MODELO DE MACHINE LEARNING EN BIG DATA PARA LA PREDICCIÓN Y RECOMENDACIÓN DE COAUTORÍA EN LAS REDES DE COLABORACIÓN FORMADO POR INVESTIGADORES PERUANOS

Cesar Daniel Rodríguez Castro, Cesar Risco Chavez y Luis Ernesto Huanay Palomino

Abstract

En este artículo se expone una investigación basada en la creación de un sistema capaz de identificar el comportamiento de investigadores registrados en la plataforma REGINA. Se aplicará técnicas de procesamiento y análisis en Python, para la generación de información a partir de la interacción entre nodos (investigadores) y aristas (trabajos académicos).

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 crear conocimientos avanzados de predicción y recomendación sobre la misma área.

Los datos solo permanecen registrados y no son capaces de crear información relevante para los usuarios (investigadores) con el fin de mejorar su producción a nivel científico y predecir/recomendar colaboraciones con potenciales investigadores. Por este motivo, este trabajo consiste en construir un sistema con tales cualidades de procesamiento de datos.

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 las interacciones entre los investigadores en la forma de coautorías. En este artículo, identificamos y caracterizamos las redes de coautoría académica peruana de investigadores registrados en la plataforma REGINA utilizando las propiedades de grafos.

Para lograr el objetivo, (a) crear un algoritmo para extraer la información de los perfiles de cada investigador en la plataforma, (b) crear otro algoritmo para identificar coautorías de manera automática de acuerdo a la información bibliográfica de cada trabajo de investigación, y (c) generar la red de autores-investigaciones para analizar las interacciones.

Este estudio caracterizó a las redes de coautor para obtener una comprensión profunda de las estructuras de la red y el comportamiento entre los investigadores. En este estudio, evaluamos 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 generar una red a partir de los 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.

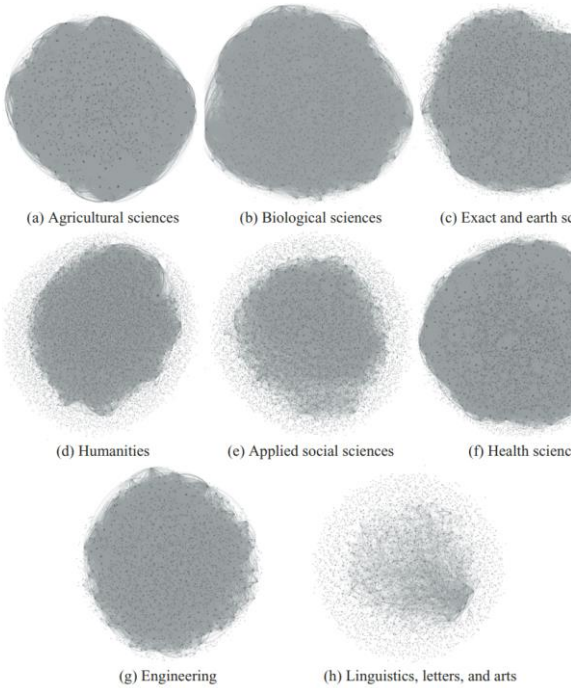

Figura 1 Redes de coautoría pertenecientes a investigadores asociados a cada área de conocimiento principal. Las redes se crearon considerando publicaciones producidas entre 1990 y 2010. No se muestran nodos aislados.

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.

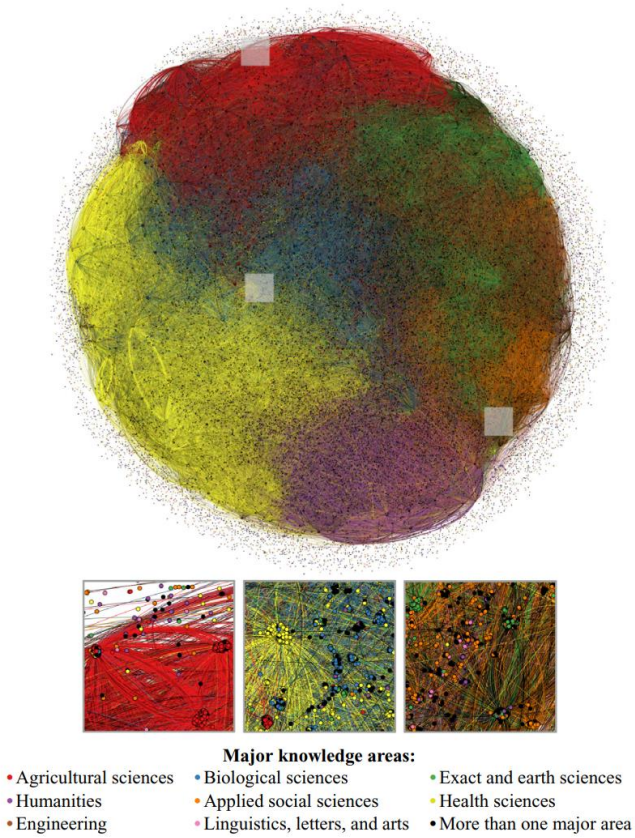

Figura 2 Red de coautoría de investigadores obtenida de la Plataforma Lattes. Los investigadores están representados por nodos; Las condecoraciones están representadas por aristas. Las redes se crearon considerando publicaciones producidas entre 1990 y 2010. No se muestran nodos aislados. Las tres regiones magnificadas están indicadas por una caja cuadrada blanca.

IV. Metodologia

a. Obtener los datos

Se creó un algoritmo para extraer los datos de la web de REGINA.

	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)

Durante este proceso se realizó una limpieza de datos y hubo la necesidad de eliminar registros que contenían campos vacíos.

b. Creación de la red

Se utilizó la librería NetworkX.

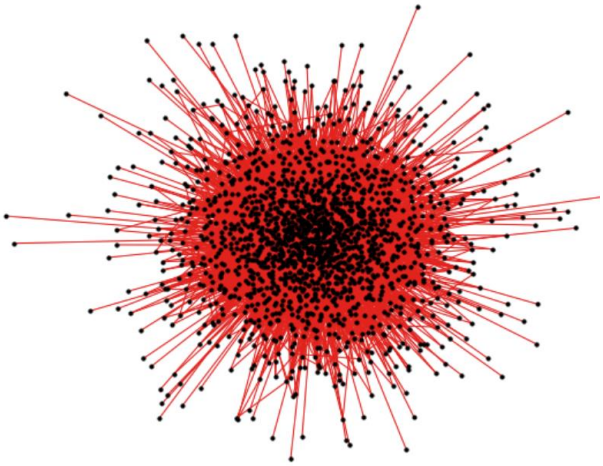

Figura 3 Red de coautoría de investigadores obtenida de la Plataforma REGINA.

Donde se tienen 1718 nodos y 18833 aristas.

```
Name:
Type: Graph
Number of nodes: 1718
Number of edges: 18833
Average degree:  21.9243
```

Así mismo, el promedio de trabajos realizados por cada investigador fue de 21. Lo cual es muy alto según una revisión previa de los perfiles. Pero puede deberse a diversos factores como: la repetición de registro en la plataforma o el algoritmo construido.

REFERENCIAS:

- Brazilian Bibliometric Coauthorship Networks. URL: https://www.researchgate.net/profile/Wendel_Dos_Santos3/publication/281585509_Mena_Chalco_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
- 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