

UNIDAD 3: ANÁLISIS DE VÍNCULOS

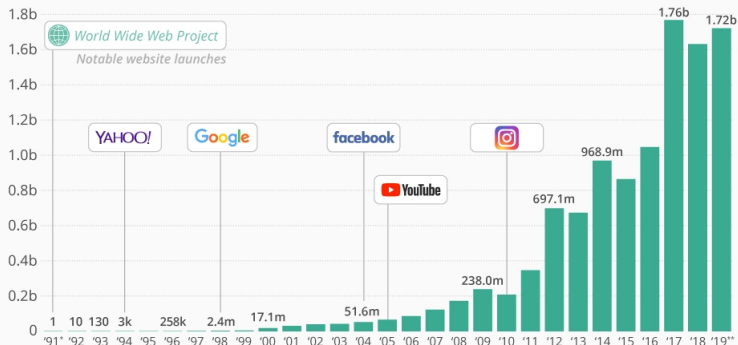
PAGE RANK SENSIBLE AL TÓPICO

Blanca Vázquez y Gibran Fuentes-Pineda

Octubre 2020

How Many Websites Are There?

Number of websites online from 1991 to 2019



"Website" is defined as a unique hostname, i.e. a name which can be resolved, using a name server, into an IP Address.



* As of August 1, 1991

** As of October 28, 2019 at 10:00 CET

Source: Internet Live Stats

- *¿Cómo podemos encontrar exactamente lo que queremos en la web de una manera rápida y eficiente?*
- Cada buscador necesita clasificar las páginas web, ¿cómo?
 - Entre más grande el valor de rango ...
 - ¿significa que la página tiene más contenido?
 - ¿significa que tiene más palabras de consulta frecuentes?
 - ¿significa que es más importante?

- Encontrar información útil en miles, millones de páginas web es un ¡reto!.
- Casi el 90 % del tráfico hacia la mayoría de los sitios web se encuentra en los buscadores.

- Conteo
- Algoritmo de HITS (Hyperlink-Induced Topic Search)
- PageRank
- SALSA (Stochastic Approach for Link Structure Analysis)
- TrustRank

... y muchas variantes de estos...!


CRÍTICAS A LAS SOLUCIONES EXISTENTES

	HITS	PageRank
Ventajas	<ul style="list-style-type: none">- Simple e iterativo- Puntuación específica de la consulta	<ul style="list-style-type: none">- Poco costoso (en tiempo de ejecución)- Las puntuaciones se calculan utilizando el grafo completo- El algoritmo puede ser personalizable
Desventajas	<ul style="list-style-type: none">- Costoso (tiempo de ejecución)- Las puntuaciones se calculan utilizando un subgrafo a partir de todo el grafo.	<ul style="list-style-type: none">- La puntuación es independiente de la consulta- El algoritmo es propenso a manipulaciones (granjas de enlaces)

- TSPR son las siglas de Topic-Sensitive PageRank
- Propuesto por Taher H. Haveliwala, Stanford University, 2003
- Es la versión personalizada de *Page Rank*.
- En lugar de calcular un solo vector de rango, ¿por qué no calcular un conjunto de vectores de rango (uno por cada tópico)?

- Usa el proyecto *Open Directory Project* como fuente de selección de tópicos ***<http://www.dmoz.org>***
 - También conocido como *DMoz* por *directory.mozilla.org*
- Es una colección de páginas web clasificadas por humanos.
- Consta de 16 tópicos (deportes, medicina, etc.)

INTRODUCCIÓN AL ALGORITMO TSPR

 open directory project In partnership with
Aol Search.

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)


[advanced](#)

<u>Arts</u> Movies, Television, Music...	<u>Business</u> Jobs, Real Estate, Investing...	<u>Computers</u> Internet, Software, Hardware...
<u>Games</u> Video Games, RPGs, Gambling...	<u>Health</u> Fitness, Medicine, Alternative...	<u>Home</u> Family, Consumers, Cooking...
<u>Kids and Teens</u> Arts, School Time, Teen Life...	<u>News</u> Media, Newspapers, Weather...	<u>Recreation</u> Travel, Food, Outdoors, Humor...
<u>Reference</u> Maps, Education, Libraries...	<u>Regional</u> US, Canada, UK, Europe...	<u>Science</u> Biology, Psychology, Physics...
<u>Shopping</u> Clothing, Food, Gifts...	<u>Society</u> People, Religion, Issues...	<u>Sports</u> Baseball, Soccer, Basketball...
<u>World</u> Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...		

