Una aplicación de la Teoría Ergódica en la búsqueda de Google

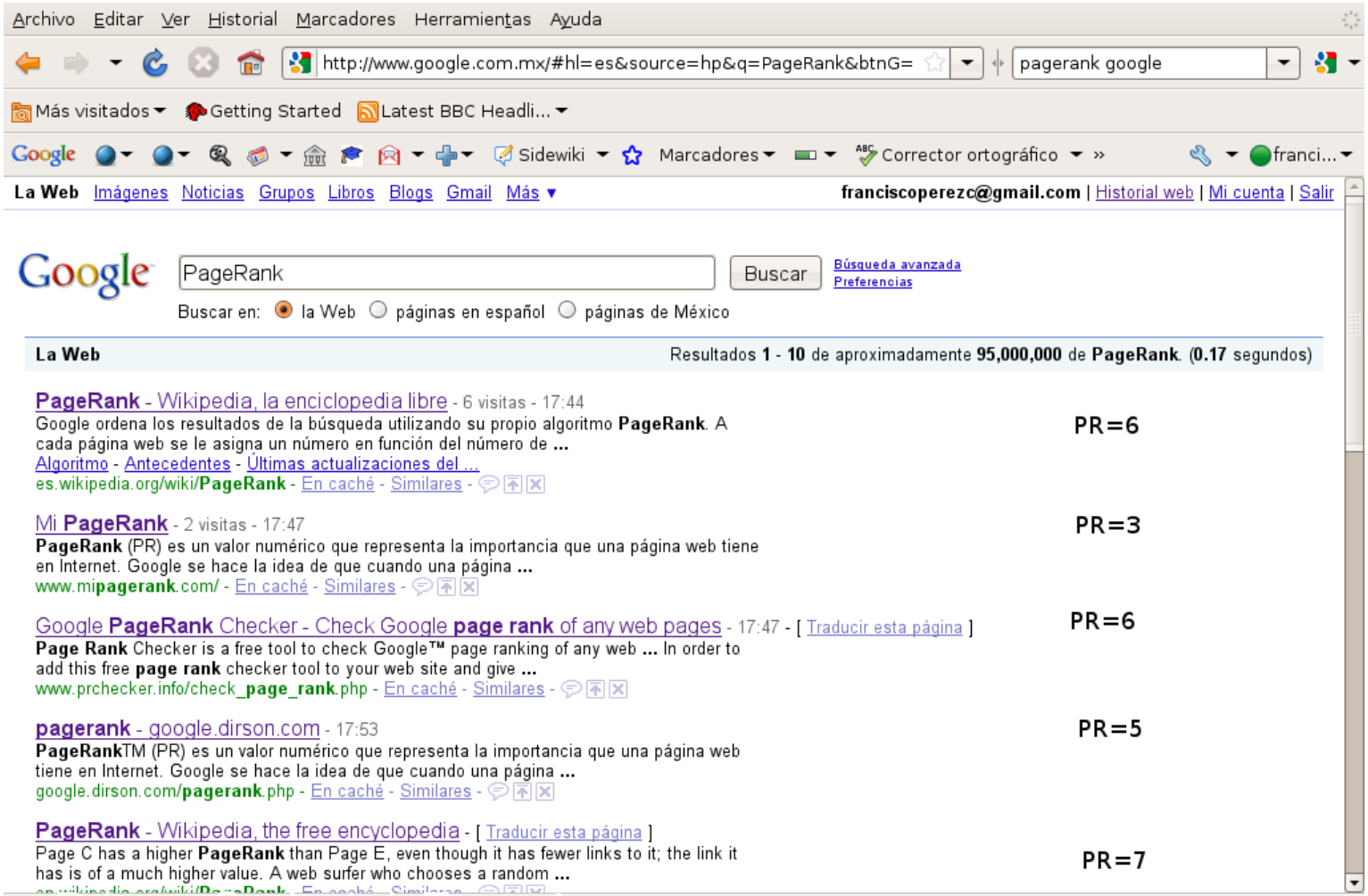
Pérez Carbajal Francisco

10 de octubre de 2009

Google


Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://www.google.com.mx/
Google
Más visitados
Getting Started
Latest BBC Headli...
La Web Imágenes Noticias Grupos Libros Blogs Gmail Más
iGoogle | Acceder
Google
México
Búsqueda avanzada
Preferencias
Herramientas del idioma
Buscar con Google
Voy a tener suerte
Buscar en: la Web páginas en español páginas de México
Programas de publicidad - Soluciones Empresariales - Todo acerca de Google - Google.com in English
©2009 - Privacidad

¿Qué es el PageRank?



