3
Extensão do Modelo de *Hubs* e *Authorities*

Passamos, neste capítulo, a estender algoritmo *HITS*, introduzindo o *XHITS*.

3.1.
O Algoritmo *XHITS*

No modelo ora proposto são adicionados dois novos conceitos às definições de *hubs* e autoridades, quais sejam, portais e novidades. Uma boa autoridade é uma página apontada por bons *hubs* e um bom *hub* é uma página que aponta para boas autoridades. Nessa extensão, temos que:

- As boas autoridades são páginas que são apontadas por bons *hubs*, por bons portais e apontam para boas novidades;
- Os bons *hubs* são páginas que apontam para boas autoridades e novidades, e são apontados por bons portais;
- Os bons portais são páginas que apontam para as boas autoridades, para bons *hubs* e para boas novidades.
- As boas novidades são páginas que são apontadas pelas boas autoridades, pelos bons *hubs* e pelos bons portais;

Formalmente, temos que

$$a_i \propto \sum_{j \to i} h_j + \alpha \sum_{j \to i} p_j + \varphi \sum_{i \to j} n_j \text{ , i = 1,2,3...,n}$$

$$h_i \propto \sum_{i \to j} a_j + \theta \sum_{j \to i} p_j + \beta \sum_{i \to j} n_j \text{ , i = 1,2,3...,n}$$

$$p_i \propto \alpha \sum_{i \to j} a_j + \theta \sum_{i \to j} h_j + \gamma \sum_{i \to j} n_j \text{ , i = 1,2,3...,n}$$

$$n_i \propto \varphi \sum_{j \to i} a_j + \beta \sum_{j \to i} h_j + \gamma \sum_{j \to i} p_j \text{ , i = 1,2,3...,n}$$

onde a_i representa o peso da autoridade, h_i representa o peso do *hub*, p_i o peso do portal e n_i o peso da novidade. Podemos observar que, se α, β, θ, φ e γ forem iguais a zero, o modelo se reduz ao modelo proposto por *Kleinberg*.

Seguindo as relações de interdependência do modelo de *Kleinberg* e passando as equações para o formato matricial, tem-se

$$a \propto A^T h + \alpha A^T p + \varphi A n$$
$$h \propto Aa + \theta A^T p + \beta A n$$
$$p \propto \alpha A a + \theta A h + \gamma A n$$
$$n \propto \varphi A^T a + \beta A^T h + \gamma A^T p$$

onde α, β, θ, φ e γ, pertencentes a R, são fatores de redução adicionais, e A a matriz de adjacência do grafo. Porém, estas quatro equações podem ser reduzidas em apenas uma sob forma matricial, data por

$$\begin{bmatrix} a \\ h \\ p \\ n \end{bmatrix} \propto \begin{bmatrix} 0 & A^T & \alpha A^T & \varphi A \\ A & 0 & \theta A^T & \beta A \\ \alpha A & \theta A & 0 & \gamma A \\ \varphi A^T & \beta A^T & \gamma A^T & 0 \end{bmatrix} \times \begin{bmatrix} a \\ h \\ p \\ n \end{bmatrix} \quad (3.1.1)$$

Uma solução simples para esse sistema é oferecida pelo autovetor associado a matriz M definida por

$$M = \begin{bmatrix} 0 & A^T & \alpha A^T & \varphi A \\ A & 0 & \theta A^T & \beta A \\ \alpha A & \theta A & 0 & \gamma A \\ \varphi A^T & \beta A^T & \gamma A^T & 0 \end{bmatrix} \quad (3.1.2)$$

Tal solução pode ser calculada pelo método da potência que, como será visto na seção 3.2, neste caso, é um bom método para se encontrar o maior autovalor e o autovetor associado.

Observamos que os parâmetros α, β, θ, φ e γ servem como um ajuste fino do método e também possuem uma semântica vinculada ao modelo. Tais

parâmetros relacionam as categorias antes descritas, aumentando ou diminuindo a influência de uma categoria na outra.

Assim, temos que

- O parâmetro α relaciona autoridade e portal;
- O parâmetro β relaciona *hub* e novidade;
- O parâmetro θ relaciona *hub* e portal;
- O parâmetro φ relaciona autoridade e novidade;
- O parâmetro γ relaciona novidade e portal.

