, and, the, in, etc.). The *Tagger* is a syntactic analyser, with some semantic capacity, that classifies words as adjectives, verbs, nouns, etc. Finally, the *2-gramm* task statistically recognizes the set of words that usually appear together (United States, artificial intelligence, machine learning, etc.).

3. *Machine Learning*: The block in charge of processing the *2-gramm* topics is selected in the NLP block. Its general operation has been explained in subsection "1.3. Machine Learning introduction". The process is as follows: 1) start with the *bag of words matrix*, 2) carry out a *factorization*, and 3) use the papers and topics *factors matrices* to process clustering of tokens and to obtain similarities.

4. *Information Visualization*: Based on the *Machine Learning* process results, we obtain useful information about the topics ranking, their clustering, the quality/quantity ratio of the publications in each topic, etc. These results are offered in various formats: tables, static graphs (suitable to print), dynamic graphs and interactive graphs (suitable for mobile and desktop). Additionally, a website is provided where these results are jointly offered.

Table I shows the resources, methods and software used to run experiments. Journals have been selected based on 1) knowledge area, 2) impact factor, and 3) universality of accepted research areas (avoiding journals that are too focused in a specific field). Table II shows some representative data from each journal.

Table I. Experimental resources

Data Mining	
Journals	See Table II
Research period	January 2017 to February 2018
Data	Keywords, index-keywords, title, abstract, #citations, affiliations
Natural Language Processing	
Tokenization	Stanford English Tokenizer
Tagger	TreeTager, Institute for natural language processing
2-gramm	Proprietary software
Machine Learning	
Factorization	LDA: JGibbLDA
Information Visualization	
Static graphs	Microsoft Access
Dynamic Graphs	d3js.org

Table II. Journals facts

Journal	Number of papers	Impact Factor	Ranking	Areas
IEEE Computational Intelligence Magazine	122	6.343	6	Art. Int.
Information Sciences	2351	4.832	7	Inf. Syst.
Artificial Intelligence	257	4.797	14	Art. Int.
Knowledge-Based Systems	1239	4.529	16	Art. Int.
Expert Systems with Applications	2340	3.928		Art. Int.
ACM Transactions on Intelligent Systems and Technology	246	3.196	26	Art. Int.
International Journal of Intelligent Systems	229	2.929	31	Art. Int.
Artificial Intelligence Review	209	2.627	38	Art. Int.
Total 6993				

"Number of papers": number of publications in the mined period (January 2017 to February 2018); "Impact Factor" in the 2017 year. "Areas": knowledge areas covered by the journal; "Ranking": best ranking in each journal area.

3. RESULTS

In this section, the following experiments are shown: a) Rankings of the Top 100 Artificial Intelligence research topics: quantity & quality results, b) Topics comparative: quantity versus quality, c) Ranking of Artificial Intelligence research areas: quantity & quality results, and d) Research areas comparative: quantity versus quality.

Information is displayed using bar graphs. The basic measures are 1) Number of published papers, 2) Average number of citations, and 3) Averaged impact factor. All of these measures refer to each topic and each research area. Table II summarizes the source journals and their quantitative and qualitative numerical values. The number of citations for each paper is obtained through data mining from Scopus.

The rankings (topics and research areas) show absolute values of the number of papers and average of citations. The impact factor has not been included in the provided set of ranking graphs, so as to not increase this paper's size. Values in the comparative graphs are normalized in the interval [0..1] so that comparisons can be made on the same scale. Therefore, the interpretation of comparative values should be performed in a relative way and not an absolute way, but relative. For example, this is correct: "Compared to the rest of the topics, the topic *Fuzzy Sets* has a high number of averaged citations regarding the number of papers published in that topic". This is not correct: "The topic *Fuzzy Sets* has more citations than the number of its published papers".

3.1. Rankings of research topics

This section shows the Top-100 *Artificial Intelligence* topics, ordered by a) Number of papers published in each topic and b) Quality of the papers published in each topic. The quality is measured as the average number of citations that are achieved in each topic. Figure 5 shows the topics ranking ordered by quantity (left side) and the topics ranking ordered by quality (right side). As can be seen in the 'x' axes, the number of papers published in each topic can reach several hundred. The average number of citations in each topic is between approximately four and fourteen.

As is depicted in Figure 5, we can find cross-sectional concepts (such as *Surveys, Benchmarking, State-of-the-art methods*, etc.), concepts related to types of Artificial Intelligence problems (such as *Classification, Forecasting, Decision Support Systems*, etc.) and concepts related to Artificial Intelligence techniques (*Genetic algorithms, Fuzzy sets, Support vector machines,* etc.). All of these concepts are reflected in the corpus of selected journals and their inclusion has been respected so that the expert researcher can determine which of those have a greater tendency of all of these concepts, without making a prior filter on each type of concept.

