ANÁLISIS DE REGINA: LA RED SOCIAL DE COLABORACIÓN FORMADO POR INVESTIGADORES PERUANOS

Cesar Daniel Rodríguez Castro, Cesar Risco Chávez

Abstract

En este artículo se expone una investigación basada en el análisis de los datos de investigadores registrados en la plataforma REGINA. Se aplicará técnicas de procesamiento y análisis en Python, para la generación de información relevante sobre el comportamiento de los miembros en general.

I. Situación actual

La red de investigadores peruanos REGINA posee gran cantidad de datos sobre sus perfiles y trabajos de investigación realizados. Sin embargo, estos datos no se están aprovechando para generar conocimientos aplicando analítica sobre ellos.

Los datos solo permanecen registrados y no son capaces de crear información relevante para los usuarios (investigadores) con el fin de tener una vista panorámica, en sentido informativo, del comportamiento de los miembros de la red social.

II. Introducción

La plataforma peruana de REGINA es un importante conjunto de datos académicos de modo currículum que posee como registros todas las actividades académicas de los investigadores peruanos asociados. La información académica en este conjunto de datos se utiliza para evaluar, analizar y documentar la producción científica de los investigadores.

Sin embargo, no se ha analizado la información sobre los investigadores en la forma de coautorías. En este artículo, identificamos, extraemos y analizamos la información de todos los miembros investigadores registrados en la red social académica peruana REGINA utilizando Python.

Para lograr el objetivo, se debe (a) crear un algoritmo para extraer y dar formato a la información de los perfiles de cada investigador en la plataforma, (b) crear otro algoritmo para realizar un pre procesamiento a los datos recolectados, y (c) generar el dashboard que refleje el análisis de los datos.

Este estudio analizó datos de todos los perfiles de investigadores registrados. Por lo tanto, analizamos información de aproximadamente 1718 investigadores asociados a 18833 trabajos de investigación registrados en la plataforma REGINA, específicamente en la tabla de trabajos de investigación SCOPUS.

III. Estado del Arte

Actualmente, la tecnología es el principal aliado para el procesamiento y análisis de datos. Una de las áreas en la cual se está tomando mucho énfasis es en el desarrollo de algoritmos o modelos capaces de realizar el análisis de grandes cantidades de datos. Es por ello que durante todo este trabajo de investigación se aplicarán algoritmos para obtener, procesar y analizar datos.

Así, una red se define como la interconexión de nodos por medio de aristas o links. Esta red es capaz de describir el comportamiento de sus componentes. Y mediante el uso de distintos recursos se puede generar mucha más información relevante sobre los mismos.

Así, las redes pueden ser muy complejas de acuerdo a su naturaleza. Por ejemplo, la red tecnológica interna de una organización en su mayoría se compone de computadoras, routers, impresoras, servidores, switches y el cable de red que los une. Esta red contiene

nodos limitados y son más simples de administrar para generar información a partir de ellos.

Por otro lado, una red de personas en una plataforma como Facebook, donde tiene un número creciente de nodos, es mucho más compleja, sin contar la cantidad de datos (amistades e interacciones en general) que genera cada usuario/nodo dentro de la misma. Por supuesto, existen redes mucho más complejas dentro del mundo biológico, pero no es tema que trata de abordar este trabajo.

De este modo, la plataforma de investigadores REGINA es una red de investigadores peruanos donde tienen un registro de sus trabajos realizados. Este tiene cierta similitud a una red social tradicional, pero las interacciones entre investigadores se ven plasmada en los trabajos realizados por coautorías.

Así, se realizará una revisión rápida de los trabajos relacionados sobre redes de investigadores/trabajos en otros países.

Brazilian Bibliometric Coauthorship Networks

Pascual J., Digiampietri L., Martins F. & Roberto C. (2015) realizaron un trabajo muy similar sobre la plataforma brasileña Lattes. Esta contiene información académica de investigadores brasileños y los trabajos realizados por cada uno de ellos.

