

Métodos Numéricos

TP2

6 de octubre de 2015

Years Later for Guillermo Vilas, He's Still Not the One



Integrante	LU	Correo electrónico
Martin Baigorria	575/14	martinbaigorria@gmail.com
Federico Beuter	827/13	federicobeuter@gmail.com
Mauro Cherubini	835/13	cheru.mf@gmail.com
Rodrigo Kapobel	695/12	rok.35@live.com.ar

Reservado para la cátedra

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

Resumen: Existe una gran polémica en torno a los rankings de la ATP entre los años 1975 y 1977. Para enriquecer la discusión, en este trabajo se decidió analizar y recomputar estos rankings utilizando PageRank. En un primer momento, se analiza la motivación original de este algoritmo y se discuten sus diferentes cualidades numéricas y métodos de calculo. Se discute la calibración del algoritmo y su robustez ante intentos de manipulación. Finalmente, concluimos que si la ATP hubiese utilizado PageRank, efectivamente Vilas hubiese sido líder del ranking de la ATP en el año 1977, convirtiéndose en el primer argentino en lograr semejante hazaña.

Keywords: PageRank, Método de la Potencia, ATP Ranking, Vilas.

Índice

1. Introduccion	3
2. PageRank	5
2.1. La motivación original: páginas web	5
2.1.1. Modelado	5
2.1.2. Propiedades	5
2.1.3. Existencia y Unicidad	6
2.2. Tenis	7
2.2.1. Modelado	7
2.2.2. Sistema de puntos y consideraciones	8
2.3. Computo: Método de la Potencia	9
2.3.1. Correctitud	9
2.3.2. Complejidad	9
2.4. Representación del grafo	9
2.4.1. Dictionary of Keys (<i>DOK</i>)	10
2.4.2. Compressed Sparse Row (<i>CSR</i>)	10
2.4.3. Compressed Sparse Column (<i>CSC</i>)	10
2.4.4. Elección de estructura	10
3. Experimentación	11
3.1. PageRank: Complejidad	11
3.2. Convergencia	12
3.3. Comparación PageRank vs In-Deg	13
3.4. Manipulación	14
3.5. Ranking ATP	14
3.5.1. Ranking ATP oficial vs. Ranking PageRank vs. Sort por diferencia de victorias/derrotas	15
4. Conclusiones	19
5. Apéndice A: Enunciado	20
6. Apéndice B: Código	25
6.1. system.cpp	25
6.2. matrix.h	33
6.3. sparseMatrix.h	39

1. Introduccion

El 25 de Mayo de 2015 el diario The New York Times publico un articulo titulado "Years Later for Guillermo Vilas, He's Still Not the One", donde se repasa el rendimiento del tenista argentino durante los años 1975/1976 y se discute el calculo del ranking de la ATP en ese momento. Aunque hoy en día Vilas es un ícono del tenis argentino, nunca logró estar en la cima del ranking de la ATP.



Figura 1: Guillermo Vilas after winning a tournament in Stockholm in 1975. A journalist has asserted that Vilas deserved to be ranked No. 1 during that year.

En 2016, un grupo de investigadores y periodistas deportivos argentinos decidieron analizar el ranking de la ATP en 1975 y 1977 para determinar si Vilas debió haber sido número 1. Dado que los rankings no se actualizaban constantemente en ese momento, los investigadores mostraron que de haberse actualizado de forma periódica, Vilas hubiese sido número 1 por durante 7 semanas entre 1975 y 1977.

Existen precedentes donde se actualizó un ranking de tenis de forma retroactiva. Este es el caso de la WTA, que determinó que Evonne Goolagong Cawley debió haber sido número 1 por dos semanas en 1976. Por esta razón el grupo de investigación argentino considera que revisar estos rankings no es un esfuerzo en vano. Cuando buscábamos los datos de la ATP entre 1975/1977, uno de los investigadores de este equipo que contactamos nos comento: "Es interesante tu decisión de indagar sobre el tema. Tal vez no estás al tanto del trabajo y lucha que estamos realizando contra la ATP, por el ranking de los 70 en el que perjudicaron a Vilas y muchísimos otros jugadores."

En ese momento, el calculo del ranking de la ATP era bastante rudimentario: "It was a system based on an average of a player's results, and it often rewarded top players who played fewer tournaments. Vilas was a workhorse, which is how he managed not to reach No. 1 in the ATP rankings in 1977, when he won the French Open, the United States Open and 14 other tournaments." [5].

Los métodos para calcular rankings no solo son relevantes para definir las posiciones de equipos y jugadores en eventos deportivos, sino que aparecen constantemente en todo tipo de situaciones donde se debe imponer algún tipo de orden. Este es el caso por ejemplo de los concursos docentes, donde se ponderan los diferentes antecedentes para decidir cual es el candidato *idoneo* para el puesto.

Otro caso sumamente relevante en cuanto algoritmos de ranqueo es el de los motores de búsqueda. Los motores de búsqueda deben encontrar alguna forma de ordenar de forma relevante los sitios web que están relacionados con una consulta. El caso icónico es el de Google con su algoritmo PageRank. Los buscadores antes de 1990 eran sumamente rudimentarios, utilizaban algoritmos de ranqueo vulnerables en el sentido que podían ser manipulados y no se explotaba gran parte de la estructura de la web. Esta fue una de las razones por las cuales una consulta no siempre devolvía resultados relevantes. Este fue el caso por ejemplo de algunos buscadores en ese momento como Yahoo! Search o AltaVista.

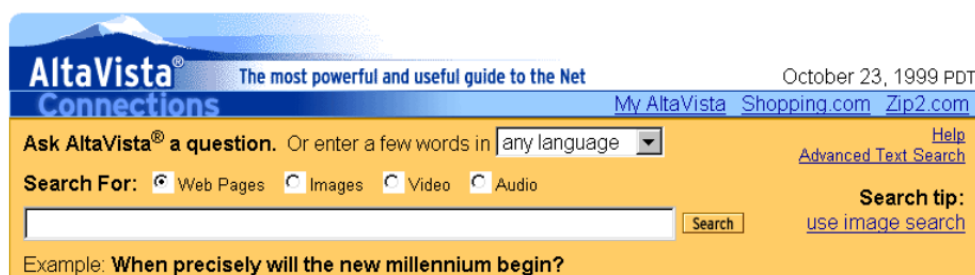


Figura 2: Sitio Web de Altavista, ano 1999.

El clásico paper de Brin y Page, “The anatomy of large-scale hypertextual Web search engine.” [3] explica brevemente el origen del motor de búsqueda de Google y del algoritmo PageRank. La idea es básicamente la siguiente, en primer lugar se implementa un crawler distribuido para poder solicitar y armar el grafo de la web. Las palabras de cada sitio son indexadas y guardadas en una base de datos. Al llegar una consulta al buscador, un programa busca la consulta en los índices de páginas. De esta forma llegamos a un conjunto de páginas que están relacionadas con la consulta. Luego, antes de devolverle al usuario los resultados, estas páginas son ordenadas utilizando el famoso algoritmo PageRank. Este algoritmo se basa en la idea de que para medir la relevancia de un sitio se puede usar como proxy la cantidad de sitios que tienen un link al mismo. Para evitar que un usuario malintencionado manipule los resultados del mismo, la relevancia otorgada por un sitio web que lineea a otro es proporcional a su propia relevancia e inversamente proporcional a la cantidad de links (o grado de salida) del mismo.

El presente trabajo práctico tendrá como objetivo implementar el algoritmo PageRank para luego utilizarlo para generar rankings de todo tipo, ya sea para ordenar la relevancia de páginas webs o generar rankings deportivos. PageRank es un algoritmo que basa su ranking en la idea abstracta del navegante aleatorio. Este problema en general se modela con Cadenas de Markov, y en ultima instancia consiste en encontrar el autovector de una matriz de transiciones, que es similar a una matriz de adyacencia en teoría de grafos. A priori esto puede sonar complicado, pero luego mostraremos que en realidad es bastante simple y elegante.

Debido a su relevancia, PageRank es un algoritmo que ha sido ampliamente estudiado en la literatura. Una muy buena introducción teórica se puede encontrar en el trabajo de Bryan y Leise [4]. Otros autores como Kamvar et al. [8] han buscado otros enfoques y métodos para poder acelerar este algoritmo. La idea es encontrar una forma eficiente de poder computar este modelo, calibrando sus diferentes parámetros de modelado y convergencia para lograr un orden relevante. Otros autores como Govan et al. han buscado modelar la matriz de transiciones para aplicar este algoritmo en eventos deportivos [7]. Es importante mencionar que en este caso la complejidad temporal de los algoritmos utilizados para calcular PageRank no tendra tanta relevancia dado que la cantidad de equipos es acotada. La forma de calibrar el modelo tambien ha sido ampicamente estudiada, en especial para el caso de las páginas web. Un estudio muy interesante es el Constantine et al. de Microsoft Research, donde utilizando los registros del toolbar de Internet Explorer y un modelo basado en la distribución Beta generalizan la matriz de transiciones para poder modelar un parámetro de teletransportación variable [6].

Una vez planteado el procedimiento para computar el PageRank, experimentaremos con la complejidad temporal de los métodos implementados y evaluaremos los diferentes parámetros a calibrar. Finalmente concluiremos si según el algoritmo PageRank y nuestro modelado Vilas efectivamente debió haber estado en la punta del ATP en periodo 1975/1977. En caso afirmativo, sin dudas nos comunicaremos con la ATP.



Figura 3: Guillermo Vilas apoya este TP.

2. PageRank

2.1. La motivación original: páginas web

El algoritmo PageRank fue ideado en un principio para buscar una medida de relevancia con el objetivo de ordenar consultas sobre el grafo de la web. El mismo explota la estructura de este grafo, y como veremos mas adelante luego de explicar la idea general del algoritmo, se calcula buscando el autovector de norma unitaria que resuelve un sistema de la forma $Px = x$ con $P \in \mathbb{R}^{n \times n}$, es decir, el autovector asociado al autovalor 1. A su vez, existen dos interpretaciones equivalentes de este modelo, que serán expuestas a continuación.

2.1.1. Modelado

El problema se modela a partir de un grafo $G(Web, Links)$ donde Web es el conjunto de sitios web y $Links$ es la cantidad de conexiones entre sitios. Consideremos que toda página web $u \in Web$ esta representada por un vértice y la relación entre páginas por un link con una arista. Una representación posible del grafo es mediante matrices de adyacencia. Definimos la matriz de adyacencia o conectividad $W \in \{0,1\}^{n \times n}$ de forma tal que $w_{ij} = 1$ si la página j tiene un link a la página i y $w_{ij} = 0$ en caso contrario. Por lo tanto, la cantidad de páginas a las que la página u apunta ($d_{out}(u)$) se puede calcular como $n_j = \sum_{i=1}^n w_{ij}$.

2.1.2. Propiedades

Sea x_j el puntaje asignado a la página o vértice $j \in Web$ y sea otra página $u \in Web$. La idea es buscar una medida de relevancia que cumpla con las siguientes propiedades:

- La relevancia de todo sitio web es positiva.
- La relevancia de un sitio web debe aumentar a medida que mas sitios únicos lo apuntan.
- La relevancia derivada de otro sitio web debe depender de su propia relevancia. Es decir, es mas valioso que me linkee un sitio relevante que uno no relevante. En caso de no cumplirse esta propiedad, el ranking seria fácilmente manipulable al permitir que un usuario cree muchos sitios que linkeen a uno para darle relevancia.
- La relevancia de todos los sitios web debe sumar uno. De esta manera estamos ante una distribución de probabilidad de los sitios. Esto nos permite a su vez utilizar muchos teoremas estudiados en procesos estocásticos. Mas adelante veremos que al interpretar esto mediante Cadenas de Markov existe una interpretación directa: la relevancia se puede ver como la proporción del tiempo total que un usuario pasa en ese sitio.

Por lo tanto, estamos buscando una medida de relevancia tal que la importancia obtenida por la página u obtenida por el link de la página v sea proporcional a la relevancia de v e inversamente proporcional al grado de v . El aporte del link de v a u entonces es $x_u = x_v/n_v$. Luego, sea $L_k \subseteq Web$ el conjunto de páginas que tienen un link a la página k . Por lo tanto, la relevancia total de un sitio sera:

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}, \quad k = 1, \dots, n. \quad (1)$$

Notar que esta es de cierta manera una definición autoreferencial. La relevancia de un sitio u puede depender de la relevancia de un sitio v , y luego la de v puede depender de la de u . A priori calcular la relevancia de un sitio puede parecer sumamente complicado, pero luego veremos que al plantearlo como un sistema de ecuaciones esta dificultad per se ya no se presenta.

Definimos entonces una matriz de transición o adyacencia con pesos en las aristas $P \in \mathbb{R}^{n \times n}$ tal que $p_{ij} = 1/n_j$ si $w_{ij} = 1$ y $p_{ij} = 0$ en caso contrario. Luego, el modelo planteado en (1) para toda página web se puede expresar $Px = x$ donde $x \in \mathbb{R}^n$. Notar que esto es equivalente a encontrar el autovector de autovalor 1 tal que $x_i > 0$ y $\sum_{i=1}^n x_i = 1$. Notar que si logramos probar que bajo ciertas condiciones nuestra matriz de transición tiene autovalor 1, el signo de todos los elementos de un autovector es el mismo y la dimension del autoespacio es 1 ya tenemos un ranking valido. Esto se debe a que cualquier autovector que cumple con estas propiedades puede ser reescalado a uno de norma unitaria con $x_i \geq 0$.

