Data analysis tools for the study of scientific citations

Morales Alarcón, Cristian; Radicelli García, Ciro; Pomboza Floril, Margarita

Data analysis tools for the study of scientific citations

Herramientas de análisis de datos para el estudio de las citaciones científicas

Cristian Morales Alarcón cristianmorales18m@gmail.com
Agricommerce Cía. Ltda, Ecuador
Ciro Radicelli García cradicelli@unach.edu.ec
Universidad Nacional de Chimborazo, Ecuador
Margarita Pomboza Floril margaritapomboza@unach.edu.ec
Universidad Nacional de Chimborazo, Ecuador

Abstract: This research work made a study of scientific citations, with the aim of identifying aspects that may influence the citationsofahighereducationalinstitution.Weanalyzed 219 records of publications of the National University of Chimborazo and 10304 records of manuscripts from Ecuador. This work had a qualitative approach and a systemic design. As a result, it was found that the impact of scientific publications is reflected by the number of citations that have the documents published by the higher education institutions; in this sense, publications with larger citations are not related to the number of authors or volume of the published magazine, but they are supported by a quality research and correspond mostly to applied sciences.

Keywords: Scientific publications, analysis of data, research methodology, higher education, quality in education.

Resumen: Este trabajo de investigación realizó un estudio de citaciones científicas con el objetivo de identificar aspectos que puedan influir en las citaciones de una institución de educación superior. Se analizaron 219 registros de publicaciones de la Universidad Nacional de Chimborazo y 10304 registros de manuscritos del Ecuador. Este trabajo tuvo un enfoque cualitativo y un diseño sistémico. Como resultado se obtuvo que el impacto de las publicaciones científicas se ve reflejado por el número de citas que tienen los documentos publicados por las instituciones de educación superior; en este sentido las publicaciones con mayores citas no se encuentran relacionadas al número de autores ni alvolumen delarevistapublicadasinoaunainvestigación decalidad y corresponden en su mayoría a ciencias aplicadas.

Palabras clave: publicacionescientíficas, análisisdedatos, metodología de investigación, educación superior, calidad en la educación.

Introduction

At present,the explosion,assimilation and intensive use of knowledge has led to what has been called the knowledge society,in which the management of information,documentation and

knowledge are emerging as a strategic component in the Institutions of Higher Education (HEI).

In this sense, in the HEI of Ecuador from the year 2003, self-evaluation, evaluation and institutional recategorization processes are executed, directed by the Council for Evaluation, Accreditation and Quality Assurance ofHigher Education (CEAACES forits acronymin

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

1

Spanish), now Higher Education Quality Assurance Council (CACES for its acronym in Spanish), which have led to quality measurements in different areas among which are considered those related to the scientific production of knowledge of both teachers and students belonging to research groups. Here it is specifically analyzed the number of publications in journals with high global impact,the production of regional impact and the publication of books and book chapters (CEAACES, 2018), which is contemplated in the Institutional Evaluation Model of

Universities and Polytechnic Schools.

High-impact scientific publications refer to the quality indicator (Radicelli et al.,2018),which in turn is measured by the number of publications in indexed journals of the ISI Web of

Knowledge and SCImago scientific databases Journal Rank (Ganga,Paredes,& Pedraja,2015).

Cristian Morales Alarcón,Ciro Radicelli García,Margarita Pomboza Floril

In addition,the number of citations that scientific articles published in a specific journal have at a given time is considered, so in order to measure said quality, there are entities such as the Institute for Scientific Information (ISI),attached to the company Thomson Reuters,that uses the Journal Citation Report (JCR),which is nothing more than a database that presents detailed figures about publications and their citations. However,in addition to the JCR index, there are other databases that measure the quality of published documents, such as Scopus, which is attached to Elsevier, and which is mainly run by the SCImago research group of Spain (Valderrama,2012),which use the SCImago Journal and Country Rank (SJR) and the

SCImago Institutions Ranking (SIR) as indicators.

