Page Rank Explicacion Matricial

Recuperación da Información (Universidade da Coruña)

Page Rank Explicación Matricial

Tenemos 4 páginas web: **a, b, c, d**
El grafo web son los cuatro nodos que corresponden con cada página y los arcos entre nodos que vienen dados por el hecho de que haya un enlace entre las páginas.

Consideramos el grafo web: de **a** sale un enlace hacia **b** y un enlace hacia **c**, de **d** sale un enlace hacia **b** y hacia **c**, de **b** sale un enlace hacia **c** y de **c** sale un enlace hacia **b**. Pinta el grafo web para verlo más claro.

La matriz de adyacencia es:

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

La primera fila quiere decir que de a hay un enlace a b y c. La segunda fila quiere decir que de b hay un enlace a c, y así las otras dos filas.

La matriz de transición de probabilidades MTP es:

$$\begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

Lo único que hicimos es transformar cada fila en una distribución de probabilidad: la suma de los elementos de cada fila tiene que valer 1. Para ello lo que hicimos es que cada elemento de la matriz de adyacencia lo dividimos por la suma de los valores de los elementos de la fila.

Page Rank es un vector de una fila y cuatro columnas de forma que cada elemento tiene la probabilidad de la página de ser visitada. Inicialmente no sabemos nada y podemos pensar que todas las páginas tienen la misma probabilidad de ser visitadas, con lo que podemos pensar que el valor inicial de Page Rank es

$$\begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

Si A es un vector de 1 fila y n columnas y M es una matriz n x n, A es un autovector (eigenvector) principal izquierdo si se cumple que

A x M = A

Pues Page Rank es el autovector principal principal de la matriz de transición de probabilidades. Es decir tenemos que encontrar que el vector PR, Page Rank, contenga dentro los valores tal que se cumpla que

PR x MTP = PR

Este problema se puede resolver iterativamente (power method). Se parte de un valor inicial del vector PR y se multiplica por MTP y se obtiene otro valor de PR. Este nuevo valor se multiplica

por MTP y se obtiene otro valor de PR y así sucesivamente hasta que el valor de PR no cambia (o cambia menos que un valor mínimo epsilon), esto es, converge.

Pues si hacemos

$$\begin{pmatrix}0.25 & 0.25 & 0.25 & 0.25\end{pmatrix} * \begin{pmatrix}0 & 0.5 & 0.5 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0.5 & 0.5 & 0\end{pmatrix} = \begin{pmatrix}0 & 0.5 & 0.5 & 0\end{pmatrix}$$

Y si hacemos

$$\begin{pmatrix}0 & 0.5 & 0.5 & 0\end{pmatrix} * \begin{pmatrix}0 & 0.5 & 0.5 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0.5 & 0.5 & 0\end{pmatrix} = \begin{pmatrix}0 & 0.5 & 0.5 & 0\end{pmatrix}$$

Ya converge, lo cual quiere decir que ya encontramos el autovector principal izquierdo de MTP que es:

$$\begin{pmatrix}0 & 0.5 & 0.5 & 0\end{pmatrix}$$

Este resultado es intuitivo porque si observamos el grafo web veremos que hacia **a** y hacia **d** no llega ningún enlace, por tanto la probabilidad de ser visitados es cero. Los nodos **b** y **c** son análogos en el grafo web (a cada uno de ellos llegan un enlace de **a** y de **d**, y de uno al otro se intercambia un enlace), por tanto tienen la misma probabilidad de ser visitados (0.5) y ellos dos ser repartieron toda la probabilidad (el total de la fila suma 1).

El problema que tiene esto hasta aquí es que en este caso este algoritmo converge, pero hay grafos web que esto no converge. Para resolver esto se introduce el concepto de teleporting, que consiste en que que las páginas tienen una probabilidad de visita por la topología del grafo web pero otra probabilidad de visita a priori (por teleportación). Esto provoca un cambio en la MTP que se calcula de otra forma, y después el procedimiento de computar el Page Rank es el mismo.

Ejemplo con teleporting

Imaginemos ahora que tenemos un teleporting del 10% (coeficiente de 0.1). Con esto modelamos que el grafo web representado por la anterior MTP define en un 90% la probabilidad de llegar a una página pero el 10% total restante lo tienen todas las páginas a priori, es decir, aunque a ellas no llegasen enlaces. Esto quiere decir que si ese 10% (0.1) se reparte entre las 4 páginas, cada página tiene una probabilidad de llegar a ella por teleporting de 0.1/4 = 0.025, y esta probalidad no depende de los enlaces.

