Internet y el protocolo HTTP

Agentes de extracción de información en Internet

Alberto Díaz Álvarez y Francisco Serradilla García

Departamento de Sistemas Informáticos

Escuela Técnica superior de Ingeniería de Sistemas Informáticos - UPM

¿Qué es el *Hypertext Transfer Protocol* (HTTP)?

Diseñado en 1990, es la base de la comunicación de datos en la World Wide Web

- Protocolo a nivel de aplicación, sobre TCP/IP, que sigue el modelo cliente-servidor
- Su puerto estándar de comunicación asignado es el 80
- En la actualidad trabajamos con la versión HTTP/2[1] (RFC7540)
- Suficientemente abierto para no impedir el desarrollo de aplicaciones de terceros
  - Por ejemplo, agentes de información

HTTP se utiliza principalmente para transmitir recursos

- Recurso: fragmento de información identificable mediante una URL (de ahí la R)
- El tipo más común de recurso es un archivo
- También puede ser una salida generada dinámicamente (e.g. script CGI)

[1] En abril de 2016, el tráfico servido por la compañía KeyCDN superó los dos tercios en HTTP/2 frente a HTTP/1.1, la versión que este reemplazaba https://www.keycdn.com/blog/http2-statistics.

Uniform Resource Locator (URL)

Es la forma de identificar recursos en la Web, y tiene la siguiente estructura:

```
protocolo://servidor:puerto/path/hasta/el/recurso?query=con&algun=parametro
```

Por ejemplo:

```
https://www.google.com/search?q=boniato+al+horno&sclient=gws-wiz-serp
```

Si no se especifica el puerto se usa el de por defecto del protocolo (http: 80, https: 443)

Modelo cliente-servidor y el protocolo HTTP

Es un modelo de comunicación donde hay dos entidades:

- Cliente: solicita recursos al servidor
- Servidor: proporciona los recursos solicitados

Una sesión HTTP se inicia cuando el cliente envía una petición al servidor

1. El cliente establece una conexion TCP con el servidor
2. El cliente envía una petición (***request***) al servidor, y queda a la espera de la respuesta
    - Esta petición obedece a uno de los **métodos** definidos para el protocolo HTTP
3. El servidor procesa la petición y envía una respuesta (***response***) al cliente
    - Incluyendo, entre otros, un **código de estado**

Mensajes en HTTP

Toda comunicación entre dispositivos que usa HTTP se basa en dos tipos de mensaje

- **Request**: mensaje enviado por el cliente al servidor
- **Response**: mensaje enviado por el servidor al cliente

Ambos mensajes tienen caracterísiticas comunes y diferencias:

- Ambos poseen cabeceras con metainformación (*headers*)
- Ambos pueden poseer un cuerpo adicional donde se envían datos
- La request indica el método a utilizar y el recurso al que se quiere acceder
- La response indica el código de estado de la respuesta

La estructura de los mensajes difiere bastante entre los protocolos 1.1 y 2.0 de HTTP

- Lo bueno es que con las bibliotecas de Python nos abstraemos de ello

Métodos definidos para el protocolo HTTP

- **GET**: solicita un recurso al servidor
- **HEAD**: solicita los metadatos de un recurso al servidor
- **POST**: envía datos al servidor
- **PUT**: envía un recurso al servidor
- **PATCH**: envía una actualización parcial de un recurso al servidor
- **DELETE**: elimina un recurso del servidor
- **CONNECT**: establece un túnel hacia el servidor identificado por el recurso
- **OPTIONS**: solicita los métodos que el servidor soporta para un recurso
- **TRACE**: realiza una prueba de bucle de retorno de mensaje al servidor

Esta es la teoría, porque luego muchas se implementan como se quiere

Códigos de estado (*status codes*)

Los códigos de estado son números de tres dígitos que indican el estado de la respuesta

- 1xx: información
- 2xx: éxito
- 3xx: redirección
- 4xx: error del cliente
- 5xx: error del servidor

Al igual que los métodos, los códigos de estado son estándar, pero no obligatorios

Gracias