The volumes of data that are stored in databases, allow a complete processing of the information, for which it works in phases such as pre-processing, data mining itself and the post-processing of said information. In this sense, to facilitate the retrieval and delivery of information carried outby personnelwhoworkwith largevolumesofdata,such aslibrarians, the horizons have been opened towards otherprofessions that are called tocooperate,thus we now have designers systems, data providers, publishers, vendors, archivists, engineers and specialists in electronic text encoding, among others; whose opinions and experiences will allow the development of adequate interfaces to facilitate the location, manipulation,

retrieval and use of digital information.

In reference to the aforementioned,Valcárcel (2004) mentions that the "minería de datos" (or commonly called Data Mining), refers to the process of extracting knowledge from databases, with the aim of discovering anomalous and/or interesting situations, as well as trends, patterns and sequences in the data. For their part, Molina and Ribiero (2001) clarify that mining is the integration of a set of areas whose purpose is to identify knowledge obtained from databases that provide

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

2

a bias towards decision-making. Likewise, Molina (2002) indicates that data mining is a non-trivial process of valid,novel,potentially useful and understandable identification of

understandable patterns that are hidden in the data.

Thus, for this purpose, new tools have been created in order to facilitate access to the accumulation of information that is generated daily, one of the most used being text mining, which offers the possibility of exploring large amounts of non-organized texts, in addition to establishing patterns and extracting useful knowledge. Text mining then refers to the examination of a collection of documents in order to discover information that is not explicit

in the analyzed text (Nasukawa,Kawano,& Arimura,2001).

The importance of text mining lies in the effectiveness of its predictive models,which have saved time and money; as well as the improvement of the capacity to respond to the needs of the interested parties, it is thus that the use of computer tools used for the discovery and

processing of information will improve the knowledge management process.

Data analysis tools for the study of scientific citations

It is also important to mention that information constitutes, under current conditions, an economic resource highly valued not only for its intrinsic properties, but also because it allows the improvement of the use of the rest of the resources of the organizations, therefore the search ofregularitiesorpatternsfound in atext,based on machinelearningtechniques,areof great help for the discovery of knowledge that does not exist in the text,but that arises when

relating the content of several texts.

All these applications are perfectly transferable, for example, to the management of information that occurs within the libraries of the HEI, which are called upon to resize the function of the entity,both inside and outside of it. Thus,there are numerous and multiple approaches to a definition of the text mining knowledge management tool,where it is intended to use machine learning techniques, considered one of the many branches of computational linguistics, in order to find the patterns previously mentioned generally in unstructured texts such as those commonly used by organizations, such as reports, emails, meeting minutes, among others,

that is,information stored in unstructured textual form.

This work focused on analyzing scientific publications and their respective citations,considering the information registered in the Scopus database,both in the case of UNACH,and for scientific publications made in Ecuador, this considering a period of approximately 6 years (2013 to 2019), for which data analytics and text mining tools were used in order to identify aspects

that may influence citations from a HEI.

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

3

Materials and Methods

This work had a qualitative approach due to the fact that significant research areas or topics weredetermined,wheretodiscover,refineandanswertheresearchquestions,thedatacolle and analysis was first carried out. This work followed a systemic design which was also based on the CRISP-DM data mining methodology, since it highlights the use of steps which are followed in an order until the desired end is reached. In the context of this research, the

methodological process detailed below was followed:

(i) data collection and analysis: For data collection,the Scopus scientific database was used to obtain UNACH publications, as well as the scientific publications made in Ecuador, the aforementioned was done using the period from 2013 to 2019. For this purpose, only journal

articles and book chapters were considered,below,the search strings used are shown:

AFFILORG(Chimborazo) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"ip") OR LIMIT-TO (DOCTYPE, "ch")) AND (LIMIT-TO (AF-ID, "Universidad Nacional de Chimborazo" 60108604) OR LIMIT-TO (AF-ID, "National University of Chimborazo" 114160995) OR LIMIT-TO (AF-ID, "National University of Chimborazo" 118104741) OR LIMIT-TO (AF-ID, "Universidad

Nacional de Chimborazo" 119728963)).

Cristian Morales Alarcón,Ciro Radicelli García,Margarita Pomboza Floril

AFFILCOUNTRY ("Ecuador") AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "ch") OR LIMIT-TO (DOCTYPE,"ip")) AND (LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR,

2013)) AND (EXCLUDE (PUBYEAR,2020)).