Entonces podemos calcular una MTP', es decir una matriz de transición de probabilidad con teleporting de 0.1, multiplicando la antigua por 0.9 y sumándole una matriz que tienen 0.025 en todos sus elementos. Entonces, esta MTP' será igual a:

$$\begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix} \text{ x } 0.9 + \begin{pmatrix} 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0.45 & 0.45 & 0 \\ 0 & 0 & 0.9 & 0 \\ 0 & 0.9 & 0 & 0 \\ 0 & 0.45 & 0.45 & 0 \end{pmatrix} + \begin{pmatrix} 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 \end{pmatrix} = \begin{pmatrix} 0.025 & 0.475 & 0.475 & 0.025 \\ 0.025 & 0.025 & 0.925 & 0.025 \\ 0.025 & 0.925 & 0.025 & 0.025 \\ 0.025 & 0.475 & 0.475 & 0.025 \end{pmatrix}$$

La MTP' se interpreta de esta forma:
-la primera fila representa que ahora desde **a** tienes una probabilidad de 0.025 de ir a **a**, una probabilidad 0.475 de ir a **b**, una probabilidad 0.475 de ir a **c** y una probabilidad 0.025 de ira a **d**
-la segunda fila representa que ahora desde **b** tienes una probabilidad de 0.025 de ir a **a**, una probabilidad 0.025 de ir a **b**, una probabilidad 0.925 de ir a **c** y una probabilidad 0.025 de ir a **d**
y así con las otras dos filas

Es decir básicamente se restan probalidades a los enlaces que hay en el grafo web original para dárselos al teleporting.

Una vez que tenemos MTP', la forma de calcular el Page Rank es igual que antes:

$$\begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \text{ x } \begin{pmatrix} 0.025 & 0.475 & 0.475 & 0.025 \\ 0.025 & 0.025 & 0.925 & 0.025 \\ 0.025 & 0.925 & 0.025 & 0.025 \\ 0.025 & 0.475 & 0.475 & 0.025 \end{pmatrix} =$$

$$= \begin{pmatrix} 0.025 & 0.475 & 0.475 & 0.025 \end{pmatrix}$$

Y si el vector $\begin{pmatrix} 0.025 & 0.475 & 0.475 & 0.025 \end{pmatrix}$ lo multiplicamos de nuevo por la matriz MTP', obtenemos

$$\begin{pmatrix} 0.025 & 0.475 & 0.475 & 0.025 \\ 0.025 & 0.025 & 0.925 & 0.025 \\ 0.025 & 0.925 & 0.025 & 0.025 \\ 0.025 & 0.475 & 0.475 & 0.025 \end{pmatrix}$$

ya converge, por tanto hemos encontrado que el vector de PageRank es

$$\begin{pmatrix} 0.025 & 0.475 & 0.475 & 0.025 \end{pmatrix}$$

Y de nuevo el resultado es intuitivo. A las páginas **a** y **d** sólo se puede llegar por teleporting ya que a estas páginas no llegan enlaces, por eso la probabilidad que resulta para estas páginas (primer y cuarto elemento del vector) es sólo la de teleporting, es decir 0.1/4 = 0.025. Las paginas **b** y **c** tienen exactamente las mismas propiedades en la topología del grafo por tanto su valor para Page Rank es el mismo y además las componentes del vector tienen que sumar 1 porque el vector representa una distribución de probabilidad, por tanto tienen que tener cada una un valor de 0.475.

También podemos compararlo con el resultado de antes. El orden no cambia. Las páginas **b** y **c** siguen siendo más importantes que **a** y **d** porque son más centrales, pero los valores de Page Rank se ajustan.

Páginas de las que no salen enlaces. Si de una página no salen enlaces, está página es un sumidero de probabilidad para Page Rank. Existe un resultado matemático que dice que para la MTP con teleporting el algoritmo anterior converge pero siempre y cuando no se de el caso de que existan páginas de las que no salen enlaces. Si existen esas páginas pudiera converger pero no hay garantía. Esto se puede resolver de una forma muy fácil sin alterar el significado de Page Rank: suponemos que para esa página que no salen enlaces, la probabilidad de Page Rank que recibe la traslada igualitariamente a todas las otras páginas. Por tanto esto es lo mismo que suponer que si en la matriz de adyacencia tenemos una fila de ceros (que se corresponde con esa paǵina de la que no salen enlaces) basta con poner esa fila todo a unos (que se corresponde con suponer en el grafo web arcos de esa página a todas las páginas). Una vez hecho eso basta con calcular la MTP y la MTP'.