Sistema de Recomendación de Skills para la complementación de perfiles profesionales del sector IT

- Daniel Orlando Ortiz Pacheco

Introducción

En la actualidad muchas veces se describe el sector de la Informática y la Tecnología como un mercado laboral donde importa más que sabes hacer a que títulos tienes. Debido a esto, las ofertas laborales de dicho sector se enfocan cada vez más en listar el conjunto de habilidades (skills) que requiere la vacante en cuestión y menos en precisar la formación académica que debe tener el postulante.

En la gran mayoría de los casos las ofertas no precisan el nivel académico que debe tener el candidato. En otras muchas basta con ser universitario o técnico de una carrera afín, lo cual en general deja un espectro bastante amplio. Sin embargo, en estas nunca falta un listado bien detallado de la lista de habilidades que debe tener el candidato. Dicha lista no solo se limita a mencionar la habilidad en cuestión, sino que usualmente se detallan la cantidad de años de experiencia que se debe tener en dicha habilidad o si la misma es excluyente o no.

Muchas personas, en especial los recién llegados al sector, en un afán por encajar cada vez mejor en el mercado laboral, comienza una ingesta insaciable de cursos para maximizar su conjunto de habilidades. Pero este plan muchas veces no es óptimo. Aunque es verdad que mientras más habilidades domine el candidato mejor es su perfil, al final, cada una de las ofertas del mercado tiene bien definido el perfil que necesita y no seleccionara a candidato con mayor número de habilidades, sino a aquel que mejor encaje con las necesidades planteadas.

El presente trabajo analiza precisamente esta problemática. Partiendo de la base de que se puede realizar un análisis masivo y frecuente de mercado laboral del sector IT mediante el cual se pueden detectar cada uno de los skills mencionado en cada oferta en particular. Este trabajo plantea la definición de un sistema de recomendación que mediante el aprendizaje de una matriz de correlación textual entre los skills pueda sugerir a un perfil determinado cuales serían las nuevas habilidades que este debería aprender según la demandas del mercado, de forma tal que con cada habilidad nueva haga que el perfil sea más compatible con la mayor cantidad de ofertas del mercado.

PageRank, Algebra Lineal y Sistemas de Recomendaciones

PageRank es uno de los secretos detrás del éxisto de Google. Google diseñó y perfecciono este algoritmo como factor diferencial en la carrera por desarrollar el mejor motor de búsqueda posible. Él mismo contiene en su interior algunos conceptos básicos del algebra lineal, los cuales se describirán a lo algo del presente capítulo.

La forma tradicional de describir este algoritmo es la representación del problema mediante grafos. Dónde dado un grafo dirigido se quiere ordenar cada uno de sus nodos según su relevancia. Para dicha tarea este algoritmo realiza su propia interpretación de relevancia, basándoselas en la democracia y la iteractividad. Partiendo de una inicialización uniforme de las relevancias, en cada iteración se propaga por cada arista x -> y una cantidad de importancia r_x/n_x donde r_x es la relevancia de x en esta iteración y n_x el número de aristas que salen de x. Desde el punto de vista de un motor de búsqueda se puede decir que con este algoritmo se intenta favoreces a las páginas más citadas dentro del conjunto de páginas relevantes, a la vez que se le resta importancia a las referencias de las páginas con gran cantidad de citas.

Aunque la descripción anterior es bastante intuitiva, no muestra toda la teoría que existe detrás de este algoritmo y justifica porque este algoritmo converge a un vector de relevancias óptimo. En realidad, este algoritmo no tiene sus bases en la teoría de grafos, sino que se apoya en el algebra lineal. Más explícitamente, este está basado en los sistemas dinámicos, una de las principales aplicaciones del producto matriz-vector, y en la existencia y propiedades de los valores y vectores propios de una matriz cuadrada.

