Análisis y predicciones sobre una red de comunicación entre investigadores a través de correos electrónicos

Machine learning sobre grafos

Sergio Andrés Rodríguez Torres
Estudiante, maestría en informática
Escuela colombiana de ingeniería Julio Garavito
Chía, Colombia
sergio.rodriguez-tor@mail.escuelaing.edu.co

Luis Felipe Díaz Chica
Estudiante, maestría en informática
Escuela colombiana de ingeniería Julio Garavito
Ciudad de México, México
luis.diaz-c@mail.escuelaing.edu.co

***Abstract*— Predicción de nodos y arcos en una red de comunicación entre investigadores vía correo electrónico usando algoritmos de machine learning sobre grafos.**

Keywords: grafos, aprendizaje de máquina, ciencia de reder, representational learning, predicciones, node2vec, emails.

I. INTRODUCCIÓN

En el siguiente artículo buscamos analizar una red real de correos electrónicos de una institución de investigación europea, aplicando varias técnicas de network science y machine learning para encontrar características del grafo y medidas de centralidad con el fin de entender mejor la red y luego lograr predecir algunas propiedades usando *Representation Learning* con Node2Vec sobre los nodos y arcos que la forman la red.

II. CONTEXTO

Contamos con una red de correos electrónicos de una prestigiosa institución de investigación en Europa. Estos datos han sido anonimizados y representan la comunicación entre investigadores de 42 centros de investigación por medio de correos electrónicos (Hao et al., n.d.).

Igualmente se conoce a qué departamento de investigación pertenece cada uno de los investigadores. Dado que pueden existir varios emails entre un par de investigadores, esta red sólo representa si en algún momento los dos investigadores se comunicaron vía email.

Cualquier correo electrónico enviado por fuera de la res es ignorado.

III. DATOS

A. Características de la red:

Propiedad	Valor
Nodos	1005
Arcos	25571
Dirigido	✅
Fuertemente conectado	❌
Dirigido aciclico	❌
Con pesos	❌
Coeficiente de asortatividad de grados	0.0055
Agrupamiento promedio	0.3657

Se calcularon las medidas de centralidad para conocer mejor la red.

1) Centralidad de grado

Corresponde al número de aristas o lazos que posee un nodo con los demás. (Sun & Tang, 2011).

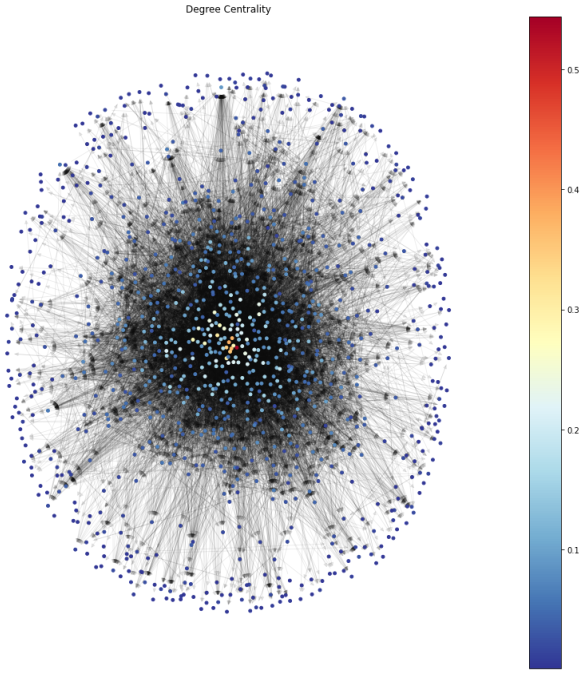


Imagen 1

2) Centralidad del vector propio

Mide la influencia de un nodo en una red y corresponde al principal vector propio de la matriz de adyacencia del grafo analizado. (Sun & Tang, 2011).

Un pequeño porcentaje de los nodos de la red tienen un nivel de influencia alta.

