DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico II

Tirate un qué, tirate un ranking...

Métodos Numéricos
Segundo Cuatrimestre de 2014

Integrante	LU	Correo electrónico
Aldasoro Agustina	86/13	agusaldasoro@gmail.com
Bouzón María Belén	128/13	belenbouzon@hotmail.com
Cairo Gustavo Juan	89/13	gjcairo@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Resumen

El resumen de no más de 200 palabras, deberá explicar brevemente el trabajo realizado y las conclusiones de los auores de manera que pueda ser útil por sí solo para dar una idea del contenido del trabajo.

Palabras claves

Las palabras claves, no más de cuatro, deben ser términos técnicos que den una idea del contenido del trabajo para facilitar su búsqueda en una base de datos temática.
♣ Wachi
♣ turroh

Índice

1. Introducción Teórica

Contendrá una breve explicación de la base teórica que fundamenta los métodos involucrados en el trabajo, junto con los métodos mismos. No deben incluirse demostraciones de propiedades ni teoremas, ejemplos innecesarios, ni definiciones elementales (como por ejemplo la de matriz simétrica).

Explicar de donde surge el querer ordenar los datos y bla...

1.1. PageRank, HITS, In-deg

El trabajo consistirá en estudiar distintos aspectos de los siguientes métodos: PageRank, HITS e In-deg.

PageRank *- Modelo del Navegante Aleatorio*

Este método consta de tres fases: explorar la web y localizar todas las páginas de acceso público; indexar los datos desde el primer paso, así se puede acceder eficientemente a palabras claves o frases relavantes; y valorar la importancia de cada una de las páginas en la base de datos. A nivel de nuestro desarrollo, sólo nos vamos a encargar de la última etapa: estimar un orden de importancia para los datos.

Teniendo un grafo dirigido, se le otorga a cada componente X_k del mismo un valor dado por la siguiente ecuación:

$$X_k = \sum_{j \epsilon L_k} \frac{X_j}{n_j}$$

Donde L_k es el conjunto de links entrantes a la página k y n_j es el número de links salientes desde la página j.

Luego, se construye una matriz (que llamaremos A), donde se encuentra por filas las respectivas ecuaciones para cada X_i como la descripta arriba.

La resolución de este método consiste en hallar el autovector con autovalor asociado 1 para la matriz A. De acuerdo al trabajo de Bryan y Leise, este cálculo se computa mediante el método de la potencia.

La matriz A cuenta con ciertas mejoras para determinados casos específicos. Por un lado, si alguna página p no tuviera ningún link saliente se considera que tiene igual probabilidad de ir a cualquiera de las otras páginas y se le otorga al vector columna p de la matriz A el valor de $\frac{1}{n}$ para cada componente. Por otro lado, existe un fenómeno denominado "*Teletransportación*"que consiste en que un navegante se mueva de una página a otra pero no mediante los links existentes, sino tipeando la URL. Para modelar de manera óptima este suceso, se reemplaza a la matriz A por la matriz M definida bajo la siguiente ecuación: *M = (1-m)A + m.S* siendo m la probabilidad de que un navegante se *teletransporte* y S una matriz cuyos valores S_{ij} tienen todos el mismo valor: $\frac{1}{n}$ representando así una matriz donde la probabilidad de ir a cualquier página del gráfico es uniforme.

El *Método de la Potencia* se realiza de manera iterativa lo cual permite reducir el tiempo de cómputo para elevar a la k la matriz M. Si tenemos en cuenta el trabajo de Kamvar, presenta una herramienta de cálculo que permite encontrar el principal autovector de M en una serie menor de pasos.

Hyperlink-Induced Topic Search (HITS)

El método planteado por Kleinberg consiste en: Dada una consulta sobre σ, nos queremos focalizar en una colección de páginas S_σ tal que sea relativamente pequeña, sea rica en páginas relevantes sobre el tema y contenga la mayoría de las autoridades más fuertes sobre el tema. Considerando autoridad a una página que tiene muchos links entrantes. Esto se realiza del siguiente modo:

a) Acorde a un parámetro *t*, se coleccionan las primeras t páginas rankeadas bajo una búsqueda basada estrictamente por texto. A este conjunto se lo llama R_σ.

b) Incrementamos el conjunto R_σ añadiendo las páginas que tienen links entrantes y salientes al mismo, formando así el conjunto S_σ. Para cada página de R_σ se permite añadir, a lo sumo *d* páginas que la apunten y *d* páginas a las cuales apunte.

c) Se eliminan de S_σ los links intrínsecos, es decir no se tienen en cuenta links que apuntan a una página del mismo dominio que la página saliente.

d) Admite hasta *m* páginas del mismo dominio apuntar a cualquier página p. Esta idea no fue utilizada por el autor.

El conjunto obtenido hasta aca lo llamamos G_σ. Nuestro trabajo asume un conjunto G_σ bien formado y comienza el trabajo desde aquí.

Se construye una matriz de adyacencia que denominaremos A, bajo la siguiente fórmula:

$$a_{ij} = \begin{cases} 1 & \exists \textit{ link desde i hasta j} \\ 0 & \textit{caso contrario} \end{cases}$$

A cada página i de la Web se le otorga un peso como Autoridad y un peso de Hub:

Peso de autoridad:

$$X_j = \sum_{i:i\to j} Y_i$$

Peso de Hub:

$$Y_i = \sum_{j:i\to j} X_j$$

In-deg

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. Desarrollo

Deben explicarse los métodos numéricos que utilizaron y su aplicación al problema concreto involucrado en el trabajo práctico. Se deben mencionar los pasos que siguieron para implementar los algoritmos, las dificultades que fueron encontrando y la descripción de cómo las fueron resolviendo. Explicar también cómo fueron planteadas y realizadas las mediciones experimentales. Los ensayos fallidos, hipótesis y conjeturas equivocadas, experimentos y métodos malogrados deben figurar en esta sección, con una breve explicación de los motivos de estas fallas (en caso de ser conocidas).