Definición Matricial de PageRank, Valores y Vectores Propios

El algoritmo PageRank se define como, sea el conjunto D = $\{x_0, x_1, \ldots, x_n\}$ de entidades a ordenar y la relación R $\in D \times D$, se define un sistema dinámico $Ax_i = x_{i+1}$, $A \in R^{nxn}$ tal que:

$$A_{ij} = \begin{cases} 0, & < x_i, x_j > \notin R \\ 1/n_i, & < x_i, x_j > \in R,\ n_i = |\{x_j| < x_i, x_j > \in R\}| \end{cases} \tag{1}$$

Definición: Sea $A \in R^{nxn}$, un número λ se dice `valor propio` de A si existe un vector llamado `vector propio` $v \in R^n/\{0\}$ tal que:

$$Av = \lambda v \tag{2}$$

Además, se garantiza que toda matrix cuadrada al menos tiene 1 valor propio. Luego por tanto el problema iterativo de PageRank en su forma matricial se reduce a la búsqueda de un vector propio asociado al valor propio 1, de esta forma tal que se cumpla que $x_{i+1} = Ax_i = x_i$

En términos más generales, la matriz A, definida de la forma anteriormente descrita, debe tener 1 como valor propio si el grafo descrito por la relación R es fuertemente conexo, en caso de que no lo sea existen transformaciones para obtener problemas equivalentes pero no es objetivo de este trabajo. Nótese que dada a una matriz A con $A_{ij} = 1/n_j$, si $< x_i, x_j > \in R$, y A_{ij} = 0 en caso contrario, entonces la columna j de A contiene n_j entradas distintas de cero, cada una de ellas igual a $1/n_j$, por tanto la columna suma por tanto 1.

Definición: Una matriz cuadrada se denomina matriz `columna-estocástica` si todas sus entradas son no negativas y las entradas de cada columna suman uno.

Proposición: Toda matriz `columna-estocástica` tiene 1 como valor propio.
Demostración: Sea $A \in R^{n \times n}$ `columna-estocástica` y $e \in R^n$ con todas las entradas iguales a 1. Dado que toda matriz y su transpuesta tienen los mismo valores propios entonces es fácil notar que $A^T e = e$ y por tanto A tiene a 1 como valor propio

Luego ya que existe el valor propio 1 para toda matriz `columna-estocástica` solo queda demostrar que el sistema lineal planteado converge al vector propio asociado a dicho valor propio, con las condiciones iniciales correctas.

Definición: La norma 1 de un vector se define como $\|v\|_1 = \sum_{i=0}^{n} |v_i|$
Proposición: Toda matriz `columna-estocástica` positiva A tiene un único vector q con componentes positivas tales que $Aq = q$ con $\|q\|_1 = 1$. El vector q puede calcularse como $q = \lim_{k \to \infty}(A^k x_0)$ para cualquier x_0 inicial con componentes positivas tales que $\|x_0\|_1 = 1$. (Proposición demostrada en [1])

Por tanto para una matrix `columna-estocástica` si el sistema dinamico comienza a partir de un vector x_0 tal que $\|x_0\|_1 = 1$ y la cantidad de iteraciones $k \to \infty$ entonces el sistema dinámico planteado converge a un vector q único el cual es vector propio de la matriz A asociado al valor propio 1

Grafo de Correlación Textual entre Habilidades (Skills)

Apoyado en la base teórica antes descrita, y siempre que se puedan detectar los distintos skills requeridos en cada una de las ofertas de empleo del sector IT, entonces se puede definir la relación:

$$requisitos\_de\_oferta_k = \{< s_i, s_j > \mid los\ skill\ s_i\ y\ s_j\ aparecen\ juntas\ en\ la\ oferta\ k\} \quad (3)$$

Sobre la cual se puede realizar el mismo análisis descritos anteriormente para obtener un vector propio de la matriz A de la relación definida como se describió en la sección anterior. Cada componente i de dicho vector se puede interpretar como la importancia del skill s_i según la demanda del mercado laboral.

Luego, dado un curriculum particular se puede realizar el mismo análisis de detección de skills. Revisando la relación definida a partir del conjunto de skills detectadas en dicho curriculum, se puede escoger el conjunto de skills que aparecen juntos a los skills detectados en las distintas descripciones de las vacantes.

Una vez tenemos la lista de los skills relevantes para dicho perfil, utilizamos el resultado del análisis PageRank realizado al mercado laboral para ordenar estos resultados en orden descendente con respecto a la importancia de cada uno de los skills según el mercado laboral.

De esta manera, cada uno de los candidatos del mercado podrían orientar su crecimiento profesional hacia skills que le proporcione un mayor porcentaje de similitud con el perfil requerido por una mayor cantidad de ofertas. Aumentando a su vez las posibilidades de obtener un nuevo empleo.

Referencias

- Kurt Bryan and Tanya Leise, THE $25,000,000,000 EIGENVECTOR THE LINEAR ALGEBRA BEHIND GOOGLE,