Grado en Ingeniería Informática

Minería Web

Curso 2024/2025

Práctica 3: Minería de Uso de la Web

Universidad de Jaén

1. Pre-procesamiento para análisis de logs

En esta primera parte de la práctica, el objetivo es realizar tareas de pre-procesamiento sobre datos obtenidos de registros web, para obtener información básica de los usuarios, las sesiones y sus duraciones, que permita un análisis exploratorio posterior y la aplicación de técnicas de minería de datos para la extracción de información interesante.

Se puede utilizar cualquier herramienta que el alumno considere adecuada para las tareas de pre-procesamiento. Por ejemplo, se puede utilizar una hoja de cálculo (como Excel), un paquete estadístico (como R o MatLab) o un lenguaje de programación como Python siempre que permita la edición avanzada de los datos y la generación de representaciones gráficas.

Para el desarrollo de la práctica se utilizará un conjunto de datos extraído de registros web, obtenidos del sitio web https://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html En esta dirección, se puede encontrar la descripción de los campos recogidos en el conjunto de datos. Es un fichero de un tamaño considerable, pues contiene 3461612 registros de uso de la web de un servidor de la NASA durante 2 meses. De este modo, el alumnno puede ver el reto que supone obtener esta información de la web.

Para desarrollar esta parte de la práctica, debéis descargar los datos y realizar los pasos de pre-procesamiento del fichero de registro web que se indican a continuación.

1.1. Carga del registro log y pre-procesamiento inicial

Se debe cargar el fichero log de la práctica en la hoja de cálculo o paquete estadístico elegido, de forma que cada campo del registro quede correctamente representado:

1. Cargar el fichero en la herramienta utilizando las opciones adecuadas para la correcta división de los datos en campos (aunque el valor esté vacío): Host remoto, Contraseña, Usuario, Fecha/Hora, Método, Página, Protocolo, Resultado y Tamaño.
2. Separar la fecha y hora de cada registro y crear una marca de tiempo que represente el número de segundos transcurridos desde una fecha de referencia concreta, como 1 de Enero de 1995.

1.2. Filtrado de datos

Filtrar los datos de acuerdo con los siguientes pasos:

1. Construir una tabla con las 10 extensiones de página más repetidas y el número de repeticiones.
2. Filtrar todas las extensiones excepto **.htm, .html, .pdf, .asp, .exe, .txt, .doc, .ppt, .xls y .xml.** Mantener los registros cuya extensión de página esté en blanco.
3. Explicar por qué se deben realizar los pasos anteriores.

1.3. De-spidering

Eliminar registros provenientes de usuarios no reales aplicando *de-spider* a los datos:

1. Construir una tabla con todos los *bots* y *crawlers*, indicando las proporciones relativas de estos elementos correspondientes a comportamiento automático.
2. Eliminar todos los *bots*, arañas y rastreadores de los datos.
3. Explicar por qué se realizan estos pasos.

1.4. Identificación de usuarios

Identificar los usuarios, de acuerdo con los siguientes pasos:

1. Comprobar si se dispone de un campo con el nombre de usuario, que puede ayudar a identificar los usuarios.
2. Comprobar si existe un campo *referrer* o si disponemos de la topología del sitio.
3. Añadir un nuevo campo o atributo para la identificación de usuarios.

1.5. Identificación de sesiones

Realizar la identificación de sesiones de la siguiente forma:

1. Utilizar las direcciones IP únicas y un umbral de *timeout* de 30 minutos, añadiendo un atributo con el identificador de sesión.
2. Mostrar el resultado con una tabla ordenada por el identificador de sesión y la marca de tiempo.

1.6. Problemas al estimar duraciones

1. Discutir la dificultad para estimar la duración de la última página de una sesión.
2. Sugerir una forma creativa de estimar la duración de la última página de una sesión.

1.7. Pre-procesamiento adicional

Si es necesario, realizar tareas adicionales de pre-procesamiento para identificar los valores perdidos, eligiendo una estrategia para su tratamiento.

2. Análisis exploratorio de datos del log

En esta segunda parte, realizaremos un análisis exploratorio de los datos obtenidos del log para calcular la duración de las sesiones, el tiempo medio por página, o información que ayude a entender cómo navegan los usuarios por la web. Para ello, se utilizarán los datos pre-procesados obtenidos en la primera parte de esta práctica.

Importante: para mejorar la visualización en algunos histogramas y diagramas de dispersión puede ser útil omitir temporalmente valores por encima de un umbral muy alto. Estos datos no deben eliminarse, sino sólo ser omitidos en los diagramas, indicando el umbral utilizado y el número y proporción de registros omitidos.

2.1. Duración de la sesión

Examinar la duración de la sesión:

1. Considerar sesiones que consten de una única visita. ¿Qué evidencia empírica tenemos respecto a la duración de esas sesiones?

