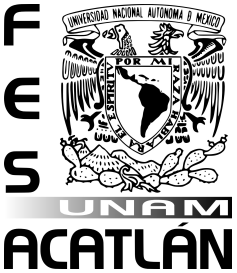


Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Ciencia de Datos

Datos Masivos

Título

- Victor Cuevas López
- Jennifer Itzel García Carrillo

Semestre 2020-2

Índice

1. Introducción

Como alumnos del séptimo semestre de la licenciatura en Ciencia de datos a punto de la búsqueda de tema para la elaboración de un trabajo de titulación, y con el ímpetu de incursionar en el área de la investigación, comunmente tenemos que recurrir a bases de datos, libros y artículos de divulgación científica.

Nuestra carrera demanda estar actualizados con los temas en boga, que están siempre en constante cambio en este nicho de las ciencias computacionales, y de cierta forma estar al tanto de la evolución de los mismos.

Un factor importante en la transformación y evolución de los tópicos de "moda"son las colaboraciones que surgen adaptando y hasta fusionando diversas áreas de estudio. Nos encontramos en una época en la que colaborar desde distintos lugares en el mundo y hasta si las partes no comparten el mismo idioma, no son barreras que lo impidan.

Es por eso que a con la técnica *Detección de Comunidades* analizaremos las comunidades formadas entre los autores de artículos de divulgación científica en el área de Ciencias de la Computación , durante los últimos 4 años.

Esto con ayuda de la extracción de los datos de los artículos pertenecientes a los 10 autores que mas escribieron durante este periodo(2017-2020). Mediante el API de DBLP (Un sitio web que posee un repositorio bibliográfico de artículos relacionados con ciencias de la computación).

2. Descripción de la Información

Como hemos mencionado la extracción de los datos se realizó mediante el servicio API que ofrece el sitio web *DBLP*. Extrayendo los datos de los artículos en los que ha colaborado un autor.

2.1. DBLP

Originalmente, en los años 80's, fue una base de datos que almacenó referencias relacionadas con programación lógica. Actualmente, DBLP lista más de un millón de artículos: aquellos pertenecientes a VLDB (Very Large Database, en inglés; una revista acerca de bases de datos con mucho contenido), artículos de la IEEE y ACM, así como también artículos científicos de distintas conferencias.

2.2. Autores semilla

Se inició con unos *autores semilla*, que son los autores que poseen más artículos en el periodo de tiempo seleccionado para analizar. A continuación una lista de aquellos personajes.

- Lei Zhang
- Wei Zhang
- Lei Wang
- Wei Wang
- Mohamed-Slim Alouini
- Zhu Han
- Mohsen Guizani
- Lajos Hanzo
- Jinde Cao
- Kim-Kwang Raymond Choo

2.3. Estructuración de los datos

Un ejemplo del dataset obtenido es el siguiente:

	key	title	year	author
0	journals/tcom/ShiMEYA18	Cooperative HARQ-Assisted NOMA Scheme in Large-Scale D2D Networks.	2018	Guanghua Yang
1	journals/nca/ZhangXPG19	An improved kernel-based incremental extreme learning machine with fixed budget for nonstationary time series prediction.	2019	Mingzhe Gao
2	journals/pieee/AkyildizPBZCWZ19	Scanning the Issue.	2019	T. Chen
3	journals/tcom/Randrianantenaina20	Interference Management in NOMA-Based Fog-Radio Access Networks via Scheduling and Power Allocation.	2020	Megumi Kaneko
4	journals/bmcbi/YangSWWLZWL18	Constructing a database for the relations between CNV and human genetic diseases via systematic text mining.	2018	Lingqian Wu

Donde:

- **key**: Llave del artículo asignada por DBLP.
- **title**: Titulo del artículo.
- **year**: Año de publicación del artículo.
- **author**: Titulo del artículo.

Se eliminaron los artículos que sólo tuvieran un autor, y aquellos autores que aparecían menos de 5 veces, para tener un grafo más condensado.

3. Modelado e Implementación

Se realizaron dos análisis para comparar en el tiempo las comunidades y sus respectivos temas principales.

Para el primer análisis se tomaron los articulos publicados en 2017 y 2018, que cumplieran con las especificaciones anteriormente mencionadas. Y para el segundo análisis se tomaron los artículos publicados en los años 2017,2018,2019 y 2020, y así tener una mejor descripción de cómo cambiaron las comunidades ya formadas en el primer lapso de tiempo con respecto al segundo lapso.

