4. El correo electrónico y la navegación web

Joaquín Seoane Pascual[1], Juan Carlos Corrales Muñoz[2] y Álvaro Rendón Gallón[2]

Los servicios especializados que se pueden proporcionar y que se desarrollan más adelante (capítulos 6 y 7) están apoyados en servicios básicos, entre los que destacan el correo electrónico, como prototipo de sistema *offline*, y la Web, como prototipo de sistema *online*. Antes de entrar a describir estos servicios es importante examinar brevemente la arquitectura cliente-servidor, que constituye el fundamento de la interacción que ocurre entre las computadoras para su realización.

La arquitectura cliente-servidor nace de la evolución de los sistemas distribuidos, descentralizando el procesamiento y los recursos, sobre todo en lo correspondiente a los servicios y la visualización de la Interfaz Gráfica del Usuario. Esto hace que ciertos servidores estén dedicados sólo a una aplicación determinada permitiendo así una ejecución más eficiente de los servicios [11]. Por otra parte, los usuarios finales pueden obtener acceso a la información en forma transparente aún en entornos multiplataforma.

La arquitectura cliente-servidor tiene tres componentes principales: el cliente, la red y el servidor (Figura 4.1). El cliente envía un mensaje solicitando un determinado servicio a un servidor (hace una petición), y éste envía uno o varios mensajes con la respuesta (provee el servicio).

El lado del cliente (*front-end*) consta de una aplicación que el usuario usa para acceder al servidor. La mayoría de dichas aplicaciones tienen una interfaz gráfica de usuario que contiene controles tales como botones, listas, casillas de selección, casillas de verificación, etc. La interfaz gráfica está compuesta por un manejador de eventos que determina qué hacer ante una petición del usuario, tal como un clic; una de las acciones puede ser enviar datos al servidor, validar los datos, realizar cálculos, abrir otra ventana, etc. El segundo componente de la arquitectura es la red, la cual permite la comunicación entre los clientes y el servidor. El rendimiento de este componente juega un papel determinante ya que controla la velocidad de transporte de las solicitudes de los clientes, y el retorno de los resultados a los mismos. Finalmente, el servidor está encargado de atender a múltiples clientes que hacen peticiones sobre algún recurso

[1]Universidad Politécnica de Madrid, España
[2]Universidad del Cauca, Colombia

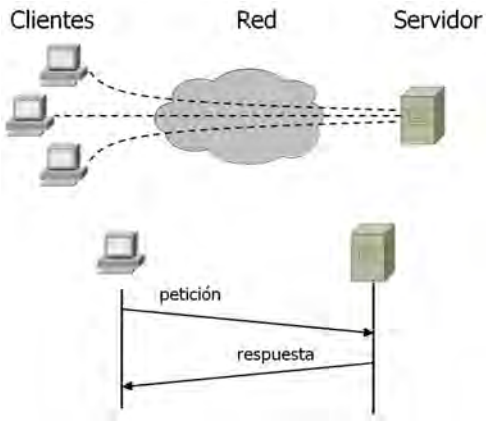

Figura 4.1.: Arquitectura cliente-servidor.

administrado por él. Al proceso servidor se le conoce también con el término *back-end* [12]. El servidor normalmente maneja la mayoría de las funciones relacionadas con las reglas del negocio y los recursos de datos. A continuación se presenta una lista de los tipos de servidor más comunes.

Servidor de Base de Datos. Este tipo de servidor proporciona servicios de acceso, gestión, administración y protección de la información (datos) a través de conexiones de red, gobernadas por unos protocolos definidos. Los usuarios acceden de modo concurrente, a través de aplicaciones cliente que pueden ser herramientas del propio sistema o aplicaciones de terceros.

Servidor de Aplicaciones. Designado a veces como un tipo de *middleware* (programa de soporte que conecta dos aplicaciones), habitualmente está ubicado entre el servidor de bases de datos y el usuario, proporcionando las funciones de la lógica del negocio y acceso a los datos de la aplicación.

Servidor Web. Básicamente sirve contenido, estático o dinámico, a los navegadores. Para suministrar el contenido el servidor carga un archivo y lo envía a través del protocolo HTTP al navegador del usuario.

Servidor *Proxy*. Se sitúa entre un programa del cliente (típicamente un navegador) y un servidor externo (típicamente un servidor web) para filtrar peticiones, mejorar el rendimiento y compartir conexiones.

Servidor de Correo. De gran importancia y uso tan extendido como el servidor web, procesa, almacena y distribuye el correo electrónico a través de las redes corporativas (LAN y WAN) y a través de Internet.

Servidor de Archivos. Permite centralizar y compartir archivos dentro de una red. En este esquema, cualquier cliente conectado a la red y con los permisos apropiados puede ver o modificar los archivos alojados en el servidor.

Servidor de Impresión. Permite a diferentes nodos compartir impresoras en la red a través de interfaces paralelo o serie y los protocolos adecuados.

Servidor *Groupware*. Permite organizar el trabajo de un grupo de usuarios, de modo que éstos puedan colaborar sin importar su localización, vía Internet o vía una Intranet corporativa.

