Análisis de Enlaces

Alvaro J. Riascos Villegas

Julio de 2021

Contenido

PageRank

- PageRank es una función que le asigna un número a cada página en la Web (o subconjunto de páginas).
- Interpretamos la Web como un grafo dirigido de páginas que tiene enlaces a otras:

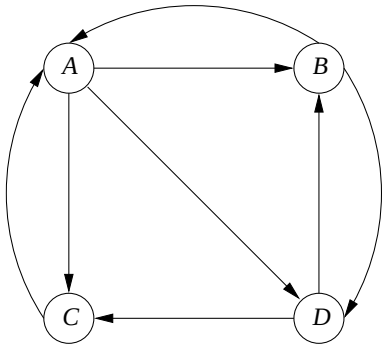

Figure 5.1: A hypothetical example of the Web

PageRank

- El modelo de un surfista que camina aleatoriamente en la web (*random surfer*) se puede representar con una matriz de transición M (el surfista puede comenzar en cualquier nodo y se mueve según los enlaces, cada uno con la misma probabilidad):

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- Obsévese que todas las entradas de M son positivas y sus columnas suman 1.
- La probabilidad de que un un surfista que comienza en alguno de los nodos con una distribución de probabilidad v (i.e, vector columna) se mueva al nodo i el siguiente periodo es: $(Mv)_i$.

PageRank

- M define una cadena de Markow.
- Un estado estacionario de la cadena de Markov es una distribución de probabilidad sobre los nodos v tal que: $v = Mv$.
- Si un surfista comienza con probabiliad v en cada nodo, el siguiente periodo estará por la misma probabilidad en cada nodo.

PageRank

- Este vector muestra cuales son los nodos más probables en donde estaría un surfista que comienza y recorre el grafo de forma aleatoria. Para ver esto, usamos un poco de teoría de cadenas de Markov.
- Para calcular el estado estacionario iteramos sucesivamente la ecuación $v_{n+1} = Mv_n$ comenzando desde cualquier distribuición v_0.
- Recordemos las condiciones para que haya convergencia.

Aperiodicidad e irreducibilidad

- Un grafo es **fuertemente conexo o irreducible** si cada par de puntos se pueden conectar con enlaces dirigidos.
- Un sunconjunto de nodos es **cerrado** si ningún enlace sale de ese subconjunto hacia un nodo que esté por fuera (i.e., un agente en el subconjunto no tiene posibilidades de visitar nodos por fuera del subconjunto).
- Un grafo es **aperiódico** si el máximo común divisor de la longuitud de todos los cíclos es 1.

Aperiodicidad e irreducibilidad

- Una condición suficiente para ser aperiódico es que $T_{ii} > 0$ para algún i (no es necesaria).

Theorem

T es convergente sí y sólo sí todo conjunto de nodos que es fuertemente conexo y cerrado, es aperiódico.

PageRank: Nodos muertos

- En la práctica cuando $50 - 80$ iteraciones son suficientes para converger con una precisión muy alta.
- El ejemplo anterior satisface las condiciones del teorema. La distribución estacionaria a la que converge es: $(\frac{3}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9})$. Ejercicio: resolver el sistema lineal y hacer la iteración.
- El siguiente ejemplo tiene un nodo muerto (la matriz **no es estocástica**: tiene una columna de ceros).
- Obsérvese que no aplica el teorema (la matriz **no es estocástica**.

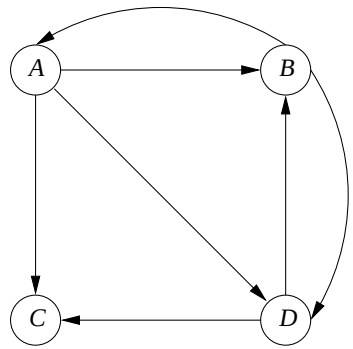

Figure 5.3: C is now a dead end

- Si se itera el estado estacionario converge a un vector de ceros: es decir, cuando un surfista llega a un nodo muerto, en el siguiente periodo desaparece!