From the "Figure 5. Quantity graph", we can see that the Top-1 topic is *Optimization*. This result confirms the current tendency towards *Artificial Intelligence* methods based on learning through optimization. Optimization methods are the basis of *machine learning*, *evolutionary algorithms*, *neural networks*, *genetic algorithms*, etc. Currently, classic *Artificial Intelligence* methods and algorithms are relegated to the last positions of the Top-100, or even disappear from the ranking. From Figure 5, the most related topics to the *Scientific Documentation* area are *Natural language processing*, *Linguistics*, *Computational linguistics*, *Text processing*, *Factorization*, *Data mining*, and *Feature extraction*; respectively: Top-43, 44, 51, 60, 84, 4, and 8.

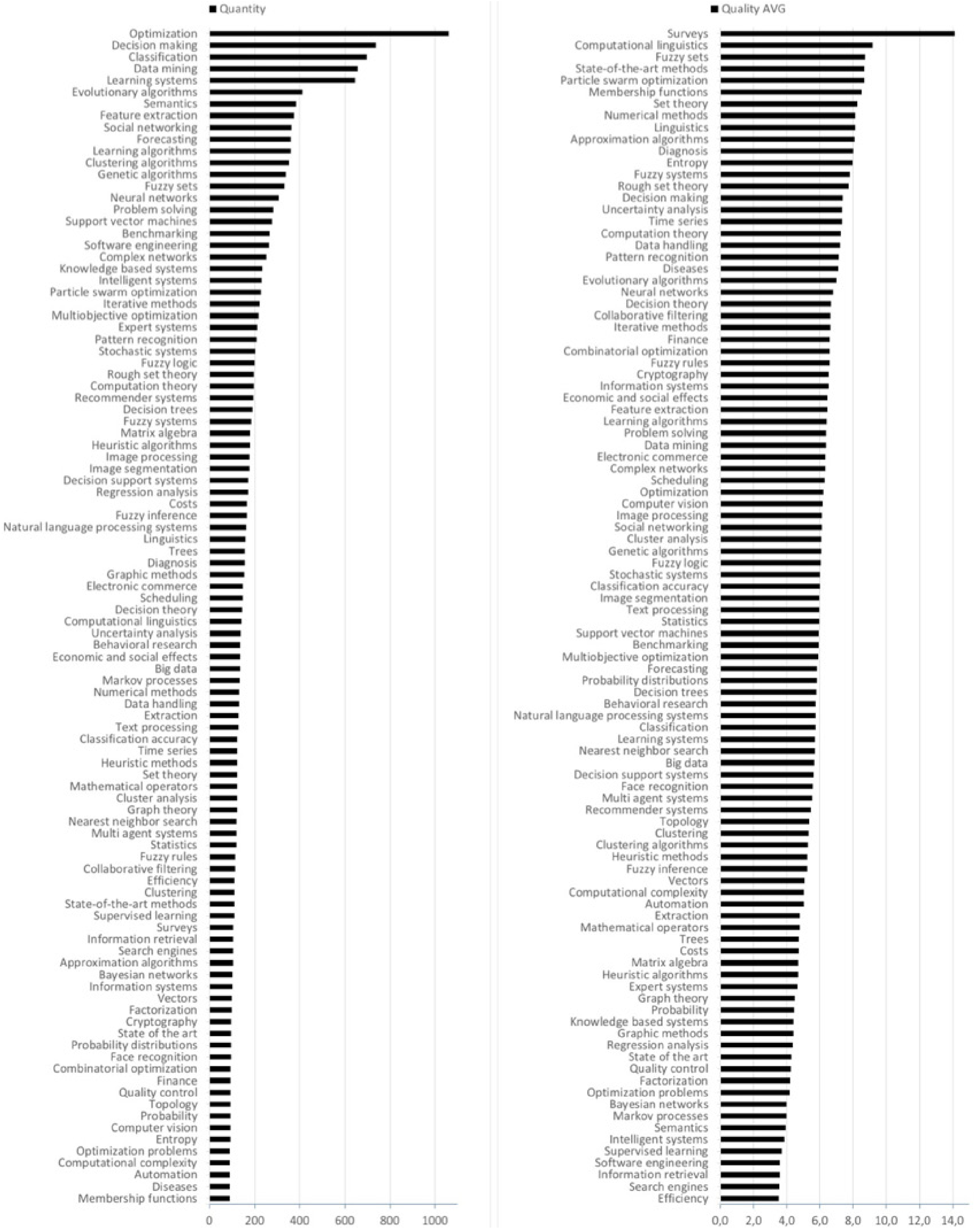
In "Figure 5. Quality graph", we have removed the topic "Optimization". It is too universal, and we want to focus the number of citations in the remainder of the topics. *Surveys* is the Top-1 topic; this topic usually joins the surveys and the reviews papers. *Surveys* is Top-1 due to the high number of citations that this type of paper usually receives.

