Ministero dell'Economia e delle Finanze

Ministero del Lavoro e delle Politiche Sociali

PEGASO
SOCIETÀ COOPERATIVA SOCIALE

MODULO:
MANUTENZIONE E DIAGNOSTICA DEL PC

Internet: come funziona

Premessa

Questa lezione contiene alcune informazioni tecniche su cos'è e come funziona Internet. Potreste pensare che non occorra conoscerle per usare la Rete, visto che, ad esempio, non occorre essere esperti di telecomunicazioni per usare il telefono. In realtà, vale la pena di fare un piccolo sforzo: Internet è ben più difficile da usare del telefono e senza queste conoscenze non andreste molto lontano.

Le basi

Cominciamo proprio dall'inizio, con una breve descrizione dei servizi telematici (i *BBS*) e delle reti di computer. Un BBS ovvero *Bulletin Board System* cioè, letteralmente, bacheca elettronica, è un programma speciale che risiede su un certo computer e consente che altri computer si colleghino a lui via telefono. Per usare il servizio telematico, l'utente deve installare nel suo computer un dispositivo di comunicazione, detto *modem*, e collegare il modem alla linea telefonica. Poi, userà un *programma di comunicazione* per connettersi al BBS. Potrà così sfruttare i servizi offerti dal BBS, che variano ovviamente di caso in caso.

Alcuni BBS consentono leggere i messaggi lasciati dagli altri utenti che hanno utilizzato il servizio in precedenza, di rispondere a questi messaggi o lasciare il proprio, o anche di riprodurre i file memorizzati sul disco del BBS. Altri forniscono servizi di tipo diverso; ad esempio consentono di giocare in linea con un altro utente di BBS a scacchi o ad altri giochi; conversare con un altro utente, cioè scrivere quello che si vuole dire alla tastiera e leggere sullo schermo la risposta quasi istantanea dell'altro utente. Nel caso dei servizi telematici specializzati si può cercare un dato in un archivio, (ad esempio, le Pagine Gialle Elettroniche, oppure il registro dei protesti cambiari), piazzare un ordine di vendita o di acquisto di beni o servizi, e nel caso dei BBS più evoluti persino vedere fotografie o carte metereologiche memorizzate sul computer remoto. I servizi telematici di tipo economico-finanziario e le *Reti Civiche* spesso evitano accuratamente di definirsi con il termine BBS, che considerano poco serio, ma il principio è lo stesso. Si tratta sempre di computer o gruppi di computer ai quali gli utenti possono collegarsi via telefono per comunicare con altri, trovare file, giocare, fare ricerche, e così via.

Ciò che Internet NON è

Internet non è un BBS e neppure un servizio telematico. Internet è una *rete di reti*. Una rete di computer è costituita da un gruppo di computer collegati che possono comunicare tra loro. Questi computer interconnessi possono mandarsi messaggi e condividere le informazioni memorizzate sui loro dischi.

Internet collega più di 20.000 di queste reti, e il loro numero è sempre in aumento. Su queste reti ci sono milioni di computer, di terminali e di utenti; secondo alcune stime si tratta di circa due milioni di computer e circa 30 milioni di utenti. Non c'è niente di strano o tecnicamente avanzatissimo nelle reti di computer; sono una tecnologia ormai molto diffusa. Si può possedere una rete a livello privato, averne due o tre come certe piccole società o addirittura possederne migliaia come alcune grandi organizzazioni. Ma Internet non è una rete: è una rete di reti. Molte reti diverse sono state unite per produrre il più numeroso gruppo di computer collegati tra loro

del mondo. Alcune di queste reti appartengono a governi, altre a Università, a imprese, a biblioteche e addirittura a scuole. La maggior parte sono negli Stati Uniti, ma parecchie si trovano in Europa (Italia compresa) e alcune si trovano oltreoceano, dall'Australia allo Zimbabwe.
Per saperne di più sulla storia di Internet, fate clic su:
http://www.isoc.org/internet-history **Errore. Il se**
Naturalmente Internet ha funzioni ben più importanti del fare comunicare gli utenti di Paesi diversi attraverso una specie di posta elettronica. Ciò che rende speciale Internet è che molti computer di ogni rete collegata agiscono come gestori di archivi. In altre parole, quando ci si collega ad Internet si ha l'opportunità di accedere a migliaia di sistemi differenti, che gestiscono archivi comunali, database universitari, cataloghi di biblioteche, messaggi di qualsiasi tipo e milioni di file, contenenti fotografie, documenti, video musicali e qualsiasi altra cosa possa essere messa in forma digitale.
Ad esempio, basta fare clic su:
http://lcweb.loc.gov
L'analogia migliore che si può utilizzare per descrivere Internet è probabilmente quella con la normale rete telefonica. Il sistema telefonico mondiale ha molti commutatori, le centrali, possedute da diverse organizzazioni, tutte collegate tra loro. Quando un utente di Milano tenta di chiamarne uno di New York, non ha bisogno di sapere quali città o quali Stati la chiamata attraverserà; è il sistema telefonico che fa tutto per lui. Le società telefoniche hanno stabilito tra loro i meccanismi di questo processo e non è necessario che l'utente sappia cosa succede. Internet funziona nello stesso modo; così come non c'è un'unica società telefonica al mondo, non c'è una singola società Internet. Nessuno possiede Internet, così come nessuno possiede la rete telefonica mondiale. Certo, ogni singola componente è posseduta da qualcuno, ma la grande Rete non è posseduta da nessuno: è un sistema che continua a funzionare grazie all'interesse comune di tutti coloro che ne fanno parte. Nel caso del telefono, le compagnie telefoniche di tutto il mondo si incontrano periodicamente e concordano qual'è il sistema di funzionamento migliore nonché i costi e i dettagli tecnici per collegare un paese all'altro. Lo stesso accade per Internet.

Perché usare Internet

Ci sono centinaia di milioni di buone ragioni per usare Internet: tutte le persone che hanno già scelto di usare la Rete e con le quali ci si potrà collegare. Naturalmente non sono tutti utenti attivi; la maggior parte della gente potenzialmente collegata raramente usa altri computer oltre a quelli della propria organizzazione e non ha idea di quello che c'è su Internet. Le potenzialità tecniche e commerciali di questa platea sono comunque enormi. Inoltre, questo numero è puramente indicativo; in realtà nessuno sa con sicurezza quante persone possano essere realmente raggiunte dalla rete. Ma c'è una ragione anche migliore per usare Internet: l'informazione che contiene. Opere letterarie, report finanziari, ricette di cucina macrobiotica, presentazioni multimediali, per arrivare fino ai più recenti programmi per la realtà virtuale: su Internet c'è tutto quello che si può desiderare, ed è in gran parte gratis! Disporre dell'informazione adeguata in tempo utile può essere addirittura vitale per un'azienda, ma ha un valore notevole anche per l'individuo. Basta conoscere l'inglese (e avere seguito questo corso!) per poter scrivere annunci, entrare in contatto con persone che condividono i propri interessi,

consultare i libri e le riviste contenuti nelle biblioteche che usano Internet, trovare i programmi adeguati alle proprie esigenze e così via.

Come collegarsi ?

Se siete già oppure pensate di entrare in Rete vuol dire che qualcuno ha provveduto o provvederà a creare tra la vostra macchina e il resto di Internet qualche tipo di collegamento. Le aziende che forniscono questo servizio si chiamano *Internet service provider*, nel seguito semplicemente *provider* o *ISP*. Un elenco completo degli Internet provider italiani si trova... su Internet, all’indirizzo
http://inews.tecnet.it.
E’ un po’ come mettere l’elenco dei concessionari auto in un posto raggiungibile solo in macchina, ma per consultarlo si può sempre ricorrere a un amico già collegato. Ci sono due modi principali per collegarsi a Internet, ciascuno dei quali ha una serie di varianti:
* Collegamenti permanenti
* Collegamenti in linea commutata
Non sempre i tecnici e gli esperti di reti usano questi termini in modo coerente; le considerazioni che seguono hanno il compito di chiarire il loro significato ed evitare confusioni.

Collegamenti permanenti

Un collegamento permanente significa che il proprio computer (o, più spesso, la rete locale di cui fa parte) è collegato direttamente a una rete di tipo*TCP/IP* che è parte di Internet. Solitamente sono i computer di proprietà delle grandi organizzazioni (aziende, Enti, Università) ad essere collegati a Internet in questo modo; i privati, invece, si limitano a collegare temporaneamente via telefono il proprio personal a un computer che a sua volta è collegato permanentemente alla rete.
La sigla *TCP/IP* significa Transmission Control Protocol/Internet Protocol, ovvero *Protocollo di controllo di trasmissione/protocollo di Internet*.
Un protocollo è l'insieme di regole che definisce come un computer deve parlare all'altro; i prossimi paragrafi tratteranno meglio questo argomento. Per ora basti sapere che qualunque sia il tipo di rete utilizzato all'interno della propria organizzazione, per collegarsi permanentemente a Internet bisogna che il proprio computer parli, almeno verso l'esterno, il linguaggio TCP/IP.