PageRank de la página de wikipedia

Archivo Editar Ver Historial Marcadores Herramientas Ayuda

http://es.wikipedia.org/wiki/PageRank

Más visitados Getting Started Latest BBC Headli...

Probar Beta Registrarse/Entrar

artículo discusión editar historial

WIKIPEDIA
La enciclopedia libre

buscar

Ir Buscar

navegación

- Portada
- Portal de la comunidad
- Actualidad
- Cambios recientes
- Página aleatoria
- Ayuda
- Donaciones

imprimir/exportar

- Crear un libro
- Descargar como PDF
- Versión para imprimir

herramientas

- Lo que enlaza aquí
- Cambios en enlazadas

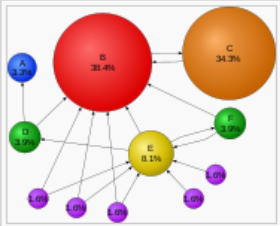
PageRank

PageRank es una marca registrada y patentada[1] por Google el 9 de enero de 1999 que ampara una familia de algoritmos utilizados para asignar de forma numérica la relevancia de los documentos (o páginas web) indexados por un motor de búsqueda. Sus propiedades son muy discutidas por los expertos en optimización de motores de búsqueda. El sistema PageRank es utilizado por el popular motor de búsqueda Google para ayudarle a determinar la importancia o relevancia de una página. Fue desarrollado por los fundadores de Google, Larry Page y Sergey Brin, en la Universidad de Stanford.

PageRank confía en la naturaleza democrática de la web utilizando su vasta estructura de enlaces como un indicador del valor de una página en concreto. Google interpreta un enlace de una página **A** a una página **B** como un voto, de la página **A**, para la página **B**. Pero Google mira más allá del volumen de votos, o enlaces que una página recibe; también analiza la página que emite el voto. Los votos emitidos por las páginas consideradas "importantes", es decir con un PageRank elevado, valen más, y ayudan a hacer a otras páginas "importantes". Por lo tanto, el PageRank de una página refleja la importancia de la misma en Internet.

Contenido [ocultar]

1 Algoritmo
  1.1 Manipulación
2 Antecedentes
3 Últimas actualizaciones del PageRank
4 Referencias
5 Bibliografía
6 Enlaces externos