En las siguientes secciones se describirán los servicios básicos ofrecidos por los servidores de correo, los servidores web y los servidores *proxy*.

4.1. Correo electrónico

El correo electrónico entre usuarios de máquinas distintas es uno de los servicios más antiguos y exitosos, solamente precedido por los sistemas de transferencia de archivos y terminal remoto. En efecto, a principos de los 70, el correo electrónico de ARPANET fué una extensión de su protocolo de transferencia de ficheros, que requería conexiones dedicadas. Mientras, el trabajo de comunicación entre máquinas Unix y MS-DOS se adaptó al uso de líneas de la red telefónica conmutada, con modems que entonces tenían una velocidad de 300 bits por segundo, muchas veces acoplados acústicamente al teléfono.

4.1.1. El correo normal de Internet

La transferencia de correo electrónico de Internet está soportada por el protocololo TCP y presupone una alta fiabilidad y un bajo coste de las conexiones. El direccionamiento está basado en el Sistema de Nombres de Dominio (DNS) y es de la forma `usuario@dominio`, donde `dominio` puede ser el nombre de una máquina concreta en la cual el `usuario` tiene cuenta, pero lo más común es que sea el dominio de una organización y que el correo se transfiera realmente a las máquinas intercambiadoras de correo de esa organización (registradas en el DNS por medio de registros MX); estas máquinas pueden pertenecer a la organización o ser propiedad de un proveedor de acceso a Internet.

En Internet existen máquinas conectadas permanentemente (como los intercambiadores de correo y los servidores de DNS) y otras que se conectan cuando interesa (los terminales o computadoras), generalmente de forma económica usando accesos ADSL, Wi-Fi, etc. Para enviar correo, los terminales generalmente usan variaciones del venerable SMTP [13] para depositar el mensaje en el servidor del destino, o más normalmente en un *smart host* cercano, que será el que haga todo lo posible por entregar el mensaje al intercambiador del destino, también usando SMTP.

El protocolo SMTP básico es muy sencillo y ha dado paso a los primeros problemas de correo no solicitado (*spam*), ya que cualquier máquina sin defensas apropiadas (*Open Relay*) podía aceptar cualquier mensaje, posiblemente con identidad falsa, y entregarlo a cualquier destinatario. Bien configurado, SMTP permite autenticar y autorizar máquinas o redes, y a usuarios, usando distintas opciones [14, 15], pero el correo no solicitado es un problema persistente que utiliza constantemente técnicas nuevas que han de ser contrarrestadas por administradores expertos. El problema se hace acuciante en una red rural de muy bajo ancho de banda (por ejemplo, vía onda corta).

Aunque un terminal o computadora puede tener un agente de transferencia de correo (MTA) *interno*[3], que se encargue de los envíos y entregas, cada vez es más frecuente que el cliente de correo electrónico o agente de usuario (MUA) hable directamente con un *smart host* externo para enviar sus mensajes por SMTP y acceder a su buzón de entrada por medio de POP [16] o, mejor, por IMAP [17], con múltiples ventajas: accesibilidad desde cualquier punto conectado del planeta, posibilidad de trabajo fuera de línea (modo desconectado), gestión de múltiples buzones, búsqueda en servidor, transferencia parcial, etc., muchas de ellas importantes a la hora de usar un enlace de bajo ancho de banda o intermitente.

Prácticamente todos los agentes de usuario funcionan así[4], ofreciendo además la posibilidad de cifrar la transferencia. En el caso de tener que transferir información sensible (por ejemplo, médica), deberían utilizarse mecanismos de cifrado y firma electrónica, como PGP [18] y S/MIME [19], además de transferir los datos por una conexión cifrada (protocolo TLS [20]). De igual forma utilizan mecanismos antispam y antivirus, que también proprocionan muchos servidores[5]. Del otro lado habrá los correspondientes servidores de correo electrónico[6].

Los MTA tienen o utilizan distintos agentes de entrega de correo (MDA), también llamados *transportes*, entre los que destacan el agente de entrega local, que sitúa los mensajes recibidos en el sistema de ficheros, el agente que transfiere el mensaje vía SMTP a otro MTA (lo estándar en Internet), y los agentes que utilizan distintas vías para transportar los mensajes en situaciones donde SMTP no es posible o deseable. Estas situaciones se presentan por ejemplo cuando sólo se dispone de redes de radio, donde puede ser más conveniente usar UUCP (ver 4.1.3.1), o de llamadas telefónicas de larga distancia, donde podría ser más eficiente usar *Fido* (ver 4.1.3.2); el agente también puede ser un fax o una impresora que escriba cartas ordinarias, muy útil en casos de guerra o aislamiento telemático, donde sí funciona el cartero (ver 4.1.3.3).

4.1.2. Webmail

Otra posibilidad de utilizar el correo, hoy día mayoritaria, es a través de una interfaz web. Aunque más incómoda para el usario avezado, y sin posibilidad de almacenar los mensajes localmente de forma organizada, es extremadamente sencilla de aprender y suficiente para el usuario normal. Además es la más accesible para el usuario viajero, ya que en muchos lugares los cortafuegos de protección siempre permiten tráfico web.

De sobra son conocidos *hotmail, yahoo, zoho* o *gmail*, todos gratuitos, con herramientas antispam mantenidas profesionalmente. No obstante debemos ser conscientes de que las interfaces web no permiten que los mensajes queden cifrados en el servidor. *zoho* y *gmail* permiten además dominios propios y acceso POP e IMAP, lo que los

[3] Son o contienen MTA: *sendmail*, *postfix*, *exim4*, *courier*, *cyrus*, etc., entre los programas libres.
[4] *evolution*, *thunderbird*, *kmail*, *sylpheed*, *seamonkey-mailnews*, etc., en modo gráfico, o *mutt* en modo texto.
[5] Por ejemplo, *spamassassin*.
[6] *dovecot*, componentes de *courier* y *cyrus*, además de las herramientas de GNU y de la Universidad de Washington.

hace muy interesantes para todo tipo de usuarios; tienen sin embargo el inconveniente de presentar propaganda en la interfaz web y, sobre todo, la incertidumbre de que nuestros mensajes están en una organización externa que los utiliza de un modo u otro para su beneficio.

Si se opta por una solución interna, hay numerosos paquetes de Webmail libres disponibles[7], la mayoría clientes de IMAP o POP.

4.1.3. Alternativas al correo de Internet

A veces no es posible o sencillo usar el correo de Internet por múltiples razones. Veamos algunas soluciones y los problemas que resuelven.

4.1.3.1. UUCP

Las líneas dedicadas eran (y son) muy caras, y una posibilidad para enviar y recibir correo electrónico era hacer uso de la red telefónica conmutada, originalmente con modems de 300 baudios. Trabajando esto surgió el protocolo UUCP, de copia de ficheros y ejecución remota *offline*, y los sistemas de transferencia de correo y noticias basados en él. Éstos requerían el conocimiento de la ruta entre las máquinas de origen y destino (por ejemplo `sehas!nmadrid!mcvax!nlima!npucp!shuc!jaime`, denominado un *bang path*). Obviamente, como los intercambios de ficheros entre vecinos se planificaban para minimizar los costes de comunicaciones, en los que el establecimiento de llamada tiene un porcentaje importante, una transferencia como esa podía tardar varios días.

En los años 80 se mantuvo un mapa de interconexiones entre máquinas de nombre único, de modo que finalmente la dirección anterior pudo convertirse en `jaime@shuc.uucp`. La no escabilidad de un sistema de nombres plano y la dificultad de mantener el mapa de red, junto con el despliegue de Internet y su sistema jerárquico de nombres, permitió ocultar los nodos UUCP detrás del intercambiador de correo más próximo conectado a Internet, obteniéndose ya direcciones como `jaime@shuc.aa.pe.ehas.org` si en el DNS hay un registro MX que diga que el intercambiador de correo para `shuc.aa.pe.ehas.org` es una máquina conectada a Internet por un lado y, directa o indirectamente, a la máquina `shuc`.

Hubo un tiempo en que UUCP se usaba básicamente con líneas telefónicas conmutadas y, durante los años 80 y 90, además de servir para potenciar la investigación y el desarrollo del software libre, sirvió para comunicar agentes de desarrollo de América Latina, África, Asia y Europa del Este, por medio de los servicios que proporcionaban organizaciones que luego formaron la APC (EcoNet/PeaceNet, GreenNet, Web, IBASE, etc). Por ejemplo, Cuba estuvo intercambiando todo su tráfico por UUCP con llamadas internacionales a Web (Canadá) y GreenNet (Reino Unido) desde 1992 hasta tener su conexión a Internet a finales de 1996, desarrollando también una compleja red UUCP interna.

[7] *squirrelmail*, *sqwebmail*, *roundcube*, etc.

Hoy día el uso de llamadas telefónicas a larga distancia tiene poco sentido, ya que en cada teléfono se dispone de acceso a internet por PPP, ADSL, etc. ¿Qué interés tiene entonces UUCP? Nos fijaremos especialmente en el espléndido paquete Taylor UUCP [21]:

1. La comunicación se puede interrumpir en cualquier momento por la caída del enlace, y reanudarse por donde iba cuando haya conectividad. Esto es muy valioso para comunicaciones costosas o poco fiables y mensajes grandes. En el mundo SMTP sólo parece soportar esto Microsoft Exchange, que es una extensión opcional de SMTP (CHUNKING y CHECKPOINT).

2. Puede transportarse encima de un enlace o transporte que asegure diversos niveles de fiabilidad, dúplex o semidúplex, de 7 u 8 bits, evitando ciertos caracteres o no, etc. Para ello soporta una buena cantidad de protocolos con nombres peculiares (g, G, i, j, y, t, e, f, v, y). Por ejemplo, Taylor UUCP viene directamente preparado para ir sobre TCP, con lo que su protocolo de transferencia de ficheros no tiene que preocuparse del ruido ni de la transparencia ni del control de flujo ni de alternar transmisión y recepción (protocolo `t`). Pero si además queremos cifrar la conversación, podemos intercalar un túnel SSH [22].

3. Y si tenemos un enlace inalámbrico fiable semidúplex, como el AX.25 [23] de los radioaficionados, emplearemos el protocolo `y`, como se ha hecho en Alto Amazonas y otros lugares del Perú rural [24].

4. Sirve para comunicar dominios enteros con un número desconocido de buzones. En efecto, la máquina conectada a Internet puede ser el intercambiador de `shuc.aa.pe.ehas.org`, una de cuyas máquinas llamará o será llamada para intercambiar mensajes encolados durante la desconexión. O puede ser el intercambiador de todo Alto Amazonas (`aa.pe.ehas.org`), todo el Perú (`pe.ehas.org`) o incluso de cualquier lugar del mundo bajo un mismo dominio administrativo (¿`ehas.org`?). ESMTP tiene la orden ETRN para hacer lo mismo, pero si el proveedor da direcciones IP dinámicas, el método es inseguro, ya que el servidor puede no saber si el que llama está autorizado o es un impostor.

5. Se puede aumentar la eficiencia comprimiendo y empaquetando la cola de mensajes pendientes, dividiendo luego el lote en trozos iguales, de tamaño apropiado al medio. Así por ejemplo, en redes como las descritas en [24] no deberán mandarse mensajes de más de 100 kB en VHF o de 10 kB en HF, siendo además muy ineficiente el envío aislado de mensajes típicos, de 2 ó 4 kB. La herramienta básica para hacerlo es BSMTP sobre UUCP [25].

4.1.3.2. Fido

Sin embargo, y en paralelo con UUCP, la gran mayoría de las comunicaciones de nodos *pobres* basados en computadoras con MS-DOS se hizo utilizando una tecnología procedente del mundo aficionado a los BBS, que desarrolló FIDONET [26], con un sistema de correo (Netmail) y grupos (Echomail) muy eficientes. Debido a esa eficiencia y tolerancia a redes malas, jugó un papel importante para coordinar la resistencia a

la guerra y ayudar a los refugiados en la ex Yugoslavia [27]. Con el paquete *ifgate* se puede conectar una red con tecnología Fido a Internet.

4.1.3.3. Transportes humanos

En zonas donde la comunicación electromagnética no es posible, el intercambiador de correo puede imprimir las cartas, que son distribuidas a mano. Basta poner como transporte, la orden de imprimir. Y si sólo tenemos un fax en destino, usaremos la pasarela a fax. Las direcciones de fax contendrán el número de fax (e.g. `34933333333@faxgw.org`), mientras que las de papel pueden tener codificado de alguna manera el lugar de destino, para encaminar el mensaje al fax o a la oficina de correos más próxima. En ambos casos la recepción requiere intervención humana. Ambas aplicaciones extremas se han utilizado en la guerra de los Balcanes [27]. También hay experiencias de utilizar al cartero para enviar los mensajes en formato digital (pendrive, CD o DVD), como en [28, 29], o que el cartero, a bordo de un vehículo, lleve los mensajes en un enrutador Wi-Fi [30] (ver también 4.1.6).

4.1.4. Listas de correo y foros

Hoy día la gente está acostumbrada a variados mecanismos de comunicación de grupos, ya sean los que ponen a su disposición las redes sociales, o los foros de discusión sobre temas diversos. La interfaz web da muchas posibilidades pero tiene también algunos inconvenientes para el usuario, siendo el más importante en nuestro caso que no es muy viable en situaciones de mala conectividad.

Las listas de correo usan el correo electrónico para participar en foros, a los que uno puede suscribirse con un mensaje especial, que sirve también para otras operaciones como borrarse o conocer a los miembros del grupo si el administrador lo permite. En principio el administrador de un grupo podría gestionarlo por correo electrónico (por ejemplo, autorizar nuevos miembros), con lo que puede mantenerse todo un sistema de conferencias sin conectividad Internet. Así funcionaban el viejo *majordomo* y *smartlist*. Hoy día *Mailman*, que es el gestor de listas por excelencia, compatibiliza una interfaz web con otra de correo, siendo la interfaz web del administrador más potente que la de correo (se supone que el administrador está en un sitio bien conectado).

También comparten esa filosofía de soportar malas conexiones los grupos de noticias de USENET, hoy en franca decadencia, que en sus inicios usaban UUCP para la transferencia de mensajes y de control.

4.1.5. Robots de correo

En zonas de mala conectividad, la gestión de la red y los equipos es un problema grave. Afortunadamente los robots de correo nos ayudan. Por ejemplo, los gestores de listas están realizados con estos robots, que no son más que direcciones especiales donde, cuando se recibe un mensaje se pasa a un programa (el robot propiamente dicho), que

no es más que un *transporte* especial. Esto puede ser usado para reiniciar remotamente un equipo, borrar ficheros sobrantes si el disco está lleno, conocer detalles del tráfico cursado, aplicaciones ejecutadas, etc. Obviamente gestionar de esta manera requiere mensajes firmados digitalmente, para obedecer sólo aquellas órdenes emitidas por un administrador autorizado.

Un proyecto de comunicación Wi-Fi en el Río Napo (Perú) utilizó este sistema un tiempo [31, 32]; no obstante, al ser una red con conectividad permanente Wi-Fi, se ha optado por algo más estándar, como se describe en 21.2.

4.1.6. Redes tolerantes al retardo

Muchos de los problemas que resuelve el correo ordinario, junto con los robots de correo, podrán resolverse de una manera más sistemática por medio de las llamadas Redes Tolerantes al Retardo (DTN) [33, 34, 35, 36, 37], una familia de protocolos orientada a *fardos de datos,* que viajan entre origen y destino saltando de nodo a nodo cuando se les presenta la oportunidad. Un ejemplo de aplicación es un conmutador móvil transportado en un vehículo, que intercambia fardos en forma inalámbrica con conmutadores fijos en los puntos que recorre.

4.2. La Web

Cuando la WWW (*World Wide Web*) fue puesta en funcionamiento por Tim Berners-Lee en los laboratorios del CERN (Centro Europeo para la Investigación Nuclear) en 1990, es poco probable que alguien hubiera imaginado el impacto que este nuevo servicio de Internet tendría sobre la sociedad moderna. Lo que simplemente empezó como un mecanismo basado en hipertexto para acceso a información, con mejores prestaciones que las ofrecidas por los servicios de entonces, el FTP (*File Transfer Protocol*) [38] y el Gopher [39], se convirtió en la aplicación estrella ("*killer application*") de Internet, contribuyendo enormemente a su popularización, al punto que el término Web, que deriva del nombre del servicio, es usado a menudo como sinónimo de Internet, que deriva de *Internet Transmission Control Protocol*, el nombre original del conjunto de protocolos TCP/IP que le sirve de base. Los clientes de navegación en la Web, o simplemente "navegadores web", eliminaron la barrera de acceso a Internet impuesta por las aplicaciones anteriores, que exigían conocimientos de la red y del uso de comandos, y pusieron Internet al alcance de todo tipo de usuarios, a través de una interfaz de muy fácil manejo.

Los pilares de la navegación web son el Lenguaje de Marcado para Hiper-Texto (HTML, *HiperText Markup Language*) [40] y el Protocolo de Transferencia de Hiper-Texto (HTTP, *HiperText Transfer Protocol*) [41]. La información se ofrece a través de documentos elaborados con HTML, llamados páginas web, que residen en un servidor, tal como se muestra en la Figura 4.2. Los usuarios disponen de una aplicación cliente, llamada navegador, que utiliza el protocolo HTTP para solicitar y obtener del servidor las páginas, e interpreta su contenido para presentar al usuario la información ofrecida.

Esta información contiene enlaces de hipertexto (hiperenlaces) que pueden conducir a otros servidores en Internet que ofrecen más información.

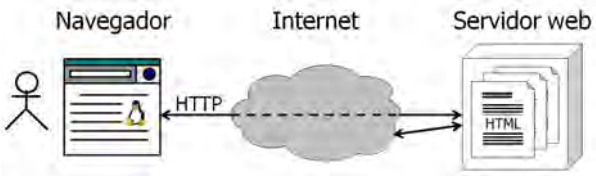

Figura 4.2.: Navegación en la Web.

4.2.1. El navegador web

El primer navegador, llamado por Berners-Lee, su autor, WorldWideWeb [42], estaba basado en texto. Aunque se convirtió en una herramienta práctica de acceso a información en Internet para los científicos del CERN y la comunidad académica, la navegación web sólo se hizo realmente popular cuando Marc Andreessen y Eric Bina, del Centro Nacional de Aplicaciones de Supercomputación (NCSA) en la Universidad del Illinois, escribieron en 1993 Mosaic [43], un navegador para el entorno de ventanas X-Window de Unix y por tanto con una interfaz totalmente gráfica. Si bien aún existen navegadores basados en texto como el Lynx[8], la mayoría de ellos presentan información de tipo multimedia (texto, imágenes, sonido, vídeo, etc.). En la actualidad, los navegadores más usados son Internet Explorer, Firefox, Chrome, Safari y Opera[9].

Una de las funciones más importantes del navegador es capturar los clics del ratón sobre los hiperenlaces de la página presentada al usuario. Cuando recibe este evento, el navegador debe cargar la página indicada en el hiperenlace, la cual utiliza una convención de nombres denominada URL (*Uniform Resource Locator*), que consta esencialmente de tres partes: el nombre del protocolo para acceder al recurso (esquema), el nombre de dominio de la máquina donde reside el recurso, y la ruta y nombre del archivo que contiene el recurso. Por ejemplo, en la URL `http://www.w3.org/standards/about.html`:

- `http`: uso del protocolo HTTP.
- `www.w3.org`: nombre de dominio del servidor web del W3C.
- `standards/about.html`: ruta y nombre de la página web.

Los protocolos que pueden incluirse en la primera parte de la URL, según el RFC 1738, se muestran en la Tabla 4.1.

La segunda parte de la URL puede incluir la dirección IP del servidor web en lugar de su nombre de dominio, y además, de manera opcional, el número del puerto TCP que atiende las peticiones, que por defecto es 80 para el protocolo HTTP.

[8] `http://lynx.isc.org/`
[9] `http://gs.statcounter.com/#browser-ww-monthly-201108-201108-bar`

Esquema	Protocolo
ftp	FTP
http	HTTP
gopher	Gopher
mailto	correo electrónico
news	noticias de USENET
nntp	noticias de USENET usando el acceso NNTP (*Network News Transport Protocol*)
telnet	acceso a una máquina remota
wais	Wide Area Information Servers
file	acceso a un archivo local
prospero	servicio de directorio Prospero

Tabla 4.1.: Esquemas cubiertos por el URL (RFC 1738).

El nombre y la ruta de la tercera parte son opcionales. Cuando no aparecen, normalmente la URL apunta a la página web principal de la organización; y cuando sólo aparece un directorio, normalmente está implícito el nombre del archivo `index.html`.

4.2.2. El servidor web

La principal función del servidor web es la de responder a las peticiones HTTP entregando las páginas web (documentos HTML) solicitadas así como los contenidos que pueden estar descritos en ella, tales como archivos multimedia, hojas de estilo y archivos de comandos.

No todos los servidores web están destinados a ofrecer información para la Internet. Cada vez es más común encontrar servidores web incorporados a diversos dispositivos para accederlos en forma remota, como por ejemplo para administrar un enrutador o una impresora, o para visualizar las imágenes de una cámara web.

Los primeros servidores web, como los que se escribieron con el cliente WorldWideWeb en el CERN y con Mosaic en el NCSA, se llamaron simplemente HTTPd (demonio HTTP). El HTTPd del NCSA fue el origen del servidor web Apache, desarrollado y mantenido por la *Apache Software Foundation*[10] como un proyecto paradigmático de la programación de código abierto (OSS, *Open Source Software*), y que a partir de 1996 ha sido el más usado; en mayo de 2011, el 63 % de los servidores web del planeta utilizan Apache[11]. Otros servidores web actualmente en uso son: Internet Information Services (IIS) de Microsoft, nginx[12], GWS de Google, lighttpd[13], iPlanet Web Server de Oracle y AOLserver[14].

[10] `http://www.apache.org/`
[11] `http://news.netcraft.com/archives/2011/05/02/may-2011-web-server-survey.html`
[12] `http://www.nginx.org/`
[13] `http://www.lighttpd.net/`
[14] `http://www.aolserver.com/`

La rápida penetración del uso de Internet ha llevado a su vez a la aparición de nuevas tecnologías que, como se ilustra en la Figura 4.3, buscan extender el uso de la Web mucho más allá de la simple navegación a través de documentos multimedia. Del lado del servidor, tecnologías como CGI (*Common Gateway Interface*) [44], PHP (*Hypertext Preprocessor*) [45], Servlets de Java [46] y ASP.NET (*Active Server Pages*) [47] permiten acceder y procesar información en bases de datos o comunicarse con aplicaciones que a su vez pueden interactuar con otras aplicaciones en red, y entregar sus resultados al usuario a través de páginas web creadas en forma dinámica.

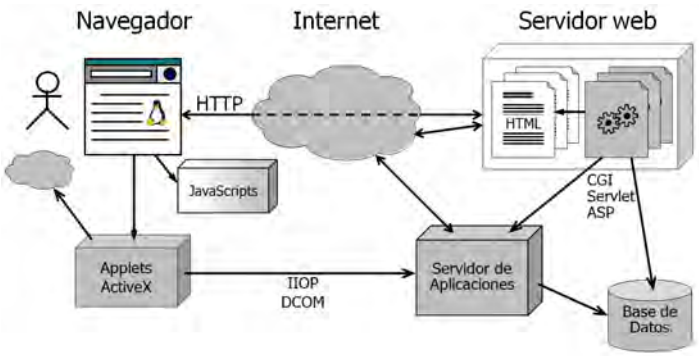

Figura 4.3.: Aplicaciones en Internet.

Del lado del cliente, se han desarrollado tecnologías como los JavaScripts [48], que permiten mejorar la presentación de la información al usuario y realizar validaciones y cálculos limitados, y los Applets [49] de Java y componentes tipo ActiveX [50] y JavaBeans [51], que son además capaces de acceder a dispositivos locales del usuario y a otras aplicaciones en la red a través de los protocolos IIOP (*Internet Inter-ORB Protocol*) [52] y DCOM (*Distributed Component Object Model*) [53].

El primer paso en la dirección de cambiar el papel de los usuarios de la Web de consumidores a productores de información, fue la introducción de los formularios en HTML 2.0. En esta nueva versión del lenguaje para producir páginas web, se incluyeron construcciones para agregarles cajas y áreas de texto, casillas de verificación (*checkbox*), botones de opción (*radio button*), listas de selección, imágenes sensibles y los botones de borrado y envío. Todos estos elementos permiten al navegador capturar información de los usuarios que luego envía al servidor web donde es almacenada y procesada.

4.2.3. Servicios de la Web

La Web fue creada inicialmente para proveer información. Un navegador sólo presentaba al usuario la información que había sido registrada el servidor web por el administrador del sitio. La introducción de los formularios permitió un mayor grado de interacción, dando lugar a la creación de un conjunto de servicios que poco a poco irían incrementando la influencia de Internet en la sociedad.

Uno de los primeros servicios ofrecidos por la Web fueron los directorios o índices temáticos de información, que presentan un catálogo de servidores web organizado por categorías. El más conocido es el directorio de Yahoo![15], y se destaca igualmente el Open Directory Project (ODP)[16], conocido también como Dmoz (*Directory Mozilla*), "el más grande y exhaustivo directorio de la Web editado por humanos", construido y mantenido por una vasta comunidad global de editores voluntarios [54].

El rápido crecimiento de sitios web hizo cada vez más difícil el mantenimiento de los directorios y la localización de información a través de los mismos, por lo que se creó para la Web un servicio que ya existía para los sitios de FTP y Gopher, que es el de los buscadores. El primer buscador completamente indexado fue WebCrawler[17], en 1994, seguido por muchos otros como Lycos, Excite, Infoseek, Altavista, y últimamente Google, que hasta marzo de 2010 fue el sitio de Internet más visitado en los Estados Unidos [55]. Los principales componentes de un buscador son [56]:

- **Rastreador (*crawler*)**. Llamado también araña o robot web, es un programa que recorre las páginas web que hay en Internet, las descarga para hacer la indexación, y sigue los hiperenlaces que contienen. Según Google, Googlebot, que es su programa de rastreo, "utiliza un proceso de rastreo algorítmico: a través de programas informáticos se determinan los sitios que hay que rastrear, la frecuencia y el número de páginas que hay que buscar en cada sitio" [57].

- **Indexador**. Es el encargado de crear el índice de búsqueda a partir de las páginas web que recibe del rastreador. En el caso de Googlebot, además de elaborar el índice de las palabras que encuentra y su ubicación en cada página, procesa la información incluida en las etiquetas (e.g. Title) y los atributos de contenido clave (e.g. ALT). Se estima que en agosto de 2011 Google indexó 47 mil millones de páginas web[18].

- **Índice de búsqueda**. Es el repositorio de datos que contiene toda la información que el buscador necesita para asociar y obtener las páginas web. La estructura de datos usada para el índice de llama "archivo invertido" y consiste en un listado de palabras en orden alfabético, donde cada palabra tiene asociada una lista de referencias a las páginas web donde aparece.

- **Motor de consulta**. Es el corazón algorítmico del buscador. Procesa la consulta del usuario en dos pasos: primero obtiene del índice de búsqueda información de las páginas web potencialmente relevantes asociadas a las palabras clave de la consulta, y luego produce una clasificación de los resultados, del más relevante hacia abajo.

- **Interfaz de búsqueda**. Presenta los resultados al usuario, una vez procesada la consulta, permitiéndole realizar nuevas consultas, navegar en la lista resultante y seleccionar páginas web para consultarlas.

[15] `http://dir.yahoo.com/`
[16] `http://www.dmoz.org/`
[17] `http://www.webcrawler.com/`
[18] `http://www.worldwidewebsize.com/`

Los servidores web también juegan un papel muy importante en la prestación del servicio de correo electrónico, mediante los Webmail (*web-based e-mail*), tal como se explicó en el apartado 4.1.2.

Otro servicio muy popular de la Web son los foros de discusión, que empezaron siendo tableros de anuncios (*bulletin board*) implementados mediante listas de correo electrónico como se describe en el apartado 4.1.4, y que han encontrado en la Web un excelente soporte. Existe una inmensa variedad de foros sobre los temas más diversos, donde los usuarios pueden formular y responder preguntas, compartir y encontrar comparaciones, participar en encuestas de opinión, y por supuesto debatir sobre algún asunto. Dos de las aplicaciones más utilizadas para la construcción de foros son vBulletin[19] y phpBB[20].

A medida que la Web se ha convertido en la fuente de información más consultada, ha crecido el interés de las empresas, entidades públicas y los propios usuarios por presentar allí su información a través de páginas y portales web. Los Sistemas de Gestión de Contenidos (CMS, *Content Management Systems*) ofrecen soporte para la administración de los contenidos presentados en un portal, permitiendo un alto nivel de interacción a administradores, editores y visitantes. Además de agilizar la difusión de contenidos multimedia, facilitan la incorporación de diversos servicios como búsqueda, foros y algunos de la Web 2.0 comentados más adelante. Entre las herramientas de CMS de código abierto más utilizadas están Drupal[21] y Joomla[22].

La creciente popularización de la Web, unida a las facilidades ofrecidas por los desarrollos tecnológicos, dieron paulatinamente lugar a la creación de servicios que tienen como común denominador una mayor participación de los usuarios en la generación de los contenidos y mecanismos más expeditos para la interactividad y la colaboración. Se ha pasado entonces de una Web donde los usuarios eran fundamentalmente consumidores de información, la Web 1.0, a otra donde los usuarios encuentran una amplia variedad de mecanismos y servicios para compartir contenidos: la Web 2.0. Entre estos servicios se destacan las bitácoras (*blogs*), las redes sociales (Facebook, LinkedIn, etc.), los editores colaborativos de contenido (*wikis*), la distribución de archivos de audio y video por suscripción (*podcast*), los marcadores sociales (i.e. Delicious[23]), el etiquetado colaborativo (*folksonomy*), la sindicación o redifusión web (basada en RSS), y los sistemas para compartir contenidos multimedia como documentos (i.e. Google docs), fotografías (i.e. Flickr), presentaciones (i.e. SlideShare) y videos (i.e. YouTube).

Se pretende acuñar también el término Web 3.0 para referirse a una nueva Web en construcción. No existe consenso sobre la definición precisa del término, pero las dos tendencias predominantes, la Web Semántica y la Internet de los objetos (IoT, *Internet of Things*), tienen como común denominador una Web donde los servicios y contenidos son compartidos también por las máquinas. La Web Semántica está orientada al uso de mecanismos como ontologías y metadatos en la descripción de los contenidos, de

[19] `https://www.vbulletin.com/`
[20] `http://www.phpbb.com/`
[21] `http://drupal.org/`
[22] `http://www.joomla.org/`
[23] `http://www.delicious.com/`

modo que los computadores puedan procesar para los humanos la creciente cantidad de información y recursos presentes en la Web. Por su parte, IoT, que también tiene muchas definiciones, hace referencia a la integración en la Web de objetos físicos y virtuales con identidades y atributos que usan interfaces inteligentes para comunicarse entre sí, con los humanos, y con el entorno, participando de manera activa en procesos de información, de negocios y sociales [58].

4.2.4. Cachés y *proxies*

Cuando un servidor web recibe una petición HTTP, en principio debe buscar en el disco duro el archivo HTML solicitado si se trata de una página estática, o incluso efectuar varios accesos al disco si se trata de una página dinámica. Estos tiempos de acceso al disco restringen notablemente el rendimiento del servidor web en términos del número de peticiones que puede atender por unidad de tiempo. Por esta razón, todos los servidores web utilizan un mecanismo de caché, guardando en memoria (por un tiempo de validez configurable) el resultado de las últimas consultas. Cuando recibe una nueva petición, averigua si la respuesta está guardada en el caché y en tal caso la recupera y la entrega sin necesidad de acceder al disco. Aunque un buen sistema de caché requiere gran cantidad de memoria y tiempo de procesamiento adicional para gestionarlo, casi siempre la ganancia en rendimiento lo justifica [59].

En términos más generales, se denomina caché al almacén de mensajes de respuesta y al sistema que controla el almacenamiento, recuperación y borrado de estos mensajes. El caché puede operar en un servidor, en un cliente o en un sistema intermedio, y su objetivo es reducir el tiempo de respuesta y el consumo de ancho de banda de la red. No todas las transacciones pueden usar el caché, y el cliente o el servidor pueden definir que ciertas transacciones sean llevadas al caché por un tiempo limitado [60].

Por su parte, se denomina *proxy* a un programa intermediario que actúa como servidor y como cliente con el propósito de enviar solicitudes a nombre de otros clientes; actúa como servidor cuando interactúa con los clientes, y como cliente cuando interactúa con el servidor. El *proxy* recibe la solicitud del cliente, la interpreta, y si es necesario reescribe el mensaje de petición antes de reenviarlo al servidor; por consiguiente, puede ser utilizado como intermediario de seguridad, haciendo las veces de portal del lado del cliente en redes con cortafuegos, o como traductor de protocolos, cuando los clientes no manejan los mismos protocolos o las mismas versiones que el servidor [60].

En función de su posición en la red, se encuentran dos tipos de *proxy*: de reenvío e inverso. Un *proxy* de reenvío (*forward proxy*) presta servicio en una red interna (LAN) para permitir a sus usuarios acceder a los servidores externos, normalmente en Internet. Una razón muy común para requerir un *proxy* de reenvío es el uso de direcciones IP privadas en la red interna; se requiere entonces la intervención del *proxy* para trasladar las direcciones IP privadas de las peticiones a direcciones IP públicas, y luego hacer el proceso inverso para las respuestas. Estos *proxies* también pueden realizar funciones de caché, para reducir el uso del ancho de banda que tiene disponible la LAN para acceder a Internet, y filtrado de conexiones, para evitar el acceso de los usuarios internos a ciertos contenidos (uso de listas negras).

Por su parte, el *proxy* inverso (*reverse proxy*) se localiza del lado del servidor web, recibiendo las peticiones de todos los clientes que desean acceder al sitio y reenviándolas a aquél. Un uso típico del *proxy* inverso es proveer a los usuarios de Internet acceso a un servidor web que se encuentra detrás de un cortafuegos, pero también puede ser usado para balanceo de carga distribuyendo las peticiones entre varios servidores, para servir de caché a un servidor con prestaciones limitadas, o para permitir la coexistencia de varios servidores web en el mismo espacio URL [61].

Existen en el mercado diversos productos *proxy* que combinan múltiples funcionalidades. Entre ellos se destaca Squid[24], con licencia GPL, que funciona como *proxy* y como caché. Hay que tener en cuenta que la mayoría de las plataformas para servidores web (Apache, IIS, etc.) ofrecen soporte para la configuración de *proxy* y caché, pero Squid sólo es un *proxy* y no puede servir páginas por sí mismo [62].

[24] `http://www.squid-cache.org/`