It is interesting to note the high number of citations that, on average, some topics related to *Scientific Documentation* receive: *Computational linguistics* (Top 2) and *Linguistics* (Top 9). It is important to emphasize that if a topic is highly positioned in "Figure 5. Quality" then papers from this topic receive, on average, many citations. It does not mean that this topic receives many citations in total because it could be that very few papers are published in that area or research. This concept can be seen in *Surveys*, which is Top-1 in quality and Top-77 in quantity.

3.2. Topics comparison

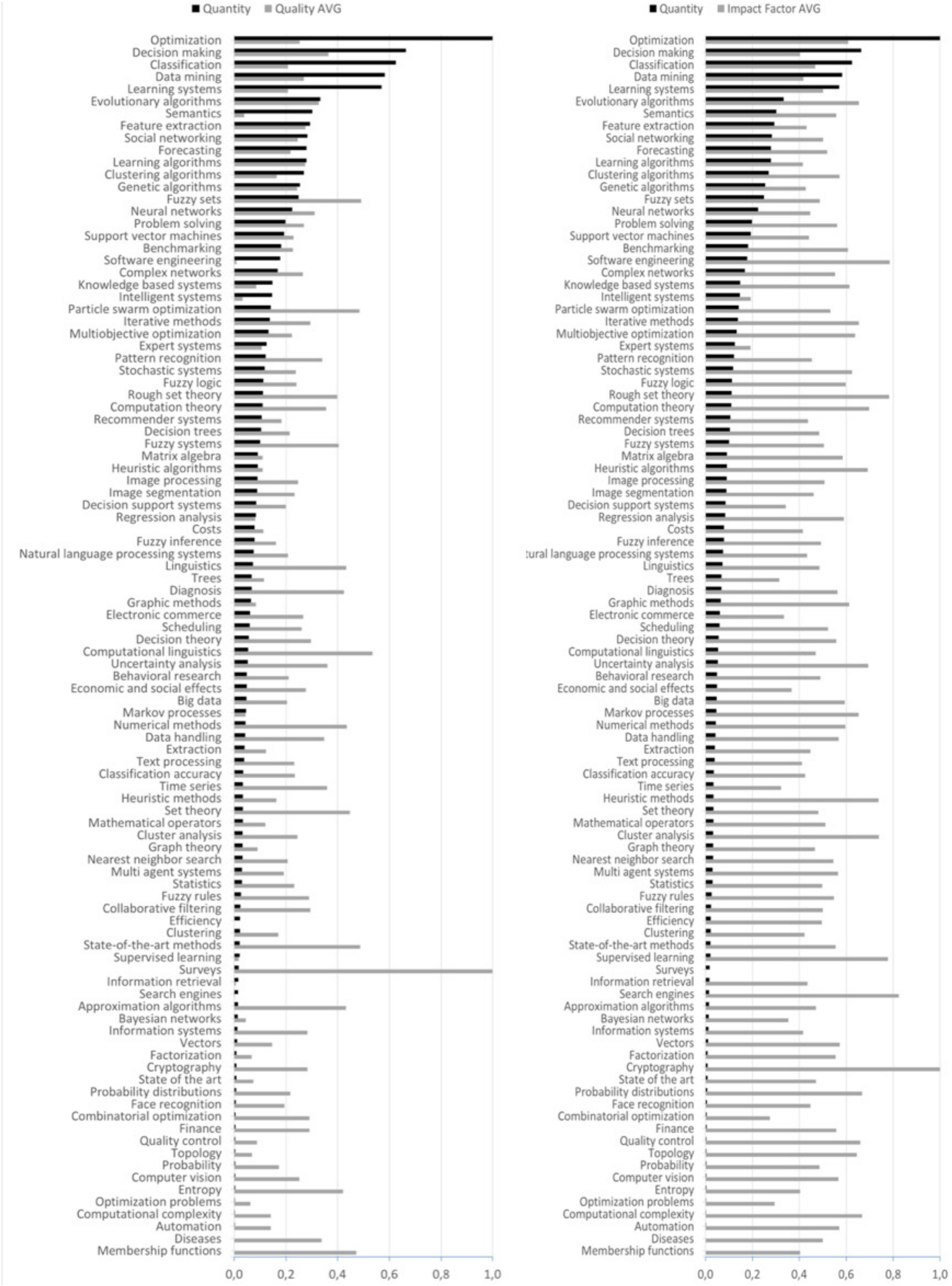
Figure 6 shows two comparative graphs. The first one (left side) shows the number of papers versus the citations average; the second one (right side) shows the number of papers versus the impact factor average. In both cases, we have included the main topic *Optimization*. The graph showing the citations proportions does not indicate a direct relationship between quantity and quality; rather, the opposite is seen: in proportion, topics with more published papers tend to receive a lower proportion of citations, and vice-versa.