Come creare un collegamento permanente

Se si è il responsabile di un'azienda che dispone di una propria rete e si desidera connetterla permanentemente a Internet, la prima cosa da fare è prendere contatto con un provider che sarà anche il punto fisico di connessione alla rete globale.
Fate clic qui per dare un’occhiata ad uno dei più veloci ed affidabili provider in Italia:
http://www.energy.it **Errore. Il segnalibro non è definito.**
Poi, bisogna collegare sulla propria rete un apparecchio detto *router* che farà da interprete, parlando in TCP/IP con il computer del provider e nel protocollo opportuno (lo stesso TCP/IP, o comunque quello usato dalla rete aziendale) con le macchine dell'azienda. Infine, occorre affittare da Telecom o da altri una linea telefonica dedicata che colleghi il router installato in azienda al computer del provider. I dettagli variano a seconda dei casi: il provider può procurare lui stesso il router oppure suggerire all'organizzazione cliente quale comprare. Poiché la linea telefonica è dedicata, è sempre attiva. Non c'è bisogno di telefonare ogni volta per raggiungere il

computer del provider; ogni utente dell'organizzazione cliente potrà collegarsi a Internet direttamente dal proprio solito posto di lavoro. Una volta collegato, potrà trasferire i file dal proprio computer agli altri computer di tutto il mono che sono collegati a Internet e viceversa. Naturalmente, questo tipo di servizio è molto caro: al costo del servizio di connessione e del router, infatti, va aggiunto quello non indifferente del noleggio della linea Telecom.
Se siete già collegati alla Rete attraverso una rete aziendale o scolastica, il vostro personal computer farà già probabilmente parte di una *rete locale* (*local area network, LAN*). Invece di un modem, il vostro personal avrà una scheda di rete che lo collega alla rete locale; il cavo che esce dalla scheda può assomigliare al cavo telefonico o al cavo coassiale usato per l'antenna della televisione. Se usate Internet dalla rete dell'ufficio o della scuola, è sicuro che anche la vostra rete è connessa a un provider; è però probabile che il collegamento non avvenga tramite un modem, ma attraverso una linea permanente a velocità maggiore (tipicamente, da 64 kbps a 2 Mbps) affittata annualmente dalla società telefonica.

Collegamenti su linea commutata

Gli utenti che si collegano a Internet da casa arrivano a un provider usando un modem che funziona sui fili telefonici normali. Presso la sede del provider vi è un altro modem convenzionale; può essere un modello industriale, oppure, se il provider vuol risparmiare, può essere come il vostro. La velocità di connessione è quella del più lento tra i due modem (nel migliore dei casi e non considerando fattori come il rumore) e va dai 2,4 kilobit al secondo (kbps) a 33,6 kbps. Un collegamento di questo tipo, detto *in linea commutata* è un collegamento che usa il protocollo *SLIP* (*Serial Line Internet Protocol*, protocollo Internet su linea seriale), *CSLIP* (*Compressed SLIP*, SLIP compresso), o *PPP* (*Point-to-Point Protocol*, protocollo punto-punto) per collegarsi a un Internet provider attraverso una normale linea telefonica e un modem.
´Connessione PPP’ significa che, telefonando al computer del provider, il vostro computer riceverà un suo numero identificativo (l’indirizzo IP) che lo renderà visibile su Internet per tutta la durata della chiamata. Tramite appositi programmi, vi sarà così possibile consultare informazioni sugli altri computer di Internet. ‘Linea commutata’ vuol dire che il collegamento usa la normale rete telefonica, il che in genere limita la velocità massima (che dipende dal tipo di modem) a 28,8 kbps. Anche questi sono collegamenti di tipo TCP/IP, come il collegamento permanente, ma sono destinati ad essere usati sulle usuali linee telefoniche, anziché su una linea dedicata. Naturalmente c'è una certa differenza in fatto di prestazioni: i trasferimenti di dati tra il computer dell'utente e gli altri saranno sensibilmente più lenti. l modem è in grado di chiamare il numero telefonico che viene comunicato dal provider. La telefonata avviene sulle linee telefoniche Telecom, le stesse che portano la voce; le tariffe applicate sono quindi quelle usuali per le conversazioni telefoniche. In alternativa, ci sono i collegamenti *ISDN* (*Integrated System Digital Network*). Quasi tutti i provider italiani offrono oltre alla ´connessione PPP in linea commutata’ anche la ´connessione PPP via ISDN’. Per usarla dovete prima chiedere a Telecom di trasformare in ISDN la vostra presa telefonica, per un costo di circa cento euro. Poi le tariffe sono di poco superiori a quelle delle chiamate tradizionali anche se la velocità.. è tutta un’altra musica. Per ora, la velocità di connessione ISDN è di 64 kbps, ma può arrivare a 128 kbps, cioè circa cinque volte più veloce di un modem tradizionale. Infine, va osservato che le connessioni PPP si possono fare anche affittando da Telecom delle linee dedicate, riservate cioè alla

connessione tra utente e provider. Ma anche le più economiche hanno costi di molti milioni all’anno e sono rivolte all’utenza aziendale.
Comunque, indipendentemente dal tipo di linea, una volta composto il numero ed effettuato il collegamento con il computer del fornitore non ci sarà più alcuna differenza tra il collegamento PPP e un collegamento dedicato; si potranno, cioè, trasferire dei dati dal proprio computer a qualsiasi altro connesso alla rete Internet e viceversa.

Vantaggi e svantaggi dei vari tipi di collegamento

Il collegamento permanente è caro, ma può essere considerato un buon investimento per le grosse società. Una volta collegati, infatti, si possono avere tanti utenti quanti sono i computer della società e la linea dedicata conviene, poiché non c'è alcun addebito legato al numero degli utenti individuali. Se ci sono molti utenti, il collegamento permanente di tutta l’azienda può risultare più economico rispetto alla scelta di fornire ad alcuni utenti VIP un proprio collegamento autonomo, specialmente se la maggior parte degli utenti "normali" usa Internet con bassa frequenza o per solo pochi minuti al giorno. Comunque, la connessione permanente resta una soluzione per pochi: i servizi specializzati sono molto cari. Il collegamento in linea commutata via modem, sebbene inferiore a quello permanente, è molto più economico. E' necessario però disporre di un modem ad alta velocità (almeno 28.800 bit al secondo) e a correzione di errore. Comunque, chi si collega da casa può stare tranquillo: ci sono nuove tecnologie che consentono agli utenti domestici di accedere a Internet a velocità superiore. Una possibilità è il servizio ISDN (Integrated Services Digital Network), che usa il collegamento telefonico esistente, ma sostituisce i modem con speciali adattatori digitali.
ISDN è già oggi una buona scelta: è accettato da un numero crescente di provider, è sicuro e non troppo costoso, sebbene costi un po’ più del normale servizio telefonico. Un altro servizio che permette collegamenti superveloci a Internet lungo le normali linee telefoniche è *ADSL* (*Asymmetric Digital Subscriber Line, Linea digitale asimmetrica*). Questo servizio prevede connessioni a velocità fino a 6 Mbps "a favore di corrente" (da Internet al vostro computer) e 640 kbps "contro corrente" (viceversa). Sebbene sia ancora lontana dall'essere onnipresente, ISDN è disponibile in Italia fin dagli anni '80 ed è la più diffusa tra le tecnologie di connessione digitale per l’utenza privata. Le reti cellulari e i sistemi di invio via satellite sono in fase di sperimentazione per il trasporto Internet; quest’ultima secondo molti esperti sarà la tecnologia del futuro.
Ma la rassegna delle tecnologie non è finita qui: l'accesso a Internet attraverso la connessione televisiva via cavo è uno strumento nuovissimo per cui si prevede un grande futuro, anche se non in Italia, dove la rete televisiva via cavo è praticamente insesistente. I sistemi di accesso a Internet tramite la rete della TV via cavo prevedono velocità di 500 kbps a 30 Mbps. Lo stesso cavo che porta il segnale della televisione viene usato per Internet: il servizio verrà fornito dalle stazioni televisive via cavo, che dovranno diventare provider.
Ad un singolo individuo, comunque, non conviene mai acquistare un collegamento permanente. Esso richiede router, linee dedicate e altri dispositivi molto costosi. Anche le spese iniziali e le tariffe annuali sono molto elevate. Se si vuole un servizio diretto in linea, basta invece pagare una (ragionevole) tariffa iniziale di collegamento. Poi si paga in proporzione all’utilizzo, secondo una tariffa oraria che è diversa durante il giorno e da mezzanotte alle otto del mattino. C'è molta varietà nelle tariffe e nelle modalità di pagamento, quindi conviene contattare il maggior numero

possibile di provider. Questo corso descrive in dettaglio gli strumenti di consultazione: bisogna verificare che il provider li metta effettivamente a disposizione.

Le arterie di Internet

Le arterie di Internet sono le linee dorsali , che sono gestite dai fornitori di Internet "all'ingrosso", i cosiddetti *National Service Provider* o *NSP*. Gli utenti si collegano agli ISP locali attraverso i modem, gli adattatori ISDN o altri strumenti, mentre gli ISP locali si connettono a loro volta alle reti di NSP come Interbusiness di Telecom, Unisource, Sprint o (nel caso degli utenti universitari) il *Gruppo Autonomo Reti Ricerca* o *GARR*. Alcuni NSP forniscono anche il servizio agli utenti finali, sebbene il loro principale interesse sia comunque la fornitura all'ingrosso. Gli ISP si collegano agli NSP attraverso linee dedicate affittate dalla Telecom. Le connessioni migliori sono le linee dette *T1* a 2 Mbps; i provider più grossi possono avere connessioni T1 multiple. Al loro interno, gli NSP usano le loro reti private, composte anch'esse di linee dedicate affittate ricorrendo a Telecom Italia per il tratto in territorio italiano e ad altre società per i tratti internazionali (in Europa, British, France e Deutsche Telecom; in America WilTel, MCI e Sprint). Queste ultime società sono proprietarie dei cavi in rame e fibra ottica e delle connessioni via satellite attraverso cui corrono insieme le normali conversazioni telefoniche e il traffico Internet. Le reti degli NSP di solito funzionano a velocità T1 localmente, ma arrivano a velocità superiori per le lunghe distanze. Le connessioni T3 e con gli Stati Uniti corrono a 44,736 Mbps. In America le cose vanno ancora meglio: la rete su fibre ottiche (OC3) di MCI funziona a155,52 Mbps (5500 volte più veloce di un modem a 28,8 kbps). ed alcune linee superveloci viaggiano a più di 600 Mbps.

Le altre reti

Accanto a Internet, ma in modo indipendente, si sono diffuse varie reti telematiche di tipo amatoriale, particolarmente interessanti per chi segue le iniziative politiche, sociali e di volontariato.

La rete amatoriale per antonomasia è *FidoNet*, che conta più di 30.000 nodi sparsi in tutto il mondo. Vi sono anche altre reti realizzate con gli stessi criteri (generalmente chiamate *FTN*, *Fidonet Technology Network*), anche se limitate ad una sola nazione o ad un determinato scopo: ad esempio in Italia oltre a FidoNet ci sono *Peacelink* (dedicata al pacifismo e alle questioni sociali), RPGNet (dedicata ai giochi di ruolo), *Cybernet* (creata dal movimento "cyberpunk"), e così via. Le reti FTN sono generalmente strutturate ad albero, ossia ogni nodo ha un nodo "padre" (*uplink*) e dei nodi "figli" (*downlink*), fino ad arrivare ai nodi terminali. Ogni sistema che fa parte della rete effettua ad orari predefiniti delle chiamate notturne su linea telefonica commutata (in modo da diminuire al massimo i costi di trasferimento) al proprio uplink, e subito dopo si predispone per ricevere le chiamate dei propri downlink, che giungeranno anch'esse ad orari predefiniti. E’ ovvio che la spesa principale che ogni singolo responsabile di nodo (universalmente chiamato *sysop*) dovrà affrontare è quella della propria bolletta telefonica. In altri termini, se voi spedite un messaggio da una nodo Fidonet di Milano ad un utente che si trova a Roma, il costo del trasferimento del vostro messaggio verrà suddiviso tra tutti i nodi che si trovano nel "percorso" tra i due sistemi. Il lettore esperto avrà già notato che ci sono alcune affinità superficiali tra Internet e le reti in tecnologia FTN, ma la differenza è sostanziale: in

FidoNet le comunicazioni avvengono tutte su linea commutata ed in modalità tipicamente "batch", ossia senza una connessione diretta e continua tra i sistemi; inoltre i sistemi FidoNet non sono gestiti da provider ma da appassionati senza fini di lucro. Un'altra differenza con Internet si può notare nell'instradamento dei messaggi: in FidoNet la topologia della rete è ad albero, come già detto, e per di più è fissa in quanto non sono previsti meccanismi automatici di instradamento alternativo in caso di indisponibilità di un nodo intermedio.
FidoNet e Internet.
Questi limiti tecnologici (soprattutto se si fa un confronto con Internet) sono largamente compensati da semplicità ed economicotà di gestione; questo ha reso FidoNet la tecnologia di elezione per la creazione di reti in molti paesi del Terzo Mondo; grazie ai punti di interconnessione tra FidoNet e Internet molte Università e istituzioni africane riescono a scambiare messaggi di posta elettronica con il mondo Internet anche se i Paesi in cui si trovano non sono connessi alla Rete.

Primi rudimenti di TCP/IP

In definitiva, attraverso un provider e il NSP a cui si rivolge, il nostro computer di casa o dell'ufficio può essere collegato a milioni di altri computer. Ma come fanno i messaggi a trovare la strada da un computer a un altro?.Per creare un collegamento a Internet è necessario disporre sul proprio computer del software di rete TCP/IP. Con le passate versioni di Windows e Mac, si trattava di un programma da installare; in Windows 95 e Unix è incorporato. TCP/IP si basa su uno schema detto a *commutazione di pacchetti*. Questo significa che ogni file inviato su Internet, dai messaggi di posta elettronica al contenuto delle pagine Web, è suddiviso in parti più piccole chiamate *pacchetti*, seguendo le regole (cioè, il protocollo) IP. Ogni pacchetto è etichettato, includendo anche l'indirizzo numerico di destinazione, detto indirizzo IP; l'indirizzo IP è formato dall'*host number* (che individua la macchina a cui va mandato il pacchetto) e un numero aggiuntivo, il *port number*, che identifica il programma ricevente tra quelli in esecuzione su quella macchina.

Gli host e port number IP

Non c'è alternativa : se si effettua un collegamento a Internet (che sia in linea commutata o meno), si ha bisogno di un *host number*. Questo numero è come il numero della carta d'identità, e serve per identificare univocamente ogni computer connesso alla rete Internet. Infatti, anche con il collegamento in linea commutata, il computer dell'utente sarà registrato come un computer Internet a tutti gli effetti . Quindi dovrà avere un numero identificativo completo e autonomo, anche se tutte le comunicazioni con la rete passeranno dalla macchina dell'utente a quella del fornitore di servizi e viceversa. Per esempio, un host number italiano valido è `145.94.50.236`. I singoli programmi in esecuzione su questa macchina saranno identificati da un altro numero, il *port number*, in modo che il computer sappia distinguere i pacchetti diretti a ciascuno di essi. I programmi che forniscono i servizi più noti hanno numeri di port standard, i cosiddetti *well-known port.* Su questo argomento torneremo brevemente nel prossimo tutorial. Se il collegamento è permanente, l’host number è attribuito una volta per tutte; altrimenti, il software TCP/IP otterrà questo numero dal computer del provider ogni volta che l'utente si collega con Internet, ricordandogli la propria esistenza. I computer speciali che collegano le varie

parti della Rete e instradano i pacchetti nelle diverse zone sono chiamati *router* . Le connessioni tra i router possono essere pubbliche (ovvero, le linee telefoniche) o private, cioè collegamenti via cavo tra i computer di una stessa società o di un campus universitario. Del loro funzionamento ci occuperemo tra poco.

I nomi di computer e di dominio

Anche se per identificarsi tra loro usano i numeri, i computer connessi a Internet vengono individuati dagli utenti umani tramite *nomi*. Il nome di un singolo computer viene spesso chiamato il suo *nome di computer*. Chi ottiene un collegamento Internet gratuito non ha molta scelta sul nome di computer: se l'utente ha accesso tramite una società o un'Università, il nome del suo computer gli verrà semplicemente comunicato dal responsabile. Ma chi acquista l'accesso da un fornitore di servizi, e sceglie il collegamento permanente, ha la facoltà di scegliere il nome del proprio computer. In certi casi, invece di usare il nome di computer del computer del provider, potrà deciderne uno proprio. Il nome di computer non è equivalente all'host number, poiché non basta a identificare univocamente una macchina; per questo il sistema assegna a ogni calcolatore anche un *dominio* .

Il *nome di dominio* serve per identificare un intero gruppo di computer collegati a Internet. I domini più grandi sono quelli corrispondenti a intere nazioni; il dominio italiano, ad esempio, si chiama `it`; quello tedesco de, mentre francesi e inglesi hanno rispettivamente i domini `fr` e `uk`. Ovviamente, un dominio grande può comprenderne altri più piccoli. L'insieme dei computer dell'Università di Milano, per esempio, sta nel sottodominio `unimi` di `it` . Ma questo sottodominio ha a sua volta dei sottodomini; uno di essi è chiamato `crema` perché comprende le macchine del Polo Didattico di Crema. Il nome completo di quest'ultimo sottodominio, in realtà, è `crema.unimi.it` perché occorre elencare, oltre a quello corrente, anche tutti i livelli superiori. Il computer usato per inviare questo corso ha nome `weblab` e dunque il suo nome completo è `weblab.crema.unimi.it`. E’ a questo nome completo che corrisponde univocamente un host number, cioè `159.149.70.70`. Ogni volta che si digita un nome, viene interpellato un apposito programma, il *Domain Name System* , per eseguirne la traduzione in numero.

Per quanto riguarda i nomi di dominio, gli Stati Uniti costituiscono un caso a parte, perché non esiste un dominio `.us`. La rete Internet negli Stati Uniti è tanto grande e complessa che sono stati definiti vari domini separati. Il più noto è `edu`: il gruppo nazionale delle istituzioni educative. La maggior parte delle Università e delle scuole americane fanno parte di questo dominio. Il dominio `com` comprende tutte le organizzazioni commerciali, mentre `mil` è quello delle installazioni militari (da solo, contiene centinaia di migliaia di computer). Detto questo, bisogna osservare che l'utente finale non deve quasi mai disporre di un proprio dominio. Questo è consigliabile solo alle grandi organizzazioni che vogliono connettere a Internet tutte le macchine della propria rete di calcolatori privata. Comunque, quando il fornitore assegna all'utente un nome di dominio, deve registrarlo presso l'*InterNIC* (*Inter Network Information Center, Centro di Informazione Reti*), o meglio presso l'Ente delegato a livello nazionale (in Italia, il *GARR*); un'operazione che richiede almeno una decina di giorni. Ciò assicura l'unicità del nome e la sua associazione ai numeri che identificano i computer usati dal richiedente.

I router: il cuore della Rete

I router sono macchine collegate a due o più reti, che hanno il compito di far passare i pacchetti da una rete all'altra in modo da avvicinarli alla loro destinazione (i pacchetti diretti a una macchina collegata alla stessa rete del mittente arrivano a destinazione senza bisogno dei router). Un tipico router può smistare 10.000 pacchetti al secondo; un router di alta qualità può avere una capacità teorica di 200.000 pacchetti al secondo. I router di Internet hanno un compito semplice: inoltrare i pacchetti che ricevono alla loro destinazione, passandoli da router a router. Per inoltrare i pacchetti fino all'ultima fermata si crea una catena di router, ognuno dei quali sa l'indirizzo del successivo sulla Rete grazie a tabelle costantemente aggiornate.
Ogni riga di queste tabelle è una coppia `(porzione di) indirizzo del destinatario - indirizzo del router di inoltro`.
Ad esempio la coppia
`159.- 145.94.50.236`
dice al router che la detiene che tutti i pacchetti il cui indirizzo comincia per `159` vanno mandati al collega router `145.94.50.236`. Si noti che il router di inoltro deve trovarsi sulla stessa rete del mittente o su una delle reti a cui è collegato il router precedente, in modo che questi sappia raggiungerlo sulla base del suo indirizzo. Il router di inoltro poi provvederà a mandare i pacchetti al destinatario o a un altro router.... e così via. L’instradamento viene fatto pacchetto per pacchetto. Ricordiamo che quando si invia un messaggio di posta elettronica o una richiesta di connessione a un server Web, il messaggio viene diviso in pacchetti e trasmesso. In alcuni casi, i diversi pacchetti possono seguire percorsi differenti , perchè il computer d'inoltro nella tabella di un router può variare, ad esempio a seguito di cambiamenti del traffico; il messaggio viene ricomosto sulla macchina di destinazione.
Sembra difficile? La normale rete telefonica funziona circa nello stesso modo, tenendo conto che le centrali telefoniche usano i numeri telefonici invece degli indirizzi di rete. Ad esempio, `0039` è il prefisso telefonico per l'Italia , `02` è quello per Milano e `824` individua la zona metropolitana di sud-ovest. C'è però una differenza: quando si digita un numero telefonico in quella zona, la rete telefonica crea all'inizio della chiamata un circuito tra gli apparecchi dei due interlocutori e la conversazione fluisce tra i due punti. Al contrario, la rete IP può inviare indipendentemente ciascun pacchetto. Quindi, almeno in linea di principio, un dato pacchetto potrebbe non seguire lo stesso percorso su Internet di quello prima o di quello dopo. In questo modo, se una linea cade o un collegamento è troppo occupato, i router cooperano per trovare un percorso alternativo per inviare i dati. Questo processo è trasparente agli utenti, a parte l'attesa che impone loro. Ogni router su Internet deve essere quindi in grado di aiutare a instradare i pacchetti a uno dei milioni di computer dotati di indirizzi IP. Anche se il router è un vero e proprio computer (e molto sofisticato!), non è possibile che le sue tabelle contengano le informazioni d'inoltro per i milioni di computer presenti su Internet. Tutto quello che fanno i router è capire se un pacchetto IP deve essere inviato all'interno di una delle reti a cui sono essi stessi collegati. Se così non è, indirizzano il pacchetto a un altro router che forse ne sa di più. Per eseguire quest'operazione, i router si scambiano regolarmente i dati via Internet; del loro funzionamento ci occuperemo in dettaglio nel prossimo tutorial.

