OPEN ACCESS
Vietnam Journal of Computer Science
Vol. 7, No. 4 (2020) 391–415

DOI: 10.1142/S2196888820500220

World Scientific
www.worldscientific.com

Analysis of Internal and External Academic Collaboration in an Institution Through Graph Theory

Jared D. T. Guerrero-Sosa*, Víctor H. Menéndez-Domínguez†,
María-Enriqueta Castellanos-Bolaños‡ and Luis F. Curi-Quintal§

Facultad de Matemáticas
Universidad Autónoma de Yucatán
Mérida, México
**jared.guerrero@correo.uady.mx*
†*mdoming@correo.uady.mx*
‡*enriqueta.c@correo.uady.mx*
§*cquintal@correo.uady.mx*

Received 16 January 2020
Accepted 2 May 2020
Published 2 September 2020

This paper presents an analysis of scientific collaboration through graph theory, based on a previous study focused on the collaborative work of researchers within an institution. This proposal also exposes the representation of inter-institutional collaboration of research groups, combining graph theory and data mining. The state of the art relates the concepts of scientific production, digital repositories, interoperability between repositories, the law of Open Science in Mexico, the theory of graphs and their use in previous studies for the analysis of scientific collaboration, and the definition of research groups in Mexico. Furthermore, the methodology uses elements of knowledge extraction for data mining, involving recovery, processing and visualization. Results present the collaboration status at the Universidad Autónoma de Yucatán, internally and externally, by the research groups. Internally, 22 groups were found and each researcher collaborates with six other professors within the institution, on average. In addition, consolidated research groups are those with the highest level of production and collaboration at national and international level, compared to the groups with less consolidation.

Keywords: Scientific collaboration; graph theory; digital repositories; external collaboration; research groups; interinstitutional collaboration.

1. Introduction

Nowadays, a common activity in scientific production is working in groups (collaborative work). Scientific collaborations involve, among other things, the definition of tasks and the reasons to collaborate, number of constant collaborators and personal

† Corresponding author.

factors.[1] On a personal level, researchers collaborate to gain experience, visibility and prestige in the scientific community; to progress scientifically in a more agile way; solve bigger problems; receive incentives by diverse evaluating institutions,[2] among other factors. It is established that various institutions grant economic resources and recognition to scientific work in order to motivate their researchers and generate more projects that solve problems in society through science, technology and innovation. Several studies involve more than one area of knowledge, so scientific collaborations become a necessity.

In addition, part of the prestige of a university or research center implies the relationship, collaboration and dissemination of work with experts from external institutions, whether national or foreign. Institutions that are responsible for evaluating group or individual production, in addition to conventional indicators (impact factor, citation index, among others) consider collaboration with external institutions important. Some evaluating instances in Mexico are the National System of Researchers (SNI) and the Program for Professional Teacher Development (PRODEP).

In Mexico, academic bodies fulfill the function of research groups, in addition to teaching work. The importance of each academic body is measured by the degree of consolidation, which is granted according to the production and collaboration of the academic body.

Several studies have been carried out on the state of scientific collaboration in countries or geographical areas, for example, the scientific production in Peru was analyzed based on the journals indexed by the Institute for Scientific Information (ISI) considering a fixed interval of time of 10 years (2000–2009), highlighting the characteristics of the production about Medicine (that is, a specific area), the production for geographical areas in Peru and collaborations with national and foreign institutions, both represented with graphs but the graph theory is not applied.[3]

On the other hand, the scientific production in Latin America was analyzed using as source the Web of Science considering a fixed interval of time (1975–2004).[3,4] Some of the main results in the work are: (1) the high levels of collaboration in small countries, (2) the effort intra-regional is focused on researchers about Biology, Medicine, Physics and Chemistry and (3) the collaboration between Brazil and Argentina is focused on Medicine and Astronomy. The work is based on Web of Science, a recognized worldwide database. This allows evaluating scientific production that has a high level of relevance in the scientific community. The authors did not use a graph to represent the collaborations.

There are studies that have been carried out on the state of scientific collaboration in specific institutions. There is a work-focused in the Universidad Nacional Autónoma de México (UNAM) using as source Web of Science and a fixed interval of time (1999–2002).[5] Some of the main results are: (1) 60% of the collaborations in Science involved only national institutions, (2) 87% of the collaborations in Social Sciences, Arts and Humanities involved only national institutions, (3) papers published in collaboration with foreign institutions were most cited than the published

with only national collaborations and (4) there were significant levels of collaboration between the UNAM research centers and local institutions, represented with a weighted graph. Using the graph theory can be a useful tool to describe and analyze those collaborations.

On the other hand, there is a work about the scientific production in the Universidad Autónoma de Yucatán (UADY).[6] The work extracted the production from Scopus and Web of Science in a fixed interval of time (1900–2016). The aim was the growth of the production along the years and the scientific impact using quantitative bibliometric theory. It highlights that part of the increasing production in the UADY is due to its researchers that are in the SNI. The work presents two graphs: one for the researchers and the topics of the two first works associated with UADY, and another one to represent the first internal collaborations in UADY. This work only uses the graphs for representation, but it does not use the quantitative elements of graph theory.

A previous stage of the project proposes a methodology for the extraction, filtering and storage of scientific production, and through graph theory, it is possible to identify members of an institution (university or research center) that have mutually collaborated.[7]

The previous work and this paper presents a model for the extraction of scientific production from two data sources: Scopus, a database of abstracts and citations of high-impact publications, considering the problem of the existence of more than one profile for the same investigator; and some open access repositories that offer the metadata of their stored products through the OAI-PMH protocol. This model retrieves and analyzes the scientific production of one specific institution.

It is necessary to have in a database some information about the researchers: full name of the researcher, faculty or area of knowledge where they belong and the institution that granted them the last academic degree.

For the extraction from Scopus, this model uses the provided APIs by this repository: Author Searching-Scopus (its aim is to search interfaces associated with Scopus-based Author profiles[8]) and Scopus Search API (its aim is to search information about Scopus abstracts[9]). Both require parameters to execute the queries and their values are taken from a database of researchers. The procedure is described as follows:

(1) Search in the Scopus searcher (online) the institution whose scientific production will be retrieved.
(2) In the affiliation profile, retrieve the Affiliation ID associated with the institution.
(3) If the institution has more than one affiliation profile, retrieve all the Affiliation IDs for the institution.
(4) Through the Author Searching-Scopus, the model retrieves all the author profiles of every researcher using as parameters the full name and the Affiliation IDs for the institution. If this query did not return any result, try again using as

parameters the full name and the institution that granted them the last academic degree.

(5) The model analyzes all the author profiles. All of them contains information as the name, the affiliation, the area of knowledge and the identifier in Scopus (Scopus ID). For every author profile, if the affiliation matches with the institution to evaluate or with the institution that granted them the last academic degree and the area of knowledge matches with the faculty of the researcher, then the information of the author profile must be stored in a database. It is necessary to consider the Scopus IDs for the next steps.

(6) To retrieving the scientific production associated with the researchers, the model creates a query through the Scopus Search API using as parameters the Scopus IDs of every researcher and the Affiliation ID of the institution which will be evaluated. In this way, the production made by a researcher as a member of another institution is excluded. The model stores in a database the metadata of every publication.