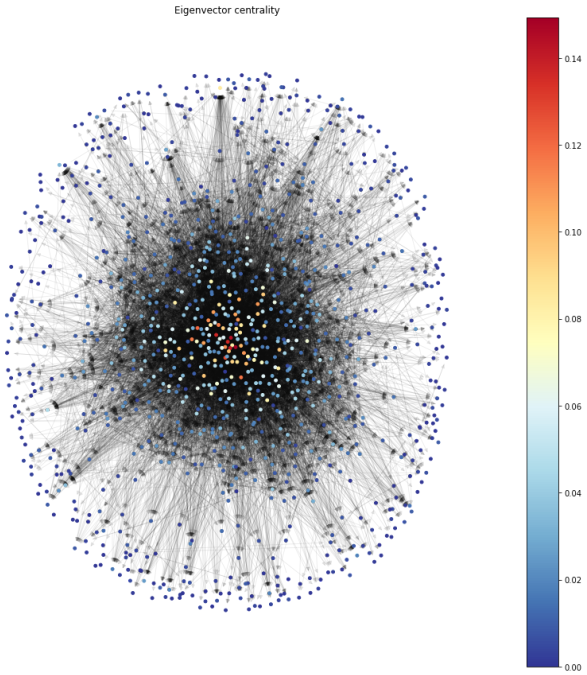


Imagen 2

3) Centralidad de cercanía

Se basa en calcular la suma o el promedio de las distancias geodésicas (o longitudes de los caminos más cortos) desde un nodo hacia todos los demás. (Sun & Tang, 2011)

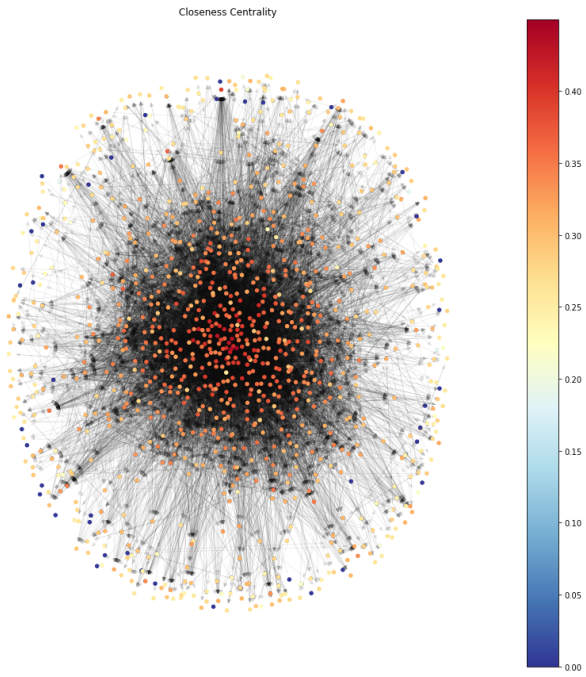


Imagen 3

4) Centralidad de intermediación

Cuantifica la frecuencia o el número de veces que un nodo se encuentra entre las geodésicas o caminos más cortos de otros nodos. (Sun & Tang, 2011)

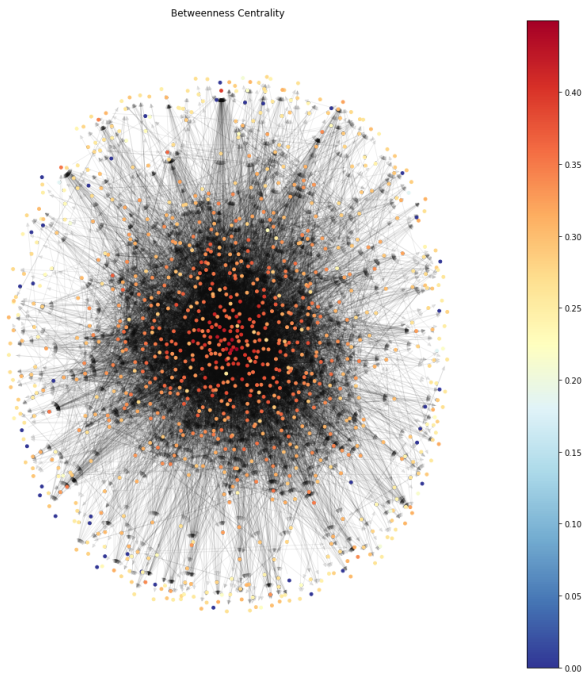
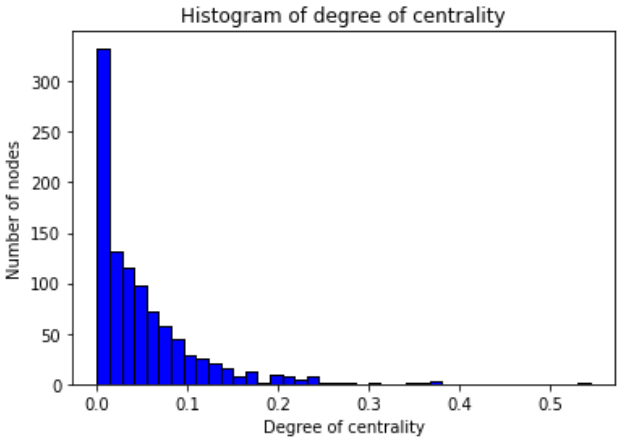