2.1.3. Existencia y Unicidad

Bryan y Leise [4] analiza y prueba las condiciones bajo las que podemos garantizar que:

- La matriz de transición tiene autovalor 1.
- La dimension del autoespacio asociado al autovalor 1 es 1. Es deseable que el ranking asociado a una matriz de transición sea único.
- El signo de todos los elementos del autovector asociado al autovalor 1 es el mismo.

Veamos bajo que condiciones nuestra matriz de transición cumple con estas propiedades:

Definición Una matriz cuadrada se llama estocástica por columnas si todos sus elementos son positivos y la suma de cada columna es igual a 1.

A partir de esta definición se puede probar la siguiente proposición:

Proposición 2.1 *Toda matriz estocástica por columnas tiene a 1 como autovalor.*

Esto significa que si no existen **dangling nodes**, es decir, vértices con $d_{out} = 0$, podemos garantizar que nuestra matriz de transición es estocástica por columnas.

Notar que bajo las condiciones actuales no podemos garantizar que si existe el autoespacio asociado al autovalor 1, el mismo tenga dimension 1. Intuitivamente, esto se debe a que el grafo de la web puede tener varias componentes conexas. ¿Como comparamos sitios web que no están relacionados? Justamente la relación, ya sea directa o indirecta mediante transitividad me da algún tipo de relación de orden. Al no tener una relación de orden entre dos sitios web bien definida, es razonable que existan múltiples autovectores, es decir, rankings. Esto se puede ver claramente en la página 4 del paper de Bryan y Leise [4].

Por lo tanto, la idea es básicamente buscar algún tipo de transformación relevante de la matriz de transición que me permita garantizar que no voy a tener **dangling nodes** y ademas que solo tenga una componente conexa, es decir, que el grafo resultante sea conexo. Definimos la siguiente matriz de transición, donde $v \in \mathbb{R}^{n \times n}$, con $v_i = 1/n$ y $d \in \{0, 1\}^n$, $d_i = 1$ si $n_i = 0$ y $d_i = 0$ en caso contrario, como:

$$\begin{aligned} D &= vd^t \\ P_1 &= P + D. \end{aligned}$$

De esta manera, en caso de tener una página web que es un **dangling node**, le asignamos un link con probabilidad o peso uniforme a todos los otros sitios web $u \in Web$. Una interpretación equivalente es tomar a la matriz de transiciones como la matriz que describe una Cadena de Markov, donde el link pesado representa la probabilidad de dirigirse de una página a la otra. Por lo tanto, esta transformación se puede interpretar como que existe una probabilidad uniforme de ir de uno de estos sitios a cualquiera de la web. Luego el famoso **navegante aleatorio** simplemente recorre el grafo utilizando estas probabilidades.

Tambien podemos considerar la posibilidad de que el navegante aleatorio se dirija a una página web que no esta linkeada a la página a la que esta actualmente. Este fenómeno se conoce como teletransportación. Para incluirlo al modelo, tomemos un numero $c \in [0, 1]$ y transformemos la matriz de transiciones de la siguiente manera, donde $\bar{1} \in \mathbb{R}^n$ es un vector tal que todos sus componentes valen 1:

$$\begin{aligned} E &= v\bar{1}^t \\ P_2 &= cP_1 + (1 - c)E, \end{aligned}$$

Notar que en caso de tener $c = 1$, estamos en la matriz de transición sin teletransportación. Por otro lado, si $c = 0$ estamos en el caso donde solo hay teletransportación y no importa la estructura del grafo de la web al momento de calcular el ranking.

Notar que la elección de este parámetro c no es trivial. En general, este parámetro se calcula haciendo análisis empíricos sobre los datos de navegadores y toolbars. Un trabajo interesante sobre este tema es el de Constantine et al. de Microsoft Research, donde utilizando los registros del toolbar de Internet Explorer y un modelo basado en la distribución Beta generalizan la matriz de transiciones para poder modelar un parámetro de teletransportación variable [6]. Para este trabajo, simplemente utilizaremos la elección original constante del trabajo de Page & Brin [3] y fijaremos $\alpha = 0,85$.

Esta nueva matriz de transición, dado que es estocástica por columnas y no tiene **dangling nodes**, nos garantiza que la dimension del autoespacio generado por el autovector de autovalor 1 es unitaria. Solo nos falta mostrar que todo autovector tiene todos sus elementos del mismo signo. Es facil probar la siguiente proposicion:

Proposición 2.2 *Si la matriz M es positiva y estocástica por columnas, entonces todo autovector en $V_1(M)$ tiene todos sus elementos positivos o negativos.*

Por lo tanto, ya probamos la existencia del autovector de norma 1 asociado al autovalor 1 de la matriz de transición transformada. El siguiente lema nos garantiza su unicidad. Su respectiva demostración se encuentra nuevamente en la página 7 del paper de Bryan y Leise [4].

Lemma 2.3

Si M es positiva y estocástica por columnas, entonces $V_1(M)$ tiene dimension 1.

Proposición 2.4 *Sea M una matriz real de $n \times n$ positiva y estocástica por columnas, y V el subespacio de \mathbb{R}^n que consiste de aquellos vectores v tales que la suma de sus componentes sea 0, entonces $Mv \in V$ y $\|Mv\|_1 \leq \alpha \|v\|_1$ donde $\alpha = \max_{1 \leq j \leq n} |1 - 2\min_{1 \leq i \leq n}| < 1$.*

2.2. Tenis

2.2.1. Modelado

El modelo GeM es un método de ranqueo basado en PageRank para rankear equipos en ligas deportivas planteado por Govan et al. en "Generalizing Google's PageRank to Rank National Football League Teams" [7]. Utilizaremos el mismo para analizar los rankings del ATP entre 1975 y 1977. A continuación explicaremos como los autores modelan la matriz de transiciones que utiliza el algoritmo. Antes de comenzar con las definiciones, vale aclarar que para mantener la consistencia de la notación utilizada en este paper trabajaremos sobre la transpuesta de la matriz que se usa en el paper de Govan et al.

Se representa una temporada como un grafo dirigido con n nodos. Cada nodo representa un participante o equipo y cada partido entre dos participantes representa una arista desde el perdedor al ganador igual a la diferencia de puntos obtenidos por cada participante en el partido.

La matriz A^t de adyacencias queda definida de la siguiente manera, donde cada w_{ij} representa la diferencia positiva del puntaje de cada participante en ese enfrentamiento.

$$A^t = \begin{cases} w_{ij} & \text{si el participante } j \text{ pierde con el participante } i \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (2)$$

En caso de que un equipo pierda mas de una vez en el mismo torneo, será la suma de las diferencias de cada partido entre i y j . Esta es la mayor generalización en cuanto a PageRank.

Luego se define la matriz H de la siguiente forma:

$$H = \begin{cases} w_{ij} / \sum_{k=1}^n A_{kj}^t & \text{si hay un link de } j \text{ a } k \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (3)$$

Finalmente definimos G a la matriz resultante de:

$$G = \alpha(A^t + ua^t) + (1 - \alpha)ev^t \quad (4)$$

Donde $0 < \alpha < 1$, v es un vector de probabilidades, a es tal que a_j es 1 si la columna j de H es el vector nulo y a_j es 0 en en cualquier otro caso. u es un vector de probabilidad de $n \times 1$. e se define como un vector de $n \times 1$ con todas sus entradas igual a 1. Luego, el vector que contendrá los puntajes de cada equipo será π tal que: $G\pi = \pi$.

Cada entrada H_{ij} de H se puede interpretar como la probabilidad de que el participante j pierda contra el participante i . Para los participantes invictos un simple ajuste que se propone es elegir un vector u donde todas las entradas son $1/n$. Esto significa cambiar la probabilidad de los invictos a que puedan perder contra cualquiera de los otros participantes (incluido si mismo) con probabilidad uniforme. El modelo básico utiliza α de la misma forma que en PageRank y v como vector de personalización del sistema. Una simple elección puede ser $v = (1/n)e$. Aunque v ofrece mucha más flexibilidad. Podría usarse con el resultado estadístico de un ranking previo, aumentando así las probabilidades de que cierto participante gane el presente torneo. La elección del α determina la importancia de la matriz de personalización ev^t . A diferencia del caso de las paginas web, la elección de este parámetro puede ser difícil de medir empíricamente y en última instancia debe ser elegida por el organizador del torneo. En nuestro caso, decidimos elegir $\alpha = 0,85$.

2.2.2. Sistema de puntos y consideraciones

Para procesar los rankings, obtuvimos todos los partidos realizados entre 1975 y 1977 de todas las competencias a nivel internacional de tenis que sirven para clasificar al ATP ¹. (Australian Open, Indianapolis, Roma, Roland Garros, US Open, Wimbledon, Davis Cup, y muchos más) junto con los rankings de cada año. Uno a mitad de año y otro a finales de año.

Los datos están completos a nivel enfrentamientos y participantes, pero carecen de los puntajes que cada jugador obtuvo por cada partido ganado, por avanzar de ronda y/o ganar un torneo. Esta es una parte importante a aclarar del sistema de rankings de la ATP. Se obtienen puntos por cada partido ganado, ronda superada y torneo ganado, los cuales además varían por torneo en importancia, siendo los Grand Slam los de mayor jerarquía. Además, en esa época el sistema estaba basado en el puntaje promedio obtenido por el jugador.

Utiliza un sistema de defensa de puntos que fue variando con los años. En aquella época, la defensa se hacía por año. Es decir, los puntos realizados durante un año, se defienden al año siguiente.

Igualmente esto no supone una limitación grave para computar los rankings y comprobar si Guillermo Vilas fue realmente o no 1ro en al menos uno de esos años.

Además, dada la naturaleza del comportamiento de PageRank, no solo importa a cuantos contrincantes derrotó un participante dado, si no a quienes derrotó. Ésta es la primera consideración a tener en cuenta y una carta a favor que será utilizada por el algoritmo. Dado que no disponemos del sistema de puntuación real y no sabemos a priori cuantos partidos jugó exactamente cada participante, es esencial que además de la puntuación que definamos por cada partido, un jugador obtenga un extra por el rango del rival derrotado. Esto mismo se puede interpretar como que ese jugador ganó un partido importante, dado que los partidos importantes son los de instancias decisivas (cuartos, semifinales o una final), y generalmente, al que solo llegan los mejores jugadores.

Dadas estas consideraciones podemos definir un sistema de puntos con el que poder computar partidos. El mismo es muy sencillo. 3 puntos al ganador y 1 punto al perdedor.

Solo utilizaremos el ranking de fin de año, totalizando por los puntos acumulados en el mismo, obteniendo así un ranking por cada año.

Veremos luego en la experimentación, como este sistema resuelve el cálculo de los rankings de manera apropiada y explicaremos que es lo que sucede en cada año computado observando algunos detalles importantes y comparándolos con los rankings oficiales de la ATP para luego concluir si Vilas fue o no 1ro en algún momento entre 1975 y 1977 según PageRank.

¹Los datos los obtuvimos de <https://github.com/JeffSackmann/tennis.atp>, un repositorio con todos los rankings historicos de la ATP. La veracidad de los mismos es discutible, pero fue lo mejor que conseguimos dado que los investigadores que contactamos no nos quisieron proveer los datos: "Lamentablemente, toda esa información está hoy en pleno ida y vuelta legal con la ATP y no puedo bajo ningún concepto filtrarla."

2.3. Computo: Método de la Potencia

Habiendo definido la matriz de transiciones como la matriz P_2 de la sección anterior, debemos hallar el vector de relevancias x que define en x_j la relevancia de la j -ésima página $\forall 0 \leq j \leq n$. Debido a la definición recursiva de este vector, x debe satisfacer que $P_2x = x$ o lo que es equivalente, que x sea un autovector asociado al autovalor 1 (i.e. $x \in V_1(P_2)$). Para este tipo de problemas es que contamos con el método de la potencia, un algoritmo iterativo que nos devuelve una aproximación lineal al x deseado.

2.3.1. Correctitud

La idea detrás de este algoritmo consiste en generar a partir de un vector inicial x_0 , una secuencia $x_k = \frac{P_2x_{k-1}}{\|P_2x_{k-1}\|}$ que aproxime al autovector q (en nuestro caso el vector x de relevancias) asociado al autovalor 1. Siguiendo esta propuesta elegimos nuestro x_0 de manera que satisfaga que todas sus componentes sean positivas y $\|x_0\|_1 = 1$, definiéndolo además como $x_0 = q + v$, en donde q es el único vector unitario que satisface que $q \in V_1(P_2)$ y que posee todas sus componentes positivas; y v siendo un vector tal que la suma de sus componentes sea 0. De esta manera, dado que P_2 es una matriz positiva y estocástica por columnas, y x_0 un vector unitario con todas sus componentes también positivas, resulta que P_2x_0 también es un vector unitario con todas sus componentes positivas. Esto nos facilita la secuencia de la siguiente forma $x_k = P_2x_{k-1} = P_2^kx_0$, pues desde un principio $x_1 = P_2x_0$ ($\|P_2x_0\| = 1$); luego $x_k = P_2^kx_0 = P_2^kq + P_2^kv = q + P_2^kv \implies P_2^kv = P_2^kx_0 - q$. Ahora bien, por la desigualdad de la proposición 2.4, $\|P_2^kx_0 - q\|_1 = \|P_2^kv\|_1 \leq \alpha^k\|v\|_1$, haciendo tender $k \rightarrow \infty$, $\alpha \rightarrow 0$, de lo que se deduce que $P_2^kx_0 - q \rightarrow 0$ o equivalentemente $P_2^kx_0 \rightarrow q$, es decir al x buscado.