Desta forma, com base na relação circular entre as classificações, o algoritmo iterativo, que permite encontrar os *hubs* e as autoridades, foi estendido adicionando-se dois novos operadores e modificando-se os já existentes. Cada página i possui quatro pesos não negativos associados, um para autoridade ($a^{<i>}$), um para *hub* ($h^{<i>}$), um para o portal ($p^{<i>}$) e um para a novidade ($n^{<i>}$).

Em face do exposto, quatro operadores foram definidos para atualizar os pesos de autoridade (I), os de *hubs* (O), os de portal (P) e os de novidade (N):

- I : $a^{<p>} \leftarrow \sum_{q:(q,p)\in E} h^{<q>} + \alpha \sum_{q:(q,p)\in E} p^{<q>} + \varphi \sum_{r:(p,r)\in E} n^{<r>}$;
- O : $h^{<p>} \leftarrow \sum_{q:(p,q)\in E} a^{<q>} + \theta \sum_{r:(r,p)\in E} p^{<r>} + \beta \sum_{q:(p,q)\in E} n^{<q>}$;
- P : $p^{<p>} \leftarrow \alpha \sum_{q:(p,q)\in E} a^{<q>} + \theta \sum_{q:(p,q)\in E} h^{<q>} + \gamma \sum_{q:(p,q)\in E} n^{<q>}$;
- N : $n^{<p>} \leftarrow \varphi \sum_{q:(q,p)\in E} a^{<q>} + \beta \sum_{q:(q,p)\in E} h^{<q>} + \gamma \sum_{q:(q,p)\in E} p^{<q>}$.

Tais operadores representam claramente a relação de interdependência entre *hubs*, autoridades, portais e novidades.

Para encontrar o valor de equilíbrio entre os pesos, são aplicados, alternadamente, os operadores I, O, P e N até que a convergência seja alcançada, ou seja, os valores a, h, p e n das páginas se tornam inalterados com a iteração do algoritmo.

A seguir, apresentamos o pseudocódigo do algoritmo de iteração:

```
XIterate(G,k, α, β, θ, φ, γ)

(G : coleção de n páginas com hyperlinks )
(k : número de iterações do algoritmo )
(α : parâmetro que relaciona autoridade e portal)
(β : parâmetro que relaciona hub e novidade)
(θ : parâmetro que relaciona hub e portal)
(φ : parâmetro que relaciona autoridade e novidade)
(γ : parâmetro que relaciona novidade e portal)

Seja w o vetor (1,1,1...1) Є R^n
a_0:=w
h_0:=w
p_0=w
n_0=w
Para i := 1,2..k
Aplicar I em (a_{i-1}, h_{i-1}, p_{i-1}, n_{i-1}, α, φ), obtendo a`_i
Aplicar O em (a_{i-1}, h_{i-1}, p_{i-1}, n_{i-1}, β, θ), obtendo h`_i
Aplicar P em (a_{i-1}, h_{i-1}, p_{i-1}, n_{i-1}, α, θ, γ), obtendo p`_i
Aplicar N em (a_{i-1}, h_{i-1}, p_{i-1}, n_{i-1}, β, φ, γ), obtendo n`_i
Normalizar a`_i, obtendo a_k
Normalizar h`_i, obtendo h_k
Normalizar p`_i, obtendo p_k
Normalizar n`_i, obtendo n_k
Retornar (a_k , h_k , p_k , n_k)
```

Por fim, basta ordenar as coordenadas do vetor a retornado pelo algoritmo *Xiterate* para expor as páginas de maior autoridade. O mesmo se aplica aos vetores h, p e n, para as páginas de maior *hub*, as de maior portal e as de maior novidade.

3.2. Convergência do *XHITS*

Primeiramente, vamos observar algumas características da matriz M definida por

$$M = \begin{bmatrix} 0 & A^T & \alpha A^T & \varphi A \\ A & 0 & \theta A^T & \beta A \\ \alpha A & \theta A & 0 & \gamma A \\ \varphi A^T & \beta A^T & \gamma A^T & 0 \end{bmatrix}$$

Uma vez que a matriz A é uma matriz de adjacência, quadrada, e possui apenas valores iguais a zero e um, a matriz M possui apenas valores reais e também é quadrada. Adicionalmente, podemos facilmente verificar que M é uma matriz simétrica.