PageRank: *Spider traps*

- El siguiente ejemplo satisface las condiciones del teorema pero arroja un resultado anti intuitivo (e.g., spider trap)
- Si se itera, el estado estacionario se va concentrar en el estado C.

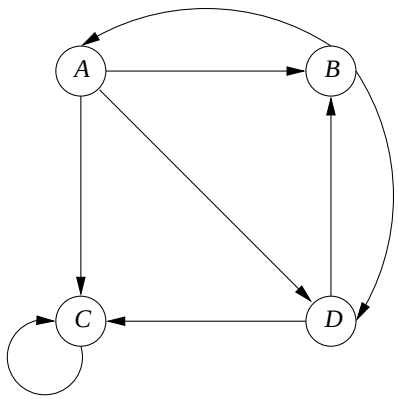

Figure 5.6: A graph with a one-node spider trap

PageRank: Componetes fuentes

- Es un conjunto de nodos que no reciben ninguna visita y tienen por lo menos un enlace que apunta hacia afuera del conjunto.
- Intutivamente, una caminata aleatoria en una red con un conjunto de nodos de este tipo tendria a asignarles probabilidad cero de ser visitados. Probablemente una sub estimación de la importancia de los nodos.

PageRank: tributación y teletransporte

- Suponemos que en cada etapa de la caminata, el surfista elige con probabilida β si utiliza la estructura del grafo para continuar navegando o con probabilidad $(1 - \beta)$ salta de forma aleatoria a cualquier otro estado todos con la misma probabilidad.

$$v_{n+1} = \beta M v_n + (1 - \beta)\frac{\mathbf{1}}{n}$$

  donde β es un número cercano a 1.
- β puede interpetarse como la fracción de páginas sucesoras que considera el surfista para el próximo paso.

PageRank: tributación y teletransporte

- Obsérvese que modificación evita las *spider traps*, los nodos muertos y componentes fuentes.
- Sin embargo, cuando hay nodos muertos el vector al que se coverge no es distribucion de probabilidad sobre todos los nodos, pero es distinto de cero.

PageRank: tributación y teletransporte

- Consideremos de nuevo el ejemplo del *spider trap*.

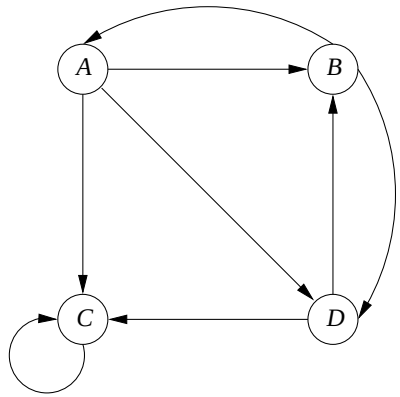

Figure 5.6: A graph with a one-node spider trap

- Si se itera la ecuación de PageRank modificada se converge a $(\frac{15}{148}, \frac{19}{148}, \frac{95}{148}, \frac{19}{148})$
- El nodo C logra un porcentaje importante de visitas del surfista pero bastantes menos que sin el teletransportado.

PageRank en la práctica

- Lo primero que hace un buscador es, dados unos términos de busqueda primero elige un conjunto de páginas que contengan esos términos y sobre este conjunto de palabras se hace PageRank.
- Típicamente el score de PageRank tiene un peso importante pero se consideran tambien otros criterios (lugar y frecuencia donde aparecen los términos en cada página, etc.).

Contenido

PageRank y Centralidad en Grafos

- Considere la siguiente definición de centralidad (Prestigio de Katz) de un nodo i, $p(i)$:

$$p(i) = \sum_j g_{ij} \frac{p(j)}{deg(j)} \tag{1}$$

- En un grafo dirigido podría tomarse el grado de enlaces salientes.
- Sea $\hat{g}_{ij} = \frac{g_{ij}}{deg(j)}$
- Entonces $\hat{g}_{ij}$ define una matriz estocástica: $\sum_i \hat{g}_{ij} = 1$
- Luego esta definición de centralidad es equivalente al problema de PageRank.