Help build the largest human-edited directory of the web

Copyright © 2013 Netscape

5,292,737 sites - 99,943 editors - over 1,020,828 categories



Open Directory Project

- Supongamos que creamos un vector uno para cada tópico usando Page Rank.
- Si se pudiera determinar / saber ¿cuál de estos tópicos son de interés para el usuario?, entonces :
 - Se podría usar el vector de Page Rank de ese tópico cuando se clasifiquen las páginas por relevancia | búsqueda.

Su formulación es similar a la de Page Rank:

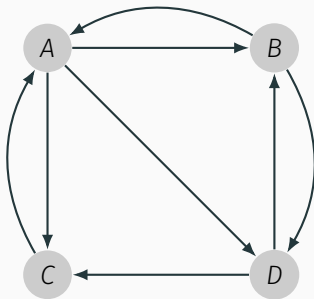
$$v' = \beta Mv + (1 - \beta)e_S/|S|$$

dónde:

- β es la probabilidad de elegir un vínculo de forma aleatoria
- M es la matriz de adyacencia
- v es el vector de Page Rank
- S indica la páginas que pertenecen a cierto tópico
- e_S es un vector que tiene 1s en los componentes S y 0s en el resto.
- $|S|$ es el tamaño del conjunto S .

EJEMPLO

Calcular el Page Rank sensible al t3pico, donde $\beta = 0.8$ y $S = \{B, D\}$

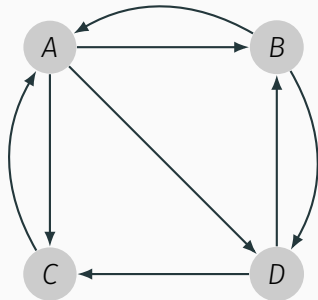


EJEMPLO

$$v' = \beta Mv + (1 - \beta)e_s/|S|$$

Paso 1: matriz de adyacencia

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$



$$v' = \beta Mv + (1 - \beta)e_s/|S|$$

Paso 2: matriz de adyacencia * β

$$\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} * 0.8 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix}$$

$$v' = \beta Mv + (1 - \beta)e_S/|S|$$

Paso 3: resolver $(1 - \beta)e_S/|S|$

$$(1 - 0.8) \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} / 2 = \frac{1}{5} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} / 2 = \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$v' = \beta Mv + (1 - \beta)e_S/|S|$$

Unimos los resultados previos

$$v' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

Cálculo de las primeras iteraciones: t_0

$$t_0 = \begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix}$$

Recordemos, solo aplica en los nodos del conjunto S

Cálculo de las primeras iteraciones: t_1

$$t_1 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

$$t_1 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

EJEMPLO

Cálculo de las primeras iteraciones: t_1

$$t_1 = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}$$

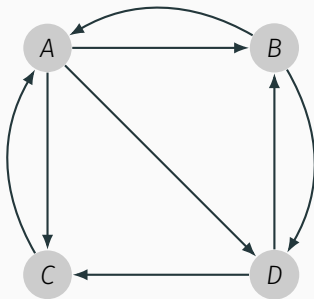
$$t_1 = \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix} = \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix}$$

Iteraciones

$$t_0 = \begin{bmatrix} 0/2 \\ 1/2 \\ 0/2 \\ 1/2 \end{bmatrix}, t_1 = \begin{bmatrix} 1/5 \\ 3/10 \\ 1/5 \\ 3/10 \end{bmatrix}, t_2 = \begin{bmatrix} 42/150 \\ 41/150 \\ 25/150 \\ 41/150 \end{bmatrix}, t_3 = \begin{bmatrix} 62/250 \\ 71/250 \\ 46/250 \\ 71/250 \end{bmatrix} \dots \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/210 \end{bmatrix}$$

EJERCICIO

Calcular el Page Rank sensible al t3pico, d3nde $\beta = 0.9$ y $S = \{A, C\}$



Para integrar, es necesario tener en cuenta:

1. Decidir sobre los tópicos para crear vectores de Page Rank especializados
2. Encontrar una manera de determinar el tópico o los tópicos que sean más relevantes
3. Usar los vectores de Page Rank de esos tópicos para responder la consulta del usuario.

¿CÓMO IDENTIFICAR LOS TÓPICOS?

- Permitir que el usuario los seleccione (usando un menú)
- Inferir los tópicos usando:
 - Las búsquedas previas del usuario.
 - La información del usuario (marcadores, Facebook).

- Ejemplo: las palabras **sarampión** y **gol** aparecen frecuentemente en las páginas web:
 - **Sarampión** — — — $> T_{medicina}$
 - **Gol** — — — $> T_{deportes}$
- Supongamos que identificamos las palabras más frecuentes de cada página.
- Supongamos que tomamos un conjunto de páginas especializadas de un cierto tópico, y extraemos las palabras más frecuentes.

