Universitat Politècnica de Catalunya

Facultad de Informática de Barcelona

PageRank

Búsqueda y Análisis de Información Masiva

Autores

Nicolas Llorens

Ricardo Lopez

21/11/2024

Contents

1 Informe

1.1 Implementación y principales dificultades

Aunque el algoritmo de PageRank no es complejo, la principal dificultad de esta práctica ha sido el uso de estructuras de datos eficientes para mantener un coste computacional asequible $O(n(logn))$ al haber usado diccionarios. Dado que la matriz de probabilidad es densa, el almacenamiento y cálculo se puede volver costosos para redes mas grandes. Una posible mejora sería utilizar matrices dispersas para optimizar la ejecucion.

Aun asi, creemos que el factor clave ha sido el uso de diccionarios, que aunque la lectura de datos es algo mas costosa, debido a la cantidad de consultas que hace el algoritmo, y estas se hacen en $O(logn)$ gracias a esta estructura de datos.

1.2 Analisis de los resultados

Aunque el factor tipico de "teleportacion" es comúnmente de 0.15, hemos hecho dos pruebas, una con 15% de probabilidad y otra con 5%. Ya que PageRank esta pensado para puntuar paginas web, y creemos que tiene mas sentido un factor de teleportacion mas alto al *surfear* la web, que al escoger aeropuerto.

1.2.1 Convergencia

Hemos hecho una comparativa de la convergencia entre diferentes factores de teleportacion para una muestra representativa de aeropuertos.