Contenido

Resúmenes de Textos: LexRank

- El objetivo es resumir varios documentos que hablan de un mismo tema (pero el tema es desconocido)
- LexRank se basa en la centralidad las frases en un grafo de frases (i.e., PageRank).
- Referencia: LexRank: Graph-based Lexical Centrality as Salince in Text Summarization. Journal of Artificial Intelligence Research (2004).
- La técnica también puede ser utilizada para tareas como: clasificación de entidades nombradas, adjunción de proposciones a frases.
- Puede utilizarse para la construcción de *features*.

Resúmenes de Textos: LexRank

- El objetivo es resumir varios documentos que hablan de un mismo tema (pero el tema es desconocido)
- LexRank se basa en la centralidad las frases en un grafo de frases (i.e., PageRank).
- Referencia: LexRank: Graph-based Lexical Centrality as Salince in Text Summarization. Journal of Artificial Intelligence Research (2004).
- La técnica también puede ser utilizada para tareas como: clasificación de entidades nombradas, adjunción de proposciones a frases.
- Puede utilizarse para la construcción de *features*.

Resúmenes de Textos: LexRank

- El objetivo es resumir varios documentos que hablan de un mismo tema (pero el tema es desconocido)
- LexRank se basa en la centralidad las frases en un grafo de frases (i.e., PageRank).
- Referencia: LexRank: Graph-based Lexical Centrality as Salince in Text Summarization. Journal of Artificial Intelligence Research (2004).
- La técnica también puede ser utilizada para tareas como: clasificación de entidades nombradas, adjunción de proposciones a frases.
- Puede utilizarse para la construcción de *features*.

Resúmenes de Textos: LexRank

- El objetivo es resumir varios documentos que hablan de un mismo tema (pero el tema es desconocido)
- LexRank se basa en la centralidad las frases en un grafo de frases (i.e., PageRank).
- Referencia: LexRank: Graph-based Lexical Centrality as Salince in Text Summarization. Journal of Artificial Intelligence Research (2004).
- La técnica también puede ser utilizada para tareas como: clasificación de entidades nombradas, adjunción de proposciones a frases.
- Puede utilizarse para la construcción de *features*.

Resúmenes de Textos: LexRank

- El objetivo es resumir varios documentos que hablan de un mismo tema (pero el tema es desconocido)
- LexRank se basa en la centralidad las frases en un grafo de frases (i.e., PageRank).
- Referencia: LexRank: Graph-based Lexical Centrality as Salince in Text Summarization. Journal of Artificial Intelligence Research (2004).
- La técnica también puede ser utilizada para tareas como: clasificación de entidades nombradas, adjunción de proposciones a frases.
- Puede utilizarse para la construcción de *features*.

Resúmenes de Textos

- Existen dos grandes formas de hacer resúmenes:
  1. Extraer subconjuntos de frases de los documentos.
  2. Abstracta, en la que se parafrasea las frases (e.g., resúmenes humanos).
- Solamente en la primera forma se ha avanzado considerablemente con modelos automáticos.

Resúmenes de Textos

- Existen dos grandes formas de hacer resúmenes:
  1. Extraer subconjuntos de frases de los documentos.
  2. Abstracta, en la que se parafrasea las frases (e.g., resúmenes humanos).
- Solamente en la primera forma se ha avanzado considerablemente con modelos automáticos.

Medidas de la Importancia de Frases: Basadas en medidas de centralidad

- Todas la propuestas a continuación se basan en el concepto de prestigio en redes.
- La idea es construir una red de frases (nodos) de todos los documentos. Donde los enlaces reflejan qué tan similares son dos frases.
- La hipótesis es que las frases que son más centrales en este grafo son las más relevante.
- Para esto tenemos que definir similaridad entre frases y segundo la centralidad de una frase dada la similaridad a otras frases.

Medidas de la Importancia de Frases: Basadas en medidas de centralidad

- Todas la propuestas a continuación se basan en el concepto de prestigio en redes.
- La idea es construir una red de frases (nodos) de todos los documentos. Donde los enlaces reflejan qué tan similares son dos frases.
- La hipótesis es que las frases que son más centrales en este grafo son las más relevante.
- Para esto tenemos que definir similaridad entre frases y segundo la centralidad de una frase dada la similaridad a otras frases.