El objetivo que tuvieron fue de identificar las redes de coautoría a través de listas de publicaciones obtenidas de la plataforma usando las propiedades de grafos. Así mismo, caracterizar las redes de coautoría para comprender la estructura y dinámica de la red generada. Esto con el fin de comprender el comportamiento de los investigadores brasileños en cada área de conocimientos en la cual publican sus trabajos.

Así, el trabajo que ellos realizaron fue enfocado en la interacción entre investigadores y las veces que participaron en la realización de un trabajo académico.

Los resultados que ellos obtuvieron fueron los siguientes:

- Cada área de conocimiento mostro un comportamiento diferente entre los investigadores.
- Los investigadores muestran un carácter interdisciplinario.
- El área lingüística, letras y artes presento una alta transitividad.
- Las áreas de ciencias y humanidades tienes comportamientos casi constantes.
- Por otro lado, ciencias agrícolas, humanidades, ciencias de la salud, lingüística, letras y artes tienen comportamientos oscilatorios que puedan indicar una dependencia de otras áreas.
- Las ciencias biológicas es la única área que tiene más colaboraciones interdisciplinares.
- Entre otros.

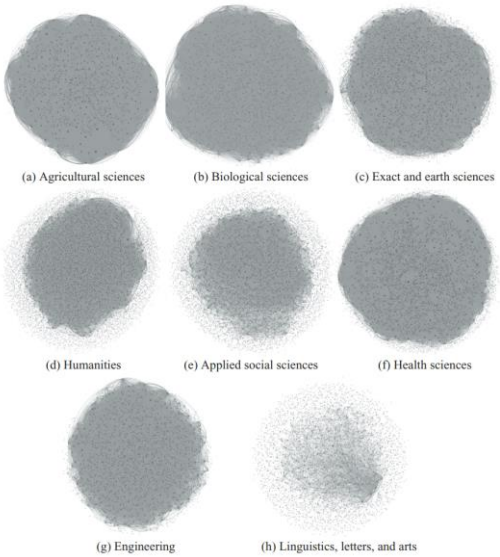


Figura 1 Redes de coautoría pertenecientes a investigadores asociados a cada área de conocimiento principal. Las redes se crearon considerando publicaciones producidas entre 1990 y 2010. No se muestran nodos aislados.

Según la Figura 1 se puede observar el comportamiento de los investigadores de cada área de conocimiento. Las redes muestran alto grado de concentración en los nodos para cada área de manera independiente. Sin embargo, esto no se cumple en tal magnitud en las áreas de Lingüística, letras y artes.

Identificación del grafo de genealogía académica de investigadores: Un enfoque basado en la Plataforma Lattes

Damaceno R., Rossi1 L., Mena J. (2017) proponen en su trabajo de investigación un algoritmo capaz de identificar de manera automática grafos genealógicos académicos en los datos registrados en la plataforma Lattes. Este algoritmo también puede identificar la importancia jerárquica de los nodos, lo que facilitaría el análisis genealógico.

Tabela 1. Estatísticas referentes aos dados considerados e ao grafo resultante

Currículos disponíveis na Plataforma Lattes	5 102 445	100, 00%
Currículos considerados neste trabalho (doutores)	272 165	5, 33%
Vértices no grafo obtido	334 952	–
Arestas no grafo obtido	300 696	–
Componentes conexas no grafo obtido	37 444	–
Vértices na maior componente conexa	178 767	53, 4%
Arestas na maior componente conexa	181 731	60, 4%
Vértices isolados (doutores sem orientador nem orientados)	4 842	1, 45%
Densidade do grafo	$2{,}68 \times 10^{-6}$	–
Grau médio do grafo	1, 79	–
Grau máximo do grafo	130	–

Los datos con los que trabajaron fue de 272 mil prospectos de investigadores Ph.D. registrados en la plataforma.