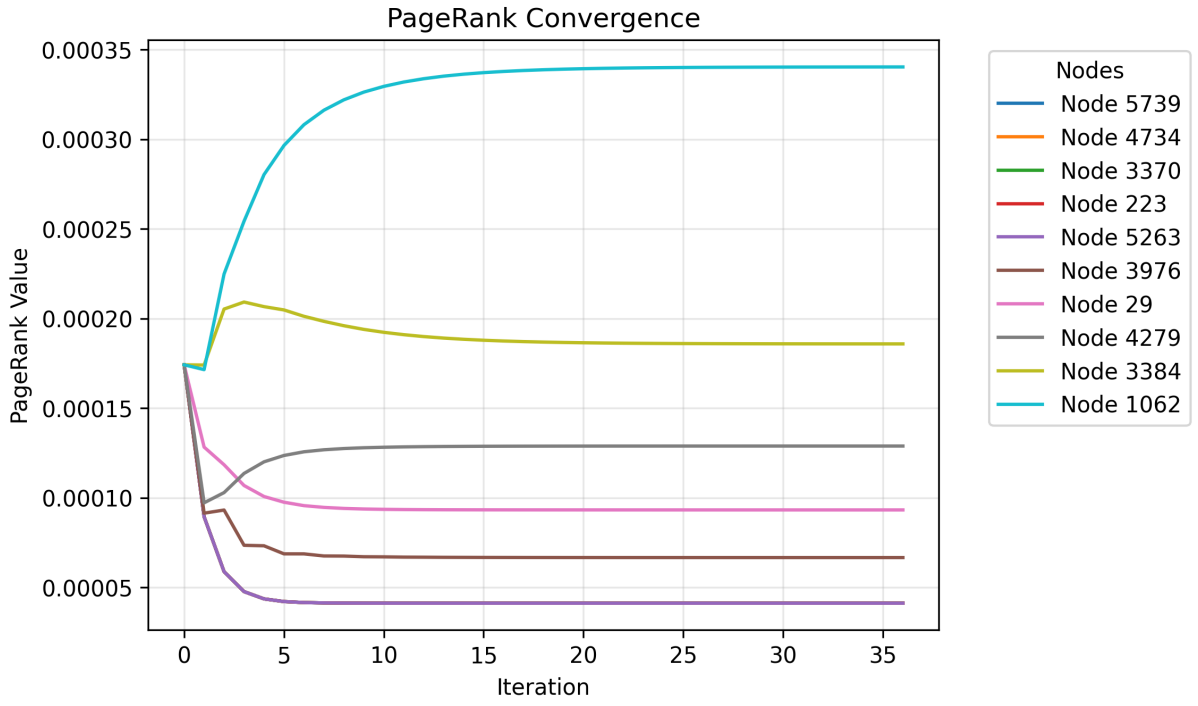

Figure 1: Factor de teleportación de $\alpha = 0.15$

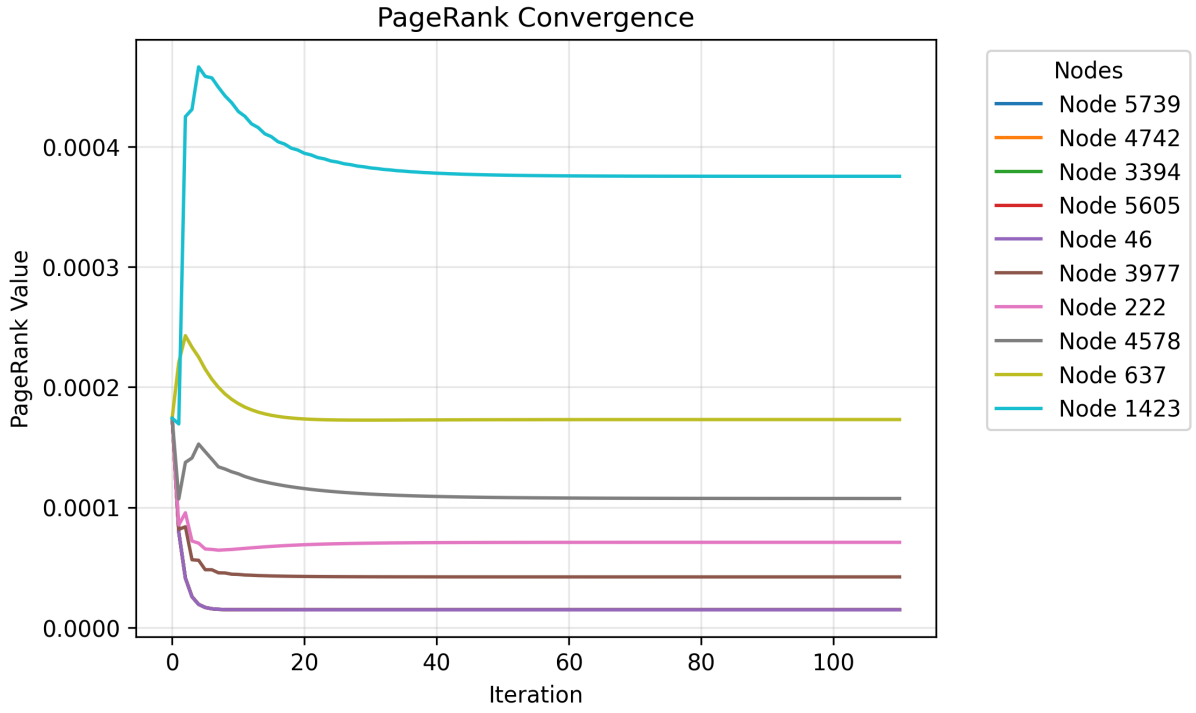

Figure 2: Factor de teleportación de $\alpha = 0.05$

Como vemos, en el caso de la figura 1 la grafica es mucho mas suave aproximandose a los valores de convergencia, y vemos tambien que hay muchas menos iteraciones.

También vemos que el *threshold* de $\epsilon = 10^{-6}$ quizas es demasiado lo que provoca muchas iteraciones demasiado similares. Para nuestro caso está bien ya que tenemos tiempo de ejecución muy moderados, pero si tuviesemos el mismo comportamiento para una muestra más grande y quisiesesmos mejorar el tiempo de ejecución podriamos aumentar el *threshold*

1.2.2 Distribucion de PageRanks

También queríamos ver si con un factor de amortiguación diferente los resultados de PageRank son los mismos. Aunque en las figuras 1 y 2 ya podemos observar una pista de que no será así. Hemos graficado la frequencia en escala logaritmica para que el grafico sea mas entendible.