2.3.2. Complejidad

En cuanto a complejidad temporal es evidente que el costo del algoritmo radica en cuantas operaciones elementales nos conlleva calcular P_2^kx , en especial si se requiere realizar una gran cantidad de iteraciones. Recordando como la definimos, $P_2 = cP_1 + (1-c)E$ donde E es la matriz con todos sus elementos iguales a $\frac{1}{n}$ y P_1 es una matriz positiva y estocástica por columnas, de lo que podemos concluir que todos los elementos de P_2 son estrictamente positivos (si $c \in (0,1)$). Esta conclusión no es muy alentadora, pues sólo tener que hacer P_2x tiene una complejidad espacial y temporal de $\Theta(n^2)$. No obstante, sabiendo que $\|x\|_1 = 1$, resulta que $P_2x = (cP_1 + (1-c)E)x = cP_1x + (1-c)Ex = cP_1x + (1-c)e$ en donde e ahora es un vector con todos sus elementos iguales a $\frac{1}{n}$, luego la atención se concentra en el costo de P_1x . Esta última matriz ya no sólo no tiene todos sus elementos estrictamente positivos, sino que suele tratarse de una matriz esparsa, ahorrándonos varias multiplicaciones (las que corresponderían a los elementos iguales a cero) aventajando en velocidad a la anterior situación; a su vez podremos aprovechar la redundancia de ceros para utilizar alguna estructura más eficiente a la hora de almacenar la matriz de la actual iteración. En conclusión dándonos una complejidad temporal total de $\mathcal{O}(k * n^2)$, pero un caso promedio notablemente menor.

2.4. Representación del grafo

Ya hemos demostrado las condiciones necesarias para poder obtener el autovector asociado al autovalor dominante de una matriz de Markov. Ahora debemos proceder a calcular el mismo. Para esto, tenemos que tener en cuenta las cualidades del sistema y el método de resolución del algoritmo. Recordemos que en general, el grafo que representa la web tenderá a ser desconexo y muy grande, es decir, que podrán existir dos o más rankings diferentes. Por lo tanto la matriz de transiciones puede ser muy esparsa e inclusive puede suceder que una página no tenga links de salida, dando lugar a dangling nodes. Para solucionar estos inconvenientes, con lo visto anteriormente disponemos de dos soluciones. Para los dangling nodes, la solución consiste en sumar una columna con probabilidad $1/n$ a la columna de ceros, esto en sí, se puede interpretar como la probabilidad de navegación aleatoria que previamente describimos. Aunque con esto no solucionamos el problema de la esparsidad de la matriz en sí y el de poder tener más de un ranking diferente. Para esto último, se agregó la matriz de probabilidad de teletransportación.

Dada esta definición, la matriz de transiciones resultante no es esparsa. Para sistemas muy grandes, esto puede resultar contraproducente a la hora de obtener el autovector asociado, dado que la complejidad espacial y temporal aumenta considerablemente con la cantidad de información representada en la matriz. Sin embargo existe un resultado que podremos utilizar para mejorar la eficiencia del algoritmo en términos de complejidad temporal y espacial. El mismo se basa en la idea de Kamvar et al. [8, Algoritmo 1] para el cálculo del autovector. Este resultado nos permite utilizar la matriz original de transiciones sin modificar en lo absoluto, pero sí cambiando su representación, valiéndonos de una buena estructura para almacenar las entradas de la misma.

Las cualidades de la matriz hacen que sea razonable intentar pensar en una forma de representar solo las entradas que no sean ceros, y dado que la matriz suele ser esparsa, la misma contendrá muchos ceros que podrían no ser representados. Para esto optamos por una entre las 3 siguientes estructuras de representación:

2.4.1. Dictionary of Keys (*DOK*)

Consiste en un diccionario que mapea pares de fila-columna a la entrada. No se representan las entradas nulas. El formato es bueno para gradualmente construir una matriz esparsa en orden aleatorio, pero pobre para iterar sobre valores distintos de cero en orden lexicográfico. Uno construye típicamente una matriz en este formato y luego se convierte en otro formato más eficiente para su procesamiento.

2.4.2. Compressed Sparse Row (*CSR*)

Pone las entradas no nulas de las filas de la matriz en posiciones de memoria contiguas. Suponiendo que tenemos una matriz dispersa no simétrica, creamos vectores: uno para los números de punto flotante (*val*), y los otros dos para enteros (*col_ind*, *row_ptr*). El vector *val* almacena los valores de los elementos distintos de cero de la matriz, de izquierda a derecha y de arriba hacia abajo. El vector *col_ind* almacena los índices de columna de los elementos en el vector *val*. Es decir, si $val(k) = a_{ij}$ entonces $col_ind(k) = j$. El vector *row_ptr* almacena los lugares en el vector *val* que comienza y termina una fila, es decir, si $val(k) = a_{ij}$ entonces $row_ptr(i) \leq k \leq row_ptr(i+1)$. Por convención, se define $row_ptr(n+1) = nnz$, en donde *nnz* es el número de entradas no nulas en la matriz. Los ahorros de almacenamiento de este enfoque es significativo. En lugar de almacenar elementos n^2 , solamente necesitamos $2nnz + n$ lugares de almacenamiento.

Veamos con un ejemplo como seria la representacion:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 6 & 0 & 0 \end{pmatrix}$$

Es una matrix de 4x4 con 4 entradas no nulas. Luego:

$$\begin{aligned} val &= [5 \ 8 \ 3 \ 6] \\ row_ptr &= [0 \ 0 \ 2 \ 3 \ 4] \\ col_ind &= [0 \ 1 \ 2 \ 1] \end{aligned}$$

2.4.3. Compressed Sparse Column (*CSC*)

La idea es análoga a *CSR*, pero la compresión se hace por columnas es decir, si *CSR* comprime *A*, *CSC* comprime A^t

Sobre la matriz definida para *CSR*, con *CSC* obtenemos lo siguiente:

$$\begin{aligned} val &= [5 \ 8 \ 6 \ 3] \\ col_ptr &= [0 \ 1 \ 3 \ 4 \ 4] \\ row_ind &= [1 \ 1 \ 2 \ 3] \end{aligned}$$

Todos los resultados anteriores permiten evitar representar valores nulos. El motivo de nuestra elección se debe a que *CSR* ofrece una buena representación de las filas de la matriz y es más eficiente a la hora de hacer operaciones del tipo $A \cdot x$ (matriz-vector) que es lo que nos interesa en el método de la potencia que realiza PageRank. *CSC* en cambio, es efectiva para el producto $x \cdot A$ (vector-matriz) dado que la misma ofrece una mejor representación de las columnas. En contra partida, tanto *CSR* como *CSC*, no permiten construcción incremental aleatoria, que si ofrece *DOK*, es decir, que cambios a la esparsidad de la matriz son costosos. En general están pensadas para ser estáticas, pero esto no es un inconveniente en nuestro caso, dado que no se realizaran cambios en la esparsidad de la matriz durante el proceso.

En el presente trabajo utilizaremos la idea de Kamvar et al. [8, Algoritmo 1] para el calculo del autovector valiendonos de nuestra estructura de representación elegida y compararemos los resultados con el algoritmo estandar para mostrar que al final de cuentas, si el sistema es muy grande y esparso, puede resultar muy beneficioso en términos de complejidad espacial y temporal.

2.4.4. Elección de estructura

En el caso de los eventos deportivos la elección de estructura no es tan relevante dado que la cantidad de equipos es acotada. Sin embargo, en el caso de las paginas web este factor es sumamente relevante dado que el uso de memoria de no representar el grafo de forma adecuada puede ser muy grande. De todas estas representaciones posibles, para este trabajo optamos por *CSR* dado que es una representación eficiente para el producto de matrices esparsas, necesario para calcular de forma rápida el método de la potencia.

3. Experimentación

3.1. PageRank: Complejidad

En el caso de PageRank, la complejidad del algoritmo esta dictada por el tamaño del grafo que representa la red, con lo cual vamos a estudiar el tiempo de ejecución en función de dicho factor.

Para generar las instancias tomamos igual cantidad de nodos y de ejes, la primera tomando valor 30, cada instancia sucesiva aumentamos la cantidad de nodos en 15, hasta llegar a las 150 iteraciones. Esto lo hacemos utilizando un $\alpha = 0,85$ como factor de teletransportacion y un $\epsilon = 10^{-5}$ como factor de convergencia. En la matriz el tamaño esta dado en función de la cantidad de nodos, mientras que los ejes son utilizados en el paso inicial para determinar las probabilidades iniciales, con lo cual la cantidad de ejes no tiene la misma importancia que la cantidad de nodos al calcular el método de la potencia.

El siguiente gráfico muestra los tiempos de convergencia según el tamaño, para suavizar el ruido se tomaron 5 muestras por cada instancia y se luego tomo el promedio.

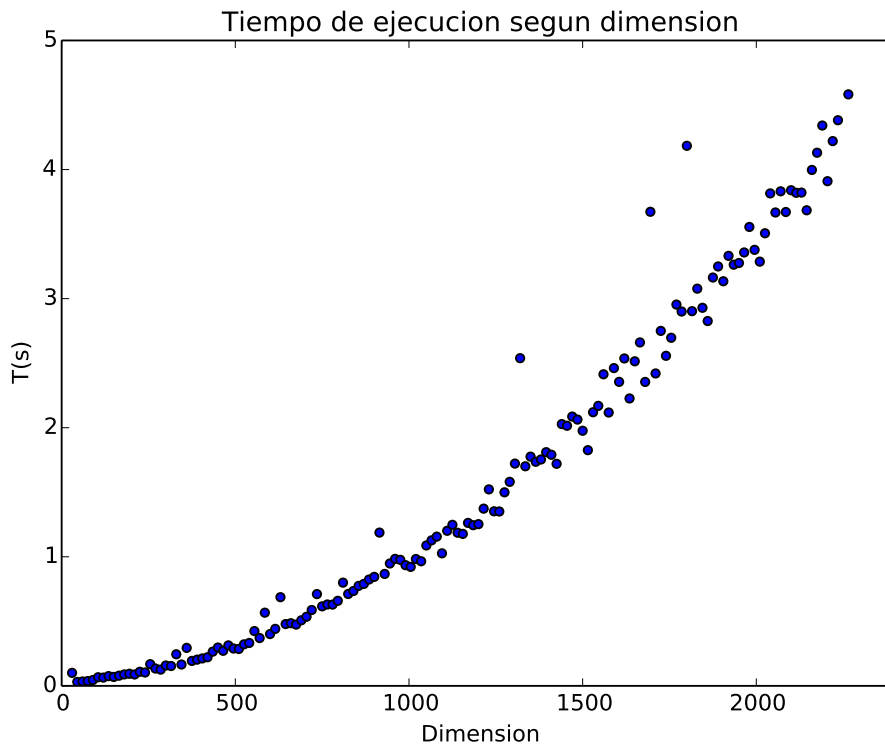


Figura 4: Tiempos de ejecución según cantidad de webs.

El método de la potencia hace k productos de matrices en $\mathcal{O}(n^3)$, donde k es la cantidad de iteraciones que toma cada instancia en converger y n es la cantidad de vértices del grafo. Como podemos observar, la tendencia es no lineal. Esto tiene sentido, dado que el producto de matrices es de orden cubico. Luego experimentaremos para ver si la dimension del grafo efectivamente tiene un gran impacto sobre la cantidad de iteraciones que necesita el método para converger. Sin embargo, a priori conjeturamos que la tendencia cubica de este gráfico esta dominada por el orden del producto entre matrices.

3.2. Convergencia

En este experimento buscamos estudiar la convergencia del método de la potencia. Para ello generamos dos instancias. La primera instancia fue generada con 2.500 vértices y 2.500 aristas. La segunda fue generada con 5000 vértices, utilizando las mismas aristas. Ambas instancias fueron generadas con un $\alpha = 0,85$. Como criterio de parada, se utilizó que la diferencia entre las normas 1 de dos vectores en sucesiones consecutivas sea menor a $\epsilon = 10^{-5}$. A continuación mostramos los resultados:

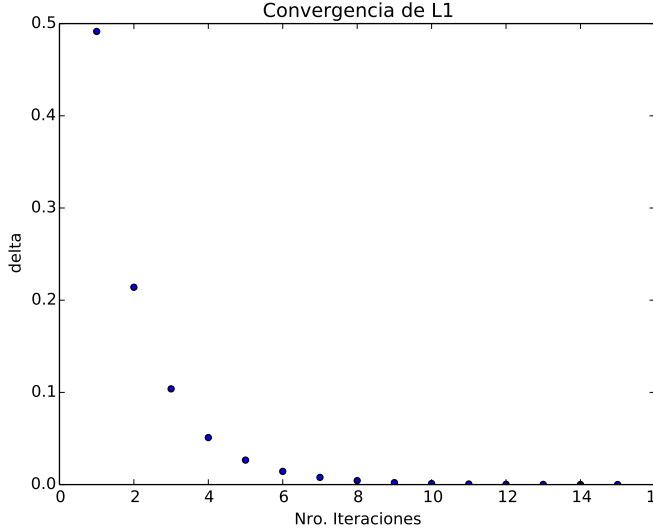


Figura 5: 2500 vértices

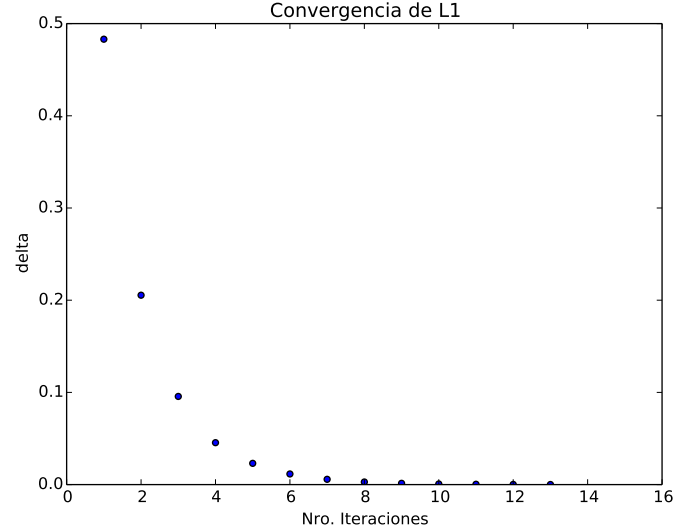


Figura 6: 5000 vértices

Como podemos observar en ambos gráficos, a medida que aumenta el número de iteraciones el delta baja de forma monótona. A su vez, en general ambos algoritmos toman aproximadamente 13 iteraciones en converger. Notar lo siguiente, que la cantidad de vértices sea mayor no implica necesariamente que el método de la potencia tome más iteraciones en converger. De hecho, podemos observar que la instancia con 5000 vértices toma 13 iteraciones, mientras que con 2500 vértices toma 15 iteraciones. Esto apoya la hipótesis del punto anterior, donde argumentamos que la complejidad del método de la potencia estaba dominada por el orden del producto entre matrices y no por el cambio en la cantidad de iteraciones necesarias para que el método converja.