The analytical and synthetic methods were also used, because the study of the information provided will allow, through the analysis of the information,to synthesize the behavior of the

study phenomenon.

(ii) Data preparation: It was carried out through an exploratory investigation,because specific aspects related to the citations were analyzed both in UNACH in particular, as well as in Ecuador in general. In addition,the necessary fields for the respective analyzes were chosen

and aspects of data quality were corrected, in terms of incomplete, missing, or erroneous data.

(iii) Preliminary descriptive analysis of the data: Descriptive research was used in order to represent the data found and observe their behavior through tables and graphs. A deductive method was also applied, because after a stage of repeated observation, analysis and classification of the

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

4

particular facts,generic computational models were obtained for future application. On the other hand,to perform the analysis of the data of the scientific publications of UNACH and Ecuador, the QlikView software was used, which is a tool that allows advanced

data visualizations,but also the mining text tools Andatos and WordStat.

(iv) Correlational analysis of the quantitative variables: For which the databases returned by

Scopus were examined.

(v) Text mining analysis: Corresponding to the data related to the citations of UNACH in

particular and of Ecuador in general.

(vi) Scientific induction: Due to the fact that when analyzing the data obtained in a particular way, methodological aspects were derived in order to increase the number of citations. Explanatory research was also used for this work,since it is intended to find the causes that originate the

phenomenon.

Results

Once the data had been collected from the Scopus scientific database,the following phases were developed: (i) transformation, where the filter and copy of the fields is carried out; (ii) cleaning,where erroneous values are eliminated and subsequently replaced;and (iii) generation,

where new variables useful for the study are generated.

Data analysis tools for the study of scientific citations

Descriptive analysis

With the databases ready forthe study,a descriptive analysis was carried out usingthe QlikView software,using a total of 219 records corresponding to the scientific manuscripts published by UNACH research staff in the period 2013 to 2019, obtaining a total of 217 journal articles
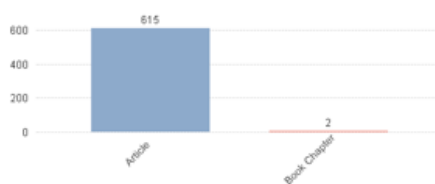
and only two book chapters.

Figure 1
Número de citaciones de la UNACH en SCOPUS
por tipo de documento. Fuente: Elaboración propia.

citationsoftheUniversity,in
contrasttothechaptersofbooksthatcorrespond only to0.32%.

Figure 1 . Number of citations of UNACH in Scopus by type of document. Source: author's own elaboration.

Figure 2 shows the history of the scientific publications produced by UNACH,where a growing trend is observed in the number of works, its

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

5

largest year of production being 2018 with 70 publications, in contrast to 2013 in which there was only one published manuscript. It should
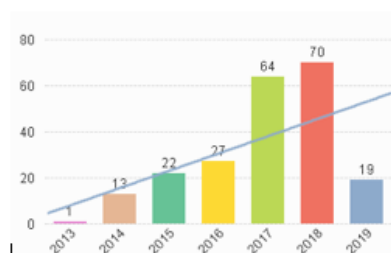be noted that so far in 2019 there are already 19 papers belonging to Scopus.

Figure 2
Histórico del número de publicaciones científicas de la UNACH en SCOPUSFuente: Elaboración propia.

Cristian Morales Alarcón,Ciro Radicelli García,Margarita Pomboza Floril

In contrast to the number of scientific publications produced by UNACH,the historical number of citations does not correspond to a growing trend. Figure 3 shows that for 2015 the number of citations reached 327,which represents the highest value in the graph; in contrast to 2013, which only had 17 citations. When analyzing this graph, it can also be inferred that the number of citations depends on the quality of the publications and not on the number of effective publications. Thus, for example, for the year 2014, with only 13 publications, 72 citations were
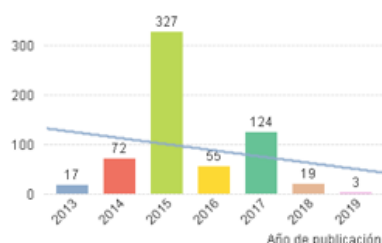obtained; while,with the 70 publications of 2018,there were only 19 citations.