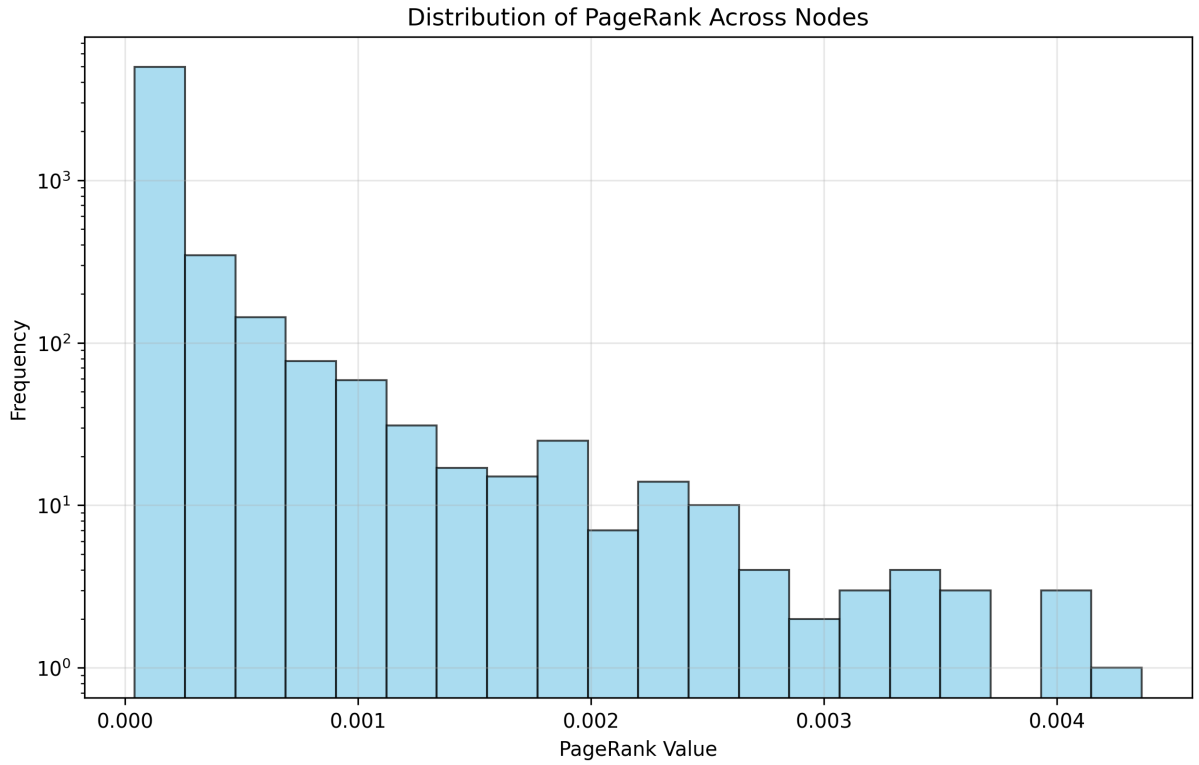

Figure 3: Factor de teleportación de $\alpha = 0.15$

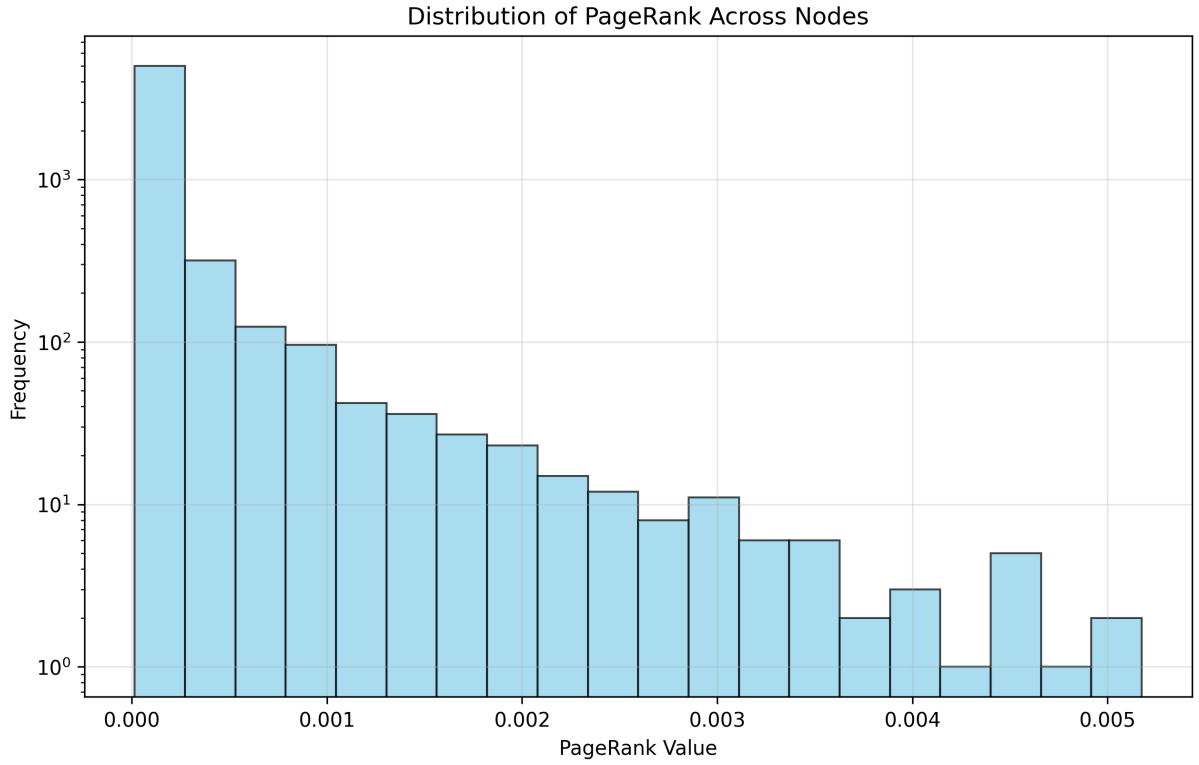

Figure 4: Factor de teleportación de $\alpha = 0.05$

Como vemos, conseguimos valores diferentes, y en el caso de $alpha = 0.05$ conseguimos valores mas hacia los extremos, es decir, más puntuaciones mas bajas y las puntuaciones mas altas un poco mas altas.

Este comportamiento que observado con un factor de teleportación de 0.05 (valores más extremos) podría deberse a que, con una red más estructurada o con hubs (aeropuertos muy conectados), PageRank depende más de las conexiones directas y esto hace que las puntuaciones se concentren en unos pocos aeropuertos. Esto es un efecto esperado cuando el factor de teleportación es bajo y tiene sentido en nuestro caso dada la naturaleza de la red.

1.2.3 Correctitud

Tambien queriamos ver que estuviese dando resultados con sentido y acabar de decidir cual seria este factor de amortiguación mas indicado para nuestra red. Por lo que queriamos graficar también que paises tenian mas importancia.