2.1. Elección de las estructuras

Luego de haber estudiado los tres tipos de estructuras dados por la cátedra: Dictionary of Keys (DOK), Compressed Sparse Row (CSR) y Compressed Sparse Column (CSC), decidimos cual elegir.

En primera instacia, consideramos la estructura *Dictionary of Keys*. La misma consiste en un diccionario con doble clave, donde cada una es fila y columna respectivamente y su significado son los elementos de la matriz distintos de cero. De esta manera, se aprovecha en términos de espacio en memoria la notable cantidad de ceros de la matriz. Contaba con la gran ventaja de que es buena para construirla incrementalmente en un arreglo esparso y además se puede trasponer de manera sencilla ya que es invertir el orden de las claves. En contraposición, para procesar los cálculos aritméticos es necesario convertirla a otro formato. Por este motivo, descartamos esta opción.

El modo de almacenamiento *Compressed Sparse Row* requiere la implementacion de tres arreglos (en nuestro caso vectores) que llamaremos val, ind_col y ptr_fila. El tamaño de los dos primeros estará dado por la cantidad de elementos distintos de cero de la matriz. Mientras que el primero (val) almacenará estos valores de izquierda a derecha y luego desde arriba hacia abajo, el segundo vector (ind_col) el número de columna para cada elemento. Es decir, el elemento almacenado en la posición i-esima del vector ind_col representa la columna correspondiente al valor almanacenado en val_i. Y el tercer vector (ptr_fila) tiene un tamaño equivalente a la cantidad de filas+1 conteniendo los índices de donde comienza cada columna.

El modo de almacenamiento *Compressed Sparse Column* cuenta también con la implementación de tres arreglos llamados: val, ind_fila, ptr_col. Val es un arreglo con todos los valores distintos de cero de la matriz, desde arriba hacia abajo y luego de izquierda a derecha. Ind_fila son los indices de fila correspondientes a esos valores. Por ultimo, ptr_col lista los indices donde comienza cada columna.

En segunda instancia, nos encontramos frente a la disyuntiva sobre si elegir el formato CSR (por filas) o CSC (por columnas) ya que no notamos un beneficio de una sobre otra.

Por ultimo, el vector ptr_fila tendra de tamaño la cantidad de filas incrementada en uno y listara los indices que indicaran los valores de val que comienzan cada fila.

Haciendo cálculos pequeños notamos que si nos situamos en el formato de *Compressed Spare* trasponer una matriz almacenada de manera CSC no es más que interpretar los mismos tres arreglos como CSC. **VA DEMOSTRACION DE ESTO???**. Fue decisión del grupo considerar el formato por defecto de la matriz el CSR (filas) y al trasponerlas sólo modificarle un bool que indique si está traspuesta y de ahora en mas leerla y considerarla como CSC (columnas). Esta decisión fue tomada luego de que **Agustín Montero nos confirmara** que estaba permitido elegir una opción de las ofrecidas y adaptarla a nuestra conveniencia, siempre que se aclararan los cambios. Por este motivo, en el algoritmo de multiplicar una matriz por un vector se diferencia la manera en que este almacenada y hace la multiplicación acorde a su manera respectiva, se incluye el pseudocódigo de este algoritmo en el *Apéndice B*.

3. Resultados y discusión

Deben incluir los resultados de los experimentos, utilizando el formato más adecuado para su presentación. Deberán especificar claramente a qué experiencia corresponde cada resultado. No se incluirán aquí corridas de máquina.
Se incluirá aquí un análisis de los resultados obtenidos en la sección anterior (se analizará su validez, coherencia, etc.). Deben analizarse como mínimo los items pedidos en el enunciado. No es aceptable decir que los resultados fueron los esperados", sin hacer clara referencia a la teorica a la cual se ajustan. Además, se deben mencionar los resultados interesantes y los casos "patologicos.[en]contrados.

4. Conclusiones

Esta sección debe contener las conclusiones generales del trabajo. Se deben mencionar las relaciones de la discusión sobre las que se tiene certeza, junto con comentarios y observaciones generales aplicables a todo el proceso. Mencionar también posibles extensiones a los métodos, experimentos que hayan quedado pendientes, etc.

5. Apéndices

5.1. Apéndice A

5.2. Apéndice B

En el apéndice B se incluirán los códigos fuente de las funciones relevantes desde el punto de vista numérico.
PONER EL PSEUDOCODIGO DE LA MULTIPLICACION DE MATRICES, PORQUE DIJE QUE ESTABA ACA. JE

5.3. Apéndice Φ

Resultados que valga la pena mencionar en el trabajo pero que sean demasiado especificos para aparecer en el cuerpo principal del trabajo podrán mencionarse en sucesivos apéndices rotulados con las letras mayúsculas del alfabeto romano. Por ejemplo: la demostración de una propiedad que aplican para optimizar el algoritmo que programaron para resolver un problema.

6. Referencias

Es importante incluir referencias a libros, artículos y páginas de Internet consultados durante el desarrollo del trabajo, haciendo referencia a estos materiales a lo largo del informe. Se deben citar también las comunicaciones personales con otros grupos.
PONER ACA LOS PAPERSSSSSSSSSSS