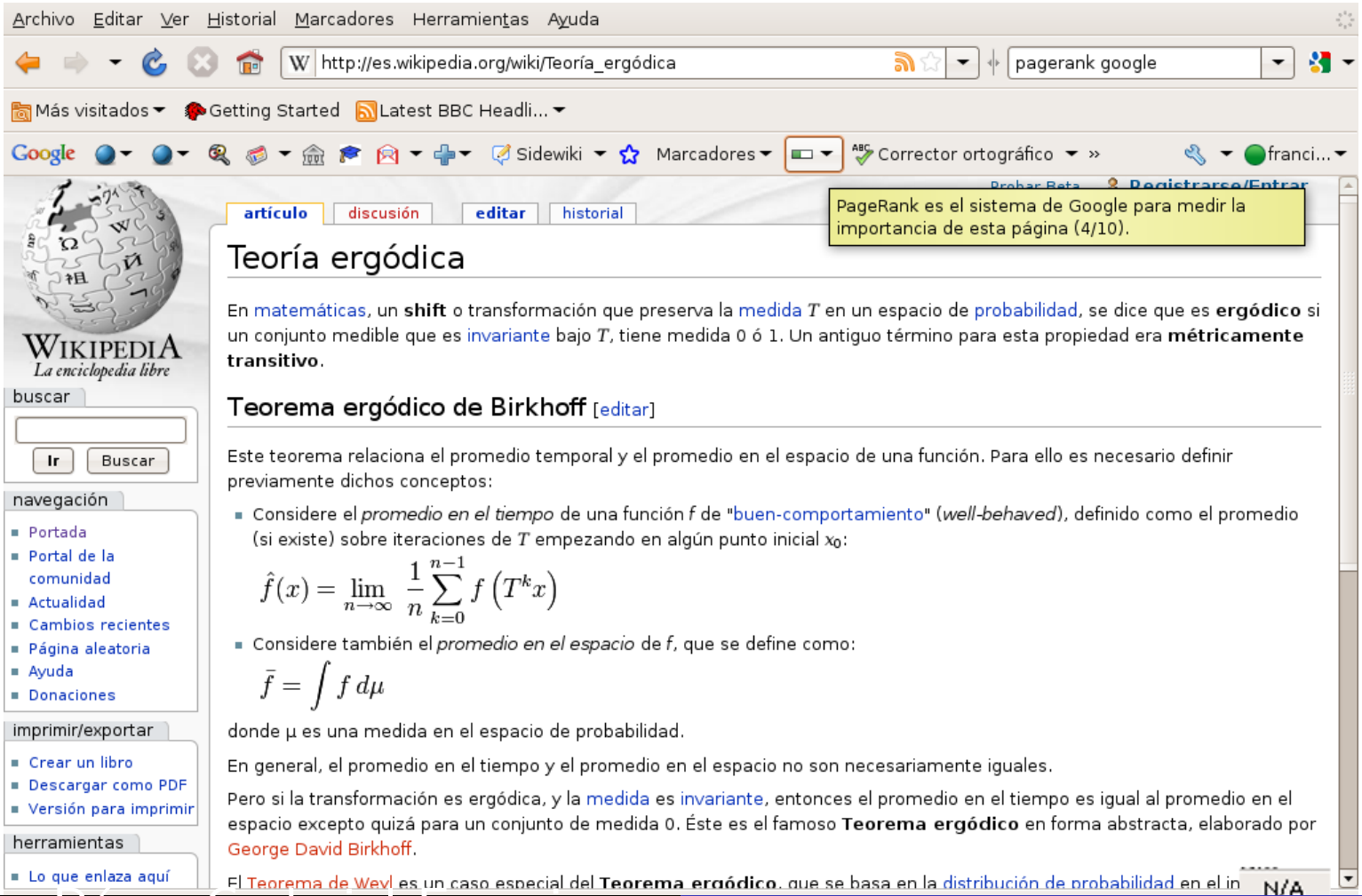


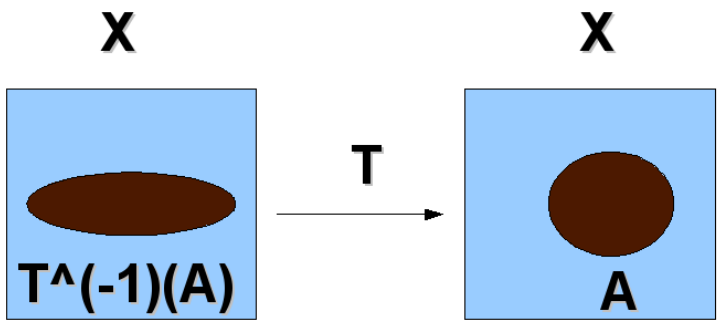
Google ordena los resultados de la búsqueda utilizando su propio algoritmo PageRank. A cada página web se le asigna un número en función del número de enlaces de otras páginas que la apuntan, el valor de esas páginas y otros criterios no públicos.

Algoritmo [editar]

6/10

¿Qué es la Teoría Ergódica?





La Teoría Ergódica es el estudio de transformaciones preservadoras de medida sobre un espacio de probabilidad.

Internet

Internet contiene una gran cantidad de información. Buscarla es como buscar un libro en una biblioteca gigantesca que no tiene catálogo.

Esto nos lleva al problema de búsqueda de información en Internet, éste hizo que aparecieran los motores de búsqueda. Uno de los más utilizados y eficaces es el de Google.

Motor de búsqueda de Google

El motor de búsqueda de Google fue inventado por Sergey Brin y Lawrence Page ambos obtuvieron su doctorado en computo científico por parte de la Universidad de Stanford, actualmente su empresa es una de las más competitivas del mercado.