Figure 5. Research topics ranking



Left side: number of published papers (quantity). Right side: citations average (quality).

Figure 6. Research topic comparison: quantity versus quality.



Left side: number of published papers versus average of obtained citations. Right side: number of published papers versus averaged impact factor. The results are normalized.

In the second (lower) half of “Figure 6. Left” there are a large number of topics that are characterized by a) being specific in the area of Artificial Intelligence, b) having small quantities, on average, of published papers; and c) receiving large quantities, on average, of citations. In the first (upper) half of the “Figure 6. Left”, the trend is the opposite. This situation can be explained, to a large extent, because our methodology assigns a variety of topics to each paper: some of these topics are specific (e.g., *clustering*, *factorization*) and others are universal (e.g., *optimization*, *classification*). In the previous example, *clustering* and *factorization* would appear in two different papers, but *classification* would appear in both papers. In this way, universal terms such as *classification* or *learning systems* will appear in many publications, while specific terms such as *fuzzy rules* or *Bayesian networks* will appear in a few publications.

Because the number of citations is not provided as an absolute value, but as an average of citations for each topic, quality results are not conditioned by quantity results. This situation explains the high citation values in the second half of “Figure 6. Left”, where each topic receives, on average, a large number of citations. The first part of the graph tends to house universal topics, coming from different types of papers, some of them providing more citations and others providing fewer citations. For example, *Recommender systems* can be considered a universal topic, which will appear in papers with the specific topics *collaborative filtering*, *factorization*, *nearest neighbours search*, etc.

In “Figure 6. Right”, the comparison between the number of published papers and the averaged impact factor is shown. As can be seen, the distribution of the impact factor is not related to the number of published papers. In addition, impact factor variations are not large because the journals selected do not have a wide variation in their impact factors. The higher impact factor proportions correspond to topics published in the journal with the highest impact factor (Table II); the topic *Cryptography* is a representative example of this situation. In the same way, the smaller impact factor proportions correspond to topics published in the journals with the lowest impact factor (Table II). *Intelligent Systems* is a representative topic example of this situation.

3.3 Rankings of research areas

From the *Research-based Information System* (Figure 4), using its *Machine Learning* subsystem, a set of research areas are obtained. Each research area is characterized by a list of topics; different areas contain a different number of topics, ranging from three to eight. Each of the resulting areas was processed to obtain a series of results classified in the same way as the topics graphs: 1) Absolute number of papers, 2) Citations average, 3) Comparative number of papers versus citations average, and 4) Comparative number of papers versus impact factor average.

Research areas have been extracted from data in the following way:

1. We have done a Machine Learning factorization process (Figure 3) from the topics/papers matrix.
2. From the factorization results, the topic’s/factors matrix is chosen (right side in Figure 3).
3. Using the topics vectors (columns in the topic’s/factors matrix), we have made a clustering: topics where the Factor (F1) is the highest factor belong to the cluster #1, topics where the Factor (F2) is the highest factor belong to the cluster #2, and so on.
4. From the topics of each cluster, we made a second clustering process by grouping topics making use of the Pearson correlation similarity measure. This similarity measure is applied to the hidden factors of each topic. This method allows us to limit the number of topics in each final cluster; we have chosen six as the maximum number of topics belonging to a cluster.

To maintain an adequate size in the figures, we have limited the number of groups to 32 in each graph. The order in which the topics of each group appear is relevant: we consider the first topics more representative than the last ones. Specifically, each topic examined is placed according to its placement, and the importance of each topic is reduced with the square root value of its position. That is, the first topic on the list retains all of its importance (it is divided by the square root of 1), the second topic reduces its importance by dividing it by the square root of 2, the fourth topic reduces its importance by half, by dividing it by the square root of 4, etc.