Figure 3
Histórico del número de citaciones de las publicaciones científicas de la UNACHFuente: Elaboración propia.

Of a total of 219 scientific publications of UNACH, 58.9 % have not been cited even once, that is, more than half of the manuscripts of the institution indexed in Scopus have not generated the expected expectation in the research community worldwide; on the other hand the remaining 41.10 %, corresponding to 90 publications have been cited at least once.
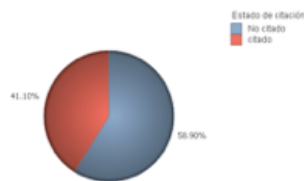
This is represented in Figure 4.

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

6

Figure 4

Porcentaje de publicaciones de la UNACH que han obtenido citaciónFuente: Elaboración propia

Data analysis tools for the study of scientific citations

Correlational analysis

Next, a correlational analysis of the quantitative variables was carried out through the RapidMiner software,which could influence the number of citations of the documents,firstly,the number of citations vs. the number of authors was considered (Figure 5), in order to find out if the self-citations could have influenced the number of citations; however, there is no relevant correlation between these two variables. Two atypical values have also been found,the first referring to the publications that stand out in relation to the rest of the manuscripts, where the correlation has been verified with the data of all the scientific publications of Ecuador (10304 records). And the second considering that the number of authors with citations of
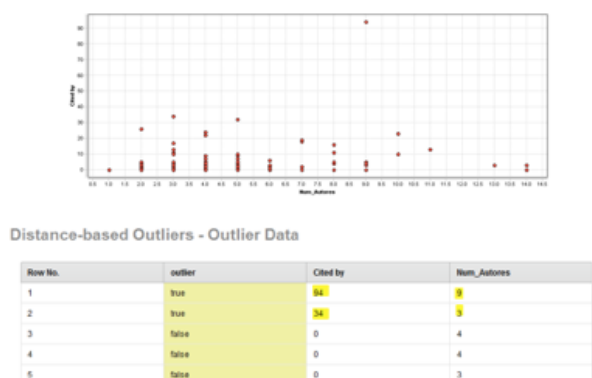
zero amounts to a figure of up to 2.314.

Figure 5

Correlación entre las variables, número de citaciones
vs número de autores. Fuente: Elaboración propia

In addition, a correlation was made between the number of citations vs. the number of the volume of the journal in which the publication was made (Figure 6). In this case, it has been considered that the second variable can affect the visibility of research papers as well as the number of citations. This correlation has been verified with the data of all the scientific publications of Ecuador, in which a relevant relationship has not been observed, due to the

fact that in a large number of volumes (9386) there are zero citations.

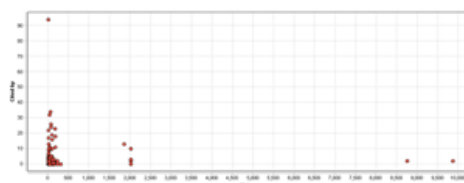Cristian Morales Alarcón,Ciro Radicelli García,Margarita Pomboza Floril

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

7

Figure 6
Correlación entre las variables, número de citaciones vs número del volumen de la revista.

Text mining analysis

Through the WordStat tool,a word cloud was generated in which it is observed that the topics on which the academic staff of UNACH have published most frequently are education and

health,followed by studies related to physical activity and computer systems.

In Figure 7 it can be seen that the topic of education tops the list with the highest number of publications, in addition to having a high number of papers that do not obtain any citation (23 manuscripts). It should be noted that only two publications have a high number of citations, 11 and 13 respectively,this out of a total of 35 citations,that is,4 % of the total scientific papers have generated 69 % of citations. Similarly, of the 219 UNACH publications analyzed,only 19 papers present citations with an index greater than 10, which represents 416 citations out of

a total of 617,that is,9 % of publications have generated 67% of citations.

	0	1	2	3	4	5	6	7	9	10	11	13	16	17	18	19	22	23	24	26	32	34	94
EDUCATION	23	1	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
HEALTH	11	0	1	2	1	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
PHYSICAL	4	1	2	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
SYSTEMS	6	1	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	

Figure 7
CrossTab de los tópicos con mayor ocurrencia de publicaciones y número de citas

Figure 8 shows the correlation between the publications (cases) and the number of citations, here values can be seen that stand out from the rest of the topics that one might think appear more frequently. In this sense, Table 1 shows the keywords that have the highest citation index and that also have to do with applied sciences. Table 2 shows the scientific publications with the highest number of citations in UNACH, while Table 3 shows the publications with the

highest number of citations in Ecuador.

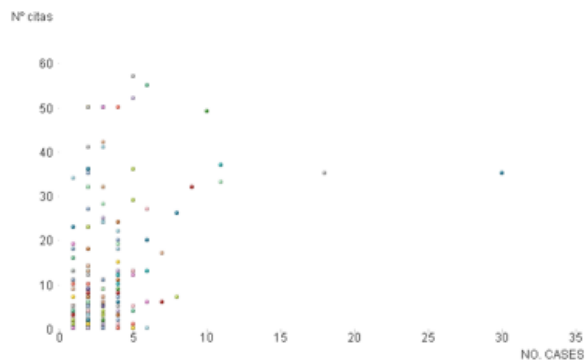Data analysis tools for the study of scientific citations

Figure 8

Tópicos encontrados con mayor citación vs el número de aparición en los casos

Table 1

Ranking of words by number of citations

Tópico	N° citas
Properties	57
Oxidative	55
Mechanical	52
Fabrics	50
Fuel	50
Microbial	50
Testing	50
Textiles	50
Natural	49
Cell	42

Topic Properties Oxidative Mechanical Fabrics
Fuel Microbial Testing Textiles Natural
Cell
Citations N° 57 55 52 50 50 50 50 50 49
42
Source: author's own elaboration.
Cristian Morales Alarcón,Ciro Radicelli García,Margarita Pomboza
Floril

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

9

Table 2

Ranking of articles in the publications of UNCH

N°	Título de la publicación	Citado por
1	Strawberry as a health promoter: An evidence based review	94
2	Development of durable cementitious composites using sisal and flax fabrics for reinforcement of masonry structures	34
3	Physical and oxidative stability of whey protein oil-in-water emulsions produced by conventional and ultra high-pressure homogenization: Effects of pressure and protein concentration on emulsion characteristics	32
4	Effects of fabric parameters on the tensile behaviour of sustainable cementitious composites	26
5	Flax and polyparaphenylene benzobisoxazole cementitious composites for the strengthening of masonry elements subjected to eccentric loading	24
6	Municipal waste liquor treatment via bioelectrochemical and fermentation (H2 + CH4) processes: Assessment of various technological sequences	23
7	Relative importance of phenotypic trait matching and species' abundances in determining plant - Avian seed dispersal interactions in a small insular community	22
8	Lipophilic antioxidants prevent lipopolysaccharide-induced mitochondrial dysfunction through mitochondrial biogenesis improvement	19
10	Single chamber microbial fuel cell (SCMFC) with a cathodic microalgal biofilm: A preliminary assessment of the generation of bioelectricity and biodegradation of real dye textile wastewater	18
11	Uncontacted Waorani in the Yasuní Biosphere Reserve: Geographical Validation of the Zona Intangible Tagaeri Taromenane (ZITT)	17

N° 1
2
3
4
5
6
7
8
10
11

Strawberry as a health promoter: An evidence-based review

Development of durable cementitious composites using sisal and flax fabrics for reinforcement of masonry structures

Physicaland oxidativestability ofwhey protein oil-in-wateremulsionsproduced by conventional and ultra high-pressure homogenization: Effects of pressure and protein concentration on emulsion characteristics

Effects of fabric parameters on the tensile behaviour of sustainable cementitious composites

Flax and polyparaphenylene benzobisoxazole cementitious composites for the strengthening of masonry elements subjected to eccentric loading

Municipal waste liquor treatment via bioelectrochemical and fermentation (H2 + CH4) processes: Assessment of various technological sequences