Una de las principales diferencias del motor de búsqueda de Google con otros motores, es la siguiente idea: *Para una búsqueda típica existen aproximadamente 10 mil páginas web, sin embargo el usuario solo revisara 30 de esas páginas*

Problema

No hace mucho cuando buscabas la palabra Internet, el primer resultado de la busqueda era una página en chino que no tenía otra palabra mas en inglés que Internet. Lo que nos lleva a la siguiente conclusión:

El orden de las páginas web que se presentan como resultado de una búsqueda es importante

Lo anterior llevo a Sergey Brin y Lawrence Page a inventar el famoso PageRank. La ideas básicas detrás de éste son las siguientes:

(a) Google interpreta un hyperlink de la página A a la página B como un voto de la página A hacia la página B

(b) No es lo mismo tener un voto de una página reconocida en el tema que un voto de una página que no tiene nada que ver con el tema.

¿Qué es una Cadena de Markov estacionaria?

Google utiliza Cadenas de Markov para el cálculo del PageRank.

Definición

Sea $(\Omega, \mathscr{F}, P)$ un espacio de probabilidad, una **Cadena de Markov** es una colección de variables aleatorias $\{X_n : n = 0, 1, \dots\}$ que toman valores en un conjunto $E = \{0, 1, \dots, N\}$ y que satisface la propiedad de Markov, es decir, que para todo $n \geq 0$ y para cualesquiera $x_0, \dots, x_{n+1} \in E$ se cumple lo siguiente:

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

Se dice que una cadena de Markov es **estacionaria u homegenea** si la probabilidad $P(X_{n+1} = j | X_n = i)$ no de pende de n para todo $i, j \in E$.

Matríz de Transición

Definición

Consideremos una Cadena de Markov estacionaria y $E = \{0, \ldots, N\}$ el conjunto donde toma sus valores. La matriz P dada por $P(i,j) = P(X_1 = j | X_0 = i)$ con $i, j \in E$, la llamaremos **matríz de probabilidades de transición**. Denotaremos a $P(i,j) = p_{ij}$. Esta matríz cumple las siguientes dos propiedades:

(a) $p_{ij} \geq 0$ para todo $i, j \in E$

(b) $\sum_{j \in E} p_{ij} = 1$ para toda $i \in E$

A una matríz A cuadrada que cumpla las dos propiedades anteriores la llamaremos **matriz estocástica**.

Problema del cálculo

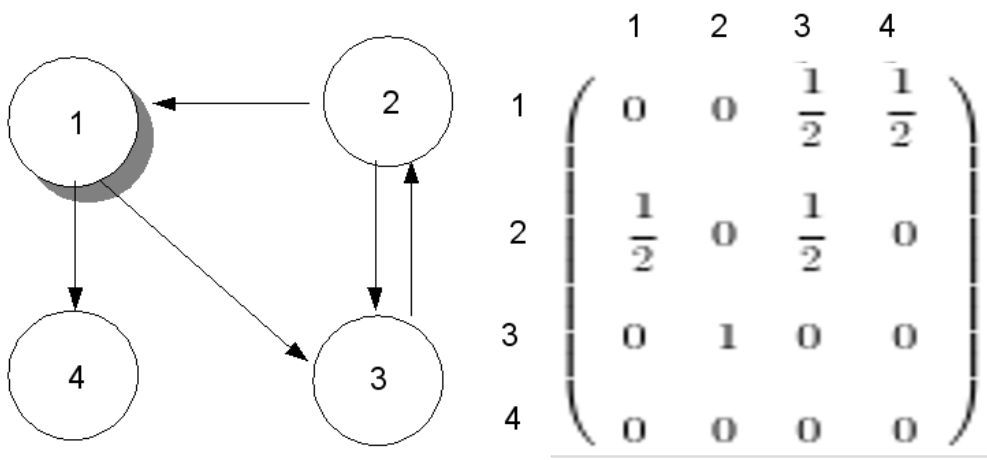
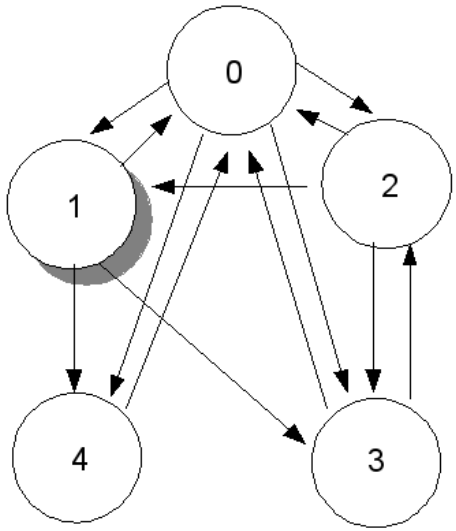


Figura: La matríz asociada a la gráfica

Para calcular el PageRank Google modela el flujo de internet asignandole una matriz.