Posteriormente, con el dataset obtenido por los cortes mencionados, se formaron los nodos del grafo principal con los nombres de los autores, y los aristas reflejaban cuando dos autores contribuían en un mismo artículo, si el artículo había sido escrito por 3 o más autores, entonces se obtenían los aristas que incluian a todos combinados de dos en dos.

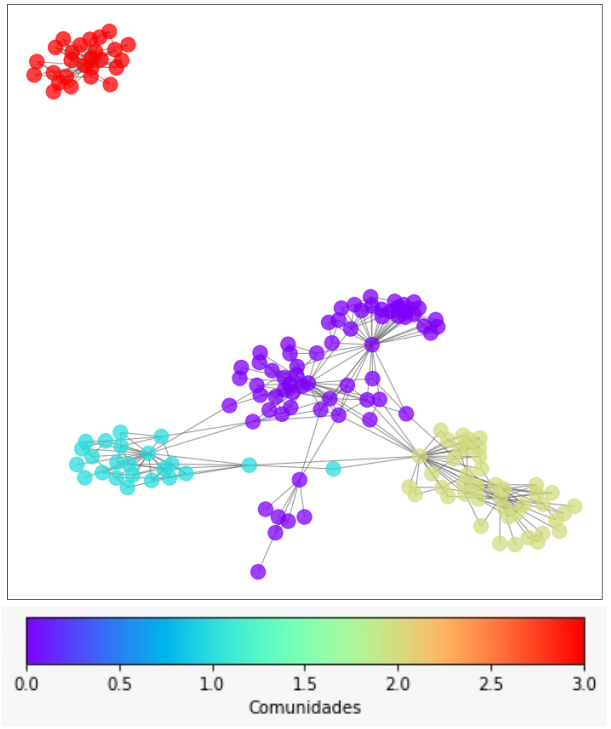
Por último, se realizó un *Análisis de Comunidades* a cada grafo principal, y para visualizar los temas que predominan en cada comunidad, se optó por visualizaciones de nubes de palabras.

4. Interpretación de resultados

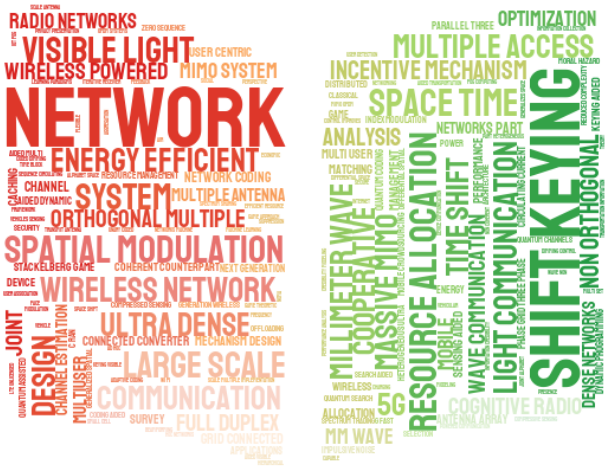
Recordemos que se realizaron dos análisis de comunidades diferentes.

4.1. 2017-2018

Durante los años de 2017 y 2018 se encontraron 4 comunidades las cuales presentaban las siguientes caracteristicas:



- **Lajos Hanzo y Zhu Han**



Comunidad identificada con el número cero, color **morado**.

Fue una comunidad representada por Lajos Hanzo y Zhu Han con 120 y 106 participaciones respectivamente, publicaron principalmente en access y tvt, la mayoria en el año de 2018.

Esta comunidad consta de 66 autores que realizarón un total 238 publicaciones, de las cuales cm fue el que realizó mas publicaciones, abordando principalmente temas de Network y Shift keying.

- **Los journals de Mohamed-Slim Alouini**

Comunidad identificada con el número 1, color **cian**.

Fue una comunidad en la cuál se encuentran completamente representada por Mohamed-Slim Alouini con una participación del 95 % en los journals de esta comunidad, además de tener una presencia mayor en el año de 2017.

Esta comunidad consta de 23 autores que realizarón un total 106 publicaciones, abordando principalmente temas de Cellular , Fading chanel y sistemas de comunicación.