For the extraction from open access repositories, the model retrieves using the OAI-PMH protocol for interoperability. The procedure is described as follows:

(1) With every URL provider of every repository, retrieve all the metadata associated with the scientific production through the verb *listRecords*.

(2) In every record the model analyzes the metadata that describe the authorship information.

(3) If some values of those metadata contain the full name of at least one researcher of the institution which will be evaluated, the model stores in a database the metadata of the publication.

For the realization of this work some previous ones were carried out, such as the analysis of the indicators for the scientific production,[10] of the libraries for the metadata harvesting through the interoperability[11] and the proposal of an indicators system for the relevance of scientific production.[12]

In the previous work, the model for retrieving, storing and processing the metadata associated with the researchers and scientific production of a specific institution was presented.[7] From the recovered data an ordinary combination of two elements (where each element is a researcher) was proposed to retrieve the scientific production associated with those two researchers. Considering the number of participations of every couple it was possible to represent the scientific collaboration through a graph, where every researcher is a node and their collaborations are the vertex. The graph is undirected and weighted. The weight of every vertex depends on the number of collaborations between the two researchers. And using the graph theory the state of the collaboration was described.

In this work, based on the stored and recovered academic production, the state of collaboration of professors from a Mexican public university is represented, in

addition to the representation of the joint work between the research groups established in the same university, adding analysis, from the perspective of data mining, to find patterns in these collaborations (national and international). Attributes that describe the production and collaboration of research groups were discretized, and through association rules, relationships between behaviors of academic bodies are described.

In this section, the introduction to the proposal of this work was presented. Section 2 exposes the related topics for the understanding of the concepts used later. Section 3 proposes the use of knowledge extraction techniques for data mining (collection, processing and visualization). Section 4 explains the graphs obtained about scientific collaboration in the UADY (Mexico), and its analysis using statistical concepts of graph theory and data mining. Finally, Sec. 6 explains the conclusions.

2. State of the Art

Scientific publications have been accessible thanks to the existence of repositories of digital resources. A digital repository is a platform that is responsible for storing, preserving and being a tool for disseminating content, and can be classified as follows: as Open Educational Resources (OER), as OER references, as OpenCourseware initiatives and as a system of learning management.[13] A characteristic of repositories is the interoperability, which is the ability to connect with other ones for the exchange and use of their data[14] using specialized protocols, such as OAI-PMH, OpenAIRE and IMS. Among the variety of publications found in the digital repositories are the original papers, the case reports, the technical notes and the pictorial essays.[15]

In Mexico, there is the Open Science Law, which establishes that anyone has free access, without making a prior payment, to the publications, products of research works stored in digital repositories of Mexican institutions and made with public resources in order to increase the accessibility of scientific research for all Mexicans through the maximum dissemination of scientific knowledge, technology and innovation.[16]

Regarding to publications, they can be indexed, in other words, they can be located in a database that is responsible for collecting citations, even through periodic evaluations. These databases uses quantitative statistical indicators (h-index, impact factor, index i10) to measure scientific productivity and with them determine the quality of publications, generally scientific journals and serialized books. Internationally, the main citation databases are two: Scopus and Web of Science. However, Google Scholar has been considered as another alternative due to the use of its own indicators for scientific productivity. There are studies that compare characteristics, strengths and weaknesses of citation databases, using as reference Scopus, Web of Science and Google Scholar.[17–20]

Something that characterizes contemporary science (from the mid-twentieth century) is the constant collaboration with scientists of homogeneous and

heterogeneous knowledge, being an advantage to complement concepts and conceive new ones. This group of people who collaborate is known as a scientific community, and they do not necessarily have to know each other face to face. Media such as telephone and Internet help the interaction between them, facilitating collaborative work.[21]

In Mexico, academic bodies fulfill the function of a research group, in addition to teaching work. In general, it is a group of professors-researchers who have one or several lines of study in common, and aim at both the application and generation of new knowledge as a result of collaborative work.[22] Academic bodies are classified into three types, and each has specific characteristics for state universities, technological institutes or technological universities. The characteristics presented below correspond to state universities, because the case of study in this work applies to academic bodies from a university of this type.

- Consolidated Academic Body (CAC). It is the maximum possible level that can be assigned to an academic body. Its characteristics are as follows: (1) the majority of its members have the maximum academic qualification to generate or apply knowledge in an innovative and independent way (PhD); (2) they have extensive experience in teaching and human resources training; (3) most of its members have the desirable profile defined by PRODEP; (4) high commitment to the institution through collaboration and scientific and academic production; (5) demonstrate an intense academic activity in congresses, seminars, workshops, etc. on a regular and frequent basis and (6) maintain an intense participation in academic exchange networks.
- Academic Body in Consolidation (CAEC). It is the intermediate level that can be assigned to an academic body. It is characterized by: (1) more than a half of its members have a doctorate; (2) have academic products with recognition due to their good quality, related to consolidated research lines; (3) at least one-third of their members have the desirable profile defined by PRODEP; (4) participate jointly in innovative research or application lines of knowledge; (5) have extensive experience in teaching and training human resources and (6) collaborate with other academic bodies.
- Academic Body in Training (CAEF). It refers to academic bodies that are in an early stage of development, created from one or more lines of research. Its characteristics are as follows: (1) members are identified; (2) at least half of members have the desirable profile defined by PRODEP; (3) research lines are clearly defined and (4) higher level related external academic bodies have been identified in order to establish collaboration.

A useful tool for representing the relation between two elements among a set of several of them is the theory of graphs. In Computer Science, a graph is a mathematical abstraction, represented as

$$G = (V, E), \tag{1}$$

where V is a set of vertices and E is a set of edges. In this way, graphs are useful for modeling relations between elements and allows the resolution of problems associated with their context and requires a less expensive process than even linear programming, and to represent them, there are three options: graphic representation, representation with an adjacent matrix, and the dictionary of the graph.[23] The graphic representation of a graph consists of presenting each vertex as a point or circle, generally. Although they can be represented by other figures, depending on what you want to graph. The relations between the vertices are represented by connecting lines.

There is a wide variety of tools for the construction and visualization of graphs, and among them are NodeXL, Gephi and Google Fusion Tables.

With regard to the representation of scientific collaborations with graph theory, either directly or indirectly, several studies have been carried out considering geographical areas, disciplines, various databases and time windows. Below are some that have obtained relevant results that confirm known facts and also discover specific cases.

There is a study about the relationship between researchers in Biomedicine, Theoretical Physics, High Energy Physics and Computer Science in the period from 1995 to 1999 using MEDLINE databases, Los Alamos e-Print Archive, SPIRES and NCSTRL.[24] This research focuses on the calculation of the distance between authors, but considers important to include the number of collaborators of the scientists, the number of papers they produce, and the probability that two collaborators of a specific researcher have collaborated. This study finds that it is very likely that the formation of scientific communities originates thanks to the existence of a common collaborator between two researchers, except in Biomedicine. In addition, this study reaffirms the high degree of collaboration among Physics researchers and emphasizes that the representation of collaboration is not specific to Computer Science, but began with Mathematics with the concept of Erdös number.