Una posible solución



$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ ,15 & 0 & 0 & \frac{,85}{2} & \frac{,85}{2} \\ ,15 & \frac{,85}{2} & 0 & \frac{,85}{2} & 0 \\ ,15 & 0 & ,85 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Figura: La matríz estocástica asociada a la gráfica

Observemos que esta matriz es estocástica.

$$P^3 = \begin{pmatrix} 0{,}1859 & 0{,}1809 & 0{,}2712 & 0{,}2261 & 0{,}1358 \\ 0{,}1086 & 0{,}2917 & 0{,}1541 & 0{,}3076 & 0{,}1381 \\ 0{,}3163 & 0{,}2013 & 0{,}2173 & 0{,}2173 & 0{,}0478 \\ 0{,}1627 & 0{,}0478 & 0{,}3708 & 0{,}2173 & 0{,}2013 \\ 0{,}3625 & 0{,}1062 & 0{,}2125 & 0{,}2125 & 0{,}1062 \end{pmatrix}$$

Figura: P es la matriz anterior

Esta nueva matriz obtenida de elevar a la tercera potencia la matríz anterior tiene todas sus entradas esctrictamente positivas.

¿Cuál es el PageRank?

Hemos visto que Google utiliza cadenas de Markov para modelar el flujo de Internet, más aún como la matriz elevada a la tercera potencia tiene sus entradas estrictamente positivas, el *Teorema de existencia y unicidad de una distribución inicial estacionaria para Cadenas de Markov* nos garantiza la existencia y únicidad de un vector $\pi = [\pi_0, \ldots, \pi_4]$ tal que $\sum_{j=0}^{4} \pi_j = 1$ y $\pi P = \pi$. Google interpreta a π_i como el PageRank de la página i.

Generalización

Supongamos que existen N páginas web, pensemos a estas N páginas como un conjunto de vértices de una gráfica G a un hyperlink de la página i a la página j como una arista del vértice i al vértice j. A los vértices los denotaremos con números enteros $k \in \{1, 2, \ldots, N\}$

Cadena de Markov

Sea $\overline{G}$, la gráfica que se obtiene de G añadiendole un vértice 0 con artistas de este vértice a todos los demás vértices y viceversa. Sean $a_{ij} = 1$ si existe una arista del vértice i al vértice j en $\overline{G}$ y $C(i) =$número de aristas que salen del vértice i, notemos que $C(i) > 0$ para toda $i \in \overline{G}$. Fijemos un parámetro $d \in (0, 1)$ (por ejemplo $d = ,85$). Sea P la matriz dada como sigue: $p_{ii} = 0$ $\forall\ i \geq 0$, para $i, j > 0$ con $i \neq j$ sean $P_{0i} = \frac{1}{N}$ y

$$P_{i0} = \begin{cases} 1 & \text{si} \quad C(i) = 1 \\ \\ 1 - d & \text{si} \quad C(i) \neq 1 \end{cases} \qquad P_{ij} = \begin{cases} 0 & \text{si} \quad a_{ij} = 0 \\ \\ \dfrac{d}{C(i) - 1} & \text{si} \quad a_{ij} = 1 \end{cases}$$

Ejemplo de P

$$P = \begin{pmatrix} 0 & \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} & \frac{1}{N} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 1-d & \frac{d}{C(2)-1} & 0 & \cdots & 0 & \frac{d}{C(2)-1} \\ 1-d & \frac{d}{C(3)-1} & \frac{d}{C(3)-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1-d & \frac{d}{C(N)-1} & \frac{d}{C(N)-1} & 0 & \cdots & 0 \end{pmatrix}$$

Existencia y Unicidad de la Solución

Esta matriz P es irreducible y aperiodica, es decir, existe una $m \in \mathbb{N}$ tal que P^n tiene todas sus entradas estrictamente positivas para toda $n \geq m$.
Por lo que existe una única distribución inicial estacionaria $\pi = [\pi_0, \ldots, \pi_N]$ tal que $\pi P = \pi$ y $\sum_{j=0}^{N} \pi_j = 1$. Google interpreta a π_i como el PageRank de la página i.

