



Solución control 2

Pregunta 2

1. ¿Cuál es el producto final del minado de *logs* de un servidor web? Muestre un ejemplo simple de este producto. 1 punto

Puede ser visto como los pageviews o las transacciones.

	<i>IP</i>	<i>Página</i>	<i>Tiempo</i>
▪ pageviews:	1.2.3.4	A	0:00
	1.2.3.4	B	0:02
	1.2.3.4	C	0:04
▪ Transacciones:	$[(p_1, w(p_1)) (p_2, w(p_2)), \dots]$		

2. Usted administra un blog y quiere implementar un sistema de recomendación de artículos. Describa una estrategia para usar los *logs* del servidor del sitio para construir una matriz de utilidad para el sistema de recomendaciones. 1 punto

Una estrategia posible puede ser, después de obtener los pageviews a partir de los logs, mapear las url a artículos del blog, con lo cual se tienen un montón de usuarios con artículos visitados. Con eso se puede poner 1 en la matriz de utilidad en las páginas vistas.

3. Usted administra un sitio web y su proveedor de *hosting* le avisa que que su sitio está usando demasiados recursos del servidor, y que quizás algún bot “descortés” esté haciendo muchas peticiones al servidor. ¿Qué heurística puede usar para detectar este bot? 1 punto

La heurística más simple es mirar el campo user-agent de los logs. Una más sofisticada puede ser, por IP, ver si hay alguna IP haciendo peticiones con una regularidad muy precisa.

4. Explique cómo sería posible sacar información de los *logs* del servidor de una página web para tomar decisiones sobre la estructura de enlaces de ésta. 1 punto

Los logs del servidor pueden mostrar cómo se mueven los usuarios en el sitio. Esto puede mostrar enlaces irrelevantes en el sitio. Por ejemplo, si muchos usuarios tienen una secuencia $A \rightarrow B \rightarrow C$, se puede recomendar poner un enlace a C en la página A.

5. Explique la diferencia entre una tienda física y una tienda en línea a la hora de decidir qué mostrar a un usuario para que compre. 1 punto

Una tienda física toma decisiones basadas en el agregado de los usuarios (típicamente se ponen los ítem más populares), mientras que una tienda en línea puede tomar decisiones basadas en el usuario. Por otro lado, las decisiones en una tienda física tienen en cuenta



el espacio de los ítem y el costo de poner y mover ítem, mientras que en una tienda en línea la decisión de qué mostrarle a un usuario es “gratis” [La prioridad de la respuesta va en el primer punto, de personalizar los ítem mostrados]

6. Explique la diferencia entre la metodología de recomendaciones basada en contenido y la técnica de filtrado colaborativo. De un ejemplo en que un sistema de filtrado colaborativo podría dar mejores recomendaciones que un sistema basado en contenido.

1 punto

Una metodología basada en contenido pasa por enganchar características de los ítem con intereses de los usuarios, y las características de los ítem son extraídas desde los ítems mismos, de alguna forma (texto, etiquetas, alguna forma de vectorización). Mientras que una metodología de filtrado colaborativo está basada en la extracción de características desde las preferencias de los mismos usuarios. Un ejemplo (bien rebuscado) puede ser que si a un usuario A le gusta un ítem I por, digamos, una característica C_1 , un sistema basado en contenido va a recomendar ítem parecidos a I , dadas las características del ítem. Sin embargo, considere que a un usuario B le gusta I por alguna característica latente (no especificada en las características del contenido), digamos C_2 . Un sistema basado en contenido no va a hacer nada diferente, mientras que un sistema basado en filtrado colaborativo comenzará a recomendarle a A ítem por la característica latente C_2 , cosa que daría más riqueza al sistema. [Aquí sirve cualquier ejemplo que muestre que los usuarios “distorsionan” el sistema de recomendaciones]