Medidas de la Importancia de Frases: Basadas en medidas de centralidad

- Todas la propuestas a continuación se basan en el concepto de prestigio en redes.
- La idea es construir una red de frases (nodos) de todos los documentos. Donde los enlaces reflejan qué tan similares son dos frases.
- La hipótesis es que las frases que son más centrales en este grafo son las más relevante.
- Para esto tenemos que definir similaridad entre frases y segundo la centralidad de una frase dada la similaridad a otras frases.

Medidas de la Importancia de Frases: Basadas en medidas de centralidad

- Todas la propuestas a continuación se basan en el concepto de prestigio en redes.
- La idea es construir una red de frases (nodos) de todos los documentos. Donde los enlaces reflejan qué tan similares son dos frases.
- La hipótesis es que las frases que son más centrales en este grafo son las más relevante.
- Para esto tenemos que definir similaridad entre frases y segundo la centralidad de una frase dada la similaridad a otras frases.

Similaridad de frases

- Cada frase se representa como un vector $N - dimensional$ donde N es el número de palabras posibles del lenguaje en consideración.
- Cada palabra i en una frase se le asocia un número en el vector $N - dimensional$: Frecuencia del término en la frase $\times$ idf (i.e., $idf = log(\frac{N}{n_i})$). Donde N es el número de documentos y n_i es el número de documentos en lo que i ocurre.
- La similitud entre dos frases se define como (el coseno entre dos vectores: idf-modified-cosine):

$$s(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2 \sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2} \qquad (2)$$

Similaridad de frases

- Cada frase se representa como un vector $N - dimensional$ donde N es el número de palabras posibles del lenguaje en consideración.
- Cada palabra i en una frase se le asocia un número en el vector $N - dimensional$: Frecuencia del término en la frase $\times$ idf (i.e., $idf = log(\frac{N}{n_i})$). Donde N es el número de documentos y n_i es el número de documentos en lo que i ocurre.
- La similitud entre dos frases se define como (el coseno entre dos vectores: idf-modified-cosine):

$$s(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2 \sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2} \qquad (2)$$

Similaridad de frases

- Cada frase se representa como un vector $N - dimensional$ donde N es el número de palabras posibles del lenguaje en consideración.
- Cada palabra i en una frase se le asocia un número en el vector $N - dimensional$: Frecuencia del término en la frase $\times$ idf (i.e., $idf = log(\frac{N}{n_i})$). Donde N es el número de documentos y n_i es el número de documentos en lo que i ocurre.
- La similitud entre dos frases se define como (el coseno entre dos vectores: idf-modified-cosine):

$$s(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2 \sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2} \qquad (2)$$

Grafo asociado a un conjunto de documentos

- Un conjunto de documentos se puede representar por una matriz de de similitud (coseno) donde cada entrada es la similitud entre la pareja de frases.
- Esta matriz se puede utilizar para representar un grafo con pesos.
- Se puede escoger un umbral para disminuir el número de elementos positivos de la matriz y también olvidarnos de que es un grafo por pesos.

Grafo asociado a un conjunto de documentos

- Un conjunto de documentos se puede representar por una matriz de de similitud (coseno) donde cada entrada es la similitud entre la pareja de frases.
- Esta matriz se puede utilizar para representar un grafo con pesos.
- Se puede escoger un umbral para disminuir el número de elementos positivos de la matriz y también olvidarnos de que es un grafo por pesos.

Grafo asociado a un conjunto de documentos

- Un conjunto de documentos se puede representar por una matriz de de similitud (coseno) donde cada entrada es la similitud entre la pareja de frases.
- Esta matriz se puede utilizar para representar un grafo con pesos.
- Se puede escoger un umbral para disminuir el número de elementos positivos de la matriz y también olvidarnos de que es un grafo por pesos.