Espacio de Probabilidad Asociado a (P, π)

Consideremos $E = \{0, 1, \ldots, N\}$ tomemos $\Omega = \prod_{-\infty}^{\infty} E = \{0, 1, 2, 3, 4\}^{\mathbb{Z}}$. Definimos una $\sigma-$álgebra de subconjuntos de Ω y una medida como sigue: Consideramos $n_1 < n_2 < \cdots < n_r \in \mathbb{Z}$ y $x_1, \ldots, x_r \in E$ fijos, denotamos:

$$Z\left(n_1, \ldots, n_r; x_1, \ldots, x_r\right) = \{\omega \in \Omega : \omega_{n_1} = x_1, \ldots, \omega_{n_r} = x_r\} \ (r \in \mathbb{N})$$

y lo llamamos un **cilindro**. Definimos a $\mathscr{F}$ como la $\sigma-$álgebra generada por todos los cilindros y

$$\widetilde{P}(Z(n_1, n_2, \ldots, n_k; x_1, x_2, \ldots, x_k)) = \pi_{x_1} p_{x_1 x_2}^{(n_2 - n_1)} \cdots p_{x_{k-1} x_k}^{(n_k - n_{k-1})}$$

donde $p_{ij}^{(r)}$ denota la componente (i, j) de la matríz P^r $(r \geq 1)$ y $\mu(Z(n; j)) = \pi_j \ \forall \ n \ \forall \ j \in E$. Finalmente extendemos a todo elemento de $\mathscr{F}$ para obtener el espacio de probabilidad $(\Omega, \mathscr{F}, \widetilde{P})$

T asociada al espacio de probabilidad

Definimos $T : \Omega \to \Omega$ poniendo $T(\omega) = \omega'$ donde $\omega'_i = \omega_{i+1}\ \forall i \in \mathbb{Z}$; "gráficamente":

$$\begin{array}{ccccccccccc} & & & & & \boxed{\text{lugar 0}} & & & & & \\ \omega & = & (\ldots & \omega_{-2}, & \omega_{-1}, & \boxed{\omega_0}, & \omega_1, & \omega_2, & \ldots) & & \\ & & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & & \downarrow T & \\ \omega' & = & (\ldots & \omega_{-1}, & \omega_0, & \omega_1, & \omega_2, & \omega_3, & \ldots) & & \end{array}$$

Figura: es decir T recorre a ω un lugar hacia la izquierda y cláramente es invertible (con $T^{-1}(\omega) = \omega'$ donde $\omega'_i = \omega_{i-1}\ \forall i \in \mathbb{Z}$).

A la cuarteta $(\Omega, \mathscr{F}, \widetilde{P}, T)$ se le conoce como el **corrimiento bilateral de Markov**.

Se puede probar que T es una transformación preservadora de medida en $(\Omega, \mathscr{F}, \widetilde{P})$ y además T es ergódica, es decir, que si $T^{-1}(F) = F$ para algún $F \in \mathscr{F}$ se tiene que $\widetilde{P}(F) = 0$ ó $\widetilde{P}(\Omega \setminus F) = 0$. También utilizando argumentos de Teoría Ergódica se puede probar el siguiente Teorema:

Teorema (Convergencia exponencial)

Sea (P, π) donde P es una matríz estocástica irreducible y aperiódica de tamao $N+1$, $\pi = [\pi_0, \ldots, \pi_N]$ un vector talque $\pi P = \pi$ y $\sum_{j=0}^{N} \pi_j = 1$, entonces existen constantes $k > 0$ y $\rho \in (0, 1)$ tales que $\left| p_{ij}^{(n)} - \pi_j \right| \leq k\rho^n$ para toda n y para todo $i, j \in \{0, \ldots, N\}$.

Donde $p_{ij}^{(n)}$ denota la componete (i, j) de la matríz P^n.

