INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence
Primavera 2016

Profesores: Juan D. Velásquez
Gaspar Pizarro V.
Prof. auxiliar: Tomás Valdivia H.

Solución control 2

Pregunta 1

1. Uno de los procesos de su empresa implica el uso de un buscador web comercial, y, para reducir los costos operacionales, deciden implementar el suyo propio. El primer paso es la implementación de un *crawler* para la recolección de las páginas. Cuando usted lo prueba ve que comienza a descargar páginas a toda velocidad, pero que después de un tiempo hay sitios que comienzan a bloquear al *crawler*, retornando páginas de error. ¿Por qué puede estar pasando esto? ¿Qué tiene que hacer el *crawler* para evitar que siga pasando? *1 punto*

   Está pasando porque el crawler *está haciendo muchas peticiones por segundo, lo cual causa que los servidores baneen a la IP del* crawler*. Una forma de evitar eso puede ser hacer que el crawler espere una cantidad de tiempo entre peticiones a un mismo dominio, de forma que baje el impacto que el* crawler *tenga en el servidor. Enfoques más "maliciosos" pueden ser cambiarse el user-agent, o introducir aleatorización en los tiempos entre peticiones para simular un usuario humano.*

2. Compute la "matriz de Google" del grafo de la figura 1 con *damping* $d = 0{,}9$ y compute una iteración del algoritmo PageRank con esta matriz. Sin tener que computar los valores finales, especule sobre la diferencia entre la relevancia de los nodos 1 y 5. ¿Cuál de estos nodos cree que es más relevante y por qué? *2 puntos*

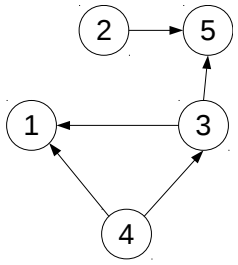

Figura 1: Un grafo de la web

Matriz de Adyacencia normalizada:

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Usando la fórmula

$$G_{ij} = \begin{cases} \frac{1-d}{n} + dA_{ji} & si\ \vec{A_j} = \vec{0} \\ \frac{1}{n} & si\ \vec{A_j} \neq \vec{0} \end{cases}$$

queda

$$G = \begin{bmatrix} 0{,}2 & 0{,}02 & 0{,}47 & 0{,}47 & 0{,}2 \\ 0{,}2 & 0{,}02 & 0{,}02 & 0{,}02 & 0{,}2 \\ 0{,}2 & 0{,}02 & 0{,}02 & 0{,}47 & 0{,}2 \\ 0{,}2 & 0{,}02 & 0{,}02 & 0{,}02 & 0{,}2 \\ 0{,}2 & 0{,}92 & 0{,}47 & 0{,}02 & 0{,}2 \end{bmatrix}$$

usando $p_0 = \begin{bmatrix} 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 \end{bmatrix}^T$ *entonces*

$$p_1 = \begin{bmatrix} 0{,}04 + 0{,}004 + 0{,}094 + 0{,}094 + 0{,}04 \\ 0{,}04 + 0{,}004 + 0{,}004 + 0{,}004 + 0{,}04 \\ 0{,}04 + 0{,}004 + 0{,}004 + 0{,}094 + 0{,}04 \\ 0{,}04 + 0{,}004 + 0{,}004 + 0{,}004 + 0{,}04 \\ 0{,}04 + 0{,}184 + 0{,}094 + 0{,}004 + 0{,}04 \end{bmatrix} = \begin{bmatrix} 0{,}272 \\ 0{,}092 \\ 0{,}182 \\ 0{,}092 \\ 0{,}362 \end{bmatrix}$$

El nodo 5 es más relevante [y bueno, así va el vector de PageRank, así que supongo que nadie va a decir que el nodo 1 es el más importante], porque, aunque tiene el mismo in-degree *que el nodo 1, el nodo 5 es apuntado por nodos que son apuntados, como el nodo 3, mientras que el nodo 1 es apuntado por nodos que no son apuntados por otros de afuera [el nodo 3 es apuntado por 4, pero también apunta a 1, entonces no cuenta como "de afuera"]. Es más, los nodos que apuntan a 1 también apuntan a otros nodos, mientras que los nodos que apuntan a 5 no apuntan a ningún nodo más. [Lo importante es el primer punto, lo de que mis apuntadores son también apuntados]*

3. Compare el tiempo de cómputo al usar el algoritmo HITS y el algoritmo PageRank para calcular la relevancia de documentos en un motor de búsqueda. ¿Dónde está la mayor parte del tiempo de cómputo al usar un modelo de relevancia como estos dos? *1 punto*

   El algoritmo HITS es dependiente de la consulta que se la haga al buscador, mientras que PageRank se computa una sola vez. Entonces usando HITS la mayor parte del tiempo de cómputo está en el uso del buscador, mientras que usando Pagerank el tiempo de cómputo está en la construcción del modelo.

4. Usted quiere encontrar gente "influyente" en un foro (como podría ser u-cursos) usando PageRank. Sin embargo, en este foro no hay relaciones explícitas entre los usuarios. Dado eso ¿cómo construiría un grafo con los datos del foro, que se pueda usar para calcular la influencia de los usuarios? *1 punto*

   Una forma de enlazar usuarios puede ser viendo sus respuestas. Por ejemplo, A sigue a B si en algún hilo A responde a B. Enfoques más detallados pueden usar un umbral para definir la relación, algo como A sigue a B si A responde a B a lo menos n veces en el foro.

5. ¿De qué forma y para qué puede servir, desde el punto de vista comercial, la identificación de comunidades en la Web? *1 punto*

La identificación de comunidades puede servir para agrupar a la gente con respecto a temas de interés (las comunidades suelen ser basadas en algún tema), con lo cual se puede hacer publicidad dirigida para la comunidad.