In another study[25] the authors conducted a study of the analysis of scientific collaboration between the Consejo Superior de Investigaciones Científicas (CSIC) and Latin America focused on various disciplines. They got a dataset from ISI (today Thomson Scientific) and Web of Science considering the publications between 2001 and 2004. In spite of having apparent disadvantages related to the language and the low representation of non-Anglo-Saxon magazines, the authors considered the registers of the names of the researchers. Based on the classification of journals in disciplines and areas, indicators of activity, impact and collaboration were obtained. According to the results obtained, Physics and Chemistry are the areas with the highest degree of collaboration, and the Social Sciences have little research work as a whole.

On the other hand, there is a method for the identification and analysis of the scientific collaboration applicable to universities and research centers in general, considering the need to facilitate the location of the most productive and relevant

areas and disciplines for the allocation of economic resources destined to projects.[26] They used the Scopus database. The methodology used is to obtain information from publications, authors and collaborative networks to analyze with elements of graph theory such as the number of nodes, density of the graph, among others.

There is a study about the interaction and the degree of collaboration between different disciplines through the analysis of social networks from a database provided by the Office of Scientific and Technology Information of the US Department of Energy using statistical techniques based on the graph theory.[27] The study considers the social network as a set of nodes (people, institutions and countries) and edges (relationships between nodes). They created two different networks, one for the publications between 1980 and 2000 and other one for 2000–2012. Among the relevant results are a strong collaboration between nuclear, energy and environmental disciplines, reinforcing the 2013 trend of studies on renewable energies.

Other studies analyze the importance of collaborative publications. An example is the analysis of publications made on a specific topic (smoking) in 79 countries,[28] and one of the most relevant results is that much of the production in United Kingdom, United States and Germany has been performed in inter-institutional collaboration. Another result is the high degree of international collaboration on the issue among United States, United Kingdom and France. However, the most interesting result is that all countries, without exception, received more citations in publications made in inter-institutional or international collaboration, than in those made without any collaboration.

Another study focused on universities of Cuba, affirms that in most of Cuban universities, international collaboration is a determining element to achieve a good performance concerning citations.[29]

On the other hand, a publication presents results of a study on visualization of international collaboration networks in order to identify the most international aspect of research from scientific publications, considering Spain, Brazil, Mexico and Cuba. It concludes that Spain is the country that collaborates with most of the countries, and the one that has more visibility in its publications due to the collaborative work internationally.[30]

3. Methodology

Figure 1 presents the architecture of the system, which proposes a solution for the retrieval of information about the researchers with their corresponding scientific production stored in digital repositories. The architecture has been designed in layers, facilitating the flow of information between them.

Each layer has a specific function. The layer of the user interface presents the information obtained by the services from the obtained data. The service layer consists of modules that work together for representing the results through the user interface and among its functions are the calculation of indicators, location of collaborations between researchers, SPARQL queries and SWRL rules for the

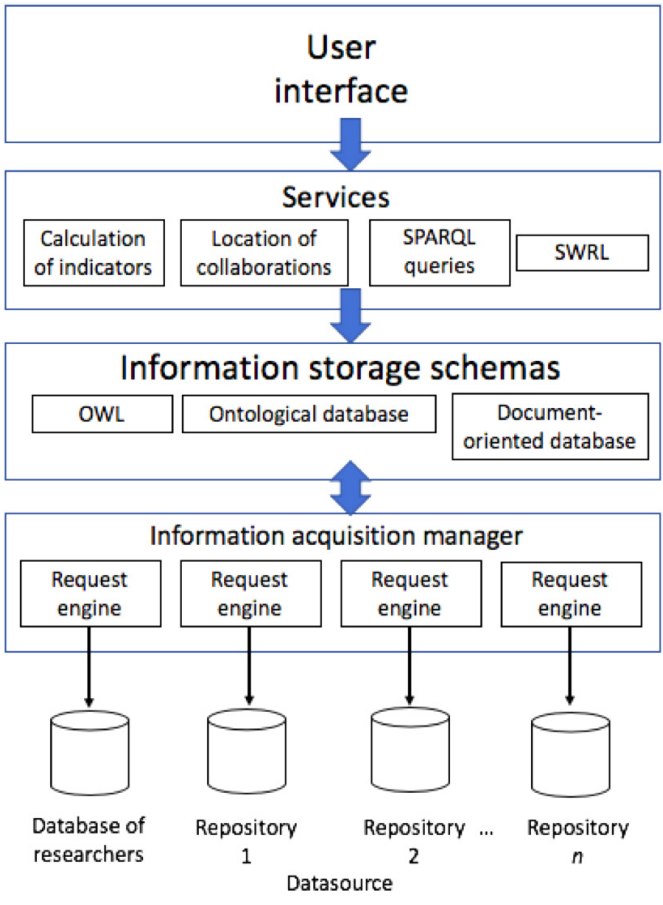

Fig. 1. System architecture.

ontological model of scientific production. The layer of information storage schemas contains the tools that are responsible for storing the information of the researchers and their production as OWL, an ontological and non-relational database. The information acquisition manager layer contains request engines that connect directly to different repositories, either through the verbs of the protocol for metadata harvesting OAI-PMH or some API, if the repository requires it. Finally, the datasources are the repositories that contain information about the academic and scientific production of the researchers.

To apply the proposed methodology, it is necessary to have two databases: (1) for researchers, considering their name, institution that granted his/her last academic degree and the area of knowledge in which they work, and (2) for academic bodies, considering their name, the names of the members, the degree of consolidation and the research line. It is necessary that members of academic bodies belong to the institution whose production will be recovered, in order to maintain the consistency

of the information. Making use of the methodology of data mining, specifically in e-learning applications, we propose one of knowledge extraction, which consists of the following stages[31]:

- Data collection. It is divided into three parts.
  (1) Researchers. Once the researchers and their related data are identified, their identifications of authors in Scopus are collected. A query is made to Scopus through its author search API,[8] which is an HTTP request, using as parameters the first and the last name of the researcher, and the institution to which he or she belongs or that granted him or her the last academic degree. The result obtained is in XML format by default, or JSON. From all the results obtained, the following are considered as identifications of authors in Scopus of a specific researcher if they meet any of the following conditions:
     — If the researcher has two last names: the first name and surnames returned by the query completely match those registered in the database; or the initials of the first name or at least the first one match the initials of the first name of the researcher registered in the database, and at least the first surname matches the one registered in the database.
     — If the researcher has only one last name: the first name and the last name returned by the query completely match those registered in the database; or the initials of the name or at least the first one match the initials of the name of the researcher registered in the database, and the surname coincides with the one registered in the database.

     In addition, each record must match the area of knowledge registered in the database, as well as the institution in which he or she belongs, or at least the institution that granted the last academic degree. Each result is stored in the document-oriented database.
  (2) Publications. For the publications indexed by Scopus, a query is made through its publication search API,[9] with an HTTP request using as parameters the Scopus keys related to each researcher and the Scopus keys associated with the institution where the researchers belongs. Among the publications to be retrieved are papers, book chapters, books and conference papers. Each result is stored in the document-oriented database. If the publication is already registered in the database, because it belongs to more than one researcher of the institution, the field of authors is updated. For publications not indexed by Scopus and available in open access repositories that use the protocol OAI-PMH, metadata are harvested with the verb *ListRecords*.
  (3) Collaborating Institutions. From the production recovered from Scopus, information on each external institution that has participated in the publications is consulted through Scopus' API. For each institution, database stores its name, the city and the country of residence.

- Data processing. It is divided into two parts.