Ejemplo: Textos y matriz adyacencia

		refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq “did not end” and that Britain is still “ready, prepared, and able to strike Iraq.”
10	d5s2	In a gathering with the press held at the Prime Minister’s office, Blair contended that the crisis with Iraq “will not end until Iraq has absolutely and unconditionally respected its commitments” towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

Ejemplo: Grafo

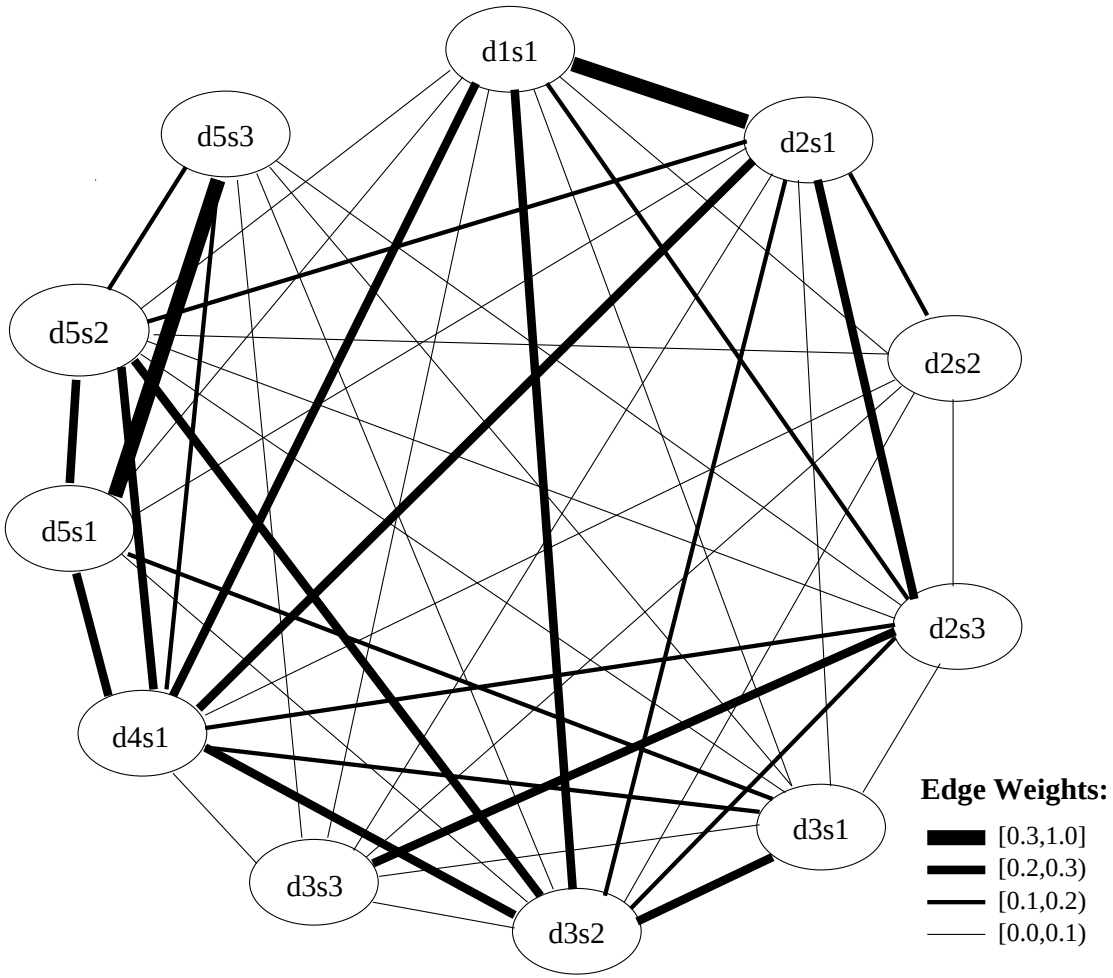

Figure 2: Weighted cosine similarity graph for the cluster in Figure 1.

Centralidad de frases

- La aplicación de esta forma de calcular la centralidad de cada nodo al grafo de similitud entre frases se denomina LexRank.
- Una versión alterntiva se basa en la versión del grafo de similitud por pesos.

Centralidad de frases

- La aplicación de esta forma de calcular la centralidad de cada nodo al grafo de similitud entre frases se denomina LexRank.
- Una versión alterntiva se basa en la versión del grafo de similitud por pesos.