INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE
IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Control 2

El siguiente control está pensado para ser respondido en un plazo de 2.5 horas (2 horas y 30 minutos). Recuerde poner su nombre en todas las hojas. Si lo desea puede utilizar las hojas de enunciado para responder. Tiene permitido el uso de calculadora, sin embargo esto excluye a Smartphones y dispositivos similares.

Pregunta 1

Responda las siguientes preguntas de manera clara y precisa.

a) (1 pto.) Explique, descriptivamente, el funcionamiento de los algoritmos *HITS* y *PageRank* para la indexación de páginas web.
b) (1 pto.) ¿De qué forma y para qué puede servir, desde el punto de vista comercial, la identificación de comunidades en la Web?
c) Explique cómo funciona el algoritmo de sesionización basado en *Network Flow Model*.
d) (1 pto.) Explique en qué consiste la personalización de la Web, desde el punto de vista algorítmico.
e) (1 pto.) Explique, descriptivamente, la metodología utilizada para la identificación de las *Web Site Keywords*.
f) (1 pto.) Explique en qué consiste la aplicación de la neurociencia al estudio del comportamiento del usuario en la Web.

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Hoja Respuesta Pregunta 1

Nombre:

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Hoja Respuesta Pregunta 1

Nombre:

Profesores: Juan D. Velásquez.
Pablo Tapia.

Pregunta 2

El 17 de noviembre de 2014 se realizará una nueva versión del CyberMonday Chile, evento que en algunos casos cuadruplica el nivel de ventas mensuales de los participantes por sus canales de venta online[1]. Tal demanda acarrea la posibilidad de sobrecargar los servidores, por lo que las empresas afiliadas están realizando esfuerzos para asegurar que el servicio se mantenga activo. Usted ha sido contratado por la línea aérea LAMA como consultor de Web Mining, específicamente para determinar el comportamiento que tienen los usuarios en su website durante el día del evento. Los requerimientos consisten en realizar un análisis del comportamiento de los posibles clientes de la compañía a través de su portal para poder comprenderlos de mejor forma.

a) (1 pto.) Explique por qué resulta relevante comprender la forma en que los usuarios navegan su sitio web, y qué conclusiones se podrían obtener de este estudio, desde el punto de vista del WUM y el WSM.
b) (1 pto.) El administrador del sitio le entrega el log de peticiones que ingresan al Web Server del portal. Identifique las sesiones agrupando las entradas del registro. Especifique sus supuestos y parámetros. Debe detallar además los posibles recorridos (por ID del log) y los supuestos en base a ellos. Base sus respuestas en la tabla 1 de la siguiente hoja.
c) (1 pto.) Mencione 5 distintos tipos de logs que pueden generar ruido a la hora de realizar el proceso de sesionización.
d) (1 pto) Suponga que posee los datos de las sesiones (bien identificadas) del último año del portal. Usted no cuenta con información adicional sobre las sesiones a priori. Indique dos modelos que permitan trabajar los datos para encontrar posibles clasificaciones en este caso, detallando qué es necesario para cada modelo. ¿Qué tipo de conclusiones se podrían obtener a partir de este estudio?
e) (1 pto.) Suponga ahora que un experto asigna etiquetas a un porcentaje de las sesiones de acuerdo a características relevantes consideradas por él y la empresa. Indique qué tipo de modelos ocuparía en este caso, qué es importante respecto de las tabulaciones del experto y de ejemplos de las etiquetas que podría asignar el experto.
f) (1 pto.) Si el vector de comportamiento del usuario está caracterizado por la secuencia de páginas visitadas y el tiempo gastado en ellas, ¿qué supuestos se están haciendo en esta definición? Bajo este mismo escenario, ¿qué potenciales errores tiene esta aproximación?

