Telematics and Informatics 35 (2018) 168–185

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

A fast method for identifying worldwide scientific collaborations using the Scopus database

Francisco G. Montoya*, Alfredo Alcayde, Raúl Baños, Francisco Manzano-Agugliaro

CeiA3, Department of Engineering, University of Almería, Carretera de Sacramento s/n, 04120 Almería, Spain

A R T I C L E I N F O

Keywords:
Data mining methods
Database management
Information retrieval
Scientific collaboration
Network analysis

A B S T R A C T

Science is essential for human prosperity because social and technological advances often depend on scientific advances. Science is living a golden era characterized by a rapidly growing number of researchers worldwide exploring different disciplines and research fields. Keeping in mind that funding is limited, many researchers are encouraged to establish new collaborations with individuals or groups of researchers. Furthermore, the funding bodies use increasingly complex criteria to determine the researchers and projects to be supported. In this regard, the analysis of scientific collaboration networks can help to determine the main areas of specialization of universities and research centres, as well as the type of internal and external collaborations of their researchers. This paper presents an advanced method for analysing scientific collaboration networks at universities and research institutions. This method is based on automatically obtaining bibliographic data from scientific publications through the use of the Scopus Database API Interface, which are then analysed using graph visualization software and statistical tools. This model has been validated through the analysis of a real university, and the results show that it is possible to determine in a fast way and with high reliability the main research lines of an institution as well as the structure of the collaboration network. The method opens new perspectives for the study of scientific collaboration networks because it can be applied at different levels of detail, from small research groups to large academic and research centres, and over different time frames.

1. Introduction

Now, as never before, the scientific community has the material and human resources to carry out relevant research activities in multiple disciplines and research fields. According to some studies, world scientific production doubles every nine years Noorden (2014). Thanks to the development of Information and Communication Technologies (ICT), it is now possible to access most of the scientific literature through the Internet (Asadi and Dahlan, 2017; Ravishankar, 2013), which allows transferring knowledge effectively. However, this process also leads to increasing competition between researchers (Whitley, 2003) and higher quality standards that increase the demands on research productivity (Guthrie et al., 1993). These changes encourage researchers, especially those in the early stages of their careers, to establish collaborative relationships with other researchers from their own or another institution (Lewis et al., 2012). The dynamism of these collaborations makes the analysis of the scientific collaboration network a topic of great interest (Shafiq et al., 2015; Shen et al., 2010; Aron, 2009), both for researchers and for funding agencies (Laudel and Gläser, 2014).

Research activities are funded by public institutions or private organizations, which invest many millions of dollars to provide

* Corresponding author.
E-mail addresses: pagilm@ual.es (F.G. Montoya), aalcayde@ual.es (A. Alcayde), rbanos@ual.es (R. Baños), fmanzano@ual.es (F. Manzano-Agugliaro).

http://dx.doi.org/10.1016/j.tele.2017.10.010
Received 18 August 2017; Received in revised form 11 October 2017; Accepted 15 October 2017
Available online 18 October 2017
0736-5853/

technical and human resources to guarantee that the established objectives will be reached (McNutt, 2014). However, evaluating scientific quality is a notoriously difficult problem that has no standard solution (Seglen, 1997). In fact, the evaluation criteria used to assign these budgets are usually based on scientometric indicators retrieved through web search engines such as Scopus, Google Scholar, etc., which include the number and quality of publications by researchers based on journal and author metrics. Nevertheless, these evaluation criteria do not explicitly take into account the relationships between researchers, even though the globalization of science in the modern world due to collaborations between researchers of the same or different institutions has become an important issue to analyse. To respond to this demand, this paper presents a method to automatically determine the structures of collaborations between researchers from different institutions and disciplines and to provide a fast and reliable method for determining the most important specialty areas of a given institution. Collaboration in publications is used as a criterion because, according to different studies, it is more important for knowledge transfer than patents (Ajay and Henderson, 2002). This method is based on programming scripts to automatically obtain bibliographic data from scientific publications through the use of the Scopus Database API Interface (Elsevier, 2017) according to different criteria (affiliation, specialty area, department, etc.). This information is then adapted for processing using parsing techniques and the refinement of structured text (Elmagarmid et al., 2007). Subsequently, the collaboration network is represented using specialized graph visualization software (Beck et al., 2016) and analysed using statistical metrics.

The rest of the paper is organized as follows: Section 2 offers an overview of previous investigations into this topic, including some studies that have analysed the collaboration networks among researchers. It also provides a description of the scientific databases and the possibilities offered by the Scopus Database API Interface to retrieve bibliographic data. Section 3 presents the method proposed in this article, which is validated in Section 4 through the analysis of a medium-sized university. The main conclusions of this research are provided in Section 5, while the implications and limitations of this investigation are presented in Section 6.

2. Background

2.1. Scientific collaborations and research merit

During the last few decades, we have been witnessing the development of scientometrics (Canas-Guerrero et al., 2013; Canas-Guerrero et al., 2014), informetrics (Rojas-Sola and de San-Antonio-Gomez, 2010; Singh et al., 2015) and bibliometrics (Canas-Guerrero et al., 2014; Rojas-Sola and Aguilera-Garcia, 2015). These disciplines, with different nuances, try to establish procedures and metrics for evaluating journal quality and the scientific output of researchers based on journal and author metrics (Franceschet, 2010; Ingwersen, 2014; Kaushal and Jeschke, 2013; Cao et al., 2016). Journal metrics are often based on the journal impact factor, which is measured using Thomson's Journal Citation Reports (JCR), Scimago Journal & Country Rank (SJR), etc. Among the author metrics, the total number of citations (Szymanski et al., 2012), as well as the h-index (Hirsch, 2005), or variations of it (Bornmann et al., 2008; Papavlasopoulos et al., 2010) stand out. However, these metrics do not take into account other aspects that are of great importance, such as the structure of collaborative relationships between researchers. In this regard, some studies have analysed scientific collaborations based on co-authoring (Uddin et al., 2013; Bhattacharyya and Bandyopadhyay, 2015). The co-authoring, author ordering, average number of cited references, self-citing, etc. have attracted the interest of researchers because they are criteria to determine the structures of social networks among scientific researchers (Fuchs, 2017). These collaboration networks have been studied in different research fields, including computer science, power networks, and social networks (Aparicio-Martinez et al., 2017; Penni, 2017), and regressions (Ebadi and Schiffauerova, 2015), graphs, and pattern searches (Newman, 2004) are some of the statistical tools used for analysis.

In recent years, Social Network Analyis (SNA) has become one of the key strategies for investigating social structures through the use of networks and graph theory (Hoppe and Reinelt, 2010). However, only a few studies have applied SNA to analyze the relationships between academic and research institutions. Abbasi et al. (2011) presented a theoretical model that applying measures often used in SNA explore collaboration (co-authorship) networks of scholars. Zheng et al. (2016) presented a review about the status of the research in SNA by analyzing a large number of papers in terms of institutional and individual contribution, citations, topic coverage, etc. of several American, British and Australian universities. In a recent paper, Schlattmann (2017) analyzed the intensity of collaboration within a German university using SNA.

2.2. Scientific databases

Currently, most researchers have access to high-quality multidisciplinary scientific databases. Some of them are open access databases (e.g., Google Scholar), while others are accessed thanks to the subscriptions made by public or private organizations for which these researchers work (e.g., Web of Science, Scopus (Ma and McGroarty, 2017)). Likewise, there are databases on specific fields or areas, such as PubMed, which includes citations for biomedical literature from MEDLINE (Kanavos et al., 2014), life science journals and online books. Different investigations have analysed the characteristics of scientific databases. The main sources of information, such as scientific databases (e.g: Scopus), search engines (e.g: Google Scholar) and social networks (e.g: academia.edu or ResearchGate), are analysed by (Asadi and Dahlan, 2017). An extensive discussion has been devoted to the advantages and disadvantages of these sources of information. For example, an important advantage of Google Scholar is that the information is updated periodically taking advantage of existing information on the Internet (Falagas et al., 2008). However, Google Scholar only includes profiles of researchers who have voluntarily discharged themselves, with numerous errors in linking documents to authors, so that many profiles of researchers have publications that do not really correspond to that author (Falagas et al., 2008). This inconvenience is partially solved by using unique and global identifiers such as the Open Researcher and Contributor ID (ORCID) (Haak et al., 2012), which assigns a unique identifier to each author, although such information must be included or linked manually. These deficiencies are limited in literature databases whose data sources are editorial, professional associations, etc. (Asadi and Dahlan, 2017). This is the case for Web of Science and Scopus, whose information is considered more reliable, although it includes a lower number of documents than other systems, such as Google Scholar (Leslie and Rensleigh, 2013). To verify the reliability and accuracy of these databases, different studies have analysed the typical errors in the information contained therein (Leslie and Rensleigh, 2013; Mongeon and Paul-Hus, 2016; Yu, 2011). For example, among the most common errors of Scopus is the existence of several authorIDs for the same individual due to their presence in different research institutions or to the fact that they used different signature formats in different publications. Thus, the reliability of any bibliometric study based on these data is limited. In the present study, these inaccuracies have been verified and have been minimized as much as possible thanks to the use of text refinement and data reconciliation tools. It is to be expected that the reliability of the information contained in Scopus will increase in the future, and as a result, the accuracy of the results compiled by the method presented here will be improved. All these databases are often oriented to allow a search for individual researchers and documents on a specific research topic. However, its use for the extraction and analysis of aggregate data has not been of interest until recent years. Fortunately, some of these databases have recently incorporated tools that allow to perform search and to collect data to develop studies of a different nature. This is the case of the Scopus Database API Interface, which is available to the public and allows obtaining raw data from the Scopus database based on different criteria. More specifically, documents in Scopus are classified under four broad subject areas: Life Sciences, Physical Sciences, Health Sciences, and Social Sciences & Humanities. These general subjects are further divided into 27 major thematic categories as well as into more than 300 specific subject categories, although a given document can be simultaneously included in different categories.

3. A framework to analyse research network collaborations

This section presents a method for analysing a large volume of data from any academic or research institution in order to determine different bibliometric indicators of individual or group researchers, as well as the collaboration networks of these researchers in terms of publications. This method, which can be of great interest for researchers, research institutions, and funding agencies, takes advantage of the features provided by the Scopus Database API Interface to automate the search of manuscripts published by authors and institutions, such that information is treated and analysed in a later stage for different specific purposes. Elsevier's Scopus is also the largest database of peer-reviewed literature in different scientific fields (Chadegani et al., 2013).

3.1. Data extraction using automated scripts

Fig. 1 shows a flowchart of the automatic information extraction script of the Scopus Database, named Research Network Bot (ResNetBOT). The operation of this bot, which allows the collection of all the data required for different analyses, can be divided into three subsequent phases:

1. Get articles data: In this phase, the bot retrieves all of the information from the publications of authors who have ever published a manuscript using a given affiliation identifier (*afid*). To perform this task, the information in each of the papers is extracted, and the "*authorid*" (unique identifier of an author in Scopus) will be stored in the database of the bot if it has not previously been registered.
2. Get authors data: From the list of researchers whose affiliation coincides with the desired one, ResNetBOT retrieves and stores in its database the information that Scopus has for each one of these authors, including the registered publications, dates of these publications, history of affiliations, h-index, number of citations, etc.
3. Get the collaboration networks: The individual information about each researcher is used to establish collaborative relationships based on the co-authoring of papers. More specifically, in this phase, the bot applies an iterative process for each of the authors of the institution, then obtains information such as the name of the institution, city, country, number of co-authors, current affiliation of these collaborators, etc.

ResNetBOT has been designed using a combination of Hypertext Preprocessor (PHP) and Bash for Linux so that the Scopus API is used according to the structure defined in (Elsevier, 2017).

3.2. Data parsing and text refining

The data obtained by ResNetBOT are structured according to the different fields of interest and saved in a set of plain text files using JavaScript Object Notation (JSON) format (JSON, 2017), see Fig. 2. The Scopus API allows requesting information with different levels of detail, which is why the bot has been programmed to request full data, then to select the most valuable information. The excerpt in listing 2 is an example of the information obtained:

Some fields of interest are "dc: identifier", which is a unique code that Scopus assigns to each author, or "document-count", which contains the number of published papers by a given author. During the data verification process, some inconsistencies have been detected, but it is a common problem in large databases due to the huge amount of information from a variety of sources they manage (Valderrama-Zurián et al., 2015; Franceschini et al., 2016). Specifically, in the results retrieved from Scopus, we have observed the following issues:

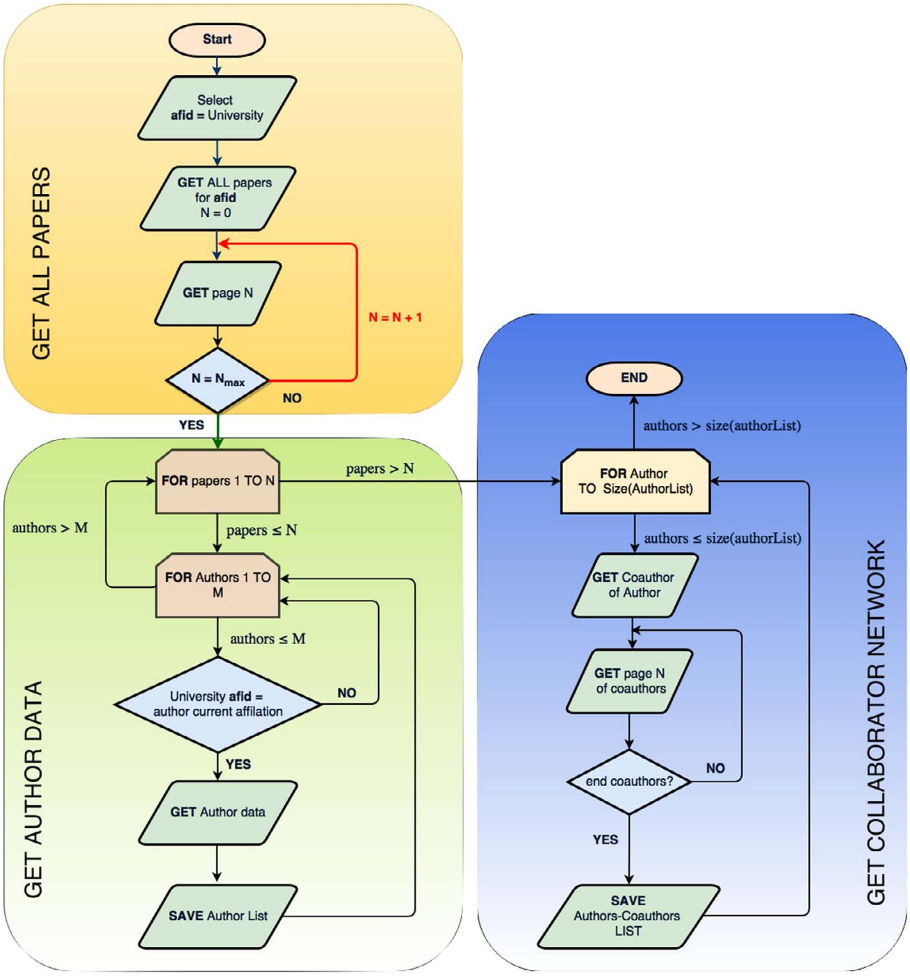


Fig. 1. Flowchart for ResNetBOT automated script.


```
{
  "author-retrieval-response": [
    {
      "@_fa": "true",
      "@status": "found",
      "coredata": {
        "prism:url": "http://api.elsevier.com/content/author/
         author_id/7102745519",
        "dc:identifier": "AUTHOR_ID:7102745519",
        "historical-identifier": [
          {
            "@_fa": "true",
            "$": "author_id:56174122300"
          }
        ],
        "eid": "9-s2.0-7102745519",
        "document-count": "4",
        "cited-by-count": "10",
        "citation-count": "10",
        .....
```


Fig. 2. JSON Scopus API retrieval.

Multiple AuthorIDs. It is possible to find authors who have several different authorID codes, where, in reality, it is the same author. This causes their scientific output and associated metrics to be unreliable and, to a large extent, devalued.

Wrong AuthorID. There are contributions that have been erroneously assigned to an author when they really are contributions from another with a very similar name/surname. This results in certain authors being assigned incorrect research areas or erroneous author metrics, which results in the scientific output of the true author being erroneously diminished.

Wrong AffiliationID. Certain authors have been mistakenly assigned to institutions at some point in their careers. This aspect is considered minor since it does not detract from the research curriculum, but it does distort its record to some extent.

Even taking into account the deficiencies found, it is possible to obtain relevant and valuable information after a process of refining the data retrieved by ResNetBOT. This depuration process is necessary since it is common to find words that express the same concepts but have been written with slight variations. For example, if we inspect the location of the University of Almeria, it is possible to find the same city written as "Almería", "Almeria", "almería", or "almeria". Therefore, it is necessary to apply refinement algorithms such as those included in the OpenRefine open-source tool (OpenRefine, 2017), which applies several algorithms based on "Key Collision Methods" and "Nearest Neighbour Methods" to refine and integrate texts with words that express the same idea but have been written with some syntactic variations (Baxter et al., 2003; Jin et al., 2003; Cavnar and Trenkle, 1994). This tool has been successfully applied in previous research papers (Montoya et al., 2014). OpenRefine is also applied to treat keywords and author names. Finally, spreadsheets are used for grouping the refined information in order to identify unique values.

3.3. Analysis and visualization of data

All the information collected by ResNetBOT and refined using OpenRefine is then saved in a database. That database is analysed using graph-based visualization and statistical tools. Graph representation is a useful tool to determine the relationships among a group of elements. A graph is composed of a set of vertex (or nodes) and edges, so that the nodes represent the elements, and the edges the relationships among these elements. The great advantage of this representation is that both nodes and edges can include specific characteristics of the elements and their relationships, respectively. In recent years, several powerful graph visualization tools have been developed. These software applications allow detailed analysis of the characteristics of graphs by allowing multiple configurations, such as modifying the sizes of the nodes and edges depending on different criteria, to grouping the nodes of the graph, to drawing them using different colours that depend on certain node characteristics, etc. In addition, these software applications often include statistical metrics that define the topological and relational characteristics among the nodes. One of the most commonly used free access tools is Gephi (Bastian et al., 2009; Szymaaski and Rzeniewicz, 2016), which has been applied to numerous projects and applications (Boden et al., 2013; Bruns, 2012; Jacomy et al., 2014; Dana and Rácek, 2015). Gephi includes a number of statistical tools, which are described in Table 1, some of them applied in other studies based on SNA (Abbasi et al., 2011).

Fig. 3 shows the flowchart of the method proposed in this paper for analysing the collaborations between researchers of a given institution. As shown, this figure integrates the three phases described above, i.e., the automatic data collection phase described in Fig. 1, the parse and text refining phases and the study of these data using graph-based visualization software applications (in this case Gephi).

4. Empirical study

This section validates the method presented in Section 3 by analysing the scientific collaborations that occur in a real-world university. With this aim, the University of Almeria (UAL) in Spain is analysed. It is a medium-sized public university that was founded in 1993. It is located in the province of Almeria in the southeast of Spain. It is a non-profit organization and currently has 34 degrees available, including 613 lecturers. Today, it has about 12,500 students, including 600 doctorate students. A total of 13 departments (135 research groups) are devoted to research and training. According to the data included in the Scopus database, this university has contributed to the development of several areas of research through a total of 2717 researchers who, at some point in

Table 1
Graph metrics and statistics analysed.

Metric	Description
Number of Nodes	Number of nodes of the network
Average Degree	The average number of edges that are adjacent to the nodes of the network
Average Clustering Coefficient	The clustering coefficient (Watts-Strogatz), when applied to a single node, is a measure of how complete the neighbourhood of a node is. When applied to an entire network, it is the average clustering coefficient over all of the nodes in the network
Average Path Length	The average graph-distance between all pairs of nodes
Connected Component	Determines the number of connected components in the network
Diameter	The maximal distance between all pairs of nodes
Graph Density	Measures how close the network is to complete. A complete graph has all possible edges and density equal to 1
Modularity	Measures how well a network decomposes into modular communities

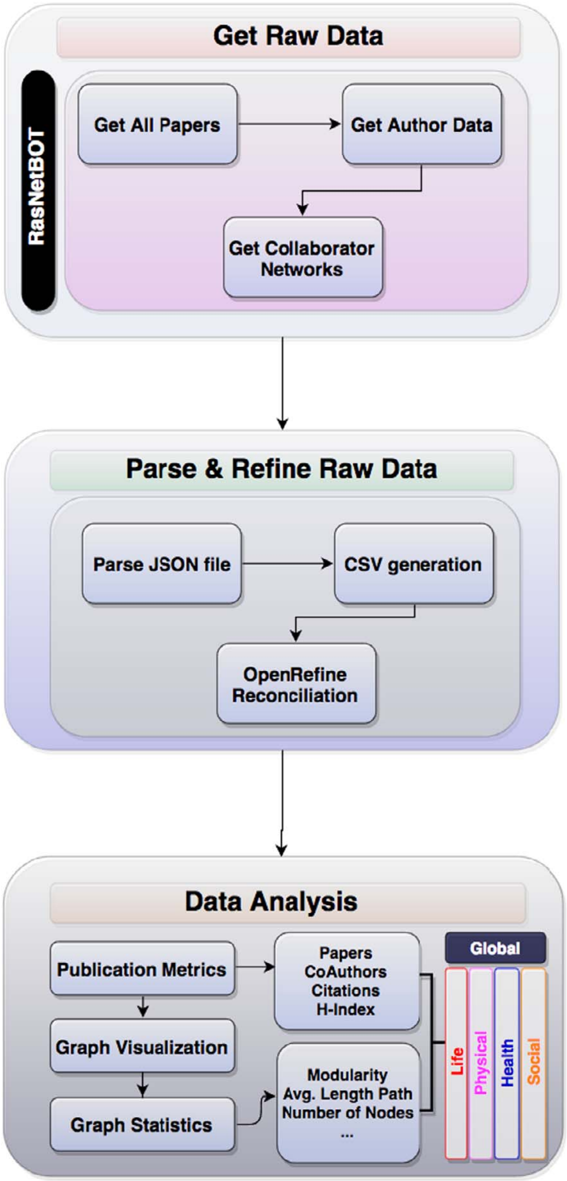


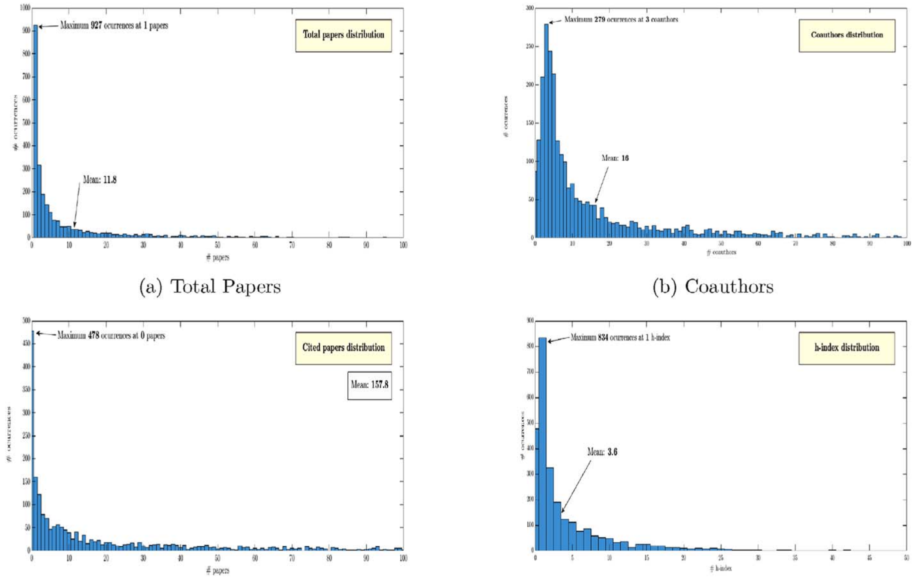
Fig. 3. Structure of the method for analysing network collaborations.

their career, have published an article with the affiliation UAL. The number of researchers who have this affiliation at the time of writing is 2150, i.e., there are 567 authors who have changed institutions or have been merged by Scopus. A total of 7174 scientific papers with this affiliation have been obtained from 1993 to 2016. These data are used to obtain general information about the scientific production of this university, as well as the collaborative relationships of the researchers affiliated to this academic institution.

According to several studies, detection, analysis and visualization of (interdisciplinary) research communities is useful to identify the research profile of an institution, as well as to support the applications of third-party funds or for establishing interdisciplinary research centres Schlattmann (2017).

4.1. General information about an institution

A general question that can be considered when analysing an academic or research institution is determining the main research areas in which that institution is specialized. This is not an easy task for those cases with a large numbers of researchers, but the method proposed in this paper is able to answer this question by automatically analysing the topics of the publications published with a certain affiliation (Baghdadi and Ranaivo-Malançon, 2011). The proposed method is able to quickly compute the scientific



(a) Total Papers (b) Coauthors

(c) Total Cites (d) h-index

Fig. 4. Structure of the method for analysing network collaborations.

production of the UAL (7174 documents), as it is described in the histograms presented in Fig. 4. Fig. 4(a) shows the frequency of how many documents are published by each author; Fig. 4(b) shows the frequencies of the number of co-authors of each author; Fig. 4(c) of shows the number of citations per document; and Fig. 4(d) displays the frequencies in terms of h-index of the authors. By applying this approach to other universities, it is possible to use these data to establish national or international rankings of scientific production.

In addition to these disaggregated data, it is possible to group the researchers by departments or according to the four subject areas that Scopus uses to classify the researchers in its database. Fig. 5 shows the average number of publications by authors affiliated with the departments of the UAL. According to these results, it is clear that the researchers included in “Physical Sciences” have a higher average number of publications (approximately 11 publications per author). This value is very interesting not only for comparing the performances of departments (or research groups) at the same institution but also for comparing departments of different universities. Furthermore, our method is able to obtain more detailed statistics about each department. For example, Table 2 shows these maximum values, average and standard deviations (S.D.) relative to the number of documents and number of co-authors, while Table 3 displays the number of citations and h-index of these departments.

Fig. 6 shows the contributions according to the Scopus subject areas: Agricultural and Biological Sciences (AGRI), Arts and Humanities (ARTS), Biochemistry, Genetics and Molecular Biology (BIOC), Business, Management and Accounting (BUSI), Chemistry (CHEM), Chemical Engineering (CENG), Computer Science (COMP), Decision Sciences (DECI), Dentistry (DENT), Earth and Planetary

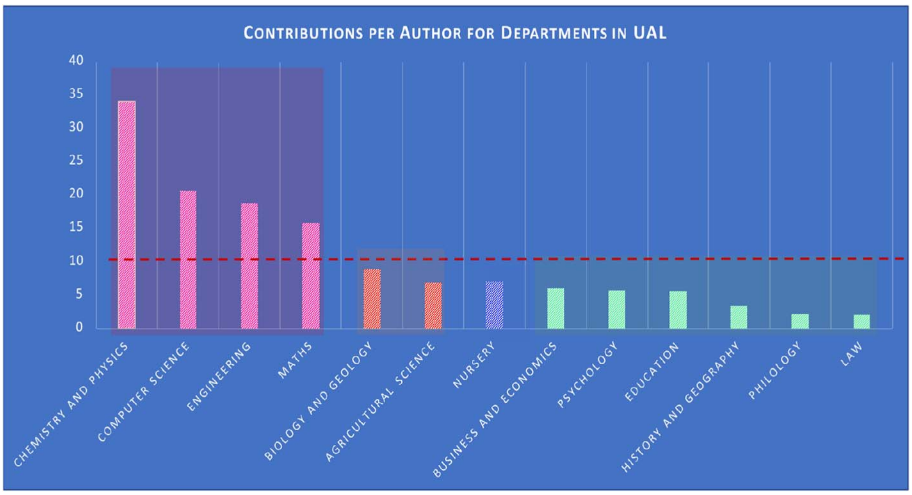


Fig. 5. Contributions per author in different departments.

Table 2
Publication statistics (papers and co-authors) of researchers affiliated to the UAL.

	Papers				Co-authors			
	Total	Average	S.D.	Max	Total	Average	S.D.	Max
Health Sciences	296	7.05	11.30	59	677	16.12	28.68	182
- Nursery	296	7.05	11.30	59	677	16.12	28.68	182
Life sciences	489	8.43	8.57	36	954	16.45	13.51	62
- Agricultural science	151	6.86	7.47	29	320	14.55	13.64	62
- Biology and geology	338	9.39	9.04	36	634	17.61	13.30	49
Physical sciences	4851	23.32	35.61	252	5595	26.90	38.07	275
- Chemistry and physics	2078	34.07	46.07	252	2709	44.41	54.62	275
- Computer science	908	21.12	37.17	195	843	19.60	26.20	140
- Engineering	1169	19.16	26.89	151	1470	24.10	28.76	136
- Maths	696	16.19	21.69	93	573	13.33	15.35	76
Social sciences & humanities	1538	5.16	9.22	77	2498	8.38	28.64	450
- Business and economics	590	6.08	10.16	77	666	6.87	14.10	122
- Education	277	5.54	12.10	70	738	14.76	63.55	450
- Geography and history	99	3.41	4.50	20	98	3.38	3.43	15
- Law	31	2.07	2.42	11	31	2.07	5.48	22
- Philology	47	2.14	1.82	7	15	0.68	1.10	5
- Psychology	494	5.81	8.88	47	950	11.18	14.11	76
Total	7174	11.84	23.72	252	9724	16.05	32.32	450

Table 3
Publication statistics (citations and h-index) of researchers affiliated to the UAL.

	Cites				h-index		
	Total	Average	St.Dev.	Max	Average	St. Dev.	Max
Health sciences	1625	38.69	102.27	528	2.00	2.54	11
⁎ Nursery	1625	38.69	102.27	528	2.00	2.54	11
Life sciences	3938	67.90	100.03	513	3.24	3.09	15
⁎ Agricultural science	957	43.50	73.42	327	2.50	2.64	10
⁎ Biology and geology	2981	82.81	110.65	513	3.69	3.26	15
Physical sciences	78,528	377.54	1014.35	11,697	6.90	8.17	63
⁎ Chemistry and physics	42,127	690.61	1542.40	11,697	11.10	10.49	63
⁎ Computer science	6121	142.35	327.74	1757	3.95	5.10	25
⁎ Engineering	25,281	414.44	917.92	5913	6.79	7.74	42
⁎ Maths	4999	116.26	187.03	779	4.14	3.97	17
Social sciences & humanities	11,549	38.76	161.88	2019	1.61	2.71	23
⁎ Business and economics	5131	52.90	163.98	1002	1.97	2.79	17
⁎ Education	2485	49.70	282.27	2019	1.48	3.26	23
⁎ Geography and history	116	4.00	6.19	20	0.76	0.93	3
⁎ Law	46	3.07	9.23	37	0.53	1.09	4
⁎ Philology	35	1.59	5.69	27	0.27	0.69	3
⁎ Physiology	3736	43.95	113.84	777	2.11	2.95	15
Total	95,640	157.82	626.92	11,697	3.60	5.80	63

Sciences (EART), Economics, Econometrics and Finance (ECON), Energy (ENER), Engineering (ENGI), Environmental Science (ENVI), Health Professions (HEAL), Immunology and Microbiology (IMMU), Materials Science (MATE), Mathematics (MATH), Medicine (MEDI), Multidisciplinary (MULT), Neuroscience (NEUR), Nursing (NURS), Pharmacology, Toxicology and Pharmaceutics (PHAR), Physics and Astronomy (PHYS), Psychology (PSYC), Social Sciences (SOCI), Veterinary (VETE). Note that some journals are assigned to different subject categories from different subject areas. As it can be seen in Fig. 6, out of the 7174 documents collected for this case study, the Scopus subareas of AGRI with 3042 documents, ENVI with 2507 documents and BIOC with 2230 documents are the specialties with the greatest number of documents published, while MULT with 87 documents, VETE with 65 documents and DENT with 10 documents are the areas in which the researchers of this university have lower scientific production. Fig. 7 shows the word cloud corresponding to the areas defined by Scopus, such that the size of the words is a function of the number of papers in each category. As shown, the areas of Agricultural and Biological Sciences, Environmental Sciences and Biochemistry are predominant in the publications of the UAL. The reason why these are important research specialties is because the province of Almeria has a favourable climate for the development of agriculture, with the largest concentration of greenhouses in the world, with an area of more than 43,000 hectares and a density of almost 5 hectares/km^2 (Clement et al., 2013).

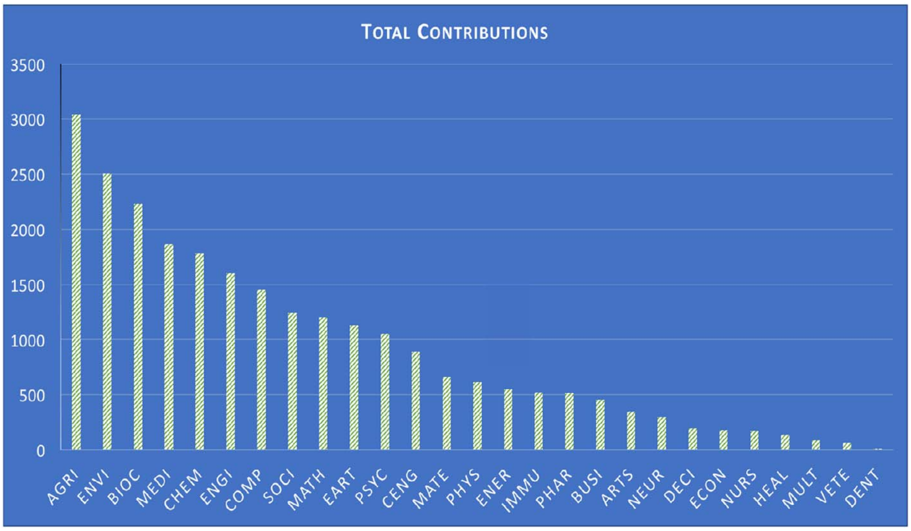


Fig. 6. Categories of the papers of researchers with affiliation to the University of Almeria.

Fig. 7. Word cloud with the specialty area of documents authored by researchers with affiliation to the University of Almeria.

Another interesting way to obtain detailed information about the scientific production and the main research lines developed in an institution comes from the analysis of the keywords in the documents published by the authors with this affiliation (Montoya et al., 2016). Here, “OpenRefine” again plays an important role because it is able to solve similarities, misspellings and to find and join clusters of words that are similar but have been written in different ways. For example, “Greenhouse”, “greenhouse” and “greenhouses” are clear examples of keywords that can be merged since the authors refer to the same concept. Fig. 8 shows the word cloud corresponding to the keywords included in the 7174 documents retrieved from Scopus with authors from the UAL. A total of 26,079 keywords have been found with this affiliation (from 1993 to 2016). After refining the data using OpenRefine, the number of keywords was reduced to 16,662, of which only 14,203 are unique keywords. Consequently, this processing is very important because the keywords have been reduced by approximately 45%. The 10 most used keywords are: Pesticides (158), Microalgae (115), Copula (80), Greenhouse (77), Spain (76), Mass spectrometry (48), Photobioreactor (48), Liquid chromatography (45), Wastewater (45), and VegeTable (41). These keywords, in turn, are part of the main areas of specialization, as shown in Fig. 6, revealing the specialization of the research carried out at this university with a greater level of detail.

4.2. Network collaborations

In addition to the general information obtained for a given institution, our method is oriented to analyse the scientific collaborations of universities and research centres, both from the internal and external viewpoint, using a graphical approach by building

Fig. 8. Word cloud with the main keywords of the papers published by researchers with affiliation to the University of Almeria.

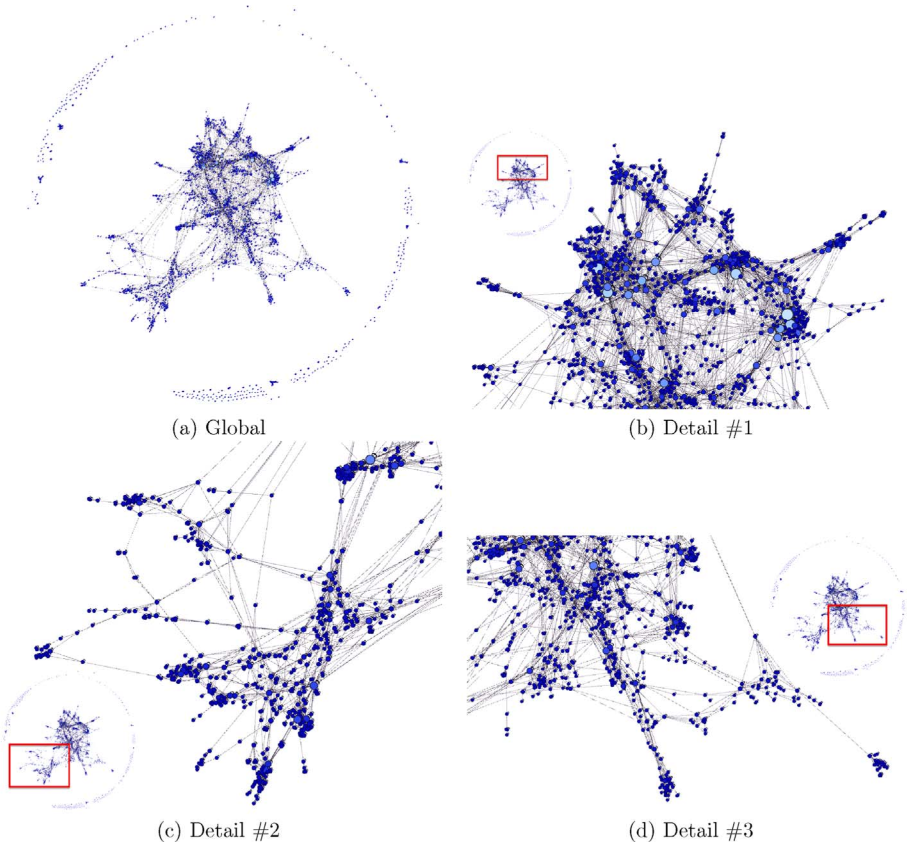


Fig. 9. Collaboration network of researchers with affiliation to the University of Almeria.

a different layout based on the ForceAtlas2 (Jacomy et al., 2014) plugin in Gephi. This visualization method builds a force directed layout by simulating a physical system in order to accommodate nodes and links in a spatial network. Nodes repel each other like charged particles, while edges attract their nodes like springs. The aim of this method is to help construct a balanced state network that facilitates the interpretation of data.

4.2.1. Internal collaborations

4.2.1.1. Global relationships. The first level of study refers to all the relationships between researchers of an academic institution. For

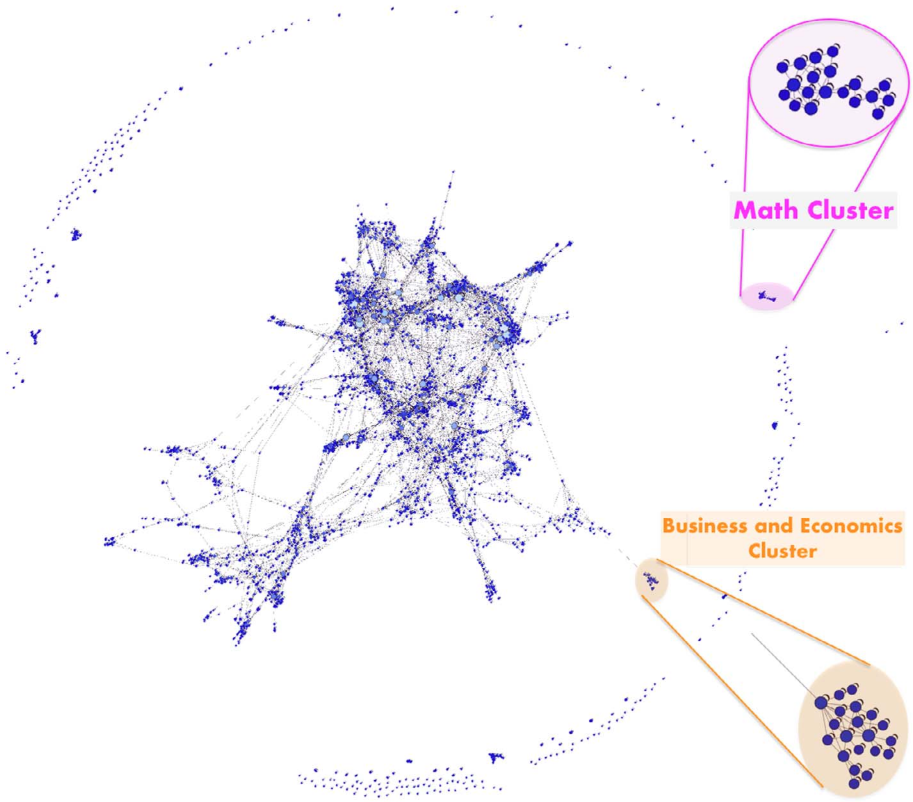
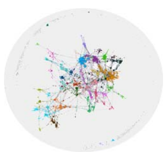
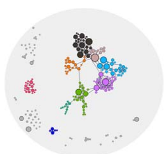
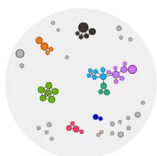
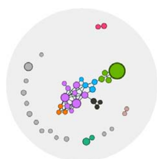
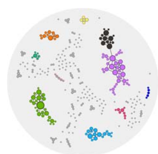


Fig. 10. Isolated clusters.

Table 4
Metrics and statistics obtained for the University of Almeria.

	University of Almeria (all researchers)	Physical Sciences	Life Sciences	Health Sciences	Social Sciences & Humanities
Number of Nodes	2717	211	60	42	300
Average Degree	12.586	6.351	3.667	4.571	3.233
Average Clustering Coefficient	0.719	0.569	0.781	0.665	0.586
Average Path Length	5.374	5.171	2.394	3.016	3.745
Connected Component	286	48	27	17	143
Diameter	14	13	5	7	11
Graph Density	0.050	0.030	0.062	0.111	0.011
Modularity	0.828	0.790	0.733	0.427	0.872
					

its development, it is considered the information collected by ResNetBOT and refined using OpenRefine. This information is shown graphically with the help of the graph visualization tool Gephi, such that each node represents a researcher, and each edge between two researchers indicate that they have co-authored at least one paper. The edges of the graph have been represented with the same thickness so as not to distort the visual representation of the relations, while the node thickness (degree) depends on the number of different co-authors that each investigator has published with. Fig. 9 shows the collaboration network obtained for the case study. Fig. 9(a) shows the existence of different clusters of researchers, with a large concentration of researchers in the central part of the graph, a lower concentration in other regions, and a set of isolated nodes disconnected from the rest of the graph. Figs. 9(b)–(d) show zoomed-in views of different regions of the graph, where it is clearly visible that there are regions with high concentrations of researchers, whereas other regions are weakly connected with the rest of the graph. The lighter the colour of a node, the larger its

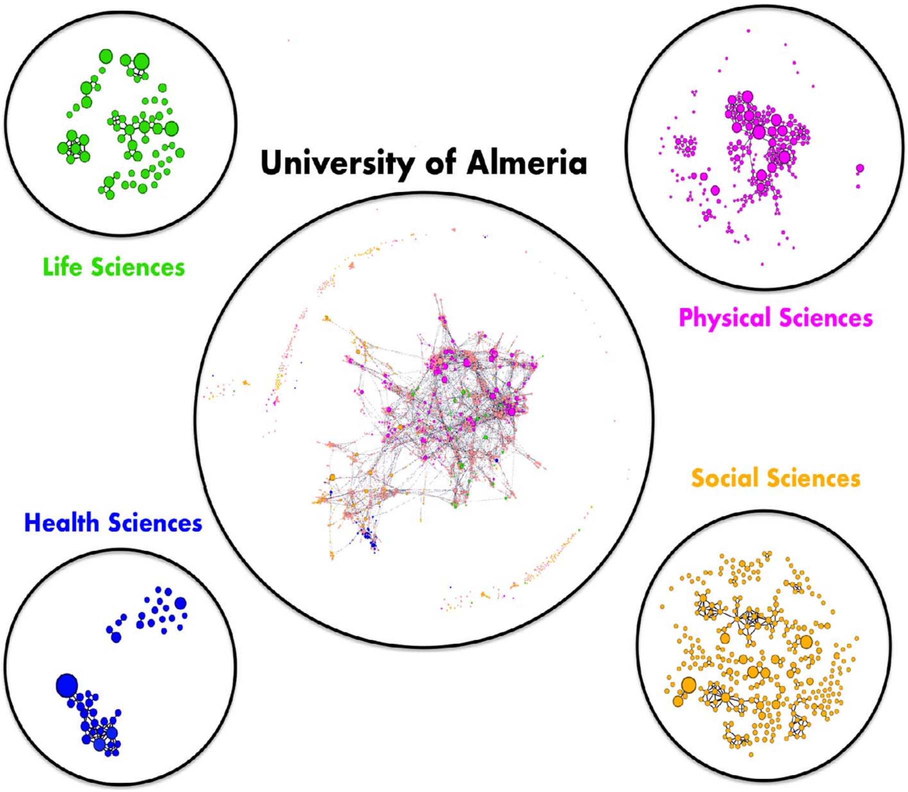


Fig. 11. Structure of collaborations by specialization areas shown as isolated clusters.

degree (number of co-authors). The nodes with low degree often correspond to two types of researchers: novice researchers who are starting their research careers, and therefore, they still do not have great relations other researchers except for their PhD supervisor (s); and experienced researchers who investigate alone or in collaboration with very few colleagues. Moreover, Fig. 10 highlights the existence of partially or totally isolated concentrations of researchers in external parts of the graph, which are related only to researchers of their discipline.

4.2.1.2. Relationships by specialty. To analyse the collaborations in different research disciplines in more detail, the information contained in the collaboration network shown in Fig. 9(a) has been disaggregated according to the four specialties into which Scopus divides the information (Physical Sciences, Life Sciences, Medical Sciences, Social Sciences & Humanities) based on the department with which the researchers are currently affiliated (according to the website of this university). Table 4 shows the numerical results related to the statistical indicators shown in Table 1 for the case of researchers affiliated with the UAL. The global data for this university are broken down by subject area based on the department with which the researchers are currently affiliated (according to the website of this university). The difference that exists between the total researchers of the UAL and the sum of those four subject areas occurs because there are many researchers (pre-doctoral or postdoctoral fellows, students, or retired lecturers) that published papers during the 24-year history of the UAL, but they are not currently part of the staff of this university.

As Fig. 11 shows, in all the disciplines, there are many researchers who collaborate with other colleagues, while in some cases, no collaboration is observed. To identify the collaborations between these four large specialties, Fig. 12 represents, in a circular form, these areas, where the arc length in the circle represents the number (proportion) of researchers, and the links between them represent the existence of collaboration (at least one paper co-authored). From this figure, it is possible to conclude that researchers grouped in Physical Sciences and in Social Sciences & Humanities have intensive inter and intra collaborations, while these collaborations are more sporadic between researchers of Life Sciences and Health Sciences. This procedure could be applied to analyse the characteristics and collaborations of any internal structure of an institution, such as departments or research groups.

4.2.1.3. Relationships of group of researchers. One aspect of great interest when analysing relationships between researchers is to determine the different structures of existing collaborations (Hinkin et al., 2007). In addition to analysing the global data of an institution and the collaboration structures of researchers grouped in specialties or departments, the method proposed here also allows analysing the characteristics and close collaborations between small groups of researchers. Fig. 13 shows some of the typical structures of collaboration that have been found in the case study:

- Clique: A group of researchers in which all collaborate with each other without exception, and therefore, they are isolated from

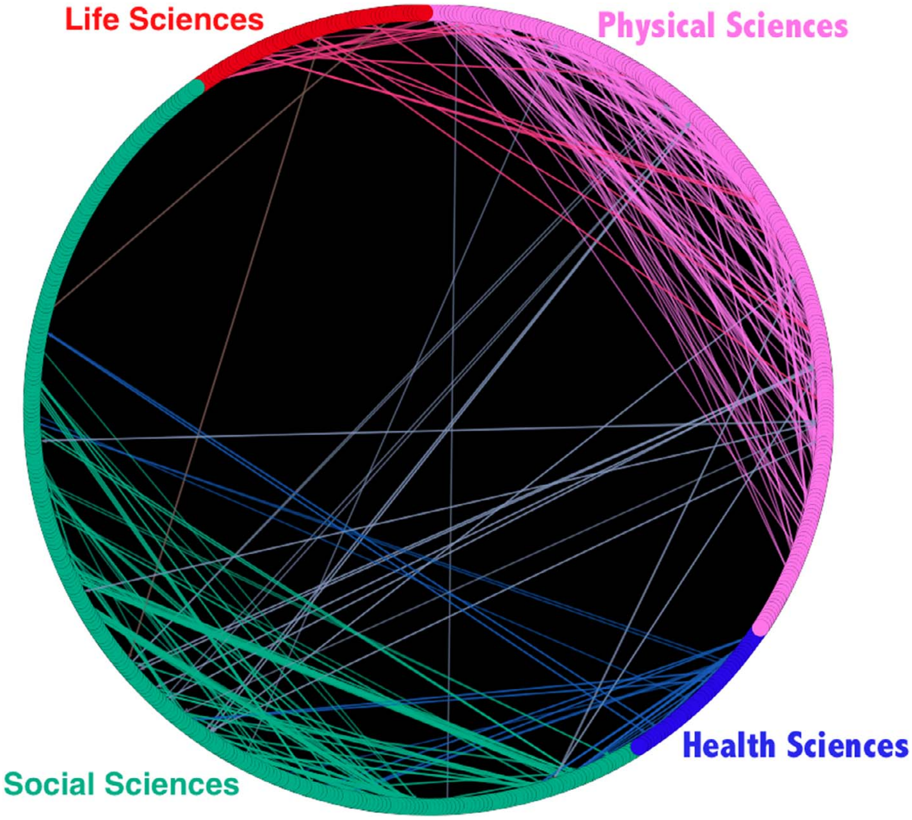


Fig. 12. Structure of collaborations within and between specialization areas shown as circular layout.

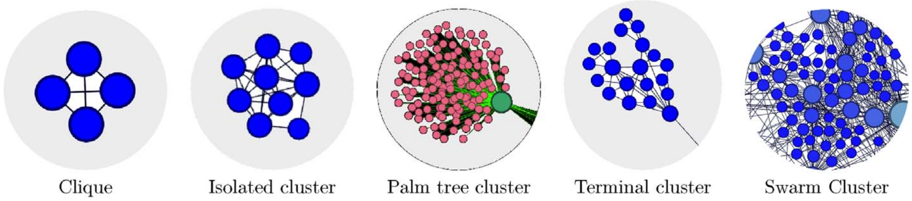


Fig. 13. Types of structures for network relationship of groups of researchers.

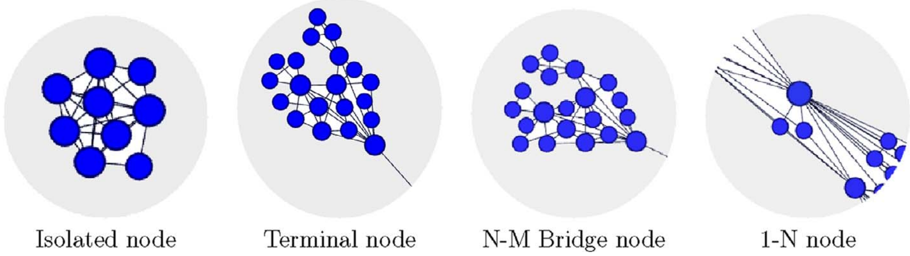


Fig. 14. Types of structures for network relationships of single researchers.

the rest.

- Isolated clusters: A group of researchers that collaborates internally but without any collaboration with researchers outside that group. In the case study, this topology is observed with some researchers enrolled in the Departments of Mathematics and Psychology.
- Palm Cluster: This structure reflects a collaboration between a single researcher of the institution with many other researchers from other institutions, which denotes high levels of external collaboration.
- Terminal cluster: A group of researchers collaborate internally, but collaboration with researchers outside that group is very

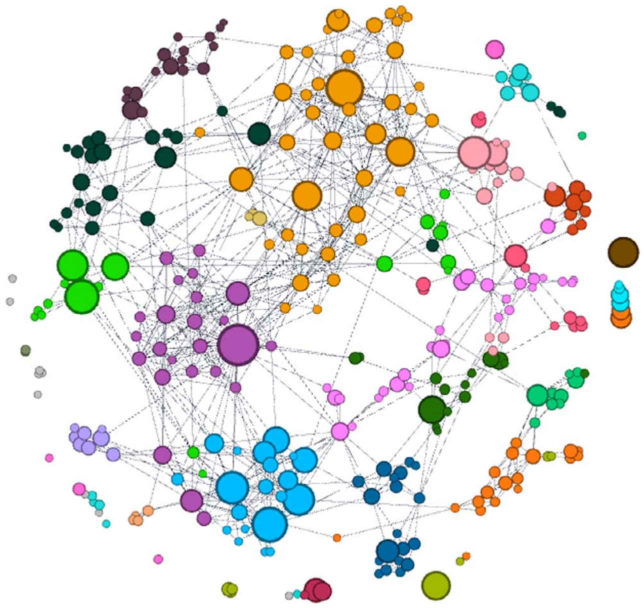


Fig. 15. Structure of collaborations by researchers with h-index greater than 10.

limited. In the case study, this typology occurs in the form of collaborating with some researchers enrolled in the Department of Business and Economics.

- Swarm cluster: Nodes are usually grouped into swarms where there is a high density of links between researchers, which indicates a high level of internal collaboration.

4.2.1.4. Relationships of individual researchers. In addition to the study of the typical structures of collaboration of groups of researchers, the different topological structures of collaboration between individual researchers are also analysed. Fig. 14 shows the typical collaborative structures typically encountered when analysing unique researchers at the UAL. As can be seen, there are isolated nodes that investigate without any collaboration; terminal nodes, which only collaborate with another researcher; bridge nodes between an investigator and N researchers ($N \geqslant 2$); and bridge nodes between N and M investigators ($N,M \geqslant 2$). This information is very useful in order to determine the level of collaboration of the researchers analysed. As Fig. 10 shows, most researchers affiliated with the UAL are N-M bridge nodes or isolated nodes. A more detailed study of these types of structures would be useful to analyse different situations, such as the possible isolation that new researchers may suffer as the funding system penalizes researchers with few publications.

It is possible to analyze the structure of collaborations of researchers with higher scientific production. A large number of author metrics have been proposed to evaluate the scientific output of researchers, being the h-index (Hirsch, 2005) one of the most widely used. Fig. 15 shows the structure of collaborations of those researchers affiliated to the UAL having an h-index greater than 10. It can be seen how social sciences areas are quite widespread in the center of the graph. In contrast, researchers investigating in life science topics are quite concentrated on one side of the graph.

4.2.2. External collaborations

Collaborations between researchers is an important issue to analyse the influence of an academic or research institution. In particular, collaborations with foreign researchers give information about the level of internationalization (Altbach and Knight, 2007; Kwiek, 2015). Fig. 16 shows the national and international collaboration of researchers at the UAL. It is observed that most of the external collaborations are with other Spanish researcher centres, but the collaboration with other European and American researchers is also significant. In addition, the method proposed here is useful to analyse the collaboration networks between researchers at different research centres (Uddin et al., 2013). First, the relationships between the researchers of this university and other Spanish universities or research centres are analysed. Subsequently, the relationships between the researchers of this university and other foreign universities or research centres are analysed.

4.2.2.1. Collaboration with national researchers. Fig. 16 shows information that is valuable for providing an idea about the national and international collaborations of researchers affiliated with the UAL, but additional information can be obtained if the network of collaborations is analysed. Thus, Fig. 17 extends the graph displayed in Fig. 9 to include the collaborations of researchers with affiliation to the "University of Almeria" (blue nodes) with external researchers from other Spanish institutions, which are displayed as red nodes. An interesting piece of information provided by this figure is that there are several concentrations that can be classified as palm clusters, which denote experienced researchers from the UAL that collaborate with many other researchers from other

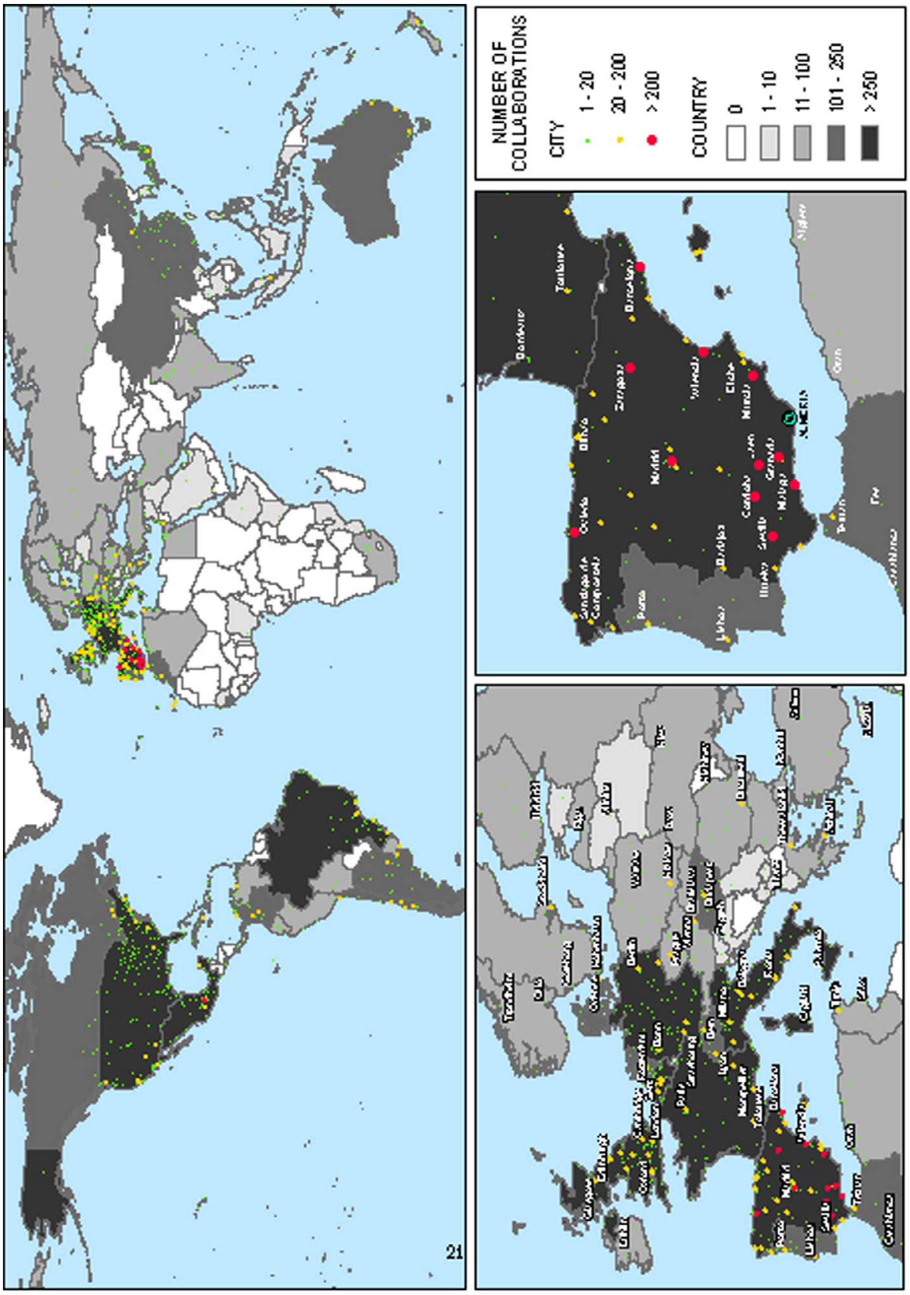


Fig. 16. Locations of co-authors of researchers with affiliation to the University of Almeria.

national institutions.

4.2.2.2. Collaboration with international researchers. Fig. 18 extends the graph displayed in Fig. 9 to include the collaborations with external researchers from foreign institutions, which are displayed as pink nodes. In the same way as Fig. 17, Fig. 18 shows the existence of a large number of palm clusters, which denote how experienced researchers from the UAL collaborate with many other researchers at other international institutions. In this case, the number of external nodes included in these palm clusters is slightly higher (these palm clusters are denser on average) than those presented in Fig. 17, which denotes that the level of external collaborations with researchers from foreign institutions is, quantitatively, more important than the external collaborations with researchers from other Spanish institutions.

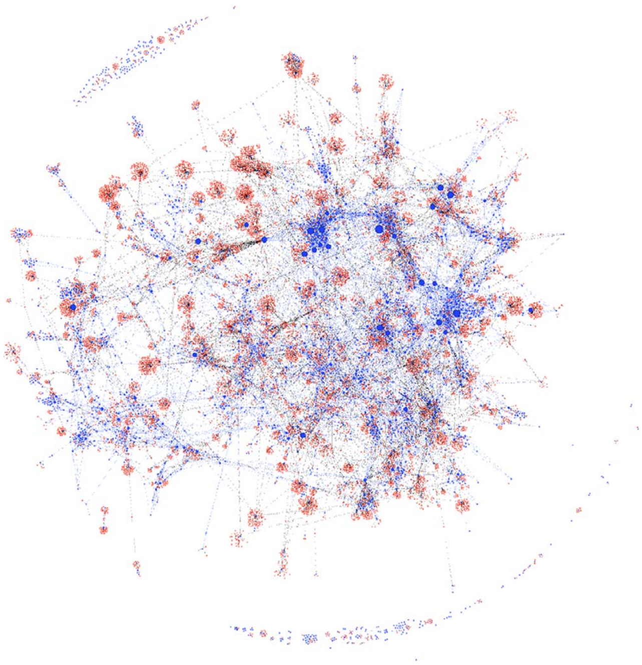


Fig. 17. External collaboration of researchers with affiliation to the University of Almeria with researchers from other Spanish institutions.

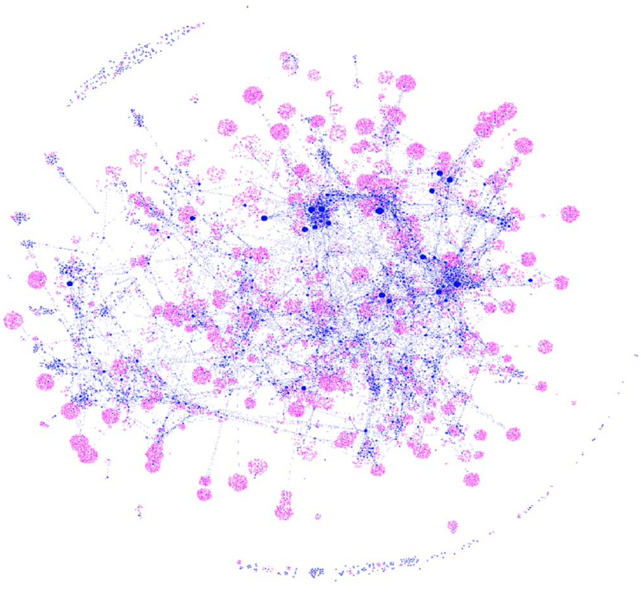


Fig. 18. External collaboration of researchers with affiliation to the University of Almeria with researchers from other Spanish institutions.

5. Conclusions

Most scientometric studies are usually based on the use of journal metrics and author metrics, whereas an analysis of how large numbers of researchers collaborate is lacking. However, the relationships among researchers play key roles in finding solutions to scientific and technological problems as well as in clarifying doubts about specific topics and increasing knowledge. This article presents a method to cover a gap that exists in the literature specialized in the systematic analysis of scientific collaborations at research institutions. Our method consists of an automated process for the extraction of large volumes of information using scripts (ResNetBOT) and the Scopus database API interface, including parsing and refinement of this information, which is later processed both to collect information using typical author metrics and to analyse the collaboration structures of these researchers using graph visualization software tools. Since the aim is to use the Scopus database for academic purposes, the method has been validated in a real case study corresponding to a medium-sized university that has 2717 researchers who have used this affiliation (613 of which form the academic staff of this university) and published a total of 7174 articles. In addition to the study of the collaboration network using specialty fields, we have also analysed the typical collaborations between group of researchers and those involving individual researchers.

6. Implications and limitations

The results show that the proposed method is a powerful tool that can be useful for funding agencies to evaluate the relationships among individual researchers, research groups or research organizations; for research managers interested in analysing how to coordinate the activities of researchers or research groups to achieve greater synergies; and for individual researchers interested in analysing potential researchers with whom to collaborate.

It is important to notice that gathering information from the Scopus database is a complex task. In particular, the number of queries is limited as well as the amount of information that can be downloaded due to the use of quotas. Having in mind that each single user can get up to 10 API keys, the analysis of a given institution using the method proposed here would require the use the API keys from different users. Moreover, they must be satisfied the terms of use and policies established by Elsevier.

Acknowledgements

This research has been supported by the Spanish Ministry of Economy and Competitiveness (project TIN2015-67020-P leaded by the University of Granada), and by the Regional Government of Andalusia (ceiA3 project) at the University of Almeria.

References

Abbasi, A., Altmann, J., Hossain, L., 2011. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. J. Informetrics 5 (4), 594–607. http://dx.doi.org/10.1016/j.joi.2011.05.007. ISSN 1751-1577.

Ajay, A., Henderson, R.M., 2002. Putting patents in context: exploring knowledge transfer from mit. Manage. Sci. 48 (1), 44–60.

Altbach, P.G., Knight, J., 2007. The internationalization of higher education: Motivations and realities. J. Stud. Int. Educ. 11 (3–4), 290–305.

Aparicio-Martínez, P., Perea-Moreno, A., Martinez-Jimenez, M., Varo, I.S.-V., Vaquero-Abellán, M., 2017. Social networks' unnoticed influence on body image in spanish university students. Telematics Inform.

Aron, D., 2009. Dynamic collaboration: a personal reflection. J. Inf. Technol. 24 (3), 214–218.

Asadi, S., Dahlan, H.M., et al., 2017. Organizational research in the field of green it: a systematic literature review from 2007 to 2016. Telematics Inform.

Baghdadi, H.S., Ranaivo-Malançon, B., 2011. An automatic topic identification algorithm. J. Comput. Sci. 7 (9), 1363.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the Third International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence (AAAI).

Baxter, R., Christen, P., Churches, T., 2003. A comparison of fast blocking methods for record linkage. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, vol. 3. ACM, pp. 25–27.

Beck, F., Burch, M., Diehl, S., Weiskopf, D., 2016. A taxonomy and survey of dynamic graph visualization. In: Computer Graphics Forum. Wiley Online Library.

Bhattacharyya, M., Bandyopadhyay, S., 2015. Finding quasi core with simulated stacked neural networks. Inf. Sci. 294, 1–14.

Boden, B., Haag, R., Seidl, T., 2013. Detecting and exploring clusters in attributed graphs: a plugin for the gephi platform. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, pp. 2505–2508.

Bornmann, L., Mutz, R., Daniel, H.-D., 2008. Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. J. Assoc. Inf. Sci. Technol. 59 (5), 830–837.

Bruns, A., 2012. How long is a tweet? mapping dynamic conversation networks on twitter using gawk and gephi. Inf. Commun. Soc. 15 (9), 1323–1351.

Cañas-Guerrero, I., Mazarron, F., Calleja-Perucho, C., Pou-Merina, A., 2014. Bibliometric analysis in the international context of the construction & building technology category from the web of science database. Constr. Build. Mater. 53, 13–25.

Cañas-Guerrero, I., Mazarron, F., Pou-Merina, A., Calleja-Perucho, C., Diaz-Rubio, G., 2013. Bibliometric analysis of research activity in the agronomy category from the web of science, 1997–2011. Eur. J. Agron. 50, 19–28.

Cao, X., Chen, Y., Ray Liu, K., 2016. A data analytic approach to quantifying scientific impact. J. Informetrics 10 (2), 471–484.

Cavnar, W.B., Trenkle, J.M., 1994. N-gram-based text categorization. Ann Arbor MI 48113 (2), 161–175.

Chadegani, A.A., Salehi, H., Yunus, M.M., Farhadi, H., Fooladi, M., Farhadi, M., Ebrahim, N.A., 2013. A comparison between two main academic literature collections: web of science and scopus databases. Asian Soc. Sci. 9 (5), 18–26.

Clement, J., Novas, N., Gazquez, J.-A., Manzano-Agugliaro, F., 2013. An active contour computer algorithm for the classification of cucumbers. Comput. Electron. Agric. 92, 75–81.

Daňa, J., Ráček, J., 2015. Analysis of czech actors' cooperation network using gephi. Inf. Technol. Pract. 33–38.

Ebadi, A., Schiffauerova, A., 2015. How to become an important player in scientific collaboration networks? J. Informetrics 9 (4), 809–825.

Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S., 2007. Duplicate record detection: a survey. IEEE Trans. Knowl. Data Eng. 19 (1), 1–16.

Elsevier, 2017. Scopus database api interface.https://dev.elsevier.com/scapis.html.

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G., 2008. Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. FASEB J. 22 (2), 338–342.

Franceschet, M., 2010. Journal influence factors. J. Informetrics 4 (3), 239–248.

Franceschini, F., Maisano, D., Mastrogiacomo, L., 2016. Empirical analysis and classification of database errors in scopus and web of science. J. Informetrics 10 (4), 933–953.

Fuchs, C., 2017. Sustainability and community networks. Telematics Inform. 34 (2), 628–639.

Guthrie, S., Wamae, W., Diepeveen, S., Wooding, S., Grant, J., 1993. Measuring Research: A Guide to Research Evaluation Frameworks and Tools. RAND Corporation, Santa Monica, CA.

Haak, L.L., Fenner, M., Paglione, L., Pentz, E., Ratner, H., 2012. Orcid: a system to uniquely identify researchers. Learned Publishing 25 (4), 259–264.

Hinkin, T., Holtom, B.C., Klag, M., 2007. Collaborative research: developing mutually beneficial relationships between researchers and organizations. Organizational Dyn. 36 (1), 105–118.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proc. Nat. Acad. Sci. 16569–16572.

Hoppe, B., Reinelt, C., 2010. Social network analysis and the evaluation of leadership networks. Leadersh. Q. 21 (3), 600–619.

Ingwersen, P., 2014. Scientific datasets: informetric characteristics and social utility metrics for biodiversity data sources. In: Library and Information Sciences. Springer, pp. 107–117.

Jacomy, M., Venturini, T., Heymann, S., Bastian, M., 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi

software. PloS One 9 (6), e98679.
Jin, L., Li, C., Mehrotra, S., 2003. Efficient record linkage in large data sets. In: Proceedings of the Eighth International Conference on Database Systems for Advanced Applications (DASFAA 2003). IEEE, pp. 137–146.
JSON, 2017. Json.http://www.json.org.
Kanavos, A., Theodoridis, E., Tsakalidis, A., 2014. A pubmed meta search engine based on biomedical entity mining. In: Proceedings – International Workshop on Database and Expert Systems Applications, DEXA. pp. 82–86.
Kaushal, S.S., Jeschke, J.M., 2013. Collegiality versus competition: how metrics shape scientific communities. Bioscience 63 (3), 155–156.
Kwiek, M., 2015. The internationalization of research in europe. A quantitative study of 11 national systems from a micro-level perspective. J. Stud. Int. Educ. 19 (4), 341–359.
Laudel, G., Gläser, J., 2014. Beyond breakthrough research: epistemic properties of research and their consequences for research funding. Res. Policy 43 (7), 1204–1216.
Leslie, S.A., Rensleigh, C., 2013. Web of science, scopus and google scholar: a content comprehensiveness comparison. Electron. Lib. 31 (6), 727–744.
Lewis, J.M., Ross, S., Holden, T., 2012. The how and why of academic collaboration: disciplinary differences and policy implications. High. Educ. 64 (5), 693–708.
Ma, T., McGroarty, F., 2017. Social machines: how recent technological advances have aided financialisation. J. Inf. Technol.
McNutt, M., 2014. The measure of research merit. Science 346, 1155.
Mongeon, P., Paul-Hus, A., 2016. The journal coverage of web of science and scopus: a comparative analysis. Scientometrics 106 (1), 213–228.
Montoya, F.G., Baños, R., Meroño, J.E., Manzano-Agugliaro, F., 2016. The research of water use in Spain. J. Cleaner Prod. 112, 4719–4732.
Montoya, F.G., Montoya, M.G., Gómez, J., Manzano-Agugliaro, F., Alameda-Hernández, E., 2014. The research on energy in Spain: a scientometric approach. Renew. Sustain. Energy Rev. 29, 173–183.
Newman, M.E., 2004. Coauthorship networks and patterns of scientific collaboration. Proc. Nat. Acad. Sci. 101 (suppl 1), 5200–5205.
Noorden, R.V., 2014. Global scientific output doubles every nine years. Mathematics & Economics Nature News Blog,http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html.
OpenRefine, 2017. Openrefine.http://openrefine.org/.
Papavlasopoulos, S., Poulos, M., Korfiatis, N., Bokos, G., 2010. A non-linear index to evaluate a journal's scientific impact. Inf. Sci. 180 (11), 2156–2175.
Penni, J., 2017. The future of online social networks (OSN): a measurement analysis using social media tools and application. Telematics Inform. 34 (5), 498–517.
Ravishankar, M., 2013. Public ICT innovations: a strategic ambiguity perspective. J. Inf. Technol. 28 (4), 316–332.
Rojas-Sola, J., Aguilera-Garcia, A., 2015. Global bibliometric analysis of the mining & mineral processing subject category from the web of science (1997–2012). Miner. Process. Extr. Metall. Rev. 36 (6), 349–369 cited By 2.
Rojas-Sola, J., de San-Antonio-Gomez, C., 2010. Bibliometric analysis of spanish scientific publications in the subject construction & building technology in web of science database (1997-2008) [analisis bibliometrico de las publicaciones cientificas españolas en la categoria construction & building technology de la base de datos web of science (1997-2008)]. Materiales de Construccion 60 (300), 143–149.
Schlattmann, S., 2017. Capturing the collaboration intensity of research institutions using social network analysis. Proc. Comput. Sci. 106, 25–31.
Seglen, P.O., 1997. Why the impact factor of journals should not be used for evaluating research. BMJ. Br. Med. J. 314 (7079), 498–502.
Shafiq, O., Alhajj, R., Rokne, J., 2015. On personalizing web search using social network analysis. Inf. Sci. 314, 55–76.
Shen, A., Lee, M., Cheung, C., Chen, H., 2010. Gender differences in intentional social action: we-intention to engage in social network-facilitated team collaboration. J. Inf. Technol. 25 (2), 152–169.
Singh, V., Perdigones, A., Garcia, J., Cañas-Guerroro, I., Mazarron, F., 2015. Analyzing worldwide research in hardware architecture, 1997–2011. Commun. ACM 58 (1), 76–85.
Szymaaski, J., Rzeniewicz, J., 2016. Identification of category associations using a multilabel classifier. Expert Syst. Appl. 61, 327–342.
Szymanski, B., De La Rosa, J., Krishnamoorthy, M., 2012. An internet measure of the value of citations. Inf. Sci. 185 (1), 18–31.
Uddin, S., Hossain, L., Rasmussen, K., 2013. Network effects on scientific collaborations. PloS One 8 (2), e57546.
Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., Aleixandre-Benavent, R., 2015. A systematic analysis of duplicate records in scopus. J. Informetrics 9 (3), 570–576.
Whitley, R., 2003. Competition and pluralism in the public sciences: the impact of institutional frameworks on the organisation of academic science. Res. Policy 32 (6), 1015–1029.
Yu, H., 2011. Selective sampling techniques for feedback-based data retrieval. Data Min. Knowl. Disc. 22 (1–2), 1–30.
Zheng, X., Le, Y., Chan, Albert P.C., Hu, Yi, Li, Y., 2016. Review of the application of social network analysis (SNA) in construction project management research. Int. J. Project Manage. 34 (7), 1214–1225. http://dx.doi.org/10.1016/j.ijproman.2016.06.005. ISSN 0263-7863.