Durante el análisis de obtuvieron los siguientes resultados:

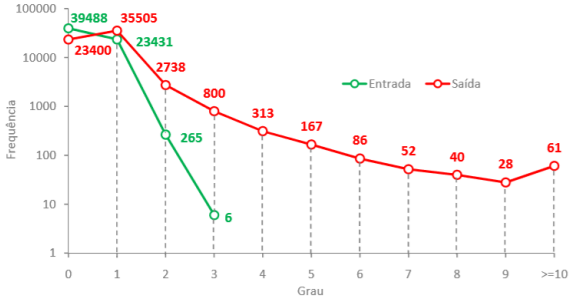


Figura 2 Cantidad de vértices por grado de entrada y salida resultantes

Existen vértices de salida con mayores grados a comparación de las entrantes. Sin embargo, la mayoría de cada uno se encuentra alrededor de los 3k y 4k de vértices en grado 0 y 1.

	Fecundidade			Descendência			Índice genealógico			Acadêmicos
	média	mediana	máximo	média	mediana	máximo	média	mediana	máximo	considerados
AGR	11,81	6	388	7,07	5	63	0,44	0	8	14,37%
BIO	14,58	5	1747	6,59	4	85	0,53	0	11	16,57%
ENG	11,07	5	474	6,62	4	129	0,43	0	9	17,87%
EXA	9,69	4	865	5,47	3	98	0,42	0	11	17,12%
HUM	13,02	5	989	6,88	4	129	0,48	0	12	14,85%
IND	7,52	1	1536	1,30	1	48	0,35	0	12	49,98%
LIN	12,84	5	975	6,71	5	119	0,47	0	12	14,16%
OUT	4,09	2	54	2,47	1	31	0,18	0	4	8,58%
SAU	11,54	4	1003	5,98	4	70	0,48	0	10	15,75%
SOC	10,75	4	877	5,70	3	93	0,40	0	8	13,50%

Figura 3 Metricas estratificadas por cada Area

Según la figura 3, los trabajos en el área "Indefinida" fueron los que más publicaciones se identificó por el algoritmo. Así mismo, ENG (INGENIERIA), EXA (CIENCIAS EXACTAS DE LA TIERRA) y SAU (CIENCIAS DE LA SALUD) fueron las 3 siguientes con mayor proporción de trabajos después de las áreas indefinidas.

Las áreas indefinidas fueron imposibles de clasificar por el algoritmo. Esto puede deberse al estado de los datos, los cuales pudieron tener campos vacíos o nulos.

IV. Metodologia

a. Obtener los datos

Se creó un algoritmo para extraer los datos de la web de REGINA. Los detalles de este procedimiento se encuentran en el anexo del código en Python.

	Titulo	Primer autor	Año de Producción	Titulo de la fuente	Fuente	Author
0	Identification of leptospira spp serogroups in...	Yaquelyn Better B.	2018	Revista de Investigaciones Veterinarias del Peru	NaN	10006
1	Draft genome sequences of enterohemorrhagic an...	Maturrano L.	2018	Genome Announcements	NaN	10006
2	Identification of leptospira spp serogroups in...	Yaquelyn Better B.	2018	Revista de Investigaciones Veterinarias del Peru	NaN	10006
3	Draft genome sequences of enterohemorrhagic an...	Maturrano L.	2018	Genome Announcements	NaN	10006
4	Standardization of a real time PCR taqman assa...	André Sedano S.	2016	Revista de Investigaciones Veterinarias del Peru	NaN	10006

Figura 3 Muestra de datos colectados (LINKS)

	AUTOR	Primer autor
0	10006	Yaquelyn Better B.
1	10015	Maturrano L.
2	10019	AndrÃ© Sedano S.
3	10021	Roberto LÃ¡zaro B.
4	10023	Siuce J.

