3
Extended Hyperlink Induced Topic Search

Neste capítulo, inicialmente, reescrevemos o algoritmo *HITS* de Jon Kleinberg (KLEINBERG, 1999; KLEINBERG et al., 1999) na forma de matrizes de blocos. Em seguida, estendemos o *HITS* para incorporar diversos outros tipos de categorias de páginas e destacamos algumas propriedades da extensão proposta, o algoritmo *XHITS*.

Introduzimos, ainda, a noção da matriz de influência, uma estrutura linear que combina o reforço mútuo de classificação e a estrutura de hiperlink.

Dois casos especiais naturalmente decorrem da nova abordagem, quais sejam:

– reforço simétrico;
– e, reforço positivo.

Por fim, provamos a convergência para o método estendido iterativo.

3.1
O Hyperlink Induced Topic Search Matricial

Como visto na seção 2.3, o algoritmo *HITS* é composto por duas partes distintas. A primeira parte é a construção de um grafo direcionado $G_\sigma = (V_\sigma, E_\sigma)$, que depende de uma consulta σ, e foi mantida inalterada na nova abordagem. A segunda é o núcleo da parte de classificação do algoritmo e trabalha diretamente no grafo direcionado, calculando os valores das autoridades e *hubs* de todas as páginas. Nessa parte começa a nova abordagem para o *HITS*.

Primeiro, denotamos por A_σ a matriz de adjacência que representa o grafo G_σ. Essa matriz é composta pelos elementos $a_{i,j}$ tal que cada elemento é igual a 1 caso haja um hiperlink da página i para a página j, e 0 caso contrário.

Os valores da autoridade e do *hub* para cada página foram representados pelos vetores a e h.

Reescrevendo o modelo HITS na forma matricial, temos

$$a \propto A_\sigma^T h$$

$$h \propto A_\sigma a$$

Esse sistema de dois conjuntos de equações lineares pode ser condensado na forma de matriz de blocos, resultando em uma única equação

$$\begin{bmatrix} a \\ h \end{bmatrix} \propto \begin{bmatrix} 0 & A_\sigma^T \\ A_\sigma & 0 \end{bmatrix} \cdot \begin{bmatrix} a \\ h \end{bmatrix} \tag{3-1}$$

A equação 3-1 fornece um meio imediato de iterativamente calcular a e h, e, para isso, é necessário garantir que os valores convergem.

De forma a examinar as questões em torno da convergência relacionada a iteração de 3-1, definimos a matriz de influência, M_σ por

$$M_\sigma = \begin{bmatrix} 0 & A_\sigma^T \\ A_\sigma & 0 \end{bmatrix} \tag{3-2}$$

É fácil verificar que M_σ é uma matriz simétrica. A iteração 3-1 é apenas uma instância do conhecido problema de extração de autovalores e autovetores. Uma forma simples e eficiente de calcular essa iteração é utilizando o Método da Potência.

De forma a revelar a estrutura de influência em M_σ foi utilizado o operador de produto direto ($\otimes$) (SEARLE, 1982) de duas matrizes, definido como

$$U \otimes V = \begin{bmatrix} u_{11}V & ... & u_{1q}V \\ ... & ... & ... \\ u_{p1}V & ... & u_{pq}V \end{bmatrix} \tag{3-3}$$

Reescrevendo M_σ a partir da utilização do produto direto, temos

$$M_\sigma = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \otimes A_\sigma^T + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \otimes A_\sigma \tag{3-4}$$

A equação 3-4 revela duas novas matrizes independentes da consulta σ,

quais sejam:

- *Forward Hyperlink Influence* F: que representa a influência das categorias que propagam valores através dos hiperlinks, conforme a equação 3-6;
- e, *Category Reinforcement Influence* B: que representa a influência das categorias que recebem valores através dos hiperlinks, conforme a equação 3-5.

$$B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{3-5}$$

$$F = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \tag{3-6}$$

$$M_\sigma = B \otimes A_\sigma^T + F \otimes A_\sigma \tag{3-7}$$

Essa nova abordagem para o *HITS* (equação 3-7) é muito importante para o modelo estendido como se mostrará na próxima seção.

3.2
O Algoritmo Extended Hyperlink Induced Topic Search

Para estender o modelo básico de *Hubs e Authorities* foram introduzidas novas categorias de páginas. Agora, ao invés de apenas duas categorias, temos u novas categorias.

Assim, em cada página i foram introduzidos u pesos representando cada uma das novas categorias e foram consolidados esses valores na matriz $(c_{iu})_\sigma$.

Sabemos que esses pesos são reforçados através dos hiperlinks e que as duas influências, F e B, estão presentes e não necessariamente são simétricas.

Sempre que a página i aponta para a página j, cada peso $(c_{jv})_\sigma$ contribui com o valor de $(c_{iu})_\sigma$ com um montante linear de $F_{uv}.(c_{jv})_\sigma$.

Similarmente, quando j aponta para i, cada peso $(c_{jv})_\sigma$ contribui para o valor de $(c_{iu})_\sigma$ com um montante linear de $B_{uv}.(c_{jv})_\sigma$.

Portanto, temos F e B com dimensões de $u \times u$, e cada uma representando até u^2 pesos de influência, tanto de recepção, quanto de propagação.