Figure 7, left side, shows the most published *Artificial Intelligence* areas from the journals and years specified in Table II and from the results obtained using the *Information System* explained in Figure 4. Areas associated with the most published papers (first positions in Figure 7) correspond to universal fields (*Intelligent Systems*, *Optimization*, *Learning Systems*, etc.). Areas with a lower number of published papers (lower positions in Figure 7) correspond to specific fields of *Artificial Intelligence* (*Recommender Systems*, *Image Processing*, *Search Engines*, etc.). Figure 7, right side, shows the ranking of *Artificial Intelligence* areas, ordered according to the average number of citations obtained in each area. The average number of citations ranges from approximately four to eight. In this graph, the first positions in the ranking do not correspond to universal fields; they correspond to specific fields (*Fuzzy Sets*, *Pattern Recognition*, *Image Processing*, etc.). Note that there are little differences between the number of citations per area since averaged citations are, at the same time, topic averaged in each group.

3.4. Research areas for comparison

This section explains the comparative quantity versus quality results in each of the research areas from Figure 7. Figure 8 (left side) compares the quantity with citation average; Figure 8 (right side) compares the quantity with the impact factor average.

In Figure 8, the left side shows a reduced number of research areas where, proportionally, the citations average is superior to the number of published papers: *Fuzzy Sets* related areas, *Image Processing*, *Search Engines* and *Recommender Systems*. It is also appropriate to consider areas contributing to a large number of published papers and simultaneously maintaining a high proportion of average citations. In this case, this occurs in the areas related to *Evolutionary Algorithms*, *Learning Algorithms,* and *Learning Systems*; we can consider these areas as research trends.

Figure 8 (right side) shows that, as in the case of topics, there is no relationship between quantity and quality in the research areas. On the other hand, there are extreme cases of interest: 1) Areas with a large number of publications in journals and with a low impact factor (topics: *Intelligent Systems*, *Expert Systems*), and 2) Areas with a small number of publications in journals and with a high impact factor (topics: *Search Engines*, *Rough Sets*, *Factorization*).

4. CONCLUSIONS

To conduct processing of *Scientific Documentation* sources it is necessary to make use of diverse methods and techniques from the following areas: data mining, natural language processing, machine learning and data visualization. The combination of these tasks provides scientific documentation for *Information System*.

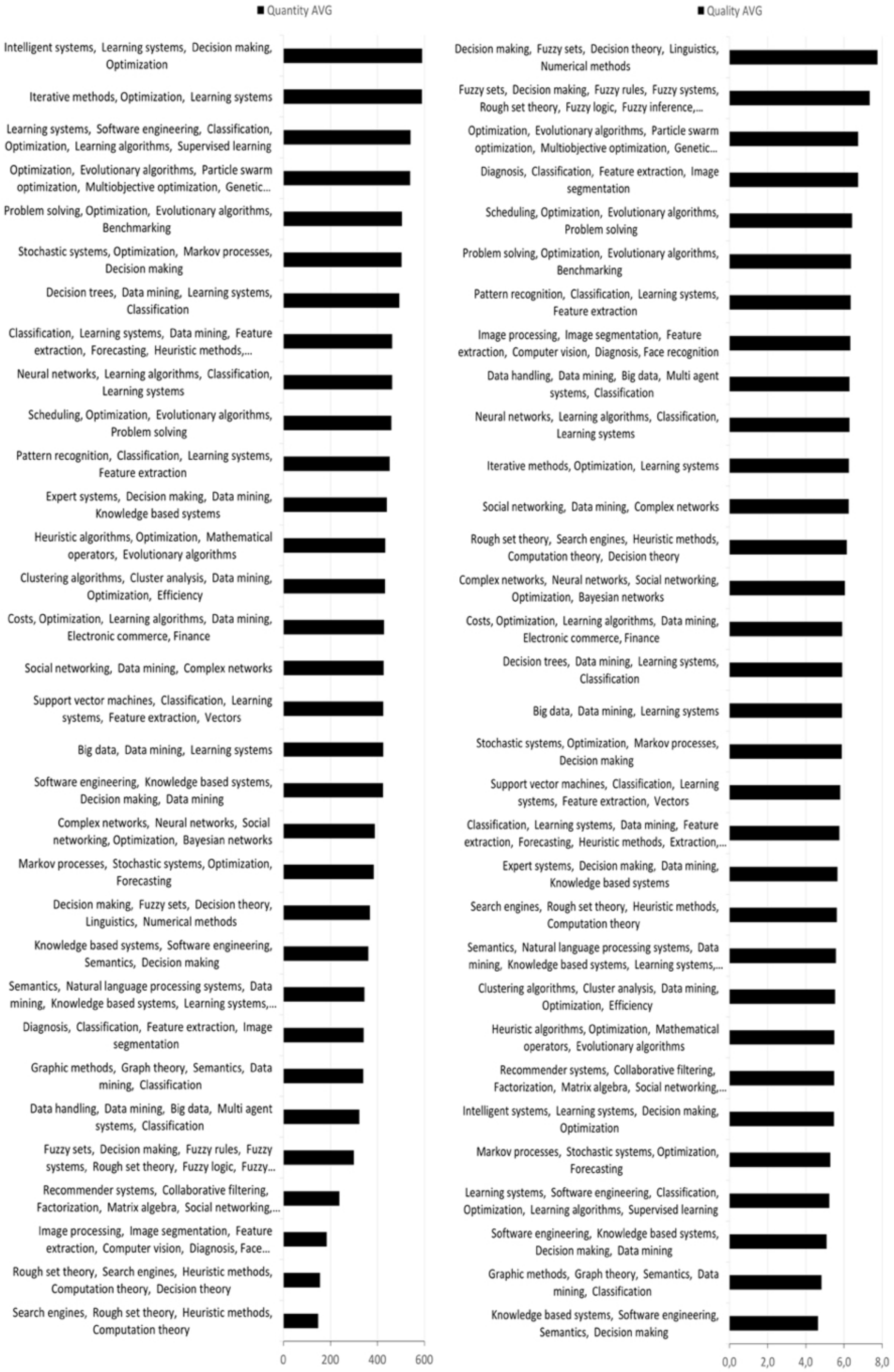
The data mining methods provide the relevant information of the published papers in the selected research area. Following the data mining stage, the natural language techniques provide semantically representative words or groups of words. The machine learning methods are able to pull out the most representative topics, relate them and create groups of topics. These groups of topics make up the research area fields.