2. De esta forma, al encontrar la duración de la sesión, necesitamos restringirnos a las sesiones que contienen más de una visita. Calcular la duración de la sesión para estas sesiones.

3. Hacer un histograma de la duración de la sesión, y un resumen estadístico que incluya media, desviación estándar, mediana, moda, mínimo y máximo.

4. ¿Crees que estos resultados subestiman o sobreestiman la verdadera duración de la sesión en todas las sesiones? ¿Por qué?

2.2. Tiempo medio por página

Calcular el tiempo medio por página.

1. Calcular el tiempo medio por página, mostrando la fórmula utilizada para derivarlo.

2. Construir un histograma del tiempo medio por página, incluyendo los estadísticos habituales. Comentar los resultados.

2.3. Eliminar comportamiento automático

Comprobar si quedan trazas de comportamiento automático:

1. Hacer una tabla con las 20 sesiones con menor tiempo medio por página. Es una forma de comprobar si hemos encontrado y eliminado todas las visitas automáticas.

2. Un tiempo medio por página menor de 0.5 segundos para una sesión indicará probablemente comportamiento automático. Eliminaremos estas sesiones, salvo, si crees que una sesión concreta está generada por usuarios reales, en cuyo caso debes indicar todas las páginas de la sesión y un argumento que avale tu hipótesis.

3. En caso necesario, actualizar los histogramas de los apartados anteriores.

2.4. Páginas visitadas

Examinar las páginas visitadas:

1. Hacer un histograma del número de visitas de página por sesión.
2. Hacer un resumen estadístico, incluyendo la media, la desviación estándar, la mediana, la moda, el mínimo y el máximo.

2.5. Relación entre visitas y duración

Explorar la relación entre las visitas de página y la duración de la sesión:

1. Utilizar un diagrama de dispersión y un modelo de regresión lineal simple, y encuentra la ecuación de regresión estimada. Superponer la línea de regresión estimada en el gráfico de dispersión.
2. Comparar la interpretación intuitiva del tiempo medio por página con la pendiente que acabas de estimar.
3. Interpretar con claridad, de forma que alguien que no sea especialista pueda entenderlo, el significado de la pendiente y el coeficiente de corte en el eje y, y si tienen sentido en este contexto.

2.6. Duración de la visita a las dos primeras páginas

Para cada sesión en la que sea posible, calcular la duración de la visita a las dos primeras páginas:

1. Hacer un histograma de la duración de la primera página.
2. Calcular los estadísticos habituales sobre la duración de ambas páginas.
3. Comparar y comentar los resultados.

2.7. Determinación del tipo de página por su extensión

Considerar cómo se pueden usar las extensiones de página para determinar si una página es de contenido o de navegación:

1. Comenzar clasificando las páginas sin extensión como correspondientes a páginas de navegación, y el resto de extensiones como páginas de contenido. Si se considera que hay una forma mejor de comenzar, se debe explicar y aplicar.
2. Comparar la duración media de página de cada una de las dos primeras páginas visitadas, separadas por navegación vs contenido.
3. Hacer un histograma normalizado de la duración media de página de cada una de las dos primeras páginas, con solapamiento de navegación vs contenido.
4. Discutir si todo esto aporta evidencias a favor de que esta forma de separar páginas de navegación y de contenido funciona o no.

2.8. Análisis de datos

Construir las siguientes tablas y gráficos, para analizar los datos obtenidos:

1. Tabla de los 20 dominios más repetidos, por número de visitas y de clics.
2. Tabla de los 7 tipos de dominio (como *.com*) más repetidos por visitas y clics.
3. Gráfico de barras de la longitud media de las visitas en 24 horas.
4. Tabla de los 10 visitantes más repetidos, por número de visitas.
5. Tabla del número de visitantes únicos, por número de visitas (del 1 al 9).
6. Tabla de las 10 páginas más visitadas, por visitas y por vistas de página.
7. Tabla de los 10 directorios más visitados, por visitas y por clics.
8. Tabla de los 10 tipos de fichero más repetidos (como *.gif*), por número de accesos.
9. Tabla de las 10 páginas de entrada más repetidas por visitas.
10. Tabla de las 10 páginas de salida más repetidas por visitas.
11. Tabla de las 10 páginas de acceso único más visitadas, por visitas.
12. Tabla de duración de las visitas en minutos (de 0 a 1 min, de 1 a 2, ...), por visitas.

3. Documentación y entrega

Se enviará un fichero ZIP que contenga lo siguiente:

- Memoria en formato PDF con los pasos llevados a cabo en cada tarea, las tablas de resultados y el análisis realizado.
- Si se usa Python, R, Excel, etc. código fuente y/o hoja de calculo que contenga el código de los estudios experimentales realizados.

4. Envío

La fecha tope de entrega será el día **16 de mayo a las 23:59** en la tarea de PLATEA habilitada a tal efecto.