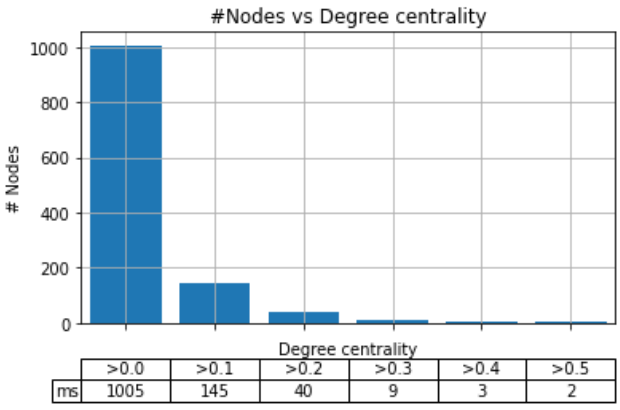
Imagen 4

B. Análisis de la red:

Podemos ver que aproximadamente un 50% de los nodos tienen una cercanía similar(color naranja, imagen 3). También la red presenta un síntoma muy natural de las redes del mundo real, la cantidad de nodos con alta influencia es baja, muy pocos investigadores se consideran influyentes [Imagen 1, Imagen 2].

Basándonos en el resultado de los grados de centralidad y en la distribución que presentan, podemos concluir que los datos de esta red pertenecen al conjunto de redes del mundo real ya que siguen una distribución *power law*.





	>0.0	>0.1	>0.2	>0.3	>0.4	>0.5
ms	1005	145	40	9	3	2

Imagen 5

IV. PROBLEMA

Se desea poder predecir el departamento al que pertenece un investigador a partir de las interacciones que ha tenido por correo con otros investigadores, a su vez se quiere poder predecir futuras interacciones entre cualquier par de investigadores.

V. SOLUCIÓN PROPUESTA

A. Predicción del departamento de un investigador:

Para la predicción del departamento de un investigador se propusieron dos soluciones usando *Representation Learning* con *node2Vec* para obtener una representación vectorial *embedding* para cada nodo del grafo, la primera solución usa aprendizaje no supervisado al medir similaridad de los vectores que representan un nodo y otra de aprendizaje supervisado con un modelo de regresión logística.

1. Similaridad

Para cada uno de los vectores calculamos la similaridad entre nodos para determinar si un par de nodos pertenecían al mismo departamento, la similaridad se calcula como la similitud del coseno entre la media simple de la proyección del vector del nodo dado y el vector de cada nodo del modelo.

El modelo se probó sobre todos los nodos de la red obteniendo más similar y evaluando si estos pertenecían al mismo departamento. En el proceso refinamos algunos parámetros del modelo para obtener mejores predicciones, nos centramos en el número de dimensiones de los vectores que representan los nodos, a continuación tenemos un gráfico con la precisión obtenida para cada dimensión.

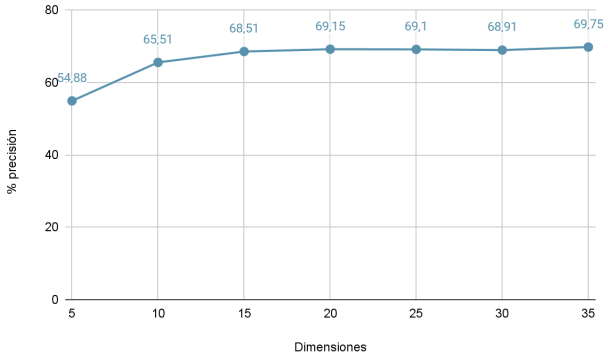


Imagen 6

2. Regresión logística

Los nodos en la representación vectorial tienden estar más cerca si describen un comportamiento similar en la red, para comprobar esto primero buscamos una forma de visualizar la representación vectorial de alta dimensionalidad en un plano de 2 dimensiones, para eso usamos el algoritmo *t-SNE* y graficamos la nueva representación en 2d del *embedding* coloreando los nodos por departamento.