Formalmente, para cada peso $(c_{iu})_\sigma$ temos

$$(c_{iu})_\sigma \;\propto\; \sum_{j\to i}\sum_{v=1}^{k} B_{uv}.(c_{jv})_\sigma \;+\; \sum_{i\to j}\sum_{v=1}^{k} F_{uv}.(c_{jv})_\sigma$$

Reescrevendo as equações na forma matricial, temos

$$C_\sigma \;\propto\; A_\sigma^T C_\sigma B^T \;+\; A_\sigma C_\sigma F^T \tag{3-8}$$

A equação 3-8 fornece um método iterativo eficiente para encontrar C_σ, e a convergência também é garantida.

3.3
Estrutura de Influência

De forma a simplificar a equação 3-8, foi introduzido o operador *vec*. Esse operador transforma uma matriz em um vetor, empilhando as colunas da matriz uma sobre a outra formando uma única coluna.

Como exemplo, suponhamos uma matriz U_{pq}, então $vec(U_{pq})$ é

$$vec\left(\left[\begin{array}{ccc} u_{11} & \dots & u_{1q} \\ & \vdots & \\ u_{p1} & \dots & u_{pq}x \end{array}\right]\right) = \left[\begin{array}{c} u_{11} \\ \vdots \\ u_{p1} \\ \vdots \\ u_{1q} \\ \vdots \\ u_{pq} \end{array}\right] \tag{3-9}$$

Portando, a equação 3-8 pode ser reescrita como

$$vec(C_\sigma) \;\propto\; vec(A_\sigma^T C_\sigma B^T + A_\sigma C_\sigma F^T)$$

e então

$$vec(C_\sigma) \;\propto\; vec(A_\sigma^T C_\sigma B^T) + vec(A_\sigma C_\sigma F^T)$$

$$vec(C_\sigma) \;\propto\; (B \otimes A_\sigma^T)vec(C_\sigma) + (F \otimes A_\sigma)vec(C_\sigma)$$

Uma vez que $vec(X+Y) = vec(X) + vec(Y)$ e $vec(XYZ) = (Z^T \otimes$

$X)vec(Y)$ são propriedades conhecidas do operador (SEARLE, 1982).

Assim, reescrevemos a equação 3-8 com a seguinte relação de autovetores

$$vec(C_\sigma) \ \propto \ M_\sigma.vec(C_\sigma) \tag{3-10}$$

onde a matriz M_σ é definida por

$$M_\sigma \ = \ (B \otimes A_\sigma^T) + (F \otimes A_\sigma) \tag{3-11}$$

é chamada de *Matriz de Influência*.

Então, a *Matriz de Influência* revela a combinação de duas fontes de mútuo reforço, F e B conforme definição da secção anterior. Este fato é particularmente útil quando são investigados aspectos teóricos do modelo.

A seguir são realçados dois casos especiais em que se tem a convergência para a iteração 3-8.

3.4
Reforço Simétrico

No caso de *Reforço Simétrico Mútuo*, temos que

$$B_{vu} = F_{uv}$$

para todo u e v. Então, temos que $B = F^T$.

Agora, a equação 3-11 pode ser simplificada para

$$M_\sigma \ = \ (F \otimes A_\sigma)^T + (F \otimes A_\sigma) \tag{3-12}$$

Por consequência, verificamos facilmente que a matriz M é simétrica nesse caso. O *Método da Potência* provê um algoritmo simples para calcular o maior autovalor e o correspondente autovetor da matriz M_σ. Portanto, a iteração 3-8 converge e generaliza a proposta inicial de Jon Kleinberg.

Finalmente, apresentamos uma proposição que caracteriza o reforço simétrico, qual seja:

Proposição Vamos assumir que $A_\sigma \neq A_\sigma^T$. Então, a matriz de influência M_σ é simétrica *se somente se* $B = F^T$.

Prova.: A suficiência da condição é objeto da discussão a seguir. Para

provar que a condição é necessária, observe que

$$M_\sigma = M_\sigma^T$$

implicando em

$$(B \otimes A_\sigma^T) + (F \otimes A_\sigma) \;=\; (B \otimes A_\sigma^T)^T + (F \otimes A_\sigma)^T$$

que é

$$(B - F^T) \otimes A_\sigma^T \;=\; (B^T - F) \otimes A_\sigma$$

A partir da definição do produto direto, seguimos que para todos os pares de páginas r e s e para os pares de graus de classificação i e j temos

$$(B_{ij} - F_{ji}).(A_{rs})_\sigma \;=\; (B_{ji} - F_{ij}).(A_{sr})_\sigma$$

Por suposição, temos um particular r e s, tal que $(A_{rs})_\sigma = 1$ e $(A_{sr})_\sigma = 0$.

Assim, obtemos

$$(B_{ij} - F_{ji}).1 \;=\; (B_{ji} - F_{ij}).0$$

Portanto, $B_{ij} = F_{ji}$ para todos os pares de graus de classificação i e j, o que completa a prova.

3.5
Reforço Positivo

Outro aspecto interessante do modelo *XHITS* se dá quando todos os elementos de B_{uv} e F_{uv} são positivos. Tal aspecto é denominado de *Reforço Positivo*.

Firme em tal premissa, como A_σ é a matriz de adjacência e possui apenas elementos que são inteiros positivos, observamos que a matriz de influência M_σ é também positiva.

Nesse caso, o teorema de *Perron-Frobenius* afirma que o maior autovalor é positivo e existe também um autovetor correspondente com os valores das

suas coordenadas também positivas. Tal assertiva é suficiente para garantir a convergência da iteração 3-8.