  (1) Researchers. To make the graph, it is necessary to represent each researcher by a vertex, and the collaborations with edges. Ordinary combinations of two elements are required because the elements of each subset are considered independently of the arrangement. Then, the number of combinations is defined by the following equation:

$$C_{n,2} = \frac{n!}{2!(n-2)!}, \tag{2}$$

  where n is the number of researchers belonging to the institution. Once the groups of combinations are obtained, a query is made to the database for retrieving the number of publications in which the researchers involved have collaborated. The results are generated as a table, where the first and second columns contain the names of the authors of the collaboration and the third column contains the number of publications in common.

  (2) Academic bodies. In order to recover the production associated with an academic body, it is necessary to find all the ordinary combinations of two elements among all the members of the group, since a publication belongs to an academic body if at least two of its members are authors. A list containing the information of all publications of the academic body will be created. Subsequently, publications associated with the pair in question are consulted in the scientific production database. If the product is not in the list, its associated information is attached (title, authors and Scopus keys of the institution of the authors).

  After recovering and storing the production of each academic body, a list is created to store other institutions that have collaborated with all academic bodies of the origin institution. Name, city and country of the external institution are retrieved through Scopus API. Each author of a publication belonging to an external institution is considered as an external collaboration. That is, if in one publication two authors belong to the same university in the United States, the academic body has two collaborations with that institution. The number of collaborations of each academic body with each external institution is counted. Results are generated as a table, where the first column contains the name of the academic body, the second column contains the name of the external institution and the third column, the tally of collaborations.

- Visualization. With the resulting data from previous stages, the table is imported into some tool for the visualization of graphs and carry out the subsequent analysis of the scientific collaboration in the evaluated institution. The graph fulfills the following characteristics:

  — Graph $G(V, E)$
  — V is a set of vertices

— E is a set of edges
— G is an undirected graph
— G is a weighted graph, and the weights of the edges are the number of publications in common between the researchers represented by the vertices

Two graphs will be created: the first will represent the internal collaboration, and the second, the external collaborations of academic bodies.

- Description of the collaboration of academic bodies. Data mining techniques were used to get a better description of the degree of external collaboration of the institution's research groups. Mining process included the following stages:

(1) Data is grouped according to similar characteristics.[32] This stage analyzes data stored in the database of academic bodies (number of members, number of external collaborations at national and international level, total production and number of collaborations), and according to classification rules, a collection of physical or abstract objects grouped in classes is displayed.
(2) Linguistic description of the groups. After grouping the data, centroid values of the attributes are analyzed and compared for each group. Based on these results, the names of each group are assigned, describing the degree of collaboration (high, medium and low).
(3) Assignment of linguistic labels of attributes. For each attribute, categories (high, medium and low) are created, depending on the maximum, average and minimum levels.
(4) Association Rules. Once labels are assigned for both groups and attributes, this stage locates rules that associate concepts from different attributes in academic bodies' database. Association and correlation are used to search for a frequent element from a large amount of information.[33,34]

4. Results

A strategy to validate our proposal consisted in the case study of the UADY, the most important institution of higher education in the southeast of Mexico. This institution has 15 faculties distributed in five campuses and one research center focused on two areas of study. As of June 1, 2018, the UADY had 78 academic bodies and 824 full-time professors. Table 1 shows the distribution of professors and academic bodies on each campus.

For the data collection phase of the methodology, we had a database of researchers from the UADY, located in Mexico, which has the names of researchers, faculty or area of knowledge where they belong, and the institution that granted them the last academic degree.

In addition, a database of academic bodies was obtained, which includes the name of the group, names of members, degree of consolidation and the research line. Scopus was consulted to retrieve the identifications of authors in this platform, because a researcher can have multiple profiles in Scopus. In order to collect the information

Table 1. Full-time professors and academic bodies by campus.

Campus	Full-time professors	Academic bodies
Architecture, habitat, art and design	34	3
Biological and agricultural sciences	84	10
Health sciences	159	13
Exact sciences and engineering	232	21
Social, economic, administrative and humanities sciences	226	21
Regional research center — social sciences unit	28	3
Regional research center — biomedical sciences unit	61	7

about the indexed publications, the query was made in Scopus using the identifications of authors and the two identifications of institution of UADY as parameters, since only the publications belonging to this institution are considered. For collecting the production not indexed by Scopus, the information was retrieved from various open access repositories that use the protocol for metadata harvesting OAI-PMH. These tasks were executed by Python scripts.

For each publication listed in Scopus, title, authors and Scopus keys of the institutions to which authors belong were retrieved and stored. From publications available in open access repositories, only titles and names of authors were retrieved and stored. From UADY professors, 2740 publications were found in Scopus and 118 in open access repositories stored in the national repository of Mexico.[35]

Results about the collaboration of researchers and academic bodies are described in Secs. 4.1 and 4.2.

4.1. *Collaboration of full-time professors*

Hypothesis: The scientific collaborations of the researchers of an institution can be retrieved and represented with a graph, and with the graph theory elements, those collaborations can be described.

All retrieved data is stored in the MongoDB database. Through another Python script, the data processing phase generates the table of scientific collaborations, which is exported in .xslx format to be later processed by the Google visualization tool, Fusion Tables, which it is useful for clarity in the representation of the graph with a considerable number of vertices. To avoid that the file has unnecessary information, such as the relations between researchers who do not collaborate, these were omitted. Gephi was used to obtain the statistics of the graph.

In this paper, 824 teachers were evaluated, of which, 390 have publications in collaboration with other professors of the same institution. Each teacher is represented by a vertex. There are 1181 edges, that represent relations of scientific collaboration. The obtained graph is presented in Fig. 2, where the size of the vertices is determined by the degree (number of collaborators) and degree with weights (number of collaborations). The largest vertex is the Researcher 670 with 14 collaborators and 174 collaborations (Fig. 3).

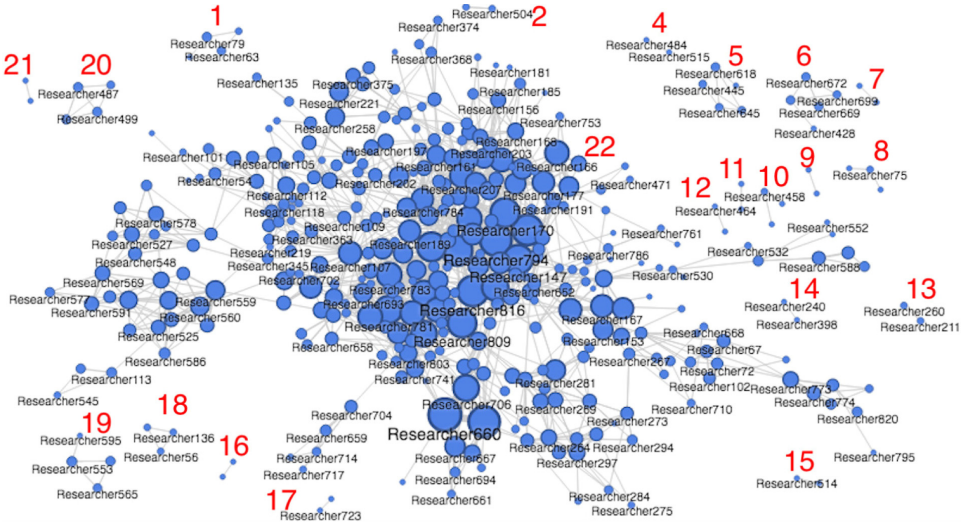

Fig. 2. Graph of representation of the scientific collaboration in UADY.