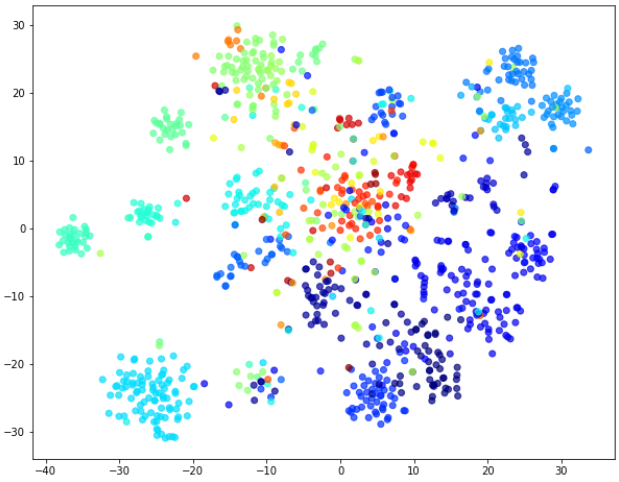


Imagen 7

Vemos una agrupación de los nodos por color, lo que nos permite usar un modelo clasificación para diferenciar los nodos en el espacio *embedding*, elegimos el algoritmo de regresión logística, dividimos los datos de forma aleatoria en dos grupos entrenamiento y prueba, entrenamos el modelos con los datos de entrenamiento y validamos su precisión con los datos de prueba, en este caso también buscamos el número de dimensiones que nos diera mejores resultados.

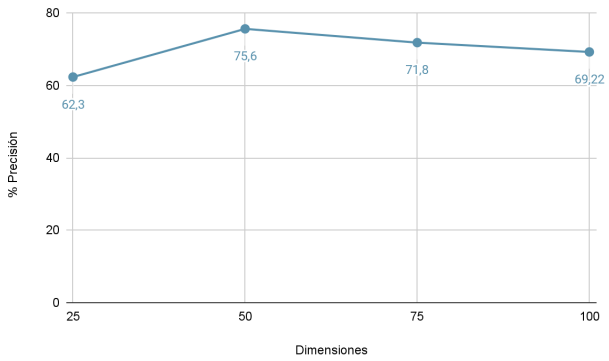


Imagen 8

B. Predicción de interacciones entre investigadores

Para la predicción de futuras interacciones entre investigadores se realizaron los siguientes pasos:

1. Preparación de los datos del grafo
2. Construcción de los embeddings
3. Construcción de los features
3. Entrenamiento de dos modelos
    a. Logistic regression
    b. Random Forest
4. Evaluación de los modelos

1. Preparación de los datos

En esta etapa se manipuló la red actual para poder disminuir la cantidad de arcos del grafo tratando de mantener la estructura original. Para esto se construyó una lista de todos los arcos del grafo que pueden ser removidos sin cambiar la estructura del grafo, esta lista se construyó por medio de un pequeño script que compara si la cantidad de componentes conectados del grafo al eliminar un arco.

Initialisation : $G :< V, E >$

for u, v in E **do**
  CopyG = G.removeEdge(u,v)
  if $(components(G) == components(CopyG)$ **then**
    removaleEdges.add(u,v)
  end if
end for
return $removaleEdges$

2. Construcción de los embeddings

Aquí se usó el algoritmo de *node2vec* para aprender representaciones de bajo nivel, se usó el grafo del paso paso 1 como entrada, los parámetros que usamos para la generación del modelo fueron:

dimensiones = 100, *longitud del camino*= 16 y *números de caminos por iteración* = 50.

3. Construcción de los features.

Los features son generados a partir de los embeddings que obtenemos con el algoritmo de *node2vec***.** Para nuestro caso en particular obtuvimos 100 features para cada arco del grafo. La obtención de los features está basada en el algoritmo de *word2vec.*

4. Entrenamiento de los modelos

Decidimos usar dos modelos de clasificación ya que nuestro problema se puede reducir a si una conexión entre dos investigadores puede o no puede existir. Los modelos utilizados fueron *LogisticRegression* y *RandomForest.*

Para la etapa de entrenamiento separamos los arcos del grafo (obtenido después de la depuración de datos) en un set de entrenamiento y en un set de prueba.

5. Evaluación de los modelos

La evaluación de los modelos se realizó con tres herramientas, la curva ROC/AUC, cross validation y la matriz de confusión, a continuación los resultados.