3.3. Comparación PageRank vs In-Deg

Consideremos el siguiente grafo que representa cierta web.

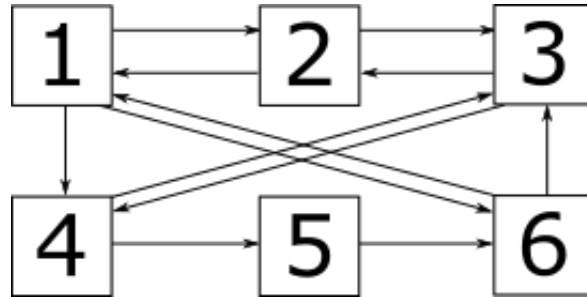


Figura 7: Caso de estudio

Computemos el ranking de esta web utilizando PageRank ($\alpha = 0,85$) y utilizando el orden por grado de entrada.

Nro. Web	PageRank	Por grado
1	0.164204	3
2	0.172456	2
3	0.237500	2
4	0.172456	2
5	0.098296	1
6	0.155089	2

Cuadro 1: PageRank vs Grado de Entrada

Como podemos observar, la web con mayor ranking en PageRank es la web número 3, mientras que al utilizar el grado como medida de relevancia la web con mayor ranking es la número 1.

Intuitivamente, notemos que el algoritmo del grado ignora el hecho de que una pagina relevante que linkea a otro sitio debe contribuirle mas a su relevancia que una que otra que no es tan relevante. Este algoritmo es claramente manipulable, dado que yo puedo crear muchos sitios fácilmente que linkeen a uno que quiero favorecer en el ranking. Aquí podemos observar nuevamente una de las ventajas de PageRank, al elegir un α suficientemente grande, el poder de manipulación que existe es muy bajo.

3.4. Manipulación

A continuación vamos a experimentar con que tan manipulable es el algoritmo PageRank a medida que va cambiando el factor de teletransportación. La idea es la siguiente, un algoritmo con un factor de teletransportación bajo pondera menos en la matriz de transición el componente de la estructura del grafo. Por lo tanto, un Search Engine que tenga este factor de teletransportación bajo va a ser sumamente manipulable, dado que puedo crear muchísimos nodos e inflar el ranking de cualquier sitio. Cuanto mayor es este factor, conjeturamos que vamos a observar que inflar cualquier sitio será mucho más costoso en términos de cantidad de sitios únicos que debo crear. De hecho, Page & Brin en su paper consideran esto y mencionan que en Google se ponderan muchos factores para evitar lo que hoy se conoce como SEO (Search Engine Optimization).

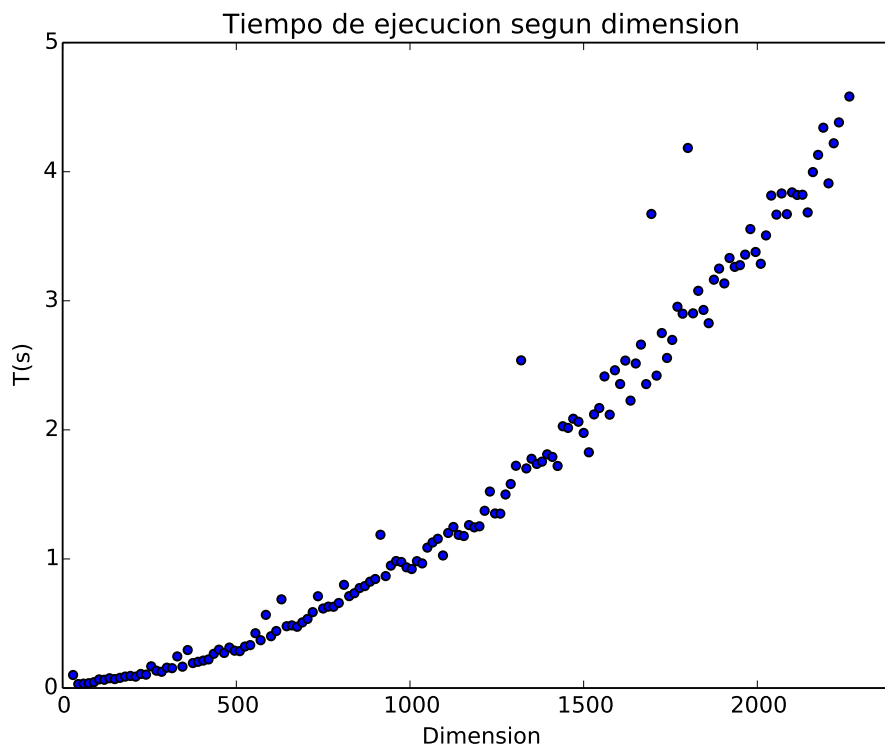


Figura 8: A medida que aumenta c , muestra como cambia el ranking.

3.5. Ranking ATP

Empezemos con el apartado que seguro el lector más esperaba de todo el t.p.. Vilas fue o no 1ro entre 1975 y 1977? Esta pregunta podemos contestarla. Pero previo a esto, necesitamos poner al lector al tanto de la situación. Empezemos viendo los rankings oficiales de la época.

En 1975, Vilas llegó a la primera final de un Grand Slam, el Roland Garros, en donde fue derrotado por el sueco Björn Borg y cuartos de final de Wimbledon.

Este es el top 10 para 1975 según la ATP, donde Guillermo Vilas se ubicó 2do:

id	nombre
173	Jimmy Connors
127	Guillermo Vilas
34	Bjorn Borg
19	Arthur Ashe
225	Manuel Orantes
210	Ken Rosewall
144	Ilie Nastase
180	John Alexander
318	Roscoe Tanner
309	Rod Laver

Cuadro 2: Ranking ATP 1975

En 1976 la ATP ubica a Vilas 6to dentro del top 10:

id	nombre
173	Jimmy Connors
34	Bjorn Borg
144	Ilie Nastase
225	Manuel Orantes
288	Raul Ramirez
127	Guillermo Vilas
1	Adriano Panatta
134	Harold Solomon
87	Eddie Dibbs
41	Brian Gottfried

Cuadro 3: Ranking ATP 1976

En 1977 Vilas se ubica segundo, por debajo de Connors. Recordemos que esto hasta el día de hoy sigue creando polémicas debido a los muy buenos resultados obtenidos por Vilas en aquel año comparados con los del estadounidense. Veamos el top 10 oficial para este año:

id	nombre
173	Jimmy Connors
127	Guillermo Vilas
34	Bjorn Borg
371	Vitas Gerulaitis
41	Brian Gottfried
87	Eddie Dibbs
225	Manuel Orantes
288	Raul Ramirez
144	Ilie Nastase
81	Dick Stockton

Cuadro 4: Ranking ATP 1977

Todos los años tienen como líder indiscutido al estadounidense Jimmy Connors.

3.5.1. Ranking ATP oficial vs. Ranking PageRank vs. Sort por diferencia de victorias/derrotas

Introducimos aquí un algoritmo de ranqueo con un criterio de ordenamiento basado en victorias/derrotas, que además, si hay empate define por diferencia de puntos, aunque para el caso de tenis y por como esta definido el sistema de puntos del mismo no tiene ninguna relevancia. Nos servirá para poder hacer un análisis cualitativo de las virtudes de PageRank sobre algoritmos más básicos y comparar estos resultados con el ranking oficial y poder obtener así alguna conclusión sobre los motivos principales de la investigación.

Veamos que obtuvimos en cada año con este algoritmo, empezando por 1975:

id	PG	PP
19	96	18
225	90	20
127	86	18
173	79	11
144	88	22
34	82	17
288	70	28
180	65	23
318	64	24
155	60	21

Cuadro 5: Ranking 1975 PG(partidos ganados) / PP(partidos perdidos)

Podemos ver que Vilas no aparece segundo, si no tercero. Connors fue desplazado al 4to lugar y en primer lugar aparece Arthur Ashe quien ocupaba el lugar que Connors ocupa ahora. Podemos ver varios cambios relacionados al top 10 oficial.

Todos ellos particularmente relacionados al hecho de que como se habia anticipado la cantidad de victorias sobre derrotas conformaria otro ranking diferente en el cual no importa exactamente que clase de victorias ha conseguido o derrotas sufrido un determinado participante. Podemos ver que Vilas con más victorias sobre derrotas que Connors, se encontraba por debajo de él en el ranking oficial. Esto se debe al sistema de puntos manejados por el ATP donde se suma más puntaje cuanto más avancemos en un torneo y cuantas más finales ganemos y por la jerarquía de ese torneo. Claramente no alcanza con ganar un solo torneo importante, para sumar se tiene que jugar.

Observemos el ranking de 1976:

id	PG	PP
173	86	9
127	83	20
144	73	15
288	91	33
225	76	19
87	83	28
34	57	12
318	71	27
380	75	32
134	70	27

Cuadro 6: Ranking 1976 PG(partidos ganados) / PP(partidos perdidos)

Vemos que Vilas avanzó del 6to lugar al 2do puesto que era ocupado por Bjorn Borg, que a su vez, fue desplazado a la 7ma posición.

Veamos que sucede en 1977:

id	PG	PP
127	126	14
41	105	22
34	71	8
173	69	15
87	77	28
371	60	15
225	65	25
134	64	27
288	61	24
380	66	30

Cuadro 7: Ranking 1977 PG(partidos ganados) / PP(partidos perdidos)

Ok. Tomemoslo con calma. Vilas aparece primero pero esto no es indicador absoluto de que la ATP cometió un error, debido a la falta de información que provee. Aunque si nos da indicios de lo que realmente pasó.

La diferencia de partidos ganados sobre perdidos con respecto a cualquier otro competidor es bastante significativa. Podemos ver que Connors fué desplazado al 4to lugar. Segundo se ubica Brian Gottfried, que durante ese año tuvo un gran desempeño, entre los que se encuentra su victoria sobre Vilas en la final del Roland Garros.

A priori...cantidad de victorias está relacionada con cantidad de puntos, pero esto no es una regla general y depende como mencionamos del sistema de puntajes. Para poder asemejar a la relevancia que la ATP le da a los torneos tenemos que utilizar un algoritmo que aproveche esa característica lo mejor posible. Para esto haremos uso del modelo GeM y analizaremos sus resultados. Esperamos que los mismos se parezcan al ranking oficial, obviamente, con ciertas variaciones.

Utilizaremos los siguientes parámetros para generar los 3 rankings:

$$c = 0.85 \quad \delta = 0.00001$$

Además, como indicamos en el diseño del sistema, usaremos una matriz de personalización uniforme, dado que no nos interesa que influyan sobre los resultados ninguna información estadística de un torneo anterior, por el simple hecho de que evaluamos a los jugadores desde cero cada año.

Avancemos sobre los resultados, comenzando con 1975:

id	puntaje
19	0.033172
34	0.030089
225	0.026483
144	0.026254
127	0.023572
173	0.021918
180	0.017379
318	0.015914
309	0.011614
210	0.011055

Cuadro 8: Ranking 1975 GeM

Vemos que Vilas se ubica en el 4to puesto, Connors fué desplazado al 5to puesto y por arriba se ubican Ashe en primer puesto y Borg en segundo lugar.

Si analizamos los partidos podremos saber que Ashe se enfrentó en una sola oportunidad a Connors y logró derrotarlo, en lo que fué la final Wimbledon. Además se enfrentó en 7 oportunidades a Borg y logró vencerlo en 4 (Wimbledon, Barcelona WCT, Dallas WCT, Munich WCT), siendo Wimbledon el más importante de los 7 enfrentamientos.

Borg obtuvo la final de Roland Garros frente a Vilas como máximo logro.

Mientras tanto Manuel Orantes hizo lo propio para obtener el 3er lugar derrotando a Vilas en las 4 oportunidades que se encontraron y derrotando a Connors en su único enfrentamiento. No es una tarea fácil determinar todas estas congruencias pero con un simple vistazo a los partidos jugados y los torneos en los que se enfrentaron parece claro que el ranking es elocuente.