Relative importance of phenotypic trait matching and species' abundances in determining plant–Avian seed dispersal interactions in a small insular community

Lipophilic antioxidants prevent lipopolysaccharide-induced mitochondrial dysfunction through mitochondrial biogenesis improvement

Single chamber microbial fuel cell (SCMFC) with a cathodic microalgal biofilm: A preliminary assessment ofthe generation of bioelectricity and biodegradation of real dye textile wastewater

Uncontacted Waoraniin theYasuníBiosphereReserve:GeographicalValidation of the Zona Intangible Tagaeri Taromenane (ZITT)

Source: author's own elaboration.

Number of citations 94

34

32

26

24

23

22

19

18

17

Table 3

Ranking of articles in publications from all over Ecuador

Nº	Título de la publicación	Nº de citas
1	The International Classification of Headache Disorders, 3rd edition (beta version)	3595
2	Trends in adult body-mass index in 200 countries from 1975 to 2014: A pooled analysis of 1698 population-based measurement studies with 19.2 million participants	1100
3	Global surveillance of cancer survival 1995-2009: Analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2)	841
4	Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016	630
5	Global conservation outcomes depend on marine protected areas with five key features	545
6	Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults	422
7	Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016	407
8	Hyperdominance in the Amazonian tree flora	407
9	Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: A cohort study	382
10	Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016	354

Nº

1

2

Article title

The International Classification of Headache Disorders,3rd edition (beta version)

Trends in adult body-mass index in 200 countries from 1975 to 2014: A pooled analysis of1698 population-based measurement studies with 19.2million participants

Number of citations

3595

1100

Data analysis tools for the study of scientific citations

Nº

3

4

5

6

7

8

9

10

Global surveillance of cancer survival 1995-2009: Analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2)

Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016

Global conservation outcomes depend on marine protected areas with five key features

Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children,adolescents,and adults

Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational,and metabolic risks orclusters ofrisks,1990-2016: A systematic analysis for the Global Burden of Disease Study 2016

Hyperdominance in the Amazonian tree flora

Uptake of pre-exposure prophylaxis, sexual practices, and HIV incidence in men and transgender women who have sex with men: A cohort study

Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016

Source: author's own elaboration.

Number of citations

841

630

545

422

407

407

382

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

12

354

Discussion

Data mining is an exploitation mechanism,consisting of the search for valuable information in largevolumesofdata,which isknown asBigData;in thissenseitsalgorithmsaimtoextract valuable information for the making of decisions (Botta-Ferret, & Cabrera-Gato, 2007), and if it is also considered that organizations have mostly unstructured data, rather than structured data,text mining is justified due to the advantages it can provide when it comes to improve

the productivity of the data obtained (Pérez,& Cardoso,2010).

Despite the fact that data mining has been used in eminently technical or business areas, accordingto the authors of this research,it is undoubted that the future ofeducation in general will have to use,and is in fact using,methods of data analysis to improve the eficiency and effectiveness of its processes. And it is precisely that one of the most important and used methods for this purpose is data mining, with the aim of making sense of the large amount

Cristian Morales Alarcón,Ciro Radicelli García,Margarita Pomboza Floril

of information that is currently stored (Riquelme, Ruiz, & Gilbert, 2006). In this sense, initially data mining was used in fields such as computing (Jaramillo,Cardona,& Fernández,2015), health (González,& Pérez,2013),business branches (Gordillo,Martínez,& Sthepens,2012), construction (Castro et al.,2014),and even on issues as specific as the detection and prevention of money laundering and the financing of terrorism. But today its use in the educational field has become very fashionable,and as they describe it Márquez,Romero and Ventura (2012),

it is a "very promising solution [...] in education" (p. 45).

When speaking then of data mining in education, there are various studies, among which are from the investigation of learning patterns (Ballesteros,Sánchez,& García,2013),going through the extraction ofstudent dropout profiles (Timarán,Calderón,&Jiménez,2013),until reachingsuch particulartopicsastheuseofdataminingfortheenrollmentprocessin private

higher education institutions (Estrada et al.,2016).