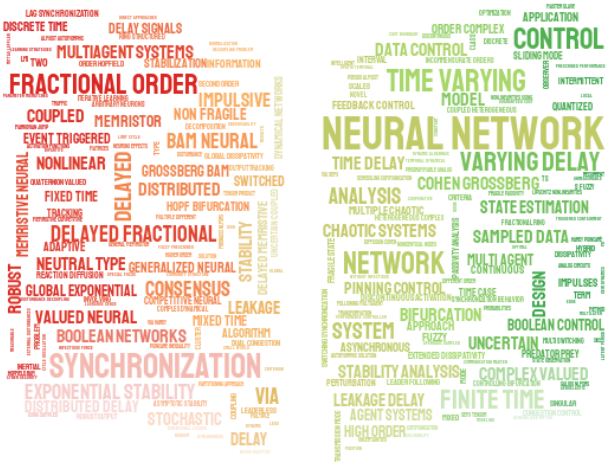
- **Tecnología**

Comunidad identificada con el número 2, color **verde**.

Fue una comunidad representada por Kim-Kwang Raymond Choo y Mohsen Guizani con una participación en 110 y 109 journals respectivamente, pero principalmente por lo topicos desarrollados como lo fueron: Data, Cloud, Internet.

Esta comunidad consta de 43 autores que realizarón un total 216 publicaciones, de las cuales cm fue el que realizó mas publicaciones.

- **Jiden Cao y colaboradores:**



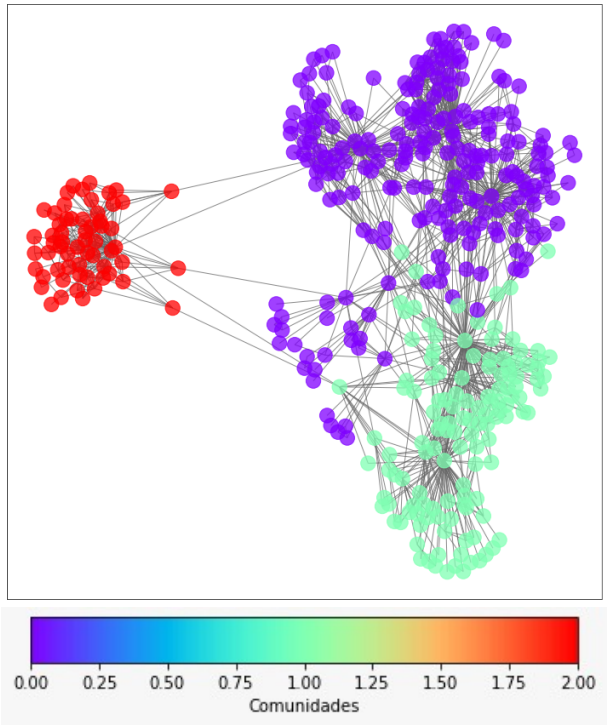
Comunidad identificada con el número 3, color **azul**.

Fue una comunidad principalmente representada por Jiden Cao con 117 articulos publicado, los principales journals en esta comunidad fueron nn y ijon, además de tener una presencia mayor en el año de 2018.

Esta comunidad consta de 26 autores que realizarón un total 117 publicaciones, abordando principalmente temas de Neural Network y Fractional order.

4.2. 2017-2020

Durante los años de 2017 al 2020 se encontraron 3 comunidades las cuales presentaban las siguientes caracteristicas:



- **Diseño de sistemas**

Comunidad identificada con el número 0, color **morado**.

Es una comunidad donde encontramos como autores principales a Mohamed-Slim Alouini,Lajos Hanzo, Zhu Han, escribiendo principalmente para access con un total de 265 publicaciones, abarcando temas principalmente de diseño de sistemas.

Esta comunidad consta de 224 autores que realizarón un total 1000 publicaciones, con actividad mayor en los ultimos años.

- **Internet de las cosas**

Comunidad identificada con el número 1, color **verde claro**.

Es una comunidad donde encontramos como autores principales a Mohsen Guizani y Kim-Kwang, escribiendo principalmente para access e iotj, con un total de 265 publicaciones y 149 respectivamente, abarcando temas principalmente de internet de las cosas.

Esta comunidad consta de 130 autores que realizarón un total 648 publicaciones, con actividad mayor en los ultimos años.

- **Jinde Cao y colaboradores**

Comunidad identificada con el número 2, color **rojo**.

Es una comunidad donde encontramos a Jinde Cao como el principal autor con una participacion mayor al 90 % de los artiuclos, con un total de 295 publicaciones abarcando temas principalmente de neural networks.

Esta comunidad consta de 60 autores que realizarón un total 296 publicaciones, con actividad mayor en los ultimos años.

5. Bibliografía

- DBLP (2021). How to use the dblp search API? Recuperado de https://dblp.org/faq/13501473.html