Veamos el ranking de 1976:

id	puntaje
173	0.033300
144	0.031773
288	0.029343
87	0.026328
41	0.025002
134	0.024189
127	0.024006
34	0.023806
225	0.020587
1	0.014804

Cuadro 9: Ranking 1976 GeM

No parece haber mucho que analizar con respecto al ranking oficial. Vilas se encuentra una posición más abajo. Connors sigue siendo líder indiscutido y las otras posiciones relativas no han cambiado demasiado. Podríamos resaltar el caso particular de Borg que no tuvo malos resultados durante ese año, destacando su final ganada contra Ilie Nastase en Wimbledon, una final perdida en el US Open contra Connors y Cuartos de final en un Roland Garros. Pero si

miramos el mismo ranking pero por diferencia de ganados/perdidos veremos que Borg se encuentra casi en la misma posición debido a la baja diferencia de partidos ganados sobre perdidos.

Concluamos con el análisis de 1977:

id	puntaje
127	0.037046
41	0.035463
34	0.029461
173	0.027319
134	0.024189
87	0.023230
81	0.021816
371	0.017483
288	0.016667
225	0.016318

Cuadro 10: Ranking 1977 GeM

Por una diferencia significativa, Vilas desplaza del primero puesto a Connors para quedarse el con el trono. Pero hay una razón bastante justificada para que esto suceda. En 1977 Vilas conquistó 17 torneos (récord todavía vigente), se consagró en Roland Garros frente a Brian Gottfried por 6-0 6-3 6-0 y el US Open frente a Jimmy Connors por 2-6 6-3 7-6 6-0, fue finalista de Australia donde cayó frente a Roscoe Tanner por 3-6 3-6 3-6, logró una seguidilla de 50 partidos consecutivos sin conocer la derrota y, durante esos doce meses, ganó 145 de los 159 encuentros que jugó (91,1%). Jimmy Connors, en 1977 no obtuvo ningún torneo de Grand Slam y ganó ocho torneos, menos de la mitad de los logrados por Vilas. Parece lógico que Connors incluso esté unos escalones más abajo. Además el ranking por partidos ganados/perdidos para los primeros 4 es idéntico. Esto no puede hacer más que confirmar que Vilas fué el primero indiscutido en 1977 por una amplia diferencia, sobre todo por el bajo rendimiento de Connors durante ese año.

Como comentamos al principio, la cantidad de partidos ganados y los rivales derrotados son factores importantes a la hora de generar el ranking con GeM. Si la ATP hubiese cambiado a este método y fuera retroactivo, hoy el número 1 indiscutido de 1977 sería sin lugar a dudas Guillermo Vilas.

4. Conclusiones

Una vez que ya este todo lo leo y escribo esto bien a los pedos, incluyendo la caratula.

5. Apéndice A: Enunciado

Métodos Numéricos
Segundo Cuatrimestre 2015
Trabajo Práctico 2



Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Ohhh solo tiran π -edras...

Contexto y motivación

A partir de la evolución de Internet durante la década de 1990, el desarrollo de motores de búsqueda se ha convertido en uno de los aspectos centrales para su efectiva utilización. Hoy en día, sitios como Yahoo, Google y Bing ofrecen distintas alternativas para realizar búsquedas complejas dentro de un red que contiene miles de millones de páginas web.

En sus comienzos, una de las características que distinguió a Google respecto de los motores de búsqueda de la época fue la calidad de los resultados obtenidos, mostrando al usuario páginas relevantes a la búsqueda realizada. El esquema general de los orígenes de este motor de búsqueda es brevemente explicado en Brin y Page [3], donde se mencionan aspectos técnicos que van desde la etapa de obtención de información de las páginas disponibles en la red, su almacenamiento e indexado y su posterior procesamiento, buscando ordenar cada página de acuerdo a su importancia relativa dentro de la red. El algoritmo utilizado para esta última etapa es denominado PageRank y es uno (no el único) de los criterios utilizados para ponderar la importancia de los resultados de una búsqueda. En este trabajo nos concentraremos en el estudio y desarrollo del algoritmo PageRank.

Por otro lado, las competencias deportivas, en todas sus variantes y disciplinas, requieren casi inevitablemente la comparación entre competidores mediante la confección de *Tablas de Posiciones* y *Rankings* en base a resultados obtenidos en un período de tiempo determinado. Estos ordenamientos de equipos están generalmente (aunque no siempre) basados en reglas relativamente claras y simples, como proporción de victorias sobre partidos jugados o el clásico sistema de puntajes por partidos ganados, empatados y perdidos. Sin embargo, estos métodos simples y conocidos por todos muchas veces no logran capturar la complejidad de la competencia y la comparación. Esto es particularmente evidente en ligas donde, por ejemplo, todos los equipos no juegan la misma cantidad de veces entre sí.

A modo de ejemplo, la NBA y NFL representan dos ligas con fixtures de temporadas regulares con estas características. Recientemente, el Torneo de Primera División de AFA se suma a este tipo de competencias, ya que la incorporación de la *Fecha de Clásicos* parece ser una interesante idea comercial, pero no tanto desde el punto de vista deportivo ya que cada equipo juega contra su *clásico* más veces que el resto. Como contraparte, éstos rankings son utilizados muchas veces como criterio de decisión, como por ejemplo para determinar la participación en alguna competencia de nivel internacional, con lo cual la confección de los mismos constituye un elemento sensible, afectando intereses deportivos y económicos de gran relevancia.

El problema, Parte I: PageRank y páginas web

El algoritmo PageRank se basa en la construcción del siguiente modelo. Supongamos que tenemos una red con n páginas web $Web = \{1, \dots, n\}$ donde el objetivo es asignar a cada una de ellas un puntaje que determine la importancia relativa de la misma respecto de las demás. Para modelar las relaciones entre ellas, definimos la *matriz de conectividad* $W \in \{0, 1\}^{n \times n}$ de forma tal que $w_{ij} = 1$ si la página j tiene un link a la página i , y $w_{ij} = 0$ en caso contrario. Además, ignoramos los *autolinks*, es decir, links de una página a sí misma, definiendo $w_{ii} = 0$. Tomando esta matriz, definimos el grado de la página j , n_j , como la cantidad de links salientes hacia otras páginas de la red, donde $n_j = \sum_{i=1}^n w_{ij}$. Además, notamos con x_j al puntaje asignado a la página $j \in Web$, que es lo que buscamos calcular.

La importancia de una página puede ser modelada de diferentes formas. Un link de la página $u \in Web$ a la página $v \in Web$ puede ser visto como que v es una página importante. Sin embargo, no queremos que una página obtenga mayor importancia simplemente porque es apuntada desde muchas páginas. Una forma de limitar esto es ponderar los links utilizando la importancia de la página de origen. En otras palabras, pocos links de páginas importantes pueden valer más que muchos links de páginas poco importantes. En particular, consideramos que la importancia de la página v obtenida mediante el link de la página u es proporcional a la importancia de la página u e inversamente proporcional al grado de u . Si la página u contiene n_u links, uno de los cuales apunta a la página v , entonces el aporte de ese link a la página v será x_u/n_u . Luego, sea $L_k \subseteq Web$ el conjunto de páginas que tienen un link a la página k . Para cada página pedimos que

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}, \quad k = 1, \dots, n. \quad (1)$$

Definimos $P \in \mathbb{R}^{n \times n}$ tal que $p_{ij} = 1/n_j$ si $w_{ij} = 1$, y $p_{ij} = 0$ en caso contrario. Luego, el modelo planteado en (1) es equivalente a encontrar un $x \in \mathbb{R}^n$ tal que $Px = x$, es decir, encontrar (suponiendo que existe) un autovector asociado al autovalor 1 de una matriz cuadrada, tal que $x_i \geq 0$ y $\sum_{i=1}^n x_i = 1$. En Bryan y Leise [4] y Kamvar et al. [8, Sección 1] se analizan ciertas condiciones que debe cumplir la red de páginas para garantizar la existencia de este autovector.

Una interpretación equivalente para el problema es considerar al *navegante aleatorio*. Éste empieza en una página cualquiera del conjunto, y luego en cada página j que visita sigue navegando a través de sus links, eligiendo el mismo con probabilidad $1/n_j$. Una situación particular se da cuando la página no tiene links salientes. En ese caso, consideramos que el navegante aleatorio pasa a cualquiera de las página de la red con probabilidad $1/n$. Para representar esta situación, definimos $v \in \mathbb{R}^{n \times n}$, con $v_i = 1/n$ y $d \in \{0, 1\}^n$ donde $d_i = 1$ si $n_i = 0$, y $d_i = 0$ en caso contrario. La nueva matriz de transición es

$$\begin{aligned} D &= vd^t \\ P_1 &= P + D. \end{aligned}$$

Además, consideraremos el caso de que el navegante aleatorio, dado que se encuentra en la página j , decida visitar una página cualquiera del conjunto, independientemente de si esta se encuentra o no referenciada por j (fenómeno conocido como *teletransportación*). Para ello, consideramos que esta decisión se toma con una probabilidad $c \geq 0$, y podemos incluirlo al modelo de la siguiente forma:

$$\begin{aligned} E &= v\bar{1}^t \\ P_2 &= cP_1 + (1 - c)E, \end{aligned}$$

donde $\bar{1} \in \mathbb{R}^n$ es un vector tal que todas sus componentes valen 1. La matriz resultante P_2 corresponde a un enriquecimiento del modelo formulado en (1). Probabilísticamente, la componente x_j del vector solución (normalizado) del sistema $P_2x = x$ representa la proporción del tiempo que, en el largo plazo, el navegante aleatorio pasa en la página $j \in \text{Web}$. Denotaremos con π al vector solución de la ecuación $P_2x = x$, que es comúnmente denominado *estado estacionario*.

En particular, P_2 corresponde a una matriz *estocástica por columnas* que cumple las hipótesis planteadas en Bryan y Leise [4] y Kamvar et al. [8], tal que P_2 tiene un autovector asociado al autovalor 1, los demás autovalores de la matriz cumplen $1 = \lambda_1 > |\lambda_2| \geq \dots \geq |\lambda_n|$ y, además, la dimensión del autoespacio asociado al autovalor λ_1 es 1. Luego, π puede ser calculada de forma estándar utilizando el método de la potencia.

Una vez calculado el ranking, se retorna al usuario las t páginas con mayor puntaje.

El problema, Parte II: PageRank y ligas deportivas

Existen en la literatura distintos enfoques para abordar el problema de determinar el *ranking* de equipos de una competencia en base a los resultados de un conjunto de partidos. En Govan et al. [7] se hace una breve reseña de dos ellos, y los autores proponen un nuevo método basado en el algoritmo PageRank que denominan GeM². Conceptualmente, el método GeM representa la temporada como un red (grafo) donde las páginas web representan a los equipos, y existe un link (que tiene un valor, llamado peso, asociado) entre dos equipos que los relaciona modelando los resultados de los posibles enfrentamientos entre ellos. En base a este modelo, Govan et al. [7] proponen calcular el ranking de la misma forma que en el caso de las páginas web.

En su versión básica, que es la que consideraremos en el presente trabajo, el método GeM (ver, e.g., [7, Sección GeM Ranking Method]) es el siguiente³:

1. La temporada se representa mediante un grafo donde cada equipo representa un nodo y existe un link de i a j si el equipo i perdió al menos una vez con el equipo j .
2. Se define la matriz $A^t \in \mathbb{R}^{n \times n}$

$$A_{ji}^t = \begin{cases} w_{ji} & \text{si el equipo } i \text{ perdió con el equipo } j, \\ 0 & \text{en caso contrario,} \end{cases}$$

donde w_{ji} es la diferencia absoluta en el marcador. En caso de que i pierda más de una vez con j , w_{ji} representa la suma acumulada de diferencias. Notar que A^t es una generalización de la matriz de conectividad W definida en la sección anterior.

3. Definir la matriz $H_{ji}^t \in \mathbb{R}^{n \times n}$ como

$$H_{ji}^t = \begin{cases} A_{ji}^t / \sum_{k=1}^n A_{ki}^t & \text{si hay un link } i \text{ a } j, \\ 0 & \text{en caso contrario.} \end{cases}$$

²Aunque no se especifica, asumimos que el nombre se debe a las iniciales de los autores.

³Notar que en artículo, Govan et al. [7] lo definen sobre la traspuesta. La definición y las cuentas son equivalentes, simplemente se modifica para mantener la consistencia a lo largo del enunciado.

4. Tomar $P = H^t$, y aplicar el método PageRank como fue definido previamente, siendo π la solución a la ecuación $P_2x = x$. Notar que los páginas sin links salientes, en este contexto se corresponden con aquellos equipos que se encuentran invictos.
5. Utilizar los puntajes obtenidos en π para ordenar los equipos.

En función del contexto planteado previamente, el método GeM define una estructura que relaciona equipos dependiendo de los resultados parciales y obtener un ranking utilizando solamente esta información.

Enunciado

El objetivo del trabajo es experimentar en el contexto planteado utilizando el algoritmo PageRank con las variantes propuestas. A su vez, se busca comparar los resultados obtenidos cualitativa y cuantitativamente con los algoritmos tradicionales utilizados en cada uno de los contextos planteados. Los métodos a implementar (como mínimo) en ambos contextos planteados por el trabajo son los siguientes:

1. *Búsqueda de páginas web*: PageRank e IN-DEG, éste último consiste en definir el ranking de las páginas utilizando solamente la cantidad de ejes entrantes a cada una de ellas, ordenándolos en forma decreciente.
2. *Rankings en competencias deportivas*: GeM y al menos un método estándar propuesto por el grupo (ordenar por victorias/derrotas, puntaje por ganado/empatado/perdido, etc.) en función del deporte(s) considerado(s).