Figura 4 Muestra de datos colectados (NODOS)

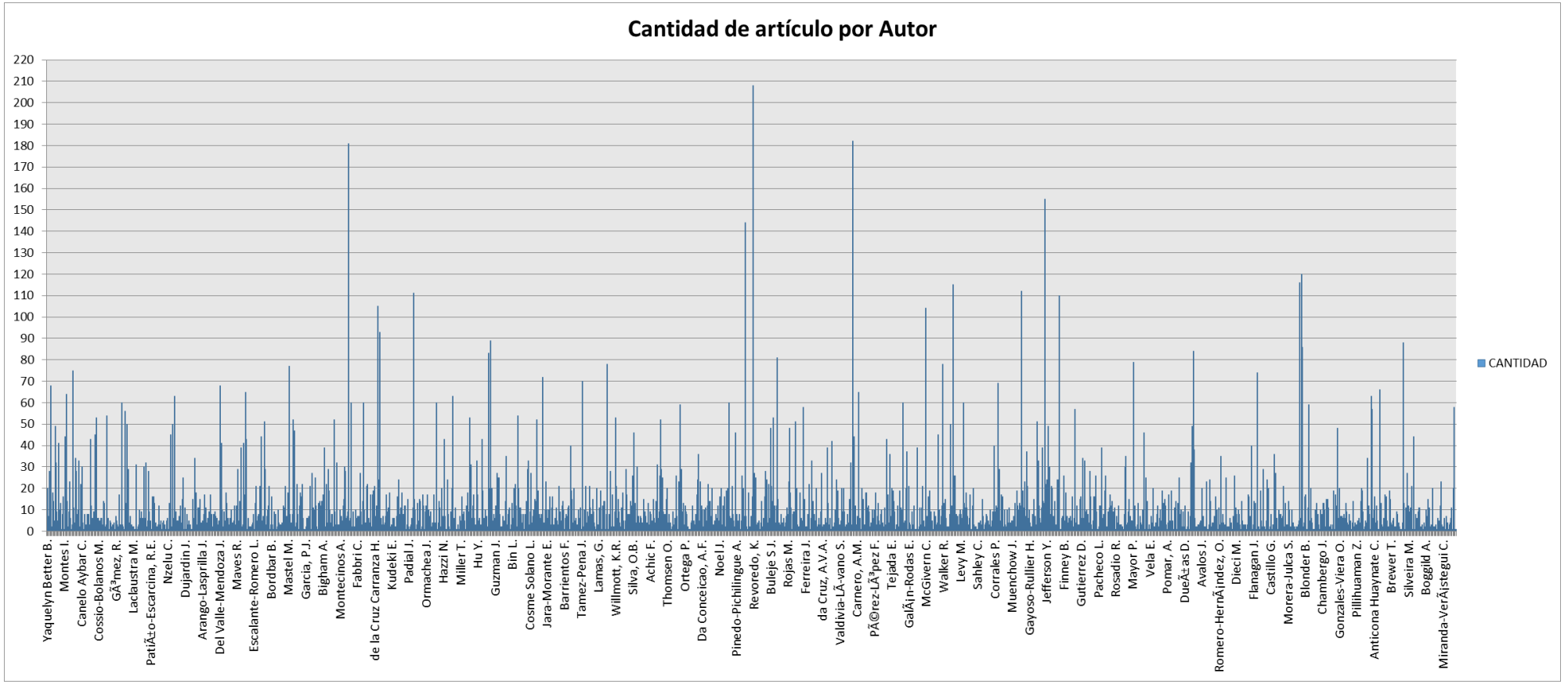
b. Pre-procesamiento de los datos

Durante este proceso se realizó una limpieza de datos respecto a caracteres que podían afectar el análisis posterior. Así mismo, hubo la necesidad de eliminar registros que contenían campos vacíos y redundantes. Los detalles de este procedimiento se encuentran en el anexo del código en Python.