Analisando o algoritmo *XHITS*, podemos observar que ele define um vetor inicial para a, h, p e n, e que, após aplicar os operadores I, O, P e N, utiliza esses valores para a próxima iteração. Então, o algoritmo reduz a equação 3.1.1 a

$$\begin{bmatrix} a_k \\ h_k \\ p_k \\ n_k \end{bmatrix} = M \times \begin{bmatrix} a_{k-1} \\ h_{k-1} \\ p_{k-1} \\ n_{k-1} \end{bmatrix} \tag{3.2.1}$$

Substituindo a vetor $\begin{bmatrix} a_k \\ h_k \\ p_k \\ n_k \end{bmatrix}$ por d_k na fórmula 3.2.1, temos

$$d_k = M \times d_{k-1} \tag{3.2.2}$$

Sendo assim, para provar que o resultado do algoritmo *XHITS* converge, basta provarmos que (d_k - d_{k-1})→ 0 conforme k cresce. Uma solução simples para esse sistema é oferecida pelo autovetor associado ao maior autovalor da matriz M. Então, vamos provar que, além do resultado do *XHITS* convergir para um determinado valor, esse é o autovetor associado ao maior autovalor da matriz M. Para tanto, exibimos o seguinte teorema e a sua respectiva prova:

Teorema 1: Seja B uma matriz real, simétrica, de ordem n e sejam λ_1, λ_2, λ_3,..., λ_n seus autovalores e u_1, u_2,..., u_n seus correspondentes auto-vetores linearmente independentes, tal que $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$. Seja X_0 apropriadamente escolhido, então as seqüências $\{X_k = [x_1^{(k)} x_2^{(k)} ... x_n^{(k)}]^T\}$ *e* { c_k } *geradas recursivamente por*

$$X_{k+1} = \frac{1}{c_{k+1}} BX_k ,$$

onde

$$c_{k+1} = \left| x_j^{(k)} \right| \quad \textit{e} \quad \left| x_j^{(k)} \right| = \max_{1 \le i \le n} \{ \left| x_i^{(k)} \right| \} ,$$

vão convergir para o autovetor dominante V_1 e o autovalor λ_1, respectivamente. Então

- $\lim_{k \to \infty} X_k = V_1$
- $\lim_{k \to \infty} c_k = \lambda_1 .$

Prova:

Uma vez que B tem n autovalores, existem n correspondentes autovetores V_j, para j = 1,2,..,n, que são linearmente independentes, normalizados, e formam a base para um espaço de dimensão n. Assim, podemos expressar o vetor X_0 como combinação linear

$$X_0 = b_1 V_1 + b_2 V_2 ... + b_n V_n .$$

Assuma que $X_k = [x_1^{(k)} x_2^{(k)} ... x_n^{(k)}]^T$ foi escolhido de forma que $b_1 \neq 0$ e que as coordenadas de X_0 são escalares tal que $\max_{1 \le i \le n} \{ \left| x_i^{(k)} \right| \} = 1$. Uma vez que $\{V_j\}_{j=1}^n$ são autovetores de B, a multiplicação BX_0 produz

$$X_1 = \frac{1}{c_1} BX_0 = \frac{1}{c_1} B[b_1 V_1 + b_2 V_2 ... + b_n V_n]$$

$$= \frac{1}{c_1} [b_1 BV_1 + b_2 BV_2 ... + b_n BV_n]$$

$$= \frac{1}{c_1} [b_1 \lambda_1 V_1 + b_2 \lambda_2 V_2 ... + b_n \lambda_n V_n].$$

Colocando λ_1 em evidência

$$X_1 = \frac{\lambda_1}{c_1}\left[b_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)V_n \right].$$