El contexto considerado en 1., en la búsqueda de páginas web, representa un desafío no sólo desde el modelado, si no también desde el punto de vista computacional considerando la dimensión de la información y los datos a procesar. Luego, dentro de nuestras posibilidades, consideramos un entorno que simule el contexto real de aplicación donde se abordan instancias de gran escala (es decir, n , el número total de páginas, es grande). Para el desarrollo de PageRank, se pide entonces considerar el trabajo de Bryan y Leise [4] donde se explica la intuición y algunos detalles técnicos respecto a PageRank. Además, en Kamvar et al. [8] se propone una mejora del mismo. Si bien esta mejora queda fuera de los alcances del trabajo, en la Sección 1 se presenta una buena formulación del algoritmo. En base a su definición, P_2 no es una matriz esparsa. Sin embargo, en Kamvar et al. [8, Algoritmo 1] se propone una forma alternativa para computar $x^{(k+1)} = P_2x^{(k)}$. Este resultado debe ser utilizado para mejorar el almacenamiento de los datos.

En la práctica, el grafo que representa la red de páginas suele ser esparso, es decir, una página posee relativamente pocos links de salida comparada con el número total de páginas. A su vez, dado que n tiende a ser un número muy grande, es importante tener en cuenta este hecho a la hora de definir las estructuras de datos a utilizar. Luego, desde el punto de vista de implementación se pide utilizar alguna de las siguientes estructuras de datos para la representación de las matrices esparsas: *Dictionary of Keys* (dok), *Compressed Sparse Row* (CSR) o *Compressed Sparse Column* (CSC). Se deberá incluir una justificación respecto a la elección que considere el contexto de aplicación. Además, para PageRank se debe implementar el método de la potencia para calcular el autovector principal. Esta implementación debe ser realizada íntegramente en C++.

En función de la experimentación, se deberá realizar un estudio particular para cada algoritmo (tanto en términos de comportamiento del mismo, como una evaluación de los resultados obtenidos) y luego se procederá a comparar cualitativamente los rankings generados. La experimentación deberá incluir como mínimo los siguientes experimentos:

1. Estudiar la convergencia de PageRank, analizando la evolución de la norma Manhattan (norma L_1) entre dos iteraciones sucesivas. Comparar los resultados obtenidos para al menos dos instancias de tamaño mediano-grande, variando el valor de c .
2. Estudiar el tiempo de cómputo requerido por PageRank.
3. Para cada algoritmo, proponer ejemplos de tamaño pequeño que ilustren el comportamiento esperado (puede ser utilizando las herramientas provistas por la cátedra o bien generadas por el grupo).

Puntos opcionales:

1. Demostrar que los pasos del Algoritmo 1 propuesto en Kamvar et al. [8] son correctos y computan P_2x .
2. Establecer una relación con la proporción entre $\lambda_1 = 1$ y $|\lambda_2|$ para la convergencia de PageRank.

El segundo contexto de aplicación no presenta mayores desafíos desde la perspectiva computacional, ya que en el peor de los casos una liga no suele tener mas que unas pocas decenas de equipos. Más aún, es de esperar que en general la matriz que se obtiene no sea esparsa, ya que probablemente un equipo juegue contra un número significativo de contrincantes. Sin embargo, la popularidad y sensibilidad del problema planteado requieren de un estudio detallado y pormenorizado de la calidad de los resultados obtenidos. El objetivo en este segundo caso de estudio es puramente experimental.

En función de la implementación, aún cuando no represente la mejor opción, es posible reutilizar y adaptar el desarrollo realizado para páginas web. También es posible realizar una nueva implementación desde cero, simplificando la operatoria y las estructuras, en C++, MATLAB o PYTHON.

La experimentación debe ser realizada con cuidado, analizando (y, eventualmente, modificando) el modelo de GeM:

1. Considerar al menos un conjunto de datos reales, con los resultados de cada fecha para alguna liga de algún deporte.
2. Notar que el método GeM asume que no se producen empates entre los equipos (o que si se producen, son poco frecuentes). En caso de considerar un deporte donde el empate se da con cierta frecuencia no despreciable (por ejemplo, fútbol), es fundamental aclarar como se refleja esto en el modelo y analizar su eventual impacto.
3. Realizar experimentos variando el parámetro c , indicando como impacta en los resultados. Analizar la evolución del ranking de los equipos a través del tiempo, evaluando también la evolución de los rankings e identificar características/hechos particulares que puedan ser determinantes para el modelo, si es que existe alguno.
4. Comparar los resultados obtenidos con los reales de la liga utilizando el sistema estándar para la misma.

Puntos opcionales:

1. Proponer (al menos) dos formas alternativas de modelar el empate entre equipos en GeM.

Parámetros y formato de archivos

El programa deberá tomar por línea de comandos dos parámetros. El primero de ellos contendrá la información del experimento, incluyendo el método a ejecutar (**alg**, 0 para PageRank, 1 para el método alternativo), la probabilidad de teletransportación c , el tipo de instancia (0 páginas web, 1 deportes), el *path* al archivo/directorio conteniendo la definición de la red (que debe ser relativa al ejecutable, o el path absoluto al archivo) y el valor de tolerancia utilizado en el criterio de parada del método de la potencia.

El siguiente ejemplo muestra un caso donde se pide ejecutar PageRank, con una probabilidad de teletransportación de 0.85, sobre la red descrita en **test1.txt** (que se encuentra en el directorio **tests/**), correspondiente a una instancia de ranking aplicado a deportes y con una tolerancia de corte de 0,0001.

```
0 0.85 1 tests/red-1.txt 0.0001
```

Para la definición del grafo que representa la red, se consideran dos bases de datos de instancias con sus correspondientes formatos. La primera de ellas es el conjunto provisto en SNAP [2] (el tipo de instancia es 0), con redes de tamaño grande obtenidos a partir de datos reales. Además, se consideran las instancias que se forman a partir de resultados de partidos entre equipos, para algún deporte elegido por el grupo.

En el caso de la base de SNAP, los archivos contiene primero cuatro líneas con información sobre la instancia (entre ellas, n y la cantidad total de links, m) y luego m líneas con los pares i, j indicando que i apunta a j . A modo de ejemplo, a continuación se muestra el archivo de entrada correspondiente a la red propuesta en Bryan y Leise [4, Figura 1]:

```
# Directed graph (each unordered pair of nodes is saved once):
# Example shown in Bryan and Leise.
# Nodes: 4 Edges: 8
# FromNodeId    ToNodeId
1    2
1    3
1    4
2    3
2    4
3    1
4    1
4    3
```

Para el caso de rankings en ligas deportivas, el archivo contiene primero una línea con información sobre la cantidad de equipos (n), y la cantidad de partidos totales a considerar (k). Luego, siguen k líneas donde cada una de ellas representa un partido y contiene la siguiente información: número de fecha (es un dato opcional al problema, pero que puede ayudar a la hora de experimentar), equipo i , goles equipo i , equipo j , goles equipo j . A continuación se muestra el archivo de entrada con la información del ejemplo utilizado en Govan et al. [7]:

```
6 10
1 1 16 4 13
1 2 38 5 17
1 2 28 6 23
1 3 34 1 21
1 3 23 4 10
1 4 31 1 6
1 5 33 6 25
1 5 38 4 23
1 6 27 2 6
1 6 20 5 12
```

Es importante destacar que, en este último caso, los equipos son identificados mediante un número. Opcionalmente podrá considerarse un archivo que contenga, para cada equipo, cuál es el código con el que se lo identifica.

Una vez ejecutado el algoritmo, el programa deberá generar un archivo de salida que contenga una línea por cada página (n líneas en total), acompañada del puntaje obtenido por el algoritmo PageRank/IN-DEG/método alternativo.

Para generar instancias de páginas web, es posible utilizar el código Python provisto por la cátedra. La utilización del mismo se encuentra descripta en el archivo README. Es importante mencionar que, para que el mismo funcione, es necesario tener acceso a Internet. En caso de encontrar un bug en el mismo, por favor contactar a los docentes de la materia a través de la lista. Desde ya, el código puede ser modificado por los respectivos grupos agregando todas aquellas funcionalidades que consideren necesarias.

Para instancias correspondientes a resultados entre equipos, la cátedra provee un conjunto de archivos con los resultados del Torneo de Primera División del Fútbol Argentino hasta la Fecha 23. Es importante aclarar que los dos partidos suspendidos, River - Defensa y Justicia y Racing - Godoy Cruz han sido arbitrariamente completados con un resultado inventado, para simplificar la instancia. En función de datos reales, una alternativa es considerar el repositorio DataHub [1], que contiene información estadística y resultados para distintas ligas y deportes de todo el mundo.

Fechas de entrega

- *Formato Electrónico*: Martes 6 de Octubre de 2015, hasta las 23:59 hs, enviando el trabajo (informe + código) a la dirección `metnum.lab@gmail.com`. El subject del email debe comenzar con el texto [TP2] seguido de la lista de apellidos de los integrantes del grupo.
- *Formato físico*: Miércoles 7 de Octubre de 2015, a las 18 hs. en la clase práctica.

Importante: El horario es estricto. Los correos recibidos después de la hora indicada serán considerados re-entrega.

6. Apéndice B: Código

6.1. system.cpp

```
1 #include <iostream>
2 #include <math.h>
3 #include <fstream>
4 #include <sstream>
5 #include <stdio.h>
6 #include <string.h>
7 #include <time.h>
8 #include <new>
9 #include <regex>
10 #include <iterator>
11 #include "sparseMatrix.h"
12
13 using namespace std;
14
15 struct dataNode {
16     //dataNode(int n) : node(n), edgesCount(0) {} // no lo pude compilar con listas
17     //de inicializacion...
18     int node;
19     int edgesCount;
20 };
21
22 struct matchesStats {
23     //matchesStats(int t) : team(t), matchesWin(0), matchesDefeat(0), pointsScored(0)
24     // , pointsReceived(0) {}
25     int team;
26     int matchesWin;
27     int matchesDefeat;
28     int pointsScored;
29     int pointsReceived;
30 };
31
32 //webs / sports
33 Matrix<double> pageRank(Matrix<double>& M, double c, double d, vector<int>&
34     nodesCount);
35 Matrix<double> enhancementPageRank(Matrix<double>& M, double c, double d, vector<int>
36     & nodesCount);
37
38 //webs
39 void in_deg(vector<dataNode>& nodesCount);
40
41 //sports
42 void basic_sort(vector<matchesStats>& stats);
43
44 //out data
45 void saveResultPageRank(FILE * pFile, Matrix<double>& data);
46 void saveResultInDeg(FILE * pFile, vector<dataNode>& data);
47 void saveResultBasicSort(FILE * pFile, vector<matchesStats>& data);
48
49 //utils
50 double uniform_rand(double a, double b);
51
52 int main(int argc, char** argv) {
```

```

52     if (argc < 3) {
53         printf("Usage %s: parametros.in salida.out \n", argv[0]);
54         return 0;
55     }
56
57     ifstream inputFile(argv[1]);
58
59     if (!inputFile.good()) {
60         printf("can't open input file.\n");
61         return 0;
62     }
63
64     FILE * outputFile = NULL;
65
66     outputFile = fopen(argv[2], "w");
67     if (outputFile == NULL) {
68         printf("can't open output file.\n");
69         return 0;
70     }
71
72     int alg = 0;
73     double c = 0;
74     int inst = 0;
75     double e = 0;
76     char testFileName[100];
77
78     string line;
79     getline(inputFile, line);
80
81     sscanf(line.c_str(), "%d %d %d %s %d", &alg, &c, &inst, testFileName, &e);
82
83     ifstream testFile(testFileName);
84     getline(testFile, line);
85
86     if (inst == 0) {
87         int nodes = 0;
88         int edges = 0;
89
90         sscanf(line.c_str(), "%d %d", &nodes, &edges);
91         cout << "nodes: " << nodes << " edges: " << edges << endl;
92
93         if (alg == 0) {
94             Matrix<double> M(nodes, nodes);
95
96             vector<int> nodesCount(nodes);
97
98             for (int i = 0; i < nodes; i++) {
99                 nodesCount[i] = 0;
100             }
101
102             int i = 0;
103             while (i < edges) {
104                 int node_from = 0;
105                 int node_to = 0;
106
107                 getline(testFile, line);
108                 sscanf(line.c_str(), "%d %d", &node_from, &node_to);
109                 //cout << "node_from: " << node_from << " node_to: " << node_to <<
                    endl;

```

```

110         nodesCount[node_from-1] += 1;
111         M(node_to-1, node_from-1) = 1;
112         i++;
113     }
114
115     Matrix<double> res = pageRank(M, c, e, nodesCount);
116     // Matrix<double> res = enhancementPageRank(M, c, e, nodesCount);
117
118     cout << "page rank result: \n" << endl;
119     res.printMatrix();
120
121     saveResultPageRank(outputFile, res);
122 } else {
123     // group algorithm webs
124     vector<dataNode> nodesCount(nodes);
125
126     for(int i = 0; i < nodes; i++){
127         dataNode nod;
128         nod.node = i+1;
129         nod.edgesCount = 0;
130         nodesCount[i] = nod;
131     }
132
133     int i = 0;
134     while(i < edges){
135         int node_from = 0;
136         int node_to = 0;
137
138         getline(testFile, line);
139         sscanf(line.c_str(), "%d %d", &node_from, &node_to);
140         //cout << "node_from: " << node_from << " node_to: " << node_to <<
141             endl;
142
143         dataNode nod = nodesCount[node_from-1];
144
145         nod.edgesCount += 1;
146
147         nodesCount[node_from-1] = nod;
148
149         i++;
150     }
151
152     in_deg(nodesCount);
153
154     saveResultInDeg(outputFile, nodesCount);
155 }
156 } else {
157     int teams = 0;
158     int matches = 0;
159
160     sscanf(line.c_str(), "%d %d", &teams, &matches);
161     cout << "teams: " << teams << " matches: " << matches << endl;
162
163     int day = 0;
164     int local = 0;
165     int visitor = 0;
166     int local_score = 0;
167     int visitor_score = 0;