- Sea $S_1, S_2 \dots S_k$ son el conjunto de palabras que definen cada tópico.
- Sea P el conjunto de palabras que aparecen en una página p .
- Calcular la medida de similitud de Jaccard entre P y en cada uno de S_i .
- Clasificar la página al tópico con mayor similitud.

INFERIR TÓPICOS BASADO EN PALABRAS

<p>BUSINESS</p> <p>“Recumbent Bikes and Kit Aircraft” www.rans.com</p> <p>www.BreakawayBooks.com java.oreilly.com/bite-size/ www.carbboom.com</p>	<p>COMPUTERS</p> <p>“GPS Pilot” www.gpspilot.com</p> <p>www.wireless.gr/wireless-links.htm www.linkstosales.com www.LiftExperts.com/lifts.html</p>
<p>GAMES</p> <p>“Definition Through Hobbies” www.flick.com/~gretchen/hobbies.html</p> <p>www.BellaOnline.com/sports/ www.npr.org/programs/wesun/puzzle/will.html www.trygve.com/</p>	<p>KIDS AND TEENS</p> <p>“Camp Shohola For Boys” www.shohola.com</p> <p>www.EarthForce.org www.WeissmanTours.com www.GrownupCamps.com/homepage.html</p>
<p>RECREATION</p> <p>“Adventure travel” www.gorp.com/</p> <p>www.GrownupCamps.com/homepage.html www.gorp.com/gorp/activity/main.htm www.outdoor-pursuits.org/</p>	<p>SCIENCE</p> <p>“Coast to Coast by Recumbent Bicycle” hypertextbook.com/bent/</p> <p>www.SiestaSoftware.com/ www.BenWiens.com/benwiens.html www.SusanJeffers.com/jeffbio.htm</p>
<p>SHOPPING</p> <p>“Cycling Clothing & Accessories for Women” www.TeamEstrogen.com/</p> <p>www.ShopOutdoors.com/ www.jub.com.au/books/ www.bike.com/</p>	<p>SPORTS</p> <p>“Swim, Bike, Run, & Multisport” www.multisports.com/</p> <p>www.BikeRacing.com/ www.CycleCanada.com/ www.bikescape.com/photogallery/</p>

Consulta: *bicycling*

Imagen tomada de: Topic-Sensitive PageRank, Haveliwala, 2002.

INFERIR TÓPICOS BASADO EN PALABRAS

computer vision	
COMPUTERS	0.24
BUSINESS	0.14
REFERENCE	0.09

gardening	
HOME	0.63
SHOPPING	0.14
REGIONAL	0.04

java	
COMPUTERS	0.53
GAMES	0.10
KIDS & TEENS	0.06

national parks	
REGIONAL	0.42
RECREATION	0.16
KIDS & TEENS	0.09

cruises	
RECREATION	0.65
REGIONAL	0.18
SPORTS	0.04

graphic design	
COMPUTERS	0.36
BUSINESS	0.23
SHOPPING	0.09

lipari	
HOME	0.19
KIDS & TEENS	0.17
NEWS	0.13

parallel architecture	
COMPUTERS	0.70
SCIENCE	0.10
REFERENCE	0.07

death valley	
REGIONAL	0.28
SOCIETY	0.14
NEWS	0.10

gulf war	
SOCIETY	0.21
KIDS & TEENS	0.18
REGIONAL	0.17

lyme disease	
HEALTH	0.96
REGIONAL	0.01
RECREATION	0.01

recycling cans	
HOME	0.42
BUSINESS	0.38
KIDS & TEENS	0.06

Define la $P[c|q]$

Imagen tomada de: Topic-Sensitive PageRank, Haveliwala, 2002.

INFERIR TÓPICOS BASADO EN PALABRAS

ARTS
Britannica Online www.britannica.com
BandHunt.com Genres (Music) www.bandhunt.com/genres.html
Artist Information (Music) www.artistinformation.com/index.html
Billboard.com (Music charts) www.billboard.com
Soul Patrol (Music) www.soul-patrol.com

HEALTH
Northern County Psychiatric Associates News www.baltimorepsych.com/news.htm
Seasonal Affective Disorder www.ncpamd.com/seasonal.htm
Women's Mental Health www.ncpamd.com/Women's_Mental_Health.htm
Wing of Madness Depression Support Group www.wingofmadness.com
Country Nurse Online www.countrynurse.com

Resultado de la búsqueda de *blues*
Imagen tomada de: Topic-Sensitive PageRank, Haveliwala, 2002.