In the correlation analysis carried out to the number of citations with respect to the number of authors and the volume of the journals, although there is no relevant relationship, said correlation allowed to discover publications that stood out with a high number of citations, which leads to think that certain areas of study and publications are more sought after for

the purposes of citation.

In addition,thetextminingcarried outon thebasisofkeywordsofUNACHpublications,allowed to corroborate that there were factors that determine the number of citations; in this

sense it could be observed that several areas had a high number of citations,especially those that

belonged to applied sciences.

Conclusions

Scientific publications are part of the educational and especially university work, where the impact of said publications is reflected by the number of citations that scientific documents published in HEI have, thus the increase in publications in UNACH and in Ecuador in general is considerable; however this increase alone is not enough,and should be accompanied by the number of citations,that is,the published works generate expectations in the world research community. In this analysis, it was possible to observe research works that have never been cited despite the fact that they were published a few years ago. Thus, the publications with the highest citations are not related to the number of authors or the volume of the published journal, but are supported by quality research of several years and correspond mostly to

applied sciences.

In addition, for an academic work to generate the desired interest, HEI in Ecuador should

implement strategies that allow the development of research supported by a well-defined

Data analysis tools for the study of scientific citations

structure, which in turn facilitates the creation of projects focused on solving the main problems

of the society and the generation of base knowledge for future research.

References

1. Ballesteros, A., Sánchez, D., & García, R. (2013). Minería de datos educativa: una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. *Revista Latinoamericana*

2. Botta-Ferret, E., & Cabrera-Gato, J. (2007). Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital.

3. Castro, A. et al. (2014). Use of Data Mining in Managing Geographical Information. *Información*

4. CEAACES. (2018). *Modelo de evaluación institucional de universidades y escuelas politécnicas 2018.* Recovered from http://uisrael. edu.ec/ wp-content/uploads/2019/03/Modelo-evaluacion-preliminar-universidades-escuelas-

5. Estrada, R. et al. (2016). Contributions to the Enrollment Process with Data Mining in Private Higher Education Institutions. *Revista Electrónica*

6. Ganga,F.,Paredes,L.,&Pedraja,L.(2015).The importance of academic publications: Some problems and recommendations to keep in mind. *Idesia,*

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

14

7. González,L.,& Pérez,Y. (2013). Spatial data mining and its application in health and epidemiology studies. *Revista Cubana de Información en Ciencias de*

8. Gordillo, J., Martínez, E., & Stephens, C. (2012). Inferring Market Strategies: Applying Data-Mining to Analysis of Financial Markets. *Computación y*

9. Jaramillo,S.,Cardona,S.,& Fernández,A. (2015). Data Mining Streams ofSocial Networks,A Tool to Improve the Library Services. *Información, Cultura y Sociedad,*

10. Márquez, C., Romero, C., & Ventura, S. (2012). Predicting of school failure using data mining techniques.

11. Molina,L.,&Ribiero,S.(2001).Descubrimientoconocimiento para el mejoramiento bovino usando técnicas de data mining.In IVCongresoCatalán de Inteligencia Artificial, Societat Catalana de Comunicació,

12. Molina, L. (2002). *Data mining: torturando a los datos hasta que confiesen.* Recovered from https:// www.uoc.edu/web/esp/art/uoc/molina1102/

13. Nasukawa, T., Kawano, H., & Arimura, H. (2001). Base technology for text mining. *Journal of Japanese Society for*

14. Pérez, M., & Cardoso, C. (2010). Minería de texto para la categorización automática de documentos.

15. Radicelli, C. et al. (2018). Análisis del ranking SCImago de universidades ecuatorianas: el caso de

16. Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de datos: conceptos y tendencias. *Revista Iberoamericana*

17. Timarán, R., Calderón, A., & Hidalgo, A. (2017). Aplicación de los árboles de decisión en la identificación de patrones de lesiones fatales por causa externa en el municipio de Pasto, Colombia. *Universidad*

18. Valcárcel,V. (2004). *Data Mining and Knowledge Discovery. Revista de la Facultad de Ingeniería Industrial,*

PDF generado a partir de XML-JATS4R por Redalyc
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

15