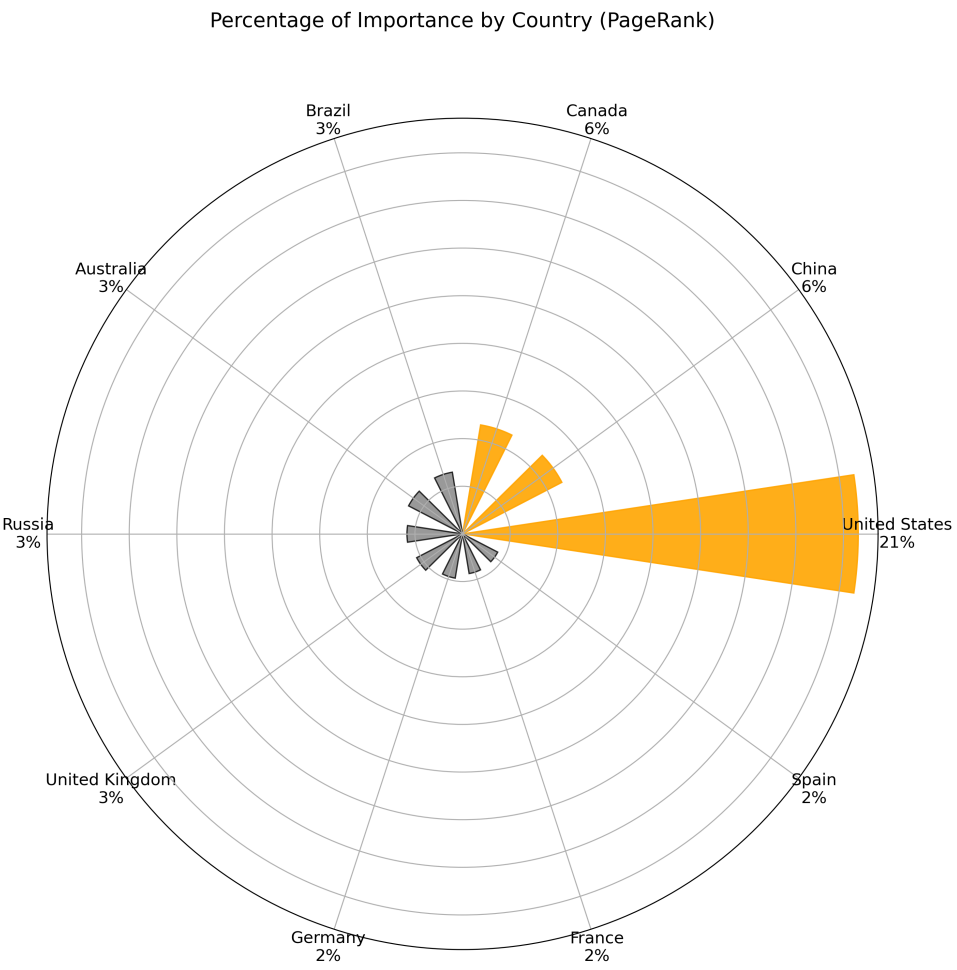

Figure 5: Factor de teleportación de $\alpha = 0.15$

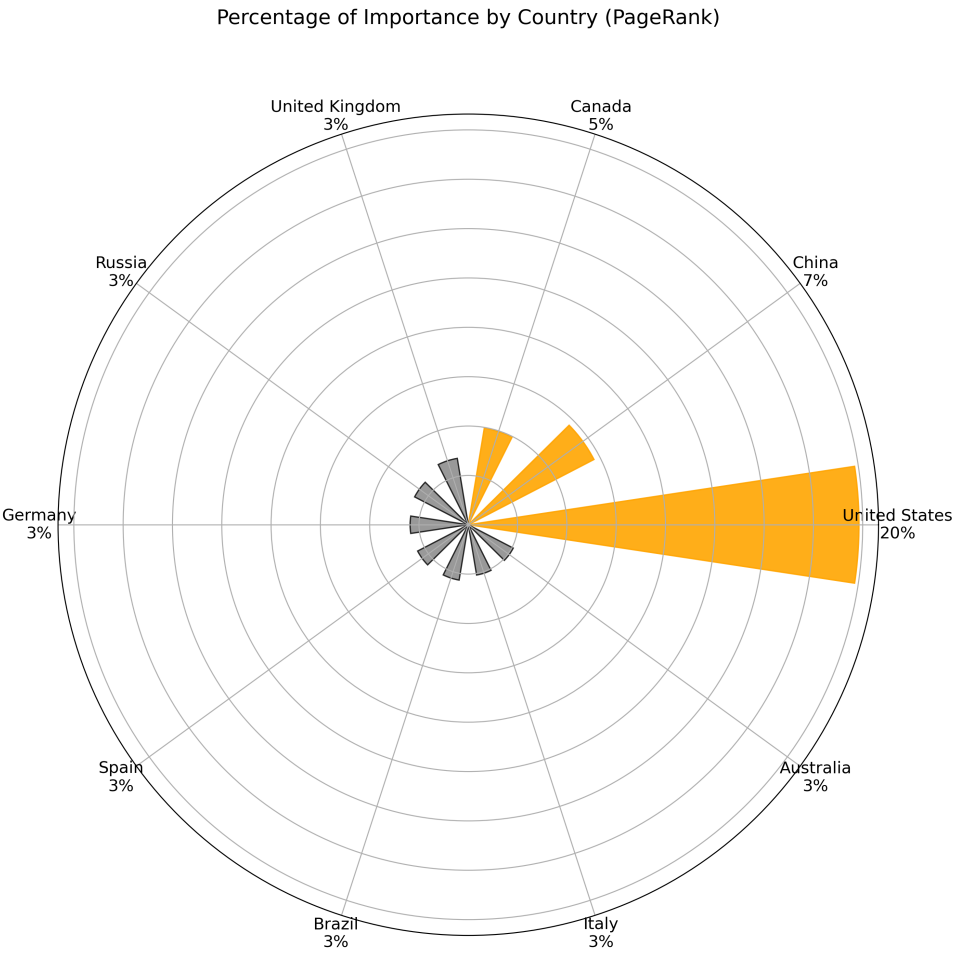

Figure 6: Factor de teleportación de $\alpha = 0.05$

Los resultados son muy similares y aunque vemos, algo mas de importancia en general en la figura **??**, USA pierda un punto de importancia, y habiendo consultado algunas fuentes (IATA y Federal Aviation Administration), esta mas cerca del 25% que del 20%. Aun asi, ambos valores vemos que funcionan de forma similar y dan resultados con mucho sentido.

2 Conclusiones

Hemos visto la importancia en la elección del tipo de estructuras de datos para obtener un algoritmo eficiente, sobretodo si se va usar para cantidades de datos grandes.

También hemos visto la importancia del ajuste de parametros tanto *epsilon* como *alpha*, y que los mejores valores para un algoritmo no existen si no se tiene en cuenta el contexto de los datos con los que se va a tratar. Ya que no es lo mismo la Web que la conexión de aeropuertos.