	Cross validation	ROC/AUC
Logistic Regression	0.6	0.67
RandomForest	0.72	0.87

LogisticRegression - Confusion Matrix

```
Accuracy: 0.6341370558375634
Precision: 0.5307331863285557
Recall: 0.6199291693496458
```

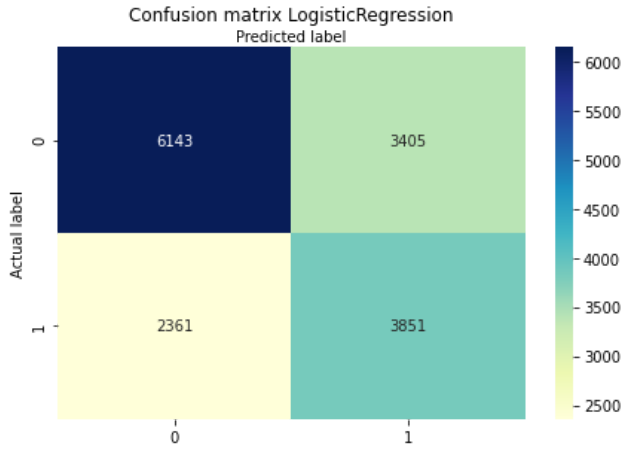


Imagen 9

RandomForest - Confusion Matrix

```
Accuracy: 0.7921319796954315
Precision: 0.8781555899021123
Recall: 0.5487765614938828
```

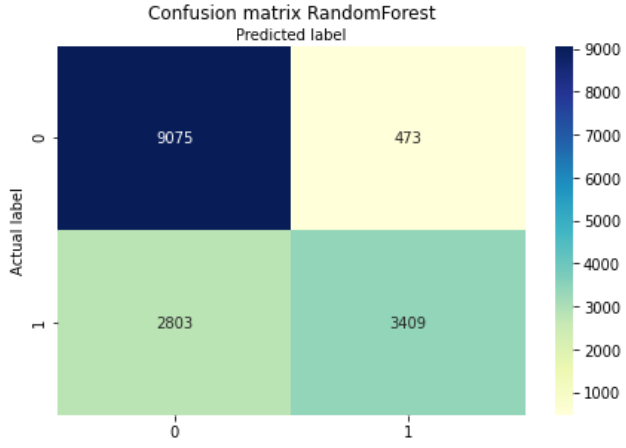


Imagen 10

VI. ANÁLISIS DE RESULTADOS

1. *Predicción de departamentos de un nodo*

La dimensionalidad de la representación vectorial del grafo juega un papel importante pero en ambos casos sigue un comportamiento similar, con valores pequeños se pierde demasiada información de la red y por lo tanto precisión, luego se aplana cerca al óptimo y si se sigue aumentando de nuevo pierde precisión.

Tanto el modelo de similaridad como el de regresión logística alcanzan una precisión similar, siendo la de este último ligeramente más alta, pero el de similaridad al no ser no-supervisado no requiere de la información de los departamentos para hacer una predicción, sin embargo la predicción es más limitada y no resuelven el mismo problema.

2. *Predicción de interacción entre los investigadores*

Es importante resaltar la gran importancia que tiene la depuración del grafo antes de comenzar a generar modelos de predicción. La granularidad y técnicas que se empleen en la preparación del grafo pueden generar un impacto significativo en los modelos y predicciones que se realicen en el futuro.

De los dos modelos usados podemos ver que RandomForest sin duda tiene un mejor desempeño en la predicción con una precisión de 87% contra *LogisticRegression* que alcanza 67%.

VII. CONCLUSIONES

Muchos escenarios de la vida real y problemas cotidianos pueden ser representados en problemas de grafos o redes; con la ayuda del poder de la inteligencia artificial, es posible usar representaciones de estos datos estructurados y predecir comportamientos sobre los mismos, ya sea como lo vimos en nuestro caso de estudio, predecir si un par de investigadores pueden llegar a trabajar juntos o simplemente predecir el departamento al que pertenece un investigador usando la cadena de correos electrónicos que ha intercambiado con otros investigadores. Problemas como este y muchos otros pueden verse beneficiados por el poder de la inteligencia artificial.

La aplicación de técnicas de machine learning e inteligencia artificial sobre grafos es sin duda uno de los campos de investigación con más fuerza en la actualidad.

REFERENCIAS

Bibliography

[1] Hao, Yin, Jure Leskovec, and David F. Gleich. n.d. “Local Higher-order Graph Clustering.” *International Conference on Knowledge Discovery and Data Mining* 23rd (2017): 5. https://doi.org/10.1145/3097983.3098069.

[2] Leskovec, J., J. Kleinberg, and C. Faloutsos. 2007. “Graph Evolution: Densification and Shrinking Diameters.” *Carnegie Mellon University* 1 (1): 41. https://doi.org/10.1145/1217299.1217301.

[3] Sun, Jimeng, and Jie Tang. 2011. “A survey of models and algorithms for social influence analysis.” *Social network data analytics* 1 (1): 177-214. https://doi.org/10.1007/978-1-4419-8462-3.