```

```

168
169     if (alg == 0){
170         // page rank sports
171         Matrix<double> M(teams, teams);
172
173         vector<int> totalAbs(teams);
174
175         for(int i = 0; i < teams; i++) {
176             totalAbs[i] = 0;
177         }
178
179         int i = 0;
180         while(i < matches){
181             getline(testFile, line);
182             sscanf(line.c_str(), "%d %d %d %d %d", &day, &local, &local_score, &
                visitor, &visitor_score);
183
184             int abs_score = abs(local_score - visitor_score);
185             if (local_score > visitor_score) {
186                 M(local-1, visitor-1) += abs_score;
187                 totalAbs[visitor-1] += abs_score;
188             } else if (visitor_score > local_score) {
189                 M(visitor-1, local-1) += abs_score;
190                 totalAbs[local-1] += abs_score;
191             }
192
193             i++;
194         }
195
196         Matrix<double> res = pageRank(M, c, e, totalAbs);
197
198         cout << "gem result: \n" << endl;
199         res.printMatrix();
200
201         saveResultPageRank(outputFile, res);
202     } else {
203         // group algorithm sports
204         vector<matchesStats> stats(teams);
205
206         for(int i = 0; i < teams; i++){
207             matchesStats teamStats;
208             teamStats.team = i+1;
209             teamStats.matchesWin = 0;
210             teamStats.matchesDefeat = 0;
211             teamStats.pointsScored = 0;
212             teamStats.pointsReceived = 0;
213             stats[i] = teamStats;
214         }
215
216         int i = 0;
217         while(i < matches){
218             getline(testFile, line);
219             sscanf(line.c_str(), "%d %d %d %d %d", &day, &local, &local_score, &
                visitor, &visitor_score);
220
221             matchesStats localStats = stats[local-1];
222             matchesStats visitorStats = stats[visitor-1];
223
224             if (local_score > visitor_score) {

```

```

225         localStats.matchesWin += 1;
226         visitorStats.matchesDefeat += 1;
227     }else if(visitor_score > local_score) {
228         localStats.matchesDefeat += 1;
229         visitorStats.matchesWin += 1;
230     }
231
232     localStats.pointsScored += local_score;
233     localStats.pointsReceived += visitor_score;
234     visitorStats.pointsScored += visitor_score;
235     visitorStats.pointsReceived += local_score;
236
237     stats[local-1] = localStats;
238     stats[visitor-1] = visitorStats;
239
240     i++;
241 }
242
243     basic_sort(stats);
244
245     saveResultBasicSort(outputFile, stats);
246 }
247 }
248
249     testFile.close();
250
251     inputFile.close();
252
253     if (outputFile != NULL) fclose(outputFile);
254
255     //M.printMatrix();
256
257     //Depending on input data, create a matrix with the input file and call rank with
        matrix a values
258
259     return 0;
260 }
261
262 Matrix<double> pageRank(Matrix<double>& M, double c, double d, vector<int>&
    nodesCount) {
263     srand(45);
264
265     int n = M.rows();
266     double dbl_n = M.rows();
267
268     int j = 0;
269     while(j < n){
270         int i = 0;
271         while(i < n){
272             if(M(i, j) != 0){
273                 M(i, j) = M(i, j) / (double)nodesCount[j];
274             }else if(nodesCount[j] == 0){
275                 M(i, j) = 1/dbl_n; // dangling node / undefeated team
276             }
277             i++;
278         }
279         j++;
280     }
281

```

```

282 Matrix<double> v(n, 1/dbl_n);
283
284 Matrix<double> E(n, n, (1 - c)*1/dbl_n); // PRE: rows == columns
285
286 //Salvo que sea c = 1 no tiene sentido usar Sparse Matrix
287 Matrix<double> A = M*c + E;
288
289 Matrix<double> x(n, 1, 1/dbl_n);
290
291 //for (int i = 0; i < M.rows(); i++) {
292 //    x(i) = uniform_rand(0, 1);
293 //}
294
295 Matrix<double> last_x(n);
296
297 double delta = 0;
298
299 do {
300     last_x = x;
301     x = A*x;
302     delta = x.L1(last_x);
303 } while (delta > d);
304
305 printf("delta is %f\r\n", delta);
306
307 return x;
308 }
309
310 Matrix<double> enhancementPageRank(Matrix<double>& M, double c, double d, vector<int
>& nodesCount) {
311     srand(45);
312
313     int n = M.rows();
314     double dbl_n = M.rows();
315
316     int j = 0;
317     while(j < n){
318         int i = 0;
319         while(i < n){
320             if(M(i, j) != 0){
321                 M(i, j) = M(i, j) / (double)nodesCount[j];
322             }
323             i++;
324         }
325         j++;
326     }
327
328     SparseMatrix<double> A(M);
329
330     SparseMatrix<double> x(n, 1/dbl_n);
331
332     //for (int i = 0; i < M.rows(); i++) {
333     //    x(i) = uniform_rand(0, 1);
334     //}
335
336     SparseMatrix<double> last_x(n);
337
338     SparseMatrix<double> v(n, 1/dbl_n);
339

```

```

340     double delta = 0;
341
342     do {
343         last_x = x;
344
345         x = A*x;
346         x = x*c;
347
348         double w = last_x.norm1() - x.norm1();
349
350         x = x + v*w;
351
352         delta = x.L1(last_x);
353     } while (delta > d);
354
355     printf("delta is %f\r\n", delta); //Deberia devolverse.
356
357     return x.descompress();
358 }
359
360 void in_deg(vector<dataNode>& nodesCount) {
361     sort(nodesCount.begin(), nodesCount.end(), [](dataNode a, dataNode b) {
362         return b.edgesCount < a.edgesCount;
363     });
364
365     cout << "IN-DEG result \n" << endl;
366     for (dataNode a : nodesCount) {
367         cout << "node: " << a.node << " points: " << a.edgesCount << "\n";
368     }
369 }
370
371 void basic_sort(vector<matchesStats>& stats) {
372     sort(stats.begin(), stats.end(), [](matchesStats a, matchesStats b) {
373         if (a.matchesWin - a.matchesDefeat != b.matchesWin - b.matchesDefeat) {
374             return b.matchesWin - b.matchesDefeat < a.matchesWin - a.matchesDefeat;
375         } else {
376             return b.pointsScored - b.pointsReceived < a.pointsScored - a.
                 pointsReceived;
377         }
378     });
379
380     cout << "basic sort result \n" << endl;
381     for (matchesStats a : stats) {
382         cout << "team: " << a.team
383             << " matches win: " << a.matchesWin << " matches defeat: " << a.matchesDefeat
384             << " points scored: " << a.pointsScored << " points received: " << a.
                 pointsReceived << "\n";
385     }
386 }
387
388 void saveResultPageRank(FILE * pFile, Matrix<double>& data) {
389     int n = data.rows();
390     int i = 0;
391
392     while(i < n){
393         fprintf(pFile, "%f\r\n", data(i));
394         i++;
395     }
396 }

```

```

397
398 void saveResultInDeg(FILE * pFile , vector<dataNode>& data) {
399     for (dataNode a : data){
400         fprintf(pFile , "%d %d\r\n" , a.node , a.edgesCount);
401     }
402 }
403
404 void saveResultBasicSort(FILE * pFile , vector<matchesStats>& data) {
405     for (matchesStats a : data){
406         fprintf(pFile , "%d %d %d %d %d\r\n" , a.team , a.matchesWin , a.matchesDefeat , a
            .pointsScored , a.pointsReceived);
407     }
408 }
409
410 double uniform_rand(double a, double b) {
411     return ((b-a)*((double)rand()/RAND_MAX))+a;
412 }

```

6.2. matrix.h

```
1  /*
2  * File:    matrix.h
3  * Author:  Federico
4  *
5  * Created on August 16, 2015, 9:54 PM
6  */
7
8  #ifndef MATRIX_H
9  #define MATRIX_H
10
11 #include <algorithm>
12 #include <math.h>
13 #include <vector>
14 #include <stdio.h>
15
16 using namespace std;
17
18 // La matriz respeta la notacion de la catedra, es decir, el primer subindice
19 // es la fila y el segundo es la columna
20
21 template<class T>
22 class Matrix {
23     public:
24         Matrix();
25         Matrix(int rows); // Columnas impllicitas (col = 1)
26         Matrix(int rows, int col);
27         Matrix(int rows, int col, const T& init);
28         Matrix(const Matrix<T>& other);
29         ~Matrix();
30
31         Matrix<T>& operator=(const Matrix<T>& other);
32         Matrix<T> operator*(const Matrix<T>& other);
33         Matrix<T>& operator*=(const Matrix<T>& other);
34         Matrix<T> operator+(const Matrix<T>& other);
35         Matrix<T>& operator+=(const Matrix<T>& other);
36         Matrix<T> operator-(const Matrix<T>& other);
37         Matrix<T>& operator-=(const Matrix<T>& other);
38
39         Matrix<T> operator*(const T& scalar);
40         Matrix<T> operator/(const T& scalar);
41
42         T& operator()(int a, int b);
43         const T& operator()(const int a, const int b) const;
44         T& operator()(int a);
45         const T& operator()(const int a) const;
46
47         //PRE are vectors
48         double L1(const Matrix<T>& other);
49         double norm1();
50
51         int rows();
52         int columns();
53
54         int rows() const;
55         int columns() const;
56
57         void printMatrix();
```

```

58
59     private:
60         vector<vector<T> > _values;
61         int _rows;
62         int _columns;
63
64 };
65
66 template<class T>
67 Matrix<T>::Matrix()
68     : _values(1), _rows(1), _columns(1)
69 {
70     _values[0].resize(1);
71 }
72
73 template<class T>
74 Matrix<T>::Matrix(int rows)
75     : _values(rows), _rows(rows), _columns(1)
76 {
77     for(int i = 0; i < rows; i++) {
78         _values[i].resize(1);
79     }
80 }
81
82 template<class T>
83 Matrix<T>::Matrix(int rows, int col)
84     : _values(rows), _rows(rows), _columns(col)
85 {
86     for(int i = 0; i < rows; i++) {
87         _values[i].resize(col);
88     }
89 }
90
91 template<class T>
92 Matrix<T>::Matrix(int rows, int col, const T& init)
93     : _values(rows), _rows(rows), _columns(col)
94 {
95     for(int i = 0; i < rows; i++) {
96         _values[i].resize(col, init);
97     }
98 }
99
100 template<class T>
101 Matrix<T>::Matrix(const Matrix<T>& other)
102     : _values(other._values), _rows(other._rows), _columns(other._columns)
103 {}
104
105 template<class T>
106 Matrix<T>::~~Matrix() {}
107
108 template<class T>
109 Matrix<T>& Matrix<T>::operator=(const Matrix<T>& other) {
110     if (&other == this)
111         return *this;
112
113     int new_rows = other._rows;
114     int new_columns = other._columns;
115
116     _rows = new_rows;

```

```

117     _columns = new_columns;
118
119     _values.resize(new_rows);
120     for (int i = 0; i < new_columns; i++) {
121         _values[i].resize(new_columns);
122     }
123
124     for(int i = 0; i < new_rows; i++) {
125         for(int j = 0; j < new_columns; j++) {
126             _values[i][j] = other(i, j);
127         }
128     }
129
130     return *this;
131 }
132
133 template<class T>
134 Matrix<T> Matrix<T>::operator*(const Matrix<T>& other) {
135     // ASUME QUE LAS DIMENSIONES DAN
136     Matrix<T> result(_rows, other._columns);
137
138     int innerDim = _columns; // Tambien podria ser other._rows
139
140     for(int i = 0; i < result._rows; i++) {
141         for(int j = 0; j < result._columns; j++) {
142             result(i,j) = 0;
143             for(int k = 0; k < innerDim; k++) {
144                 result(i,j) += _values[i][k] * other(k,j);
145             }
146         }
147     }
148
149     return result;
150 }
151
152 template<class T>
153 Matrix<T>& Matrix<T>::operator*=(const Matrix<T>& other) {
154     Matrix<T> result = (*this) * other;
155     (*this) = result;
156     return (*this);
157 }
158
159 template<class T>
160 Matrix<T> Matrix<T>::operator+(const Matrix<T>& other) {
161     // ASUME QUE LAS DIMENSIONES DAN
162     Matrix<T> result(_rows, other._columns);
163
164     for(int i = 0; i < result._rows; i++) {
165         for(int j = 0; j < result._columns; j++) {
166             result(i,j) = _values[i][j] + other(i,j);
167         }
168     }
169
170     return result;
171 }
172
173 template<class T>
174 Matrix<T>& Matrix<T>::operator+=(const Matrix<T>& other) {
175     Matrix<T> result = (*this) + other;