In addition, the obtained graph has 22 subgraphs that show collaborations that have no relation among them, or, collaboration groups, represented in Fig. 2 with numbers in red. Using the Gephi tool, interesting results were obtained about the collaboration, using elements of graph theory, which can be observed in Table 2.

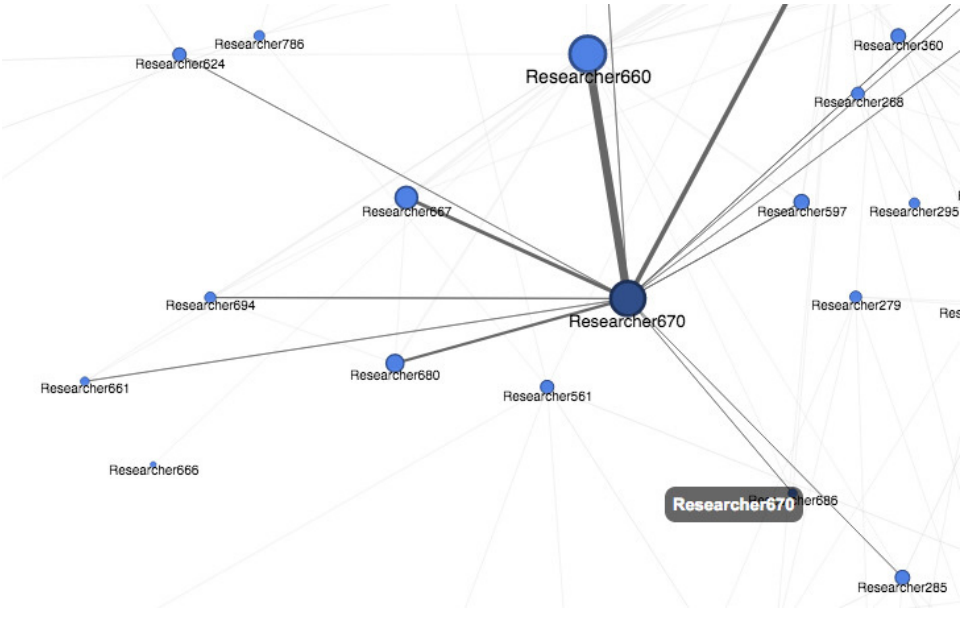

Fig. 3. The researcher with the highest degree of collaboration in UADY.

Table 2. Characteristics and statistics of the obtained graph.

Characteristic	Value
Number of vertices	390
Number of edges	1181
Average degree	6.056
Average weighted degree	18.4
Graph density	0.016

The interesting results of the representation of the collaboration through the graph and its statistics are the following:

- On average, each researcher has collaborated with six researchers belonging to the UADY, considering the average degree of the graph.
- On average, each researcher has 18 collaborations, considering the average weighted degree.
- The graph density is very low, being 0.016 a value very close to 0. If the graph was completely connected, the graph density would be 1. But, considering that the UADY has 390 collaborating researchers, it is very difficult for each researcher to collaborate with each other.
- In the UADY there are 22 collaboration groups.

Additionally, external collaborations with other universities and research centers in Mexico and with other countries were located from the researchers who collaborate internally at UADY. Table 3 shows the Mexican institutions with which UADY has the highest degree of collaboration, and Table 4 shows the countries with which there is greater collaboration. There are 499 Mexican institutions and 89 countries that had collaborated with UADY.

4.2. *Collaboration of academic bodies*

Hypothesis: The scientific collaborations of the academic bodies with national and foreign institutions can be retrieved and represented with a graph, and the elements of graph theory can describe the collaborations by the level of consolidation of those groups while the association rules allow reporting the behavior in the collaborations.

Table 3. Ranking of Mexican institutions with the greatest collaboration with UADY.

Rank	Institution	Collaborations
1	Universidad Autónoma de México	375
2	Centro de Investigación Científica de Yucatán	253
3	CINVESTAV Unidad Mérida	226
4	Centro de Investigación y de Estudios Avanzados	208
5	Instituto Politécnico Nacional	175

Table 4. Ranking of countries with the greatest collaboration with UADY.

Rank	Country	Collaborations
1	United States	1144
2	Spain	306
3	France	213
4	United Kingdom	206
5	Germany	160

Previously, all the scientific production of UADY professors found in Scopus has been stored. Since part of these publications belong to the academic bodies, a Python script was created to identify them, considering, for each academic body, all the ordinary combinations of two body members. For each pair of members their joint publications are located. Titles, authors and Scopus keys of each institution for each author are considered. Due to periodic external evaluations to academic bodies (approximately every three years), production from 2016 to November 2019 was considered. As a result, 313 publications associated to the academic bodies of UADY were found.

Moreover, tally of academic bodies in UADY with production, classified by its consolidation degree, are as follows:

- In Training: 1
- In Consolidation: 18
- Consolidated: 29

Academic bodies that collaborated with external institutions are as follows:

- In Training: 0
- In Consolidation: 14
- Consolidated: 26

Academic bodies that collaborated with Mexican institutions are as follows:

- In Training: 0
- In Consolidation: 13
- Consolidated: 25

Academic bodies that collaborated with foreign institutions are as follows:

- In Training: 0
- In Consolidation: 8
- Consolidated: 18

In addition, only academic bodies in consolidation and consolidated were considered since they are experienced research groups. Of these academic bodies, only those who

registered at least one inter-institutional collaboration were analyzed. We worked with 41 academic bodies, 26 consolidated and 15 in consolidation.

Then, names, cities and countries of each external institution that has collaborated with the academic bodies were consulted through the Scopus API, using the Scopus key of each institution as a query parameter.

A Python script generated a .csv file with the following structure: first column contains the name of the academic body, second column, the name of the collaborating external institution, and third column, the weight of the connection between both nodes, that is, the number of collaborations.

This file was used to generate a graph through the Gephi tool. In Fig. 4, each node represents an academic body or a collaborating external institution. Orange nodes represent consolidated academic bodies, blue ones to in consolidating academic bodies, purple ones to Mexican institutions, and green ones to foreign institutions. Size of each node depends on vertex degree, that is, the larger the number of collaborators of the academic body or of the institution, the bigger is its size.

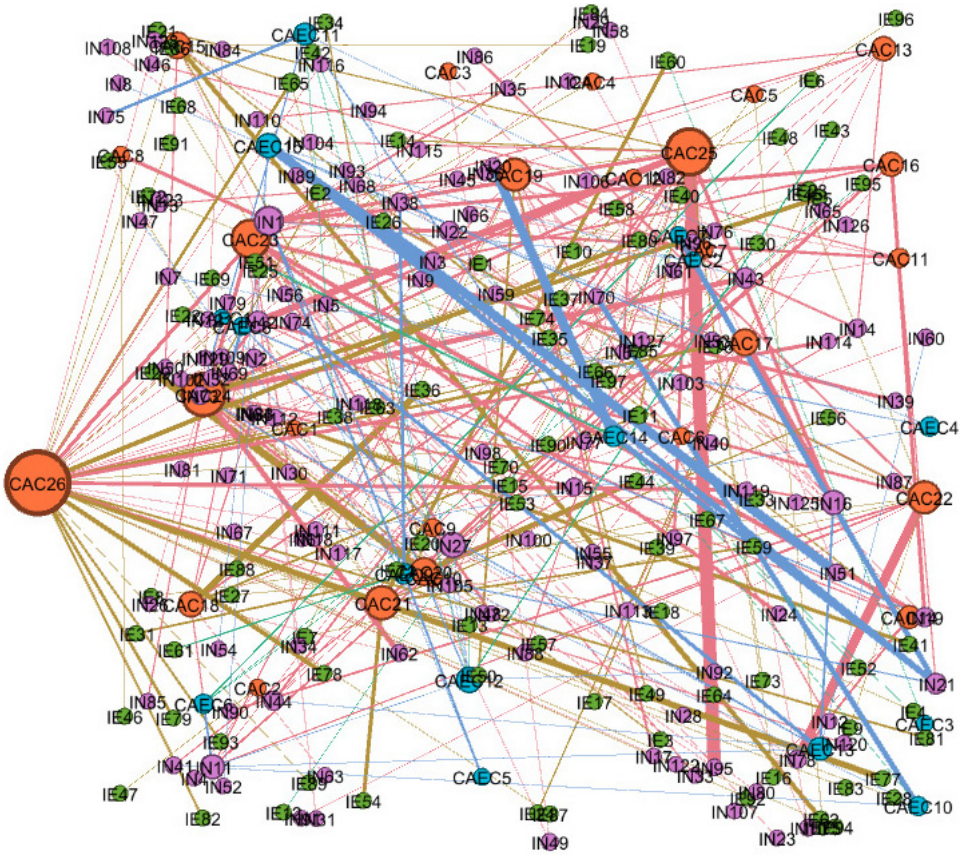

Fig. 4. Graph of representation of the scientific collaboration of the research groups in the UADY.

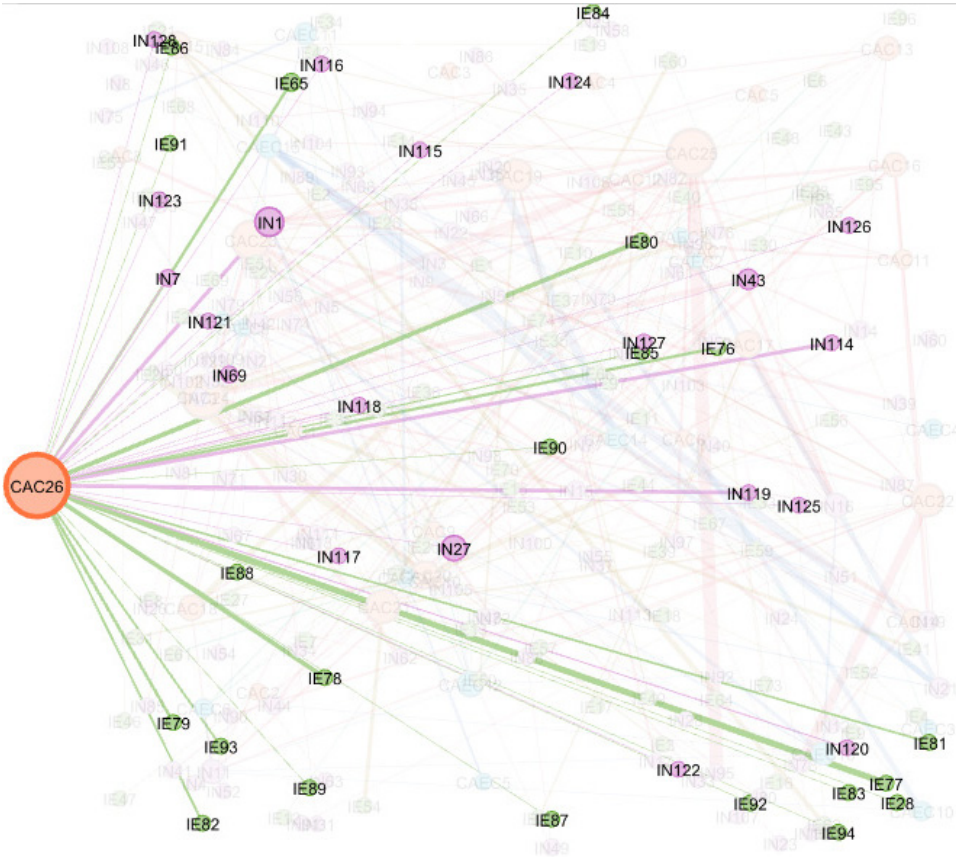

Fig. 5. The research group with the highest degree of collaboration in UADY.

The thickness of each edge depends on the weight or number of collaborations between the academic body and the external institution.

In the graph, there are 264 nodes, of which: 15 are academic bodies in consolidation, 26 are consolidated academic bodies, 126 are Mexican institutions and 96 are foreign institutions. Among all the academic bodies considered, the CAC26 has the greatest degree, since it has collaborated with 41 external institutions through 89 collaborations (Fig. 5).

Similar to the analysis of collaboration of professors in UADY through graph theory, it is possible to analyze the state of collaboration of academic bodies with other institutions (Table 5).

Observed results about collaboration of academic bodies include the following:

- On average, each academic body in consolidation has collaborated with four external institutions, and has made an average of nine collaborations with external institutions.

Table 5. Statistics for the characteristics in the graph of collaboration of academic bodies.

Characteristic	Value
Consolidated academic bodies	26
Academic bodies in consolidation	15
National collaborating institutions	126
Foreign collaborating institutions	96
Average degree of academic bodies in consolidation	4.2
Weighted average degree of academic bodies in consolidation	9.57
Average degree of consolidated academic bodies	9.8
Weighted average degree of consolidated academic bodies	21.07
Average degree of national collaborating institutions	1.65
Weighted average degree of national collaborating institutions	3.86
Average degree of foreign collaborating institutions	1.08
Weighted average degree of foreign collaborating institutions	2.17

- On average, each consolidated academic body has collaborated with nine external institutions, and has made an average of 21 collaborations with external institutions.
- Regarding to 126 Mexican external institutions that have collaborated with academic bodies, each one has collaborated on average with one academic body, three times with that academic body between 2016 and 2019.
- Regarding to 96 foreign institutions that have collaborated with academic bodies, each one has collaborated on average with one academic body, twice with that body between 2016 and 2019.

On the other hand, a .csv file was generated containing the degree of consolidation of each academic body, the number of members, the number of collaborations with national institutions, the number of collaborations with foreign institutions, the total production from 2016 to November 2019 and the total external collaborations. This file is used to analyze the collaboration of academic bodies, which is reflected in the graph presented in Fig. 4.

From this file, academic bodies were classified into groups, or clusters, and through the G-Means algorithm (which is a variant of KMeans), three types were obtained: less collaborative, moderately collaborative and most collaborative. Grouping was based on attributes: number of members, total number of collaborations (national and foreign) and total production.

Data values were discretized considering the minimum, maximum and average (Table 6) in order to get a better representation in grouping results. Each attribute is linguistically represented with the tags low, medium and high.

The resulting distribution of academic bodies in clustered groups is:

- Less collaborative: 26 bodies (15 consolidated and 11 in consolidation)
- Moderately collaborative: nine bodies (five consolidated and four in consolidation)
- Most collaborative: six consolidated bodies

Table 6. Minimum, maximum and average values for academic bodies attributes.

Attribute	Minimum	Maximum	Average
Number of members	3	11	5.02
National collaborations	0	61	12.02
Foreign collaborations	0	55	5.12
Total production	1	26	7.80
Total external collaborations	1	89	17.14

Based on discretized variables and the three level of collaboration for academic bodies, association rules with consequent (the name of the group) and without consequent.

Association rules found with consequent are as follows:

- If an academic body has low national collaborations, then it is less collaborative.
- If an academic body has low collaborations in general, then it is less collaborative.
- If an academic body has low members, then it is less collaborative.

Table 7. Top ranked Mexican institutions that collaborate with academic bodies in UADY.

Rank	Institution	Collaborations
1	Universidad Nacional Autónoma de México	36
2	CINVESTAV Unidad Mérida	26
2	Universidad Autónoma de San Luis Potosí	26
3	Universidad Juárez Autónoma de Tabasco	20
4	El Colegio de la Frontera Sur	17
5	Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias	16

Table 8. Top ranked foreign institutions that collaborate with academic bodies in UADY.

Rank	Institution	Country	Collaborations
1	Martin-Universität Halle-Wittenberg	Germany	10
2	Mohammed V University in Rabat	Moroco	8
2	University of California, Davis	United States	8
3	University of California, Riverside	United States	7
3	Universidad de Murcia	Spain	7
4	Colorado State University	United States	6
4	Arizona State University	United States	6
4	Universidad Nacional del Litoral	Argentina	6
4	University of California, Santa Cruz	United States	6
4	University of California, Irvine	United States	6
5	Baylor College of Medicine	United States	5
5	The University of British Columbia	Canada	5
5	CSIC — Instituto de la Grasa (IG)	Spain	5

Association rules found without consequent are as follows:

- If an academic body has low collaborations in general, then it has low foreign collaborations.
- If an academic body has high collaborations in general, then it has high national collaborations.
- If an academic body has low collaborations in general, then it has low national collaborations.

Based on the same information, Table 7 presents the five top ranked Mexican institutions and Table 8 the five top ranked foreign institutions that have collaborated with UADY.

5. Discussion of Results

Results explain the scientific collaboration in the UADY, which consists of 390 researchers, and each one collaborates on average with six researchers from the same institution (obtained with the average degree of the graph) and has 18 collaborations in his or her publications (obtained with the average weighted degree of the graph).

From the results obtained in the first experiment, it can be concluded that the first hypothesis has been demonstrated and therefore it can be stated that: The scientific collaborations of the researchers of an institution can be retrieved and represented with a graph, and with the graph theory elements, those collaborations can be described.

From the collaboration of academic bodies with external institutions, it is observed that only 48 of 78 academic bodies analyzed made at least one indexed publication in Scopus (61.5%). Consolidated academic bodies generate the majority of the production, with 60% of the academic bodies being productive in the period considered. Academic bodies in consolidation generate 37.5% of the production. In contrast, academic bodies in training only generate 2.08.

Most of the consolidated academic bodies collaborated with some external institution, with 89% of the bodies making some publications. However, only one of 26 consolidated academic bodies did not collaborate with a Mexican institution, while 18 of them made publications with some foreign institution.

Regarding academic bodies in consolidation, based on 18 bodies that generated publications, 14 (77%) did so in collaboration with an external institution, and only one of these 14 bodies did not published in collaboration with any Mexican institution. Respecting international collaboration, only eight academic bodies (44.44%) published in collaboration with an foreign institution.

Theory of graphs confirmed the latter findings, since it showed that the average grade for consolidated academic bodies is 9, and this value is greater than 4, the average grade for academic bodies in consolidation. In consequence, consolidated academic bodies collaborate with more external institutions.

Similarly, consolidated academic bodies have a weighted average grade greater than the weighted average grade of academic bodies in consolidation (21 and 9, respectively). Consequently, a consolidated academic body has collaborated more than a body in consolidation.

This information is visualized in the graph, where a greater presence of consolidated academic bodies is observed (the size of the node represents its degree, that is, the number of collaborators). However, an academic body in consolidation that collaborates with external institutions can easily aspire to increase its degree of consolidation, since it is one of the factors considered by PRODEP to grant the distinction.

On the other hand, grouping allows knowing the degree of collaborations of an academic body. This is a useful tool for decision making, even for future work, as a system of recommendations for scientific collaboration.

Regarding association rules, the majority focuses on the less collaborative academic bodies, since most of them are in this category, thus reinforcing the correlation that exists between the number of collaborations and the group to which it belongs, according to the classification. However, a found rule indicates that if an academic body has high level of collaborations, it has a high level of national collaborations. This rule confirms the fact that most of the collaborations are made with Mexican institutions.

An interesting result is that both the UNAM and the CINVESTAV Unit Mérida have a high collaboration with UADY, since both are top ranked institutions with greatest number of collaborations.

From the results obtained in the second experiment, it can be concluded that the second hypothesis has been demonstrated and therefore it can be stated that: The scientific collaborations of the academic bodies with national and foreign institutions can be retrieved and represented with a graph, and the elements of graph theory can describe the collaborations by the level of consolidation of those groups while the association rules allow reporting the behavior in the collaborations.

6. Conclusions

This paper presented a detailed analysis of scientific collaboration of academic bodies in an institution based on the theory of graphs applied in a previous study, which focused solely on identifying the status of collaboration within an institution. In this extended work, collaboration of research groups at the UADY with other institutions, whether Mexican or foreign, was studied and characterized.

Proposed methodology used knowledge extraction techniques in order to generate internal (among members within an institution) and external (for academic bodies) collaboration graphs, and data mining techniques (clustering and association rules) to complement collaboration analysis of research groups with national and foreign institutions.

Proposal was validated through the analysis of the status of collaboration of professors and academic bodies at the UADY. A relevant result was that members of the university are organized in 22 collaboration groups; each professor collaborates with six other professors and have worked 18 times together, on average.

In addition, of the 41 research groups of UADY analyzed, those with the highest degree of consolidation are the ones that have made the most amount of publications and are also the ones that have established the greatest number of collaborations with national and foreign institutions. Consolidated academic bodies have worked with nine external institutions on average, while academic bodies in consolidation have done so with four, on average.

Using data mining allowed finding behavioral patterns of UADYs research groups. These groups were classified in three levels: most collaborative, moderately collaborative and less collaborative, the latter being the one that contains the most of instances. Association rules mainly highlight the dependence of national collaborations in order to classify an academic body as one of the most collaborative, reinforcing a previously found result which indicates that most of the collaborations made by academic bodies are with Mexican institutions.

Acknowledgment

This work has been developed, thanks to the support by Consejo Nacional de Ciencia y Tecnología (CONACYT, Mexico), through the Grant: 853088/630948.