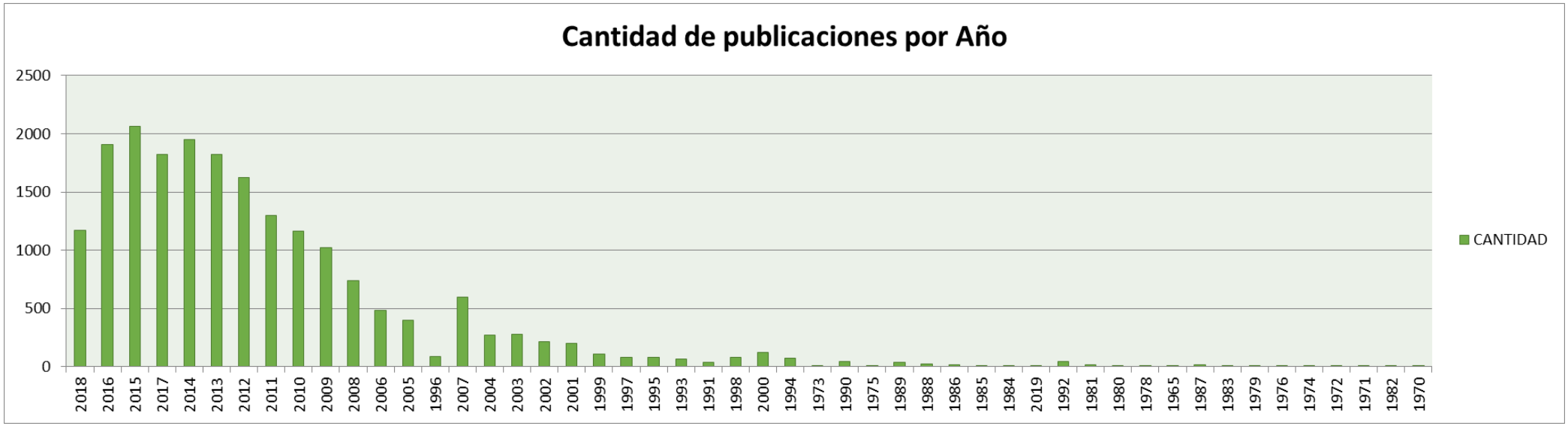
c. Análisis de los datos

Los datos ya listos para realizar el análisis fueron sometidos a diversas consultas con el fin de obtener información adicional en relación a todos los investigadores, publicaciones, fuentes, etc.

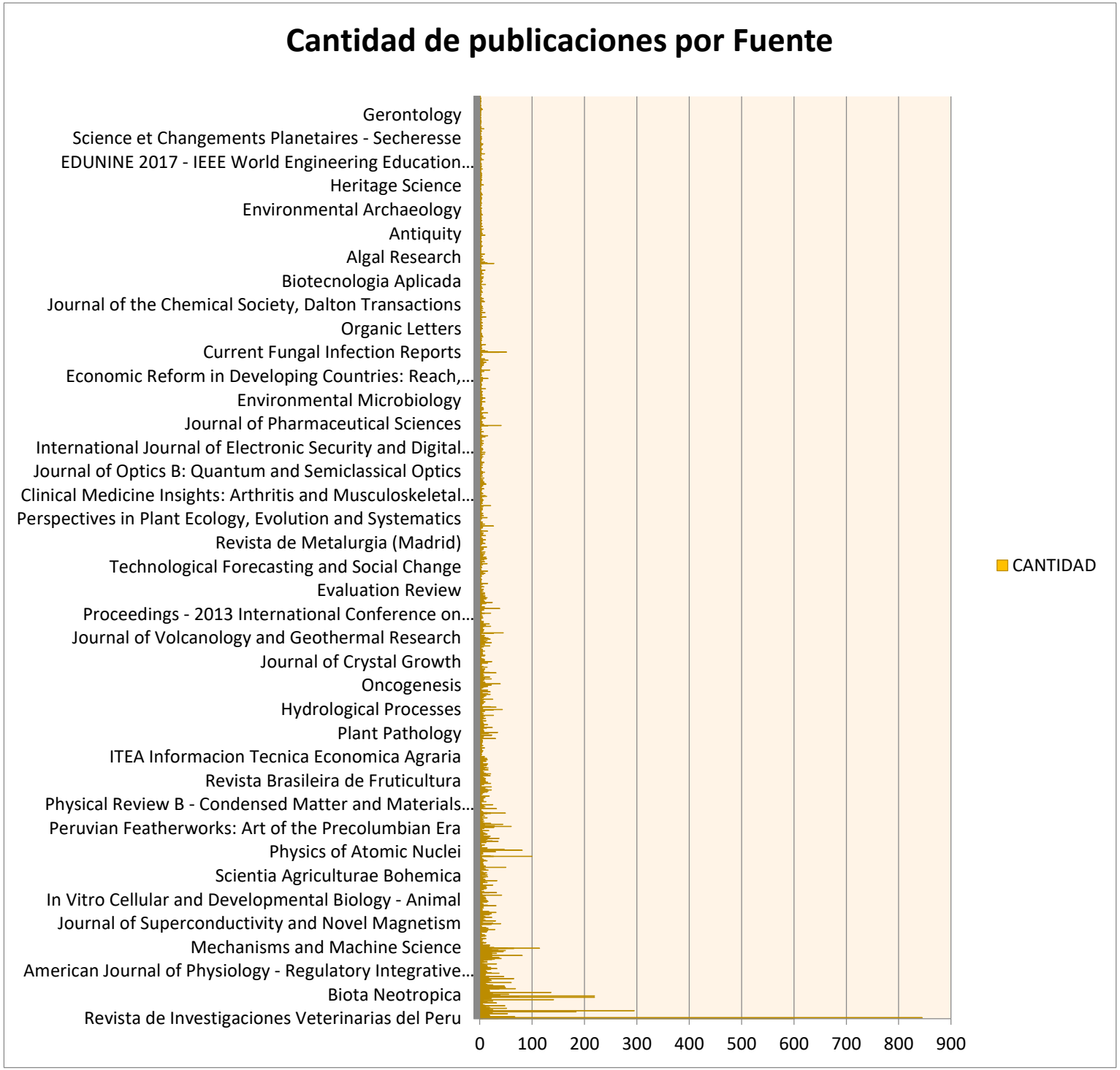
Los resultados obtenidos se muestran en las siguientes páginas:



La cantidad de artículos por investigador son en promedio aproximadamente 10. Sin embargo, existen investigadores que poseen altos índices de publicaciones con hasta más de 150 artículos en total durante los años de registro que contiene REGINA.



En los últimos años, las publicaciones de trabajos de investigación han aumentado de manera importante. Desde el año 2000, la cantidad de estos trabajos han estado creciendo en mayor magnitud hasta la fecha. Este comportamiento coincide con el avance tecnológico vivido en estas últimas décadas, lo cual tiene mucha coherencia. Así mismo, este hecho significa que este crecimiento continuará hasta que los avances en la tecnología sean muy difíciles de superar ya que estas publicaciones son contenedoras de nuevas metodologías propuestas para dar solución a problemas de nuestra sociedad en distintas áreas de conocimiento.



Las fuentes con mayores publicaciones están altamente relacionados al área biológica del conocimiento. Algunas de ellas son:

- Revista de investigaciones veterinarias del Perú
- Biota Neotropica
- American Journal of Physiology

La primera alcanzando un total de más de 800 publicaciones, muy superior al resto de fuentes dentro de los registros de REGINA.

REFERENCIAS:

- Pascual J., Digiampietri L., Martins F. & Roberto C. (2015) Brazilian Bibliometric Coauthorship Networks. URL: https://www.researchgate.net/profile/Wendel_Dos_Santos3/publication/281585509_Mena_C halco_et_al-2014-Journal_of_the_Association_for_Information_Science_and_Technology/links/55eeb78b08ae f559dc43d2f8/Mena-Chalco-et-al-2014-Journal-of-the-Association-for-Information-Scienceand-Technology.pdf
- Damaceno R., Rossi1 L., Mena J. (2017) Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na Plataforma Lattes. URL: https://www.researchgate.net/profile/Rafael_Damaceno/publication/318768717_Identificacao_do_grafo_de_genealogia_academica_de_pesquisadores_Uma_abordagem_baseada_na_Plataforma_Lattes/links/598a3c6fa6fdcc7cf91c5789/Identificacao-do-grafo-de-genealogiaacademica-de-pesquisadores-Uma-abordagem-baseada-na-PlataformaLattes.pdf?origin=publication_detail