

Análisis de Vínculos

Alondra Berzunza



Link Spam

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

Se crearon estrategias dirigidas a PageRank que usan link spam.

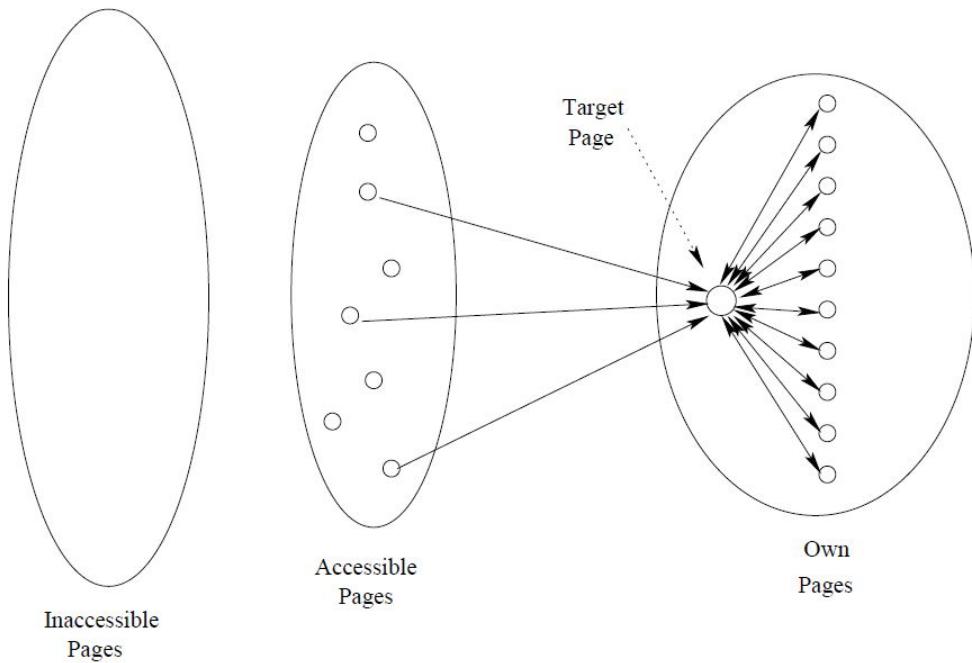
Un conjunto de páginas que tienen el propósito de incrementar el PageRank de otra página se conoce como granja de enlaces o de spam.



Granja de Enlaces

Para creadores de granjas de enlaces la Web se divide en:

- Páginas inaccesibles: las que no pueden afectar
- Páginas accesibles: las que sí pueden afectar, aún cuando no estén directamente en su control
- Páginas propias: las que están en su control.





Páginas accesibles

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



Páginas de Soporte

En una granja de enlaces 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.



Page Rank de una Granja de Enlaces

Supongamos que hay n páginas en la Web:

- Una es la página objetivo t .
- m son las páginas de soporte.
- p son las páginas accesibles.
- x la cantidad de PageRank que contribuyen las páginas accesibles.

PageRank y de t se obtiene de 3 fuentes

1. La contribución de x .
2. β veces el PageRank de cada página de soporte.
3. La cantidad de $(1 - \beta)/n$ asociada a t (despreciable).

El Page Rank de t es:



Estrategias para Combatir el Link Spam

- TrustRank: extensión de PageRank sensible al tema diseñado para que disminuya la relevancia de páginas spam.
- Spam mass: medida relacionada al impacto de link spam en la relevancia de una página.



Trust Rank

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.



Trust Rank

Fue desarrollado por la Universidad de Stanford y Yahoo! y funciona partiendo de un conjunto de sitios web "semilla" (de alta credibilidad) y midiendo la distancia en enlaces hasta otras páginas para determinar su autoridad y si son de confianza.

¿Cómo Funciona?

1. Sitios Semilla.
2. Análisis de Enlaces.
3. Detección de Spam.
4. Clasificación.



SpamMass

Se define como "la medida del impacto del spam de enlaces en el ranking de una página". El concepto fue desarrollado por Zoltán Gyöngyi y Héctor García-Molina, de la Universidad de Stanford, en colaboración con Pavel Berkhin y Jan Pedersen, de Yahoo !.

Este artículo amplía la metodología TrustRank propuesta.

Desarrollaron un núcleo de datos válidos y uno de datos incorrectos de documentos web seleccionados , a partir de los cuales midieron la masa de spam en una colección de documentos.



SpamMass

Se utilizan dos tipos de mediciones, masa absoluta y masa relativa , para comparar grupos de documentos. Cuanto mayor sea la masa, mayor será la probabilidad de que los documentos sean considerados spam.

Se utiliza un valor umbral para identificar grupos de documentos como spam. Si su masa relativa supera este umbral, los documentos se consideran spam. Se aplica un segundo umbral para los valores de PageRank de los documentos seleccionados. Solo los documentos con un PageRank alto se etiquetan como spam.



SpamMass

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



Práctica

- Descarga el [grafo de la Web de Stanford](#), programa el algoritmo de PageRank usando MapReduce y calcula las relevancias de las páginas del grafo. Para este programa considera que el vector de relevancia r cabe en la memoria principal de los nodos que ejecutan las tareas. Compara tus resultados con la función [pagerank](#) de la biblioteca [NetworkX](#).
- Punto extra: programa el algoritmo de PageRank usando MapReduce, considerando que el vector de relevancia r no cabe en la memoria principal de los nodos que ejecutan las tareas.