References

1. J. Gómez-Ferri and G. González-Alcaide, Patrones y estrategias en la colaboración científica: la percepción de los investigadores, *Rev. Esp. Doc. Cient.* **41**(1) (2018) 7.
2. D. D. B. Beaver, Reflections on scientific collaboration (and its study): Past, present, and future, *Scientometrics* **52**(3) (2001) 365–377.
3. J. M. Russell, S. Ainsworth, J. A. Del Río, N. Narváez-Berthelemot and H. D. Cortés, Colaboración científica entre países de la región latinoamericana, *Rev. Esp. Doc. Cient.* **30**(2) (2007) 180–198.
4. C. Huamaní and P. Mayta-Tristán, Producción científica peruana en medicina y redes de colaboración, análisis del Science Citation Index 2000–2009, *Rev. Peru. Med. Exp. Salud Pública* **27**(3) (2010) 315–325.
5. J. M. Russell, S. Ainsworth and N. Narváez-Berthelemot, Colaboración científica de la Universidad Nacional Autónoma de México (UNAM) y su política institucional, *Rev. Esp. Doc. Cient.* **29**(1) (2006) 56–73.
6. M. E. Luna-Morales, E. Luna-Morales and S. Luna-Morales, La UADY en la literatura científica registrada en Web of Science y Scopus: 1900–2016, *Educ. Cienc.* **7**(50) (2018) 17–29.
7. J. D. Guerrero-Sosa, V. Menendez-Domínguez, M.-E. Castellanos-Bolaños and L. F. Curi-Quintal, Use of graph theory for the representation of scientific collaboration, in *11th Int. Conf. Computational Collective Intelligence*, ed. N. T. Nguyen (Springer, Hendaye, 2019), pp. 543–554.
8. Elsevier, Author Search API, https://tinyurl.com/yxm7ugcz, accessed on 7 April 2019.

9. Elsevier, Scopus Search API, https://tinyurl.com/y644xqlc, accessed on 7 April 2019.
10. J. Guerrero Sosa, V. H. Menéndez Domínguez and M. E. Castellanos Bolaños, Indicadores de calidad en investigaciones científicas: Antecedentes, *Abstr. Appl.* **19** (2018) 6–24.
11. J. Guerrero Sosa, D. Sánchez Ferriz, V. H. Menéndez Domínguez, M. E. Castellanos Bolaños and J. Gómez Montalvo, Tools for interoperability between repositories of digital resources, in *Proc. INTED 2019*, eds. L. Gómez Chova, A. López Martínez and I. Candel Torres (IATED, Valencia, 2019), pp. 6292–6300.
12. J. Guerrero Sosa, V. H. Menéndez Domínguez and M. E. Castellanos Bolaños, Sistema de índices para valorar la calidad de la producción académica y la investigación, a partir de repositorios digitales y metadatos, in *X Conf. Conjunta Int. sobre Tecnologías y Aprendizaje*, eds. M. E. Prieto-Méndez, S. J. Pech-Campos and A. Francesa-Alfaro (CIATA.org-UCLM, Cartago, 2018), pp. 45–52.
13. X. Ochoa and E. Duval, Quantitative analysis of learning object repositories, *IEEE Trans. Learn. Technol.* **2**(3) (2009) 226–238.
14. L. F. Gómez-Dueñas, Interoperabilidad en los Sistemas de Información Documental (SID): la información debe fluir, *Códice* **3**(1) (2007) 23–39.
15. W. C. G. Peh and K. H. Ng, Basic structure and types of scientific papers, *Singapore Med. J.* **49**(7) (2008) 522–525.
16. CONACYT, Lineamientos Jurídicos de Ciencia Abierta, http://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-juridicos-de-ciencia-abierta/3828-lineamientos-juridicos-de-ciencia-abierta/file, accessed on 30 November 2017.
17. M. E. Falagas, E. I. Pitsouni, G. A. Malietzis and G. Pappas, Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses, *The FASEB J.* **22** (2007) 338–342.
18. N. Bakkalbasi, K. Bauer, J. Glover and L. Wang, Three options for citation tracking: Google Scholar, Scopus and Web of Science, *Biomed. Digital Lib.* **3** (2006) 7.
19. P. Jacso, As we may search — Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases, *Curr. Sci.* **89**(9) (2005) 1537–1547.
20. A.-W. Harzing and S. Alakangas, Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison, *Scientometrics* **106**(2) (2016) 787–804.
21. G. González-Alcaide and J. Gómez-Ferri, La colaboración científica: Principales líneas de investigación y retos de futuro, *Rev. Esp. Doc. Cient.* **37**(4) (2014) 1–15.
22. DGESU, Direccion General Educacion Superior Universitaria Inicio. http://www.dgesu.ses.sep.gob.mx/PRODEP.htm, accessed on 3 January 2020.
23. J. M. Sallán-Leyes, J. B. Fonollosa-Guardiet, V. Fernández-Alarcón and A. Suñé-Torrents, Teoría de grafos, in *Métodos Cuantitativos en Organización Industrial I*, ed. J. M. Sallán-Leyes (Edicions UPC, 2002), pp. 137–172.
24. M. E. J. Newman, The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci.* **98** (2001) 404–409.
25. D. De Filippo, F. Morillo and M. T. Fernández, Indicators of scientific collaboration between CSIC and Latin America through international databases, *Rev. Esp. Doc. Cient.* **31**(1) (2008) 66–84.
26. F. G. Montoya, A. Alcayde, R. Baños and F. Manzano-Agugliaro, A fast method for identifying worldwide scientific collaborations using the Scopus database, *Telemat. Inform.* **35** (2018) 168–185.
27. H. Bozdogan and O. Akbilgic, Social network analysis of scientific collaborations across different subject fields, *Inf. Serv. Use* **33**(3–4) (2013) 219–233.

28. J. I. de Granda-Orive, S. Villanueva-Serrano, R. Aleixandre-Benavent, J. C. Valderrama-Zurián, A. Alonso-Arroyo, F. García Río, C. A. Jiménez Ruiz, S. Solano Reina and G. González Alcaide, Redes de colaboración científica internacional en tabaquismo: análisis de coautorías mediante el Science Citation Index durante el periodo 1999–2003, *Gace. Sanit.* **23**(3) (2009) 222.e34–222.e43.
29. R. Arencibia-Jorge, E. Corera-Alvarez, Z. Chinchilla-Rodríguez and F. de Moya-Anegón, Scientific output of the emerging Cuban biopharmaceutical industry: A scientometric approach, *Scientometrics* **108** (2016) 1621–1636.
30. F. De-Moya Anegón, Z. Chinchilla-Rodríguez, B. Vargas-Quesada and A. González-Molina, Visualización de redes de colaboración internacional, in *Int. Conf. Multidisciplinary Information Sciences and Technologies, InSciT2006* (Mérida, Spain, 2006), pp. 593–597.
31. M. Prieto Méndez, V. H. Menéndez Domínguez and A. Zapata González, Data Mining Learning Objects, in *Handbook of Educational Data Mining*, eds. C. Romero, S. Ventura and M. Pechenizkly (CRC Press Editors, 2010), pp. 315–342.
32. K. Yeung and W. Ruzzo, Principal component analysis for clustering gene expression data, *Bioinformatics* **17**(9) (2001) 763–774.
33. C. Romero and S. Ventura, Educational data mining: A survey from 1995 to 2005, *Expert Syst. Appl.* **33** (2007) 135–146.
34. C. Romero and S. Ventura, Educational data mining: A review of the state of the art, *IEEE Trans. Syst. Man Cybern. Part C, Appl. Rev.* **40**(6) (2010) 601–618.
35. CONACYT, Repositorio Nacional, https://www.repositorionacionalcti.mx/, accessed on 3 January 2020.