Unidad 3: Análisis de vínculos

Lidiando con *spam*

Blanca Vázquez y Gibran Fuentes-Pineda

11 de octubre de 2022

Link spam

- PageRank hizo que las estrategias para para incrementar la relevancia a través de términos fuera mucho menos efectiva.

- Se crearon estrategias dirigidas a PageRank que usan *link spam*.

- Un conjunto de páginas que tienen el propósito de incrementar el PageRank de otra página se conoce como granja de *spam*.

Organización de una granja de *spam*

- Para creadores de granjas de *spam* la Web se divide en:
  1. Páginas inaccesibles: las que no pueden afectar
  2. Páginas accesibles: las que sí pueden afectar, aún cuando no estén directamente en su control
  3. Páginas propias: las que están en su control

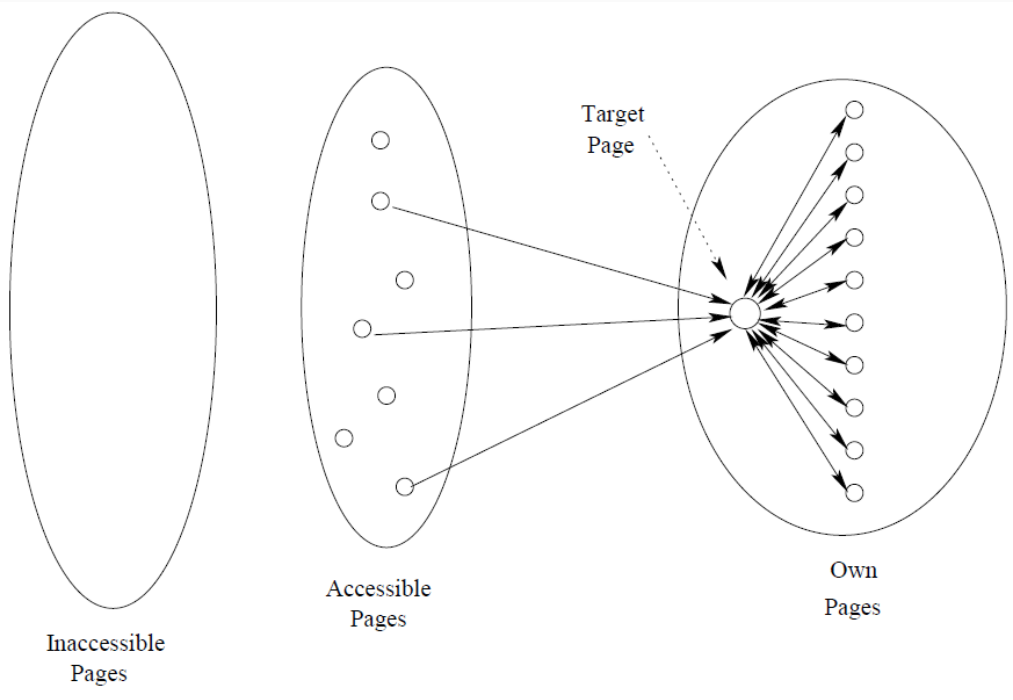

Imagen tomada de Leskovec et al. Mining of Massive Datasets, 2014.

Páginas accesibles

- Sin vínculos entrantes de páginas de fuera, las páginas de la granja ni siquiera se indizarían
- Sección de comentarios de distintos sitios pueden servir para crear estos vínculos

Páginas de soporte

- En una granja de *spam* hay:
  - Una página objetivo t a la que se busca aumentar el PageRank
  - m páginas de soporte

- Las páginas de soporte
  - Ayudan a incrementar el PageRank
  - Tienen vínculos a t y viceversa
  - En conjunto acumulan la porción del PageRank asociada a la teletransportación aleatoria

PageRank de una granja de *spam* (1)

- Supongamos que hay n páginas en la Web
  - Una es la página objetivo t
  - m son la páginas de soporte
  - p son páginas accesibles
  - x la cantidad de PageRank que contribuyen las páginas accesibles
- PageRank y de t se obtiene de 3 fuentes
  1. La contribuciǫn de x
  2. β veces el PageRank de cada página de soporte

$$\beta \cdot \left[\frac{\beta \cdot y}{m} + \frac{1-\beta}{n} \right]$$

  3. La cantidad de $\frac{1-\beta}{n}$ asociada a t (despreciable)

PageRank de una granja de *spam* (2)

- El PageRank de t es

$$y = x + \beta \cdot m \cdot \left[\frac{\beta \cdot y}{m} + \frac{1-\beta}{n}\right]$$
$$= x + \beta^2 \cdot y + \beta \cdot [1-\beta]\,\frac{m}{n}$$

- Resolviendo la ecuación para y

$$y = \frac{x}{1-\beta^2} + c \cdot \frac{m}{n}$$

donde

$$c = \beta \cdot \left[\frac{1-\beta}{1-\beta^2}\right]$$
$$= \frac{\beta}{1+\beta}$$

Ejemplo de PageRank de una granja de *spam*

- Si $\beta = 0.85$

$$\frac{1}{1-\beta^2} = 3.6$$

$$\underbrace{\frac{\beta}{1+\beta}}_{c} = 0.46$$

- Por lo tanto

$$y = 3.6 \cdot x + 0.46 \cdot \frac{m}{n}$$

Estrategias para combatir el *link spam*

- TrustRank: extensión de PageRank sensible al tópico diseñado para que disminuya la relevancia de páginas *spam*

- *Spam mass*: medida relacionada al impacto de *link spam* en la relevancia de una página.

TrustRank

- Es un PageRank sensible al tópico, donde el tópico es un conjunto de páginas confiables
  - Es muy poco probable que una página confiable tenga vínculos de salida a una página *spam*
  - Blogs y otros sitios con sección de comentarios no se consideran confiables

- Para seleccionar páginas confiables
  - Humanos examinan un conjunto de páginas y deciden cuál es confiable
  - Se toman dominios controlados

Spam mass

- Mide qué fracción de una página t proviene de *spam*
  - Se realiza calculando tanto el PageRank r como TrustRank s
- El *spam mass* de t es

$$\text{SpamMass}(t) = \frac{r - s}{r}$$

- Valores pequeños negativos o positivos de SpamMass indican que t probablemente no es una página *spam*
- Se eliminan las páginas con un alto SpamMass