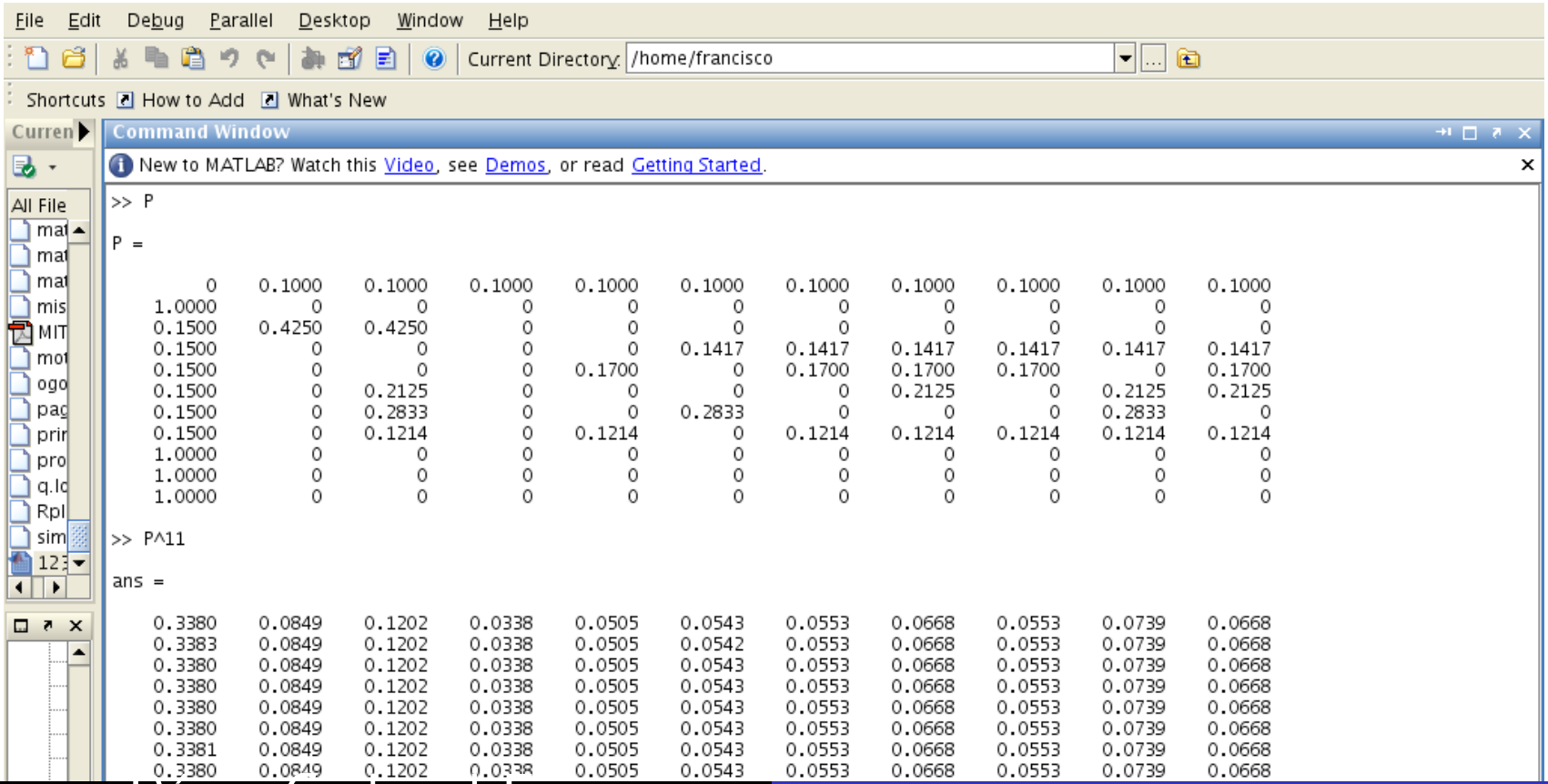
Aplicación

Aplicando el Teorema anterior a nuestra matriz P y a nuestro vector π, tenemos que para todo vector $\theta = [\theta_0, \ldots, \theta_N]$ tal que $\sum_{j=0}^{N} \theta_j = 1$ se tiene que la siguiente sucesión $\{\theta P^n\}$ converge de manera exponencial a π. Se puede obtener una buena aproximación de π utilizando $\theta = [1/(N+1), \ldots, 1/(N+1)]$ y evaluando θP^n para una n suficientemente grande.

Sin embargo, en la práctica no es tan sencillo como en la teoría. Actualmente existen alrededor de 1.7 billones de páginas web, tratar de realizar un cálculo de θP^n no es nada sencillo. Existen métodos númericos para realizar este cálculo, uno de ellos es conocido como el método de la potencia.

Ejemplo

Consideramos P dada por:



Ahora veamos la obtención de π