[1] http://www.cyberday.cl/noticias.html

ID	IP	DATE	REQUEST	STATUS	USER/AGENT
1	119.63.196.59	[2012-11-11 12:55:30]	GET /index.php/frontend/content/quienes HTTP/1.1	200	Mozilla/5.0
2	119.63.196.59	[2012-11-11 12:55:30]	GET /css/default.css HTTP/1.1	200	Mozilla/5.0
3	119.63.196.59	[2012-11-11 12:55:30]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/5.0
4	157.55.33.84	[2012-11-11 12:58:21]	GET /index.php/frontend/content/albumes HTTP/1.1	200	Mozilla/4.0
5	157.55.33.84	[2012-11-11 12:58:21]	GET /css/default.css HTTP/1.1	200	Mozilla/4.0
6	157.55.33.84	[2012-11-11 12:58:21]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/4.0
7	119.63.196.59	[2012-11-11 12:58:46]	GET /index.php/frontend/content/noticia/25 HTTP/1.1	200	Mozilla/5.0
8	119.63.196.59	[2012-11-11 12:58:46]	GET /css/default.css HTTP/1.1	200	Mozilla/5.0
9	119.63.196.59	[2012-11-11 12:58:46]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/5.0
10	119.63.196.59	[2012-11-11 12:58:46]	GET /images/noticias/noticia25.jpg HTTP/1.1	200	Mozilla/5.0
11	157.55.33.84	[2012-11-11 12:59:21]	GET /index.php/frontend/content/galeria/32 HTTP/1.1	200	Mozilla/4.0
12	157.55.33.84	[2012-11-11 12:59:21]	GET /css/default.css HTTP/1.1	200	Mozilla/4.0
13	157.55.33.84	[2012-11-11 12:59:21]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/4.0
14	157.55.33.84	[2012-11-11 12:59:21]	GET /images/galeria/thumb6.jpg HTTP/1.1	200	Mozilla/4.0
15	157.55.33.84	[2012-11-11 12:59:21]	GET /images/galeria/thumb7.jpg HTTP/1.1	200	Mozilla/4.0
16	157.55.33.84	[2012-11-11 12:59:21]	GET /images/galeria/thumb8.jpg HTTP/1.1	200	Mozilla/4.0
17	119.63.196.59	[2012-11-11 12:59:54]	GET /index.php/frontend/content/noticias HTTP/1.1	200	Mozilla/5.0
18	119.63.196.59	[2012-11-11 12:59:54]	GET /css/default.css HTTP/1.1	200	Mozilla/5.0
19	119.63.196.59	[2012-11-11 12:59:54]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/5.0
20	176.21.146.51	[2012-11-11 13:05:32]	GET /index.php/frontend/content/quienes HTTP/1.1	200	Mozilla/5.0
21	176.21.146.51	[2012-11-11 13:05:32]	GET /css/default.css HTTP/1.1	200	Mozilla/5.0
22	176.21.146.51	[2012-11-11 13:05:32]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/5.0
23	131.253.41.224	[2012-11-11 13:10:17]	GET /robots.txt HTTP/1.1	404	msnbot-media/1.1
24	131.253.41.224	[2012-11-11 13:10:17]	GET /images/albumes/imagen10.jpg HTTP/1.1	200	msnbot-media/1.1
25	200.111.12.98	[2012-11-11 13:12:21]	GET / HTTP/1.1	200	Mozilla/5.0
26	200.111.12.98	[2012-11-11 13:12:21]	GET /css/default.css HTTP/1.1	200	Mozilla/5.0
27	176.21.146.51	[2012-11-11 22:05:32]	GET /index.php/frontend/content/noticias HTTP/1.1	200	Mozilla/5.0
28	176.21.146.51	[2012-11-11 22:05:32]	GET /css/default.css HTTP/1.1	200	Mozilla/5.0
29	176.21.146.51	[2012-11-11 22:05:32]	GET /images/favicon.gif HTTP/1.1	200	Mozilla/5.0

Tabla 1. Tabla de Registro de Actividad en Sitio Web.

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Hoja Respuesta Pregunta 2

Nombre:

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Hoja Respuesta Pregunta 2

Nombre:

Profesores: Juan D. Velásquez.
Pablo Tapia.

Pregunta 3

Responda las siguientes preguntas de forma clara y precisa. Puede hacer uso de calculadora para realizar cálculos.

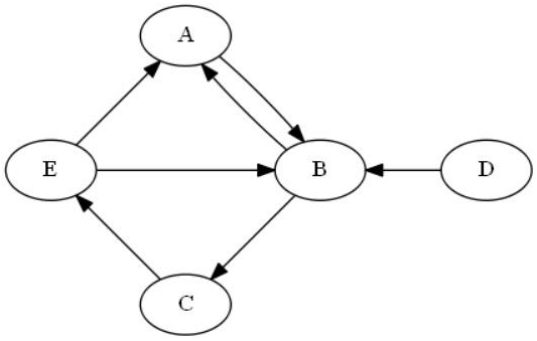

Figura 1. Grafo dirigido.

a) (4 pts.) En base al grafo de la figura 1 estime el PageRank de cada nodo. Para ello calcule los valores en 3 iteraciones. Puede utilizar la siguiente tabla para escribir sus resultados.

Nodos	PageRank por Iteración			
	0	1	2	3
A				
B				
C				
D				
E				

b) (1 pto.) Comente la siguiente frase "*PageRank es un algoritmo que permite estimar la importancia de un nodo basándose en la cantidad de arcos que lo apuntan*".

c) (1 pto.) El algoritmo de PageRank requiere que, en la iteración inicial (iteración cero), se fijen los valores de los nodos en algún número arbitrario $PR_i \in [0,1]$. Explique a qué se debe que con cualquier valor inicial, el resultado será siempre el mismo (asuma un número infinito de iteraciones). Además, indique un valor para cada nodo que podría agilizar la convergencia.

$$r_p^{(i+1)} = \frac{b}{N} + (1-b) \sum_{\forall q,\, q \to p}^{I_p} \frac{r_q^{(i)}}{outdeg(q)}$$

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Hoja Respuesta Pregunta 3

Nombre:

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 Web Intelligence

Profesores: Juan D. Velásquez.
Pablo Tapia.

Hoja Respuesta Pregunta 3

Nombre: