Recuperación de Información en Internet
Tema 2: La web
P.O.P. *Língua e usos profesionais*

Miguel A. Alonso Jorge Graña Jesús Vilares

Departamento de Computación
Facultad de Informática
Universidade da Coruña

Índice

¿Qué es la WWW?

- La World Wide Web, "telaraña mundial" o simplemente "**la web**" es un sistema de documentos de hipertexto enlazados y accesibles a través de Internet
- Con un **navegador web**, un usuario visualiza páginas web que pueden contener texto, imágenes u otros contenidos multimedia, y navega a través de ellas usando hiperenlaces
- La web fue creada alrededor de 1990 por el inglés Tim Berners-Lee y el belga Robert Cailliau mientras trabajaban en el CERN en Ginebra, Suiza
- el **W3C**, *World Wide Web Consortium*, es un consorcio internacional que intenta estandarizar la web

Funcionamiento de la web

Para visualizar una página web u otro recurso en un navegador implica:

1. Teclear el **URL** de la página en el navegador o seguir un enlace de hipertexto a esa página o recurso
2. Traducir el nombre del servidor del URL en una dirección IP usando el DNS.
3. Establecer una conexión TCP con el servidor web en esa dirección IP
4. Enviar una petición **HTTP** al servidor web solicitando el recurso
    1. Solicitar el texto **HTML**
    2. Analizar el texto por parte del navegador
    3. Realizar peticiones adicionales para los gráficos y otros ficheros que formen parte de la página.
5. Visualizar en al ventana del navegador la página o recurso solicitado

URL: Uniform Resource Locator

- Un URL es una secuencia de caracteres, de acuerdo a un formato estándar, que se usa para **nombrar recursos** por su localización
- Aunque subsumido en el concepto de URI, *Uniform Resource Identifier*, aún se utiliza ampliamente
- El formato general es:
  `esquema://servidor/ruta?consulta#fragmento`

Esquemas URL

Un URL comienza con el nombre de su esquema, seguida por dos puntos, seguido por una parte específica del esquema.
Algunos esquemas URL:

- **http** — recursos HTTP
- https — HTTP sobre SSL (cifrado)
- ftp — File Transfer Protocol
- mailto — direcciones E-mail
- ldap — búsquedas en el Lightweight Directory Access Protocol
- file — recusos disponibles en el ordenador local
- news — grupos de noticias Usenet (newsgroups)
- gopher — el protocolo Gopher
- ...

El servidor de un URL

- El **Servidor** consiste usualmente en el nombre o dirección IP de una máquina, seguido a veces de dos puntos y un número de puerto TCP. También puede incluir un nombre de usuario y una clave, para autenticarse ante el servidor.
  - www.grupolys.org
  - www.grupolys.org:80
  - pepiño@www.misitio.com
  - pepiño:sfgu5hfq@www.misitio.com:8000
- El soporte de nombres de usuario y claves ha sido dejado de lado por algunos navegadores. Actualmente se recomienda que los navegadores deben muestren el usuario y/o contraseña de otra forma que no sea en la barra de direcciones, a causa de los problemas de seguridad mencionados y porque las contraseñas no deben ser nunca mostradas como *texto claro*

La ruta de un URL

- La **ruta** es usada por el servidor de cualquier forma en la que su software lo establezca
- Generamente se usa para especificar un nombre de archivo, posiblemente precedido por nombres de directorio
- Ejemplo: /cole/cfp/cfp.html
  - cole sería el nombre de un directorio,
  - cfp sería el de un subdirectorio del anterior
  - cfp.html sería un nombre de archivo
- Cuando el nombre de archivo es vacío, el servidor web entiende que se está solicitando un archivo por defecto, generalmente index.html

La consulta de un URL

- Habitualmente no existe
- puede haber una sola pareja **parámetro=valor** o muchas de ellas separadas por &
- Las parejas parámetro-valor sólo son relevantes si el archivo especificado por la ruta no es una página web simple y estática, sino algún tipo de página generada bajo demanda, con
  - ASP, *Active Server Pages (ASP)*, una tenología de Microsoft
    http://www.udc.es/persoal/ga/pdi/concursos/consconcursos.asp?id=17
  - JSP, *JavaServer Pages*, una tecnología Java de Sun Microsytems
    http://www.mec.es/univ/jsp/plantilla.jsp?area=erasmus-mundus&id=2
  - PHP, *PHP Hypertext Pre-processor*
    http://www.phpnuke-espanol.org/modules.php?name=Search
  - CGI, *Common Gateway Interface*
    http://www.adobe.com/shockwave/download/index.cgi?P1_Prod_Version=Shoc
- Otro ejemplo conocido

  http://www.google.es/search?source=ig&hl=es&q=universidad&btnG=Buscar+con+Google

El fragmento de un URL

- La parte `#fragmento` de un URLes conocida como **identificador de fragmento**
- Se refiere a ciertos lugares significativos dentro de una página
- Se suele usar cuando una URL de una página ya cargada en un navegador permite saltar a cierto punto en una página larga.
- Ejemplos:

  http://www.grupocole.org/cole/publications.html#2000

  http://www.dc.fi.udc.es/~alonso/papers.html#integrationNLPIR

URL absolutos y relativos

- Un URL **absoluto** empieza por un esquema, seguido de dos puntos, seguido de una parte específica del esquema
- Un URL **relativo** comprende sólo la parte específica del esquema de un URL, o de algún componente de seguimiento de aquella parte
- El esquema y componentes principales se infieren del contexto en el cual aparece la referencia URL, el URL **base** del documento que contiene la referencia

HTML

- **HTML**, *HyperText Markup Language*, Lenguaje de Marcas HiperTextuales
- Es el formato estándar de las páginas web
- Es un lenguaje de marcación diseñado para **estructurar** textos y **presentar**los en forma de hipertexto
- Indica la estructura de presentación de un texto, no la forma exacta en la que visualizará.
- Diferentes navegadores puedes mostrar de forma diferente un mismo texto HTML

XHTML

- XHTML, *eXtensible Hypertext Markup Language*, Lenguaje eXtensible de Marcas HiperTextuales
- El sucesor de HTML: XHTML es la versión XML de HTML
- Su objetivo es avanzar en el proyecto del W3C de lograr una web semántica, que trata de separar claramente la información de la forma de presentarla
- XHTML indica la información que contiene un documento, dejando para hojas de estilo su aspecto y diseño en distintos medios

HTTP

- **HTTP**, *HyperText Transfer Protocol*, Protocolo de Transferencia de HiperTexto
- Es el sistema mediante el cual se envían las peticiones de acceso a una página y la respuesta con el contenido
- También puede enviar información adicional en ambos sentidos (por ejemplo, formularios con campos de texto,)
- Es un protocolo **sin estado**, no guarda información sobre conexiones anteriores
- Basado en un modelo cliente-servidor
  - Software del lado servidor: **servidor web** (Apache, Microsft IIS, Roxen, ...)
  - Software dellado cliente: **navegador web**
- **HTTPS** es la variante cifrada mediante SSL

Navegadores

- Un **navegador web** o *web browser* es una aplicación que permite recuperar y visualizar documentos HTML desde servidores web de todo el mundo a través de Internet.
- Historia

  1990 Primeros navegadores web desarrollados en el CERN
  1993 NCSA Mosaic
  1994 Netscape Navigator
  1995 Microsoft **Internet Explorer** (IE)
  1996 Opera
  1998 Mozilla (1.0 en 2002)
  2002 **Firefox** (1.0 en 2004)
  2003 **Safari**
  2008 **Chrome**

¿Cómo encontrar cosas en la red?

- En el inicio de los tiempos, había pocos sitios web, y los usuarios **navegaban** saltando de página en página
- Cuando un usuario consideraba que una página era interesante, la almacenaba como un **bookmark**
- Con el crecimiento exponencial de la web, esta forma de *buscar* semostró inoperante
- Aparecieron los sitios web especializados en encontrar/buscar sitios web:
  - **Directorios**
  - **Buscadores**

Directorios web

- Un directorio web es un tipo de sitio web que contiene un **directorio organizado** de enlaces a otros sitios web
- Las entradas del directorio se revisan manualmente
- El más importante ha sido **Yahoo!**
- Han ido perdiendo importancia gradualmente en favor de los buscadores
- Actualmente son un "complemento" de los buscadores (por ejemplo, Google Directorio)

Buscadores

- Recorren la web recopilando información sobre los contenidos de las páginas web
- Ante una solicitud, un buscador consulta su "base de datos" y presenta las páginas web que considera más relevantes, por orden de relevancia.
- Historia

  1993 Aliweb
  1994 **WebCrawler**, InfoSeek, **Lycos**
  1995 **Altavista**, Excite
  1997 Northern Light
  1998 **Google**
  1999 AllTheWeb, ...
  2004 Yahoo! Search
  2005 MSN Search
  2009 **Bing**

¿web = internet?

- No, pero desde el punto de vista del usuario, cada vez se identifican más
- Ello se debe en gran parte a que
  - los navegadores ya no se limitan a trabajar con HTTP, sino que tienden a incorporar cada vez un mayor número de protocolos
  - los sitios web tienden a incorporar cada vez más páginas dinámicas, applets, servlets, etc
- El navegador web se está convirtiendo en el interfaz estándar de acceso a aplicaciones "en red"