Our information system selects topics, provides topic rankings, detects research areas, generates research area rankings, and compares the qualities versus quantities of the topics and research areas. As an example, the *Artificial Intelligence* area is discussed. The results show a strong decline in the Artificial Intelligence classical research and a strong increase in machine learning methods. Analysing results, *optimization-based learning* is the most promising area; *evolutionary algorithms*, *learning algorithms,* and *learning systems* can be considered research trends.

This paper differentiates between universal topics and specific topics. The first topics refer to generic topics (*optimization*, *classification*, *learning systems*, etc.). The second set refers to specialized topics (*fuzzy sets*, *search engines*, etc.). Universal topics tend to have a high number of published papers, while the specific topics tend to have a high number of average citations. This same trend is noted in the research areas from the *Information System*.

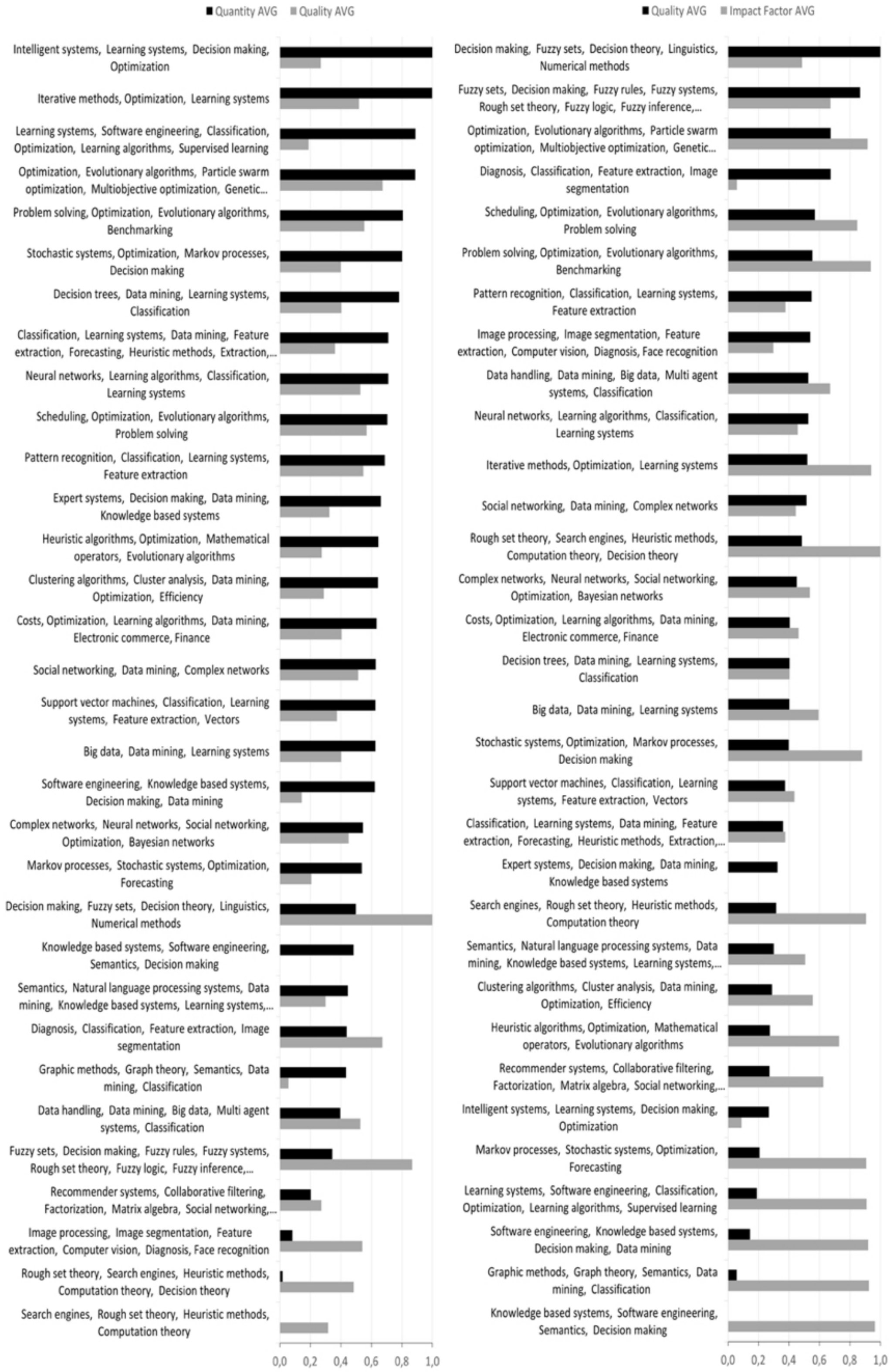
For future work, this paper provides the basis for the following: a) Repeating the process for any research area, b) Making use of different machine learning methods, c) Including the temporal component in the machine learning process, showing the research areas and topics evolution, d) Breaking down results by country, and e) Making comparisons between different research areas.

Figure 7. Research area rankings



Left side: Number of published papers (quantity). Right side: Average of obtained citations (quality)

Figure 8. Research area comparison



Left side: Number of published papers versus average of obtained citations. Right side: Number of published papers versus averaged impact factor. The results are normalized.

REFERENCES

Abramo, G.; Cicero, T.; D'Angelo, C. (2014). Are the authors of highly cited articles also the most productive ones?. *Journal of Informetrics*, 8 (1), 89-97. https://doi.org/10.1016/j.joi.2013.10.011

Aguillo, I. F.; Ortega, J.; Fernández, M.; Utrilla, A. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics*, 82 (3), 477-486. https://doi.org/10.1007/s11192-010-0183-y

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, 12 (3), 159-170. https://doi.org/10.3152/147154403781776645

Aksnes, D. W.; Sivertsen, G. (2004). The effect of highly cited papers on national citation indicators. *Scientometrics*, 59 (2), 213-224. https://doi.org/10.1023/B:SCIE.0000018529.58334.eb

Altszyler, E.; Sigman, M.; Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

Anicic, K.; Divjac, B.; Arbanas, K. (2016). Preparing ICT Graduates for Real-World Challenges: Results of a Meta-Analysis. *IEEE Transactions on Education*, 60 (3), 191-197. https://doi.org/10.1109/TE.2016.2633959

Bleu, D.M.; Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Bobadilla, J.; Bojorque, R.; Hernando, A.; Hurtado, R. (2017). Recommender systems clustering using Bayesian non negative matrix factorization. *IEEE Access*, 6, 3549-3564, https://doi.org/10.1109/ACCESS.2017.2788138

Bobadilla, J.; Ortega, F.; Hernando, A.; Gutierrez, A. (2013). Recommender Systems Survey. *Knowledge Based Systems*, 46, 109-132. https://doi.org/10.1016/j.knosys.2013.03.012

Bornmann, L.; Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: the avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics*, 5 (1), 228-230. https://doi.org/10.1016/j.joi.2010.10.009

Haustein, S.; Peters, I.; Bar-Ilan, J.; Priem, J.; Shema, H.; Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, 101 (2), 1145-1163. https://doi.org/10.1007/s11192-013-1221-3

Hayati, Z. (2009). Correlation between quality and quantity in scientific production: A case study of Iranian organizations from 1997 to 2006. *Scientometrics*, 80 (3), 625-636. https://doi.org/10.1007/s11192-009-2094-3

Hernando, A.; Bobadilla, J.; Ortega, F. (2016). A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. *Knowledge Based Systems*, 97, 188-202. https://doi.org/10.1016/j.knosys.2015.12.018

Hindle, A.; Bird, C.; Zimmermann, T.; Nagappan, N. (2015). Do topics make sense to managers and developers? *Empirical Software Engineering*, 20 (2), 479-515. https://doi.org/10.1007/s10664-014-9312-1

Karlsson, A.; Hammarfelt, B.; Steinhauer, H.J.; Nolin, J. (2014). Modeling uncertainty in bibliometrics and information retrieval: an information fusion approach. *Scientometrics*, 102 (3), 2255-2274. https://doi.org/10.1007/s11192-014-1481-6

Kaya, M.; Cetin, E.; Socery, A. (2010). Introduction to Webometrics: quantitative Web research for the ranking of world universities; research centers and hospitals. *ICEGEG-2010*, Antalya, Turkey.

Khan, B.S.; Niazi, M.A. (2017). Emerging Topics in Internet Technology: A Complex Networks Approach, *arXiv.* https://arxiv.org/abs/1708.00578v1

Lis-Gutierrez, J.P.; Gaitan-Angulo, M.; Robayo, P.V.; Aguilera-Hernandez, D.; Viloria, A. (2017). Academic production patterns in public administration: An analysis based on scopus. *Journal on Engineering and Applied Sciences*, 12 (11), 2904-2909.

Lu, K.; Cai, X.; Ajiferuke, I.; Wolfram, D. (2017). Vocabulary size and its effect on topic representation. *Information Processing and Management*, 53 (3), 653-665. https://doi.org/10.1016/j.ipm.2017.01.003

Manolopoulus, Y.; Katsaros, D. (2017). Metrics and rankings: Myths and fallacies. *Communications in computer and information science*, 706, 265-280. https://doi.org/10.1007/978-3-319-57135-5_19

Martín-Martín, A.; Orduna-Malea, E.; Ayllón, J:M.; López-Cózar, E.D. (2016). A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentación Científica*, 39 (4), e149. https://doi.org/10.3989/redc.2016.4.1405

Mingers, J.; Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246 (1), 1-19. https://doi.org/10.1016/j.ejor.2015.04.002

Mustafee, N., Katsaliaki, K., Fishwick, P., (2014). Exploring the modelling and simulation knowledge base through journal co-citation analysis. *Scientometrics*, 98 (3), 2145-2159. https://doi.org/10.1007/s11192-013-1136-z

Naili, M.; Chaibi, A.H.; Ghezala, H. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112, 340-349. https://doi.org/10.1016/j.procs.2017.08.009

Nedra, I.; Chaibi, A. H.; Ahmed, M. B. (2015). New scientometric indicator for the qualitative evaluation of scientific production. *New Library World*, 116 (11/12), 661-676. https://doi.org/10.1108/NLW-01-2015-0002

Nejati, A.; Hosseini Jenab, S.M. (2010). A two-dimensional approach to evaluate the scientific production of countries (case study: the basic sciences). *Scientometrics*, 84 (2), 357-364. https://doi.org/10.1007/s11192-009-0103-1

Orduna-Malea, E.; Martín-Martín, A.; Delgado López-Cózar, E. (2017). Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista Española de Documentación Científica*, 40 (4), e185. https://doi.org/10.3989/redc.2017.4.1500

Ortoll, E.; Canals, A.; García, M.; Cobarsí, J. (2014). Main parameters for the study of scientific collaboration in big science. *Revista Española de Documentación Científica*, 37 (4), e069. https://doi.org/10.3989/redc.2014.4.1142

Ravikumar, S.; Agrahari, A.; Singh, S.N. (2015). Mapping the intellectual structure of scientometrics: a co-word analysis of the journal Scientometrics. *Scientometrics*, 102 (1), 929-955. https://doi.org/10.1007/s11192-014-1402-8

Sun, S.; Luo, Ch.; Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36, 10-25. https://doi.org/10.1016/j.inffus.2016.10.004

Wang, S.; Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, 111 (2), 1017-1031. https://doi.org/10.1007/s11192-017-2298-x

Xue, H.J.; Dai, X.Y.; Zhang, J.; Huang, S. (2017). Deep matrix factorization models for recommender systems, *IJCAI*, pp. 3203-3209. Melbourne, Australia. https://doi.org/10.24963/ijcai.2017/447

Yazdani, K; Nedjat, S; Rahimi-Movaghar, A; Ghalichee, L; Khalili, M. (2015). Scientometrics: Review of concepts, applications, and indicators. *Iranian Journal of Epidemiology*, 10 (4), 78-88.

Yu, D.J. (2015). A scientometrics review on aggregation operator research. *Scientometrics*, 105 (1), 115-133. https://doi.org/10.1007/s11192-015-1695-2