Substituindo X_1 em

$$X_2 = \frac{1}{c_2} B X_1$$

$$= \frac{1}{c_2} B \frac{\lambda_1}{c_1}\left[b_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right) V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right) V_n \right]$$

$$= \frac{\lambda_1}{c_1 c_2}\left[b_1 B V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right) B V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right) B V_n \right]$$

$$= \frac{\lambda_1}{c_1 c_2}\left[b_1 \lambda_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right) \lambda_2 V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right) \lambda_n V_n \right].$$

Colocando novamente λ_1 em evidência

$$X_2 = \frac{\lambda_1^2}{c_1 c_2}\left[b_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^2 V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^2 V_n \right].$$

Após k iterações temos

$$X_k = \frac{1}{c_k} B X_{k-1}$$

$$= B \frac{\lambda_1^{k-1}}{c_1 c_2 \ldots c_k}\left[b_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1} V_n \right]$$

$$= \frac{\lambda_1^{k-1}}{c_1 c_2 \ldots c_k}\left[b_1 B V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} B V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1} B V_n \right]$$

$$= \frac{\lambda_1^{k-1}}{c_1 c_2 \ldots c_k}\left[b_1 \lambda_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} \lambda_2 V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1} \lambda_n V_n \right]$$

$$= \frac{\lambda_1^{k}}{c_1 c_2 \ldots c_k}\left[b_1 V_1 + b_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k} V_2 \ldots + b_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k} V_n \right].$$

Como $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$, $\left(\frac{\lambda_i}{\lambda_1}\right) < 1$ para i =2,3,...,n, temos

$$\lim_{k\to\infty} b_i \left(\frac{\lambda_i}{\lambda_1}\right) V_i = 0 \qquad \text{para i =2,3,...,n.}$$

Por conseqüência, temos

$$\lim_{k\to\infty} X_k = \lim_{k\to\infty} \frac{b_1 \lambda_1^k}{c_1 c_2 ... c_k} V_1 . \qquad (3.2.3)$$

De acordo com as premissas, X_k e V_1 são normalizados e a sua maior componente possui valor um. Uma vez que o limite do vetor do lado esquerdo de (3.2.3) será normalizado, sua maior componente terá valor um. Conseqüentemente, o limite do escalar que multiplica V_1 no lado direito de (3.2.3) existe e o seu valor valor deve ser um, ou seja

$$\lim_{k\to\infty} \frac{b_1 \lambda_1^k}{c_1 c_2 ... c_k} = 1 . \qquad (3.2.4)$$

Logo, substituindo (3.2.4) em (3.2.3)

$$\lim_{k\to\infty} X_k = V_1 .$$

Substituindo k por k-1 em (3.2.4)

$$\lim_{k\to\infty} \frac{b_1 \lambda_1^{k-1}}{c_1 c_2 ... c_{k-1}} = 1 \qquad (3.2.5)$$

e podemos reescrever o limite $\lim_{k\to\infty} \frac{\lambda_1}{c_k}$ como

$$\lim_{k\to\infty} \frac{\lambda_1}{c_k} = \lim_{k\to\infty} \frac{\dfrac{b_1 \lambda_1^k}{c_1 c_2 ... c_k}}{\dfrac{b_1 \lambda_1^{k-1}}{c_1 c_2 ... c_{k-1}}} . \qquad (3.2.6)$$

Substituindo (3.2.5) e (3.2.5) em (3.2.6)

$$\lim_{k\to\infty}\frac{\lambda_1}{c_k}=\lim_{k\to\infty}\frac{\dfrac{b_1\lambda_1^k}{c_1c_2\ldots c_k}}{\dfrac{b_1\lambda_1^{k-1}}{c_1c_2\ldots c_{k-1}}}=\frac{1}{1}=1.$$

Então a seqüência de constantes { c_k } converge para o autovalor dominante λ_1, completando a prova do teorema.

Corolário 1: O algoritmo XHITS converge.

Prova:

Podemos ver que a matriz M do *XHITS* atende as condições da matriz B. Além disso a seqüência d_k, do *XHITS*, obedece também as condições da seqüência X_k do teorema. Sendo assim, é imediata a convergência do resultado do algoritmo *XHITS* e esse valor é o autovetor associado ao maior autovalor da matriz M.