```

```

176     (*this) = result;
177     return (*this);
178 }
179
180 template<class T>
181 Matrix<T> Matrix<T>::operator-(const Matrix<T>& other) {
182     // ASUME QUE LAS DIMENSIONES DAN
183     Matrix<T> result(_rows, other._columns);
184
185     for(int i = 0; i < result._rows; i++) {
186         for(int j = 0; j < result._columns; j++) {
187             result(i,j) = _values[i][j] - other(i,j);
188         }
189     }
190
191     return result;
192 }
193
194 template<class T>
195 Matrix<T>& Matrix<T>::operator-=(const Matrix<T>& other) {
196     Matrix<T> result = (*this) - other;
197     (*this) = result;
198     return (*this);
199 }
200
201 template<class T>
202 Matrix<T> Matrix<T>::operator*(const T& scalar) {
203     Matrix<T> result(_rows, _columns);
204
205     for(int i = 0; i < result._rows; i++) {
206         for(int j = 0; j < result._columns; j++) {
207             result(i,j) = _values[i][j] * scalar;
208         }
209     }
210
211     return result;
212 }
213
214 template<class T>
215 Matrix<T> Matrix<T>::operator/(const T& scalar) {
216     Matrix<T> result(_rows, _columns);
217
218     for(int i = 0; i < result._rows; i++) {
219         for(int j = 0; j < result._columns; j++) {
220             result(i,j) = _values[i][j] / scalar;
221         }
222     }
223
224     return result;
225 }
226
227 template<class T>
228 T& Matrix<T>::operator()(int a, int b) {
229     return _values[a][b];
230 }
231
232 template<class T>
233 const T& Matrix<T>::operator()(const int a, const int b) const {
234     return _values[a][b];

```

```

235 }
236
237 template<class T>
238 T& Matrix<T>::operator()(int a) {
239     return _values[a][0];
240 }
241
242 template<class T>
243 const T& Matrix<T>::operator()(const int a) const {
244     return _values[a][0];
245 }
246
247 template<class T>
248 double Matrix<T>::L1(const Matrix<T>& other) {
249     Matrix<T> vectorSubs = *this - other;
250
251     double res = 0;
252     for(int i = 0; i < rows(); i++){
253         res = res + abs(vectorSubs(i));
254     }
255
256     return res;
257 }
258
259 template<class T>
260 double Matrix<T>::norm1() {
261     double res = 0;
262     for(int i = 0; i < rows(); i++){
263         res = res + abs(_values[i][0]);
264     }
265
266     return res;
267 }
268
269
270 template<class T>
271 int Matrix<T>::rows() {
272     return _rows;
273 }
274
275 template<class T>
276 int Matrix<T>::columns() {
277     return _columns;
278 }
279
280 template<class T>
281 int Matrix<T>::rows() const{
282     return _rows;
283 }
284
285 template<class T>
286 int Matrix<T>::columns() const{
287     return _columns;
288 }
289
290 template<class T>
291 void Matrix<T>::printMatrix() {
292     for(int i = 0; i < _rows; i++) {;
293         for(int j = 0; j < _columns; j++) {

```

```
294         printf("%f", _values[i][j]);
295         // cout << _values[i][j] << " ";
296     }
297     cout << endl;
298 }
299 cout << endl;
300 }
301
302 #endif /* MATRIX.H */
```

6.3. sparseMatrix.h

```
1  /*
2  * File:    SparseMatrix.h
3  * Author:  Rodrigo Kapobel
4  *
5  * Created on August 21, 2015, 23:05 PM
6  */
7
8  #ifndef SparseMatrix_H
9  #define SparseMatrix_H
10
11  #include <algorithm>
12  #include <math.h>
13  #include <vector>
14  #include <stdio.h>
15  #include <cmath>
16  #include "matrix.h"
17
18  using namespace std;
19
20  //CSR implementation for sparse matrix. For more information check the next link:
21  // https://en.wikipedia.org/wiki/Sparse\_matrix#Compressed\_row\_Storage\_.28CRS\_or\_CSR.29
22
23  template<class T>
24  class SparseMatrix {
25  public:
26      SparseMatrix();
27      SparseMatrix(int rows); // column vector with 0's
28      SparseMatrix(int rows, T value);
29      SparseMatrix(vector<T>& values, vector<int>& iValues, vector<int>& jValues, int
30          columns); // PRE: jValues::size == values::size & values of iValues[0..iValues
31          ::size-2] are indices of values & iValue[iValues::size-1] == values::size
32      SparseMatrix(const SparseMatrix<T>& other); // compress matrix
33      SparseMatrix(const Matrix<T>& other);
34      ~SparseMatrix();
35
36      SparseMatrix<T>& operator=(const SparseMatrix<T>& other);
37      SparseMatrix<T> operator*(const SparseMatrix<T>& other); // performs a matrix-
38          vector multiplication
39      SparseMatrix<T>& operator*=(const SparseMatrix<T>& other); // performs a matrix-
40          vector multiplication
41      SparseMatrix<T> operator+(const SparseMatrix<T>& other); // performs a vector-
42          vector sum
43      SparseMatrix<T>& operator+=(const SparseMatrix<T>& other); // performs a vector-
44          vector sum
45      SparseMatrix<T> operator-(const SparseMatrix<T>& other); // performs a vector-
46          vector subs
47      SparseMatrix<T>& operator-=(const SparseMatrix<T>& other); // performs a vector-
48          vector subs
49
50      SparseMatrix<T> operator*(const T& scalar);
51      SparseMatrix<T> operator/(const T& scalar);
52
53      T& operator()(int a, int b);
54      const T& operator()(const int a, const int b) const;
55      T& operator()(int a);
56      const T& operator()(const int a) const;
```

```

48
49     Matrix<T> descompress();
50     //PRE are vectors
51     double L1(const SparseMatrix<T>& other);
52     double norm1();
53
54     int rows();
55     int columns();
56     int rows() const;
57     int columns() const;
58     void printSparseMatrix();
59
60 private:
61     vector<T> _values;
62     vector<int> _iValues;
63     vector<int> _jValues;
64
65     int _columns;
66 };
67
68 template<class T>
69 SparseMatrix<T>::SparseMatrix()
70 : _values(1), _iValues(1), _jValues(1), _columns(1)
71 {}
72
73 template<class T>
74 SparseMatrix<T>::SparseMatrix(int rows)
75 : _values(rows), _iValues(1), _jValues(rows), _columns(1)
76 {
77     for(int i = 0; i <= rows; i++){
78         _iValues.resize(i+1, i);
79     }
80 }
81
82 template<class T>
83 SparseMatrix<T>::SparseMatrix(int rows, T value)
84 : _values(rows), _iValues(1), _jValues(rows), _columns(1)
85 {
86     for(int i = 0; i <= rows; i++){
87         _iValues.resize(i+1, i);
88         if(i < rows){
89             _values[i] = value;
90         }
91     }
92 }
93
94 template<class T>
95 SparseMatrix<T>::SparseMatrix(vector<T>& values, vector<int>& iValues, vector<int>&
96     jValues, int columns)
97 : _values(values), _iValues(iValues), _jValues(jValues), _columns(columns)
98 {}
99
100 template<class T>
101 SparseMatrix<T>::SparseMatrix(const SparseMatrix<T>& other)
102 : _values(other._values), _iValues(other._iValues), _jValues(other._jValues),
103     _columns(other._columns)
104 {}
105
106 template<class T>

```



```

105 SparseMatrix<T>::SparseMatrix(const Matrix<T>& other)
106 : _values(0), _iValues(0), _jValues(0), _columns(other.columns())
107 {
108     int new_rows = other.rows();
109     int new_columns = other.columns();
110
111     for(int i = 0; i < new_rows; i++) {
112         _iValues.resize(_iValues.size()+1, _values.size());
113         for(int j = 0; j < new_columns; j++) {
114             if (other(i, j) != 0) {
115                 _values.resize(_values.size()+1, other(i, j));
116                 _jValues.resize(_jValues.size()+1, j);
117             }
118         }
119     }
120     _iValues.resize(_iValues.size()+1, _values.size());
121 }
122
123 template<class T>
124 SparseMatrix<T>::~~SparseMatrix() {}
125
126 template<class T>
127 SparseMatrix<T>& SparseMatrix<T>::operator=(const SparseMatrix<T>& other) {
128     if (&other == this)
129         return *this;
130
131     _values = other._values;
132     _iValues = other._iValues;
133     _jValues = other._jValues;
134     _columns = other.columns();
135
136     return *this;
137 }
138
139 template<class T>
140 SparseMatrix<T> SparseMatrix<T>::operator*(const SparseMatrix<T>& other) {
141     // ASUME QUE LAS DIMENSIONES DAN
142     SparseMatrix<T> result(rows());
143
144     for (int i = 0; i < rows(); i++) {
145         for (int j = _iValues[i]; j < _iValues[i+1]; j++) {
146             result(i) += _values[j]*other(_jValues[j]);
147         }
148     }
149
150     return result;
151 }
152
153 template<class T>
154 SparseMatrix<T>& SparseMatrix<T>::operator*=(const SparseMatrix<T>& other) {
155     SparseMatrix<T> result = (*this) * other;
156     (*this) = result;
157     return (*this);
158 }
159
160 template<class T>
161 SparseMatrix<T> SparseMatrix<T>::operator+(const SparseMatrix<T>& other) {
162     SparseMatrix<T> result(rows());
163

```

```

164     for(int i = 0; i < rows(); i++) {
165         result(i) = _values[i] + other(i);
166     }
167
168     return result;
169 }
170
171 template<class T>
172 SparseMatrix<T>& SparseMatrix<T>::operator+=(const SparseMatrix<T>& other) {
173     SparseMatrix<T> result = (*this) + other;
174     (*this) = result;
175     return (*this);
176 }
177
178 template<class T>
179 SparseMatrix<T> SparseMatrix<T>::operator-(const SparseMatrix<T>& other) {
180     SparseMatrix<T> result(rows());
181
182     for(int i = 0; i < rows(); i++) {
183         result(i) = _values[i] - other(i);
184     }
185
186     return result;
187 }
188
189 template<class T>
190 SparseMatrix<T>& SparseMatrix<T>::operator-=(const SparseMatrix<T>& other) {
191     SparseMatrix<T> result = (*this) - other;
192     (*this) = result;
193     return (*this);
194 }
195
196 template<class T>
197 SparseMatrix<T> SparseMatrix<T>::operator*(const T& scalar) {
198
199     vector<T> resValues(_values.size());
200
201     for(int i = 0; i < _values.size(); i++) {
202         resValues[i] = _values[i] * scalar;
203     }
204
205     SparseMatrix<T> result(resValues, _iValues, _jValues, _columns);
206
207     return result;
208 }
209
210 template<class T>
211 SparseMatrix<T> SparseMatrix<T>::operator/(const T& scalar) {
212     vector<T> resValues(_values.size());
213
214     for(int i = 0; i < _values.size(); i++) {
215         resValues[i] = _values[i] / scalar;
216     }
217
218     SparseMatrix<T> result(resValues, _iValues, _jValues, _columns);
219
220     return result;
221 }
222

```

```

223 template<class T>
224 T& SparseMatrix<T>::operator()(int a) {
225     return _values[a];
226 }
227
228 template<class T>
229 const T& SparseMatrix<T>::operator()(const int a) const {
230     return _values[a];
231 }
232
233 template<class T>
234 Matrix<T> SparseMatrix<T>::descompress() {
235     Matrix<T> result(rows(), _columns, 0);
236     for (int i = 0; i < rows(); i++) {
237         for (int j = _iValues[i]; j < _iValues[i+1]; j++) {
238             result(i, _jValues[j]) = _values[j];
239         }
240     }
241
242     return result;
243 }
244
245 template<class T>
246 double SparseMatrix<T>::L1(const SparseMatrix<T>& other) {
247     SparseMatrix<T> vectorSubs = *this - other;
248     double res = 0;
249     for (int i = 0; i < rows(); i++){
250         res = res + abs(vectorSubs(i));
251     }
252
253     return res;
254 }
255
256 template<class T>
257 double SparseMatrix<T>::norm1() {
258     double res = 0;
259     for (int i = 0; i < rows(); i++){
260         res = res + abs(_values[i]);
261     }
262
263     return res;
264 }
265
266 template<class T>
267 int SparseMatrix<T>::rows() {
268     return _iValues.size() - 1;
269 }
270
271 template<class T>
272 int SparseMatrix<T>::columns() {
273     return _columns;
274 }
275
276 template<class T>
277 int SparseMatrix<T>::rows() const {
278     return _iValues.size() - 1;
279 }
280
281 template<class T>

```

```
282 int SparseMatrix<T>::columns() const{
283     return _columns;
284 }
285
286 template<class T>
287 void SparseMatrix<T>::printSparseMatrix() {
288     Matrix<T> descompressedMatrix = descompress();
289     descompressedMatrix.printMatrix();
290 }
291
292 #endif /* SparseMatrix_H */
```

Referencias

- [1] Datahub. <http://datahub.io>.
- [2] Stanford large network dataset collection. <http://snap.stanford.edu/data/#web>.
- [3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.
- [4] Kurt Bryan and Tanya Leise. The linear algebra behind google. *SIAM Review*, 48(3):569–581, 2006.
- [5] Christopher Clarey. Years later for guillermo vilas, he’s still not the one. *The New York Times*.
- [6] Paul G. Constantine, Abraham Flaxman, David F. Gleich, and Asela Gunawardana. Tracking the random surfer: Empirically measured teleportation parameters in pagerank. In *WWW. Association for Computing Machinery, Inc.*, April 2010.
- [7] Angela Y. Govan, Carl D. Meyer, and Russell Albright. Generalizing google’s pagerank to rank national football league teams. In *Proceedings of SAS Global Forum 2008*, 2008.
- [8] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the 12th international conference on World Wide Web, WWW ’03*, pages 261–270, New York, NY, USA, 2003. ACM.