Se qualcosa va storto

In definitiva, lo scopo dei router è capire il miglior percorso per inviare i pacchetti da un computer all’altro. Quando cambiano le condizioni, a causa di un malfunzionamento o di una

congestione delle linee, i router della Rete modificano quasi istantaneamente le loro tabelle di inoltro. Questa capacità autonoma di autoapprendimento fa sì che sulla Rete in cui l'instradamento effettivo di singolo pacchetto non sia noto a priori.
Questa adattabilità è una potenziale fonte di problemi. I router, infatti, si scambiano via Internet gli aggiornamenti per le loro tabelle, e ciò può dar luogo a un fenomeno chiamato *flapping* in cui una rete compare e scompare continuamente facendo sì che tutti i router del circondario si diano da fare per comunicarsi l'un l'altro che è scomparsa o riapparsa. Internet si difende da questo fenomeno attraverso un processo di "punizione" delle reti instabili diffondendo gli aggiornamenti alle tabelle d'instradamento tanto più lentamente quanto più la rete è instabile. I grandi provider possono anche rifiutarsi di collegarsi alle reti che causano dei problemi.

I server

Tutte le informazioni su Internet hanno origine all'interno dei server, che sono dei computer come gli altri, collegati in permanenza alla rete; l’unica differenza è che su di essi girano particolari programmi (detti anch’essi *server*) che hanno il ruolo di fornire i dati a chi li richiede. Qualsiasi computer direttamente connesso a Internet può fornire informazioni, assumendo così i panni del server. Storicamente, i server erano principalmente macchine con il sistema operativo Unix (di marca DEC, Hewlett-Packard, IBM, Silicon Graphics, Sun e altri), ma ora sono sempre più usati i sistemi Macintosh e Windows NT. Esistono anche molti server Web che sfruttano il sistema operativo gratuito Linux, disponibile anche su personal computer economici. I siti Internet più grandi girano su macchine Unix per sfruttare il loro sistema operativo evoluto e i processori più potenti. Ad esempio il server Alta Vista, di cui parleremo nei prossimi capitoli, gestisce giornalmente milioni di collegamenti e gira su un computer parallelo con quattro CPU e molti gigabyte di RAM. I server eseguono del software specializzato per ogni tipo di applicazione Internet, tra quelle che vedremo in seguito, il World Wide Web, Gopher, i gruppi di discussione delle news di Usenet e la posta elettronica. Ogni organizzazione presente su Internet deve poi ospitare il *DNS* (*Domain Name System*) per la traduzione dei nomi simbolici in indirizzi di rete di cui abbiamo parlato in precedenza. Il DNS consente agli utenti di specificare i nomi, invece di usare gli indirizzi IP numerici.

La rete internet

1 La rete Internet

Il Dipartimento della Difesa degli Stati Uniti, alla fine degli anni ’60, pose l’attenzione sul problema dell’affidabilità delle telecomunicazioni. Il presupposto era quello di avere un sistema di comunicazione e di controllo militare in grado di rimanere operativo anche in caso di distruzione, quasi totale, delle linee telefoniche. Inoltre tale sistema doveva avere anche il compito di permettere a coloro che svolgevano ricerche militari, per conto del Ministero della Difesa, di scambiarsi dati ed informazioni. Nel 1964 un ricercatore della Rand Corp., di nome Paul Baran, propose una rete di comunicazione tra computer che non aveva alcun nodo centrale, nessuna autorità governante ed in grado di funzionare anche con collegamenti totalmente inaffidabili. Nello schema di Baran ogni messaggio era suddiviso in sotto unità, dette pacchetti, capaci di viaggiare senza un tragitto rigidamente prestabilito tra una rete di computer collegati. Il primo collegamento fu realizzato nel 1969 con il protocollo NCP (Network Control Protocol), tra quattro nodi dell’ARPA (Advanced Research Projects Agency) e nacque come ARPAnet, il predecessore di Internet. Nel 1973 fu realizzata la prima connessione internazionale sulla rete ARPAnet con l’Inghilterra e la Norvegia. Solo nel 1982 l’ISO/OSI (Intenational Standard Organization/Open System Interconnection) riuscì ad uniformare tutte le topologie di rete, che nel frattempo si erano create come la rete delle organizzazioni militari denominata MILNET, la rete degli utenti UUCP, la rete delle ricerche scientifiche CSNET e così via, facendo nascere Internet che utilizza un unico protocollo comunemente chiamato TCP/IP (Trasmission Control Protocol/Internet Protocol). Il TCP ha il compito di scomporre alla sorgente i messaggi in una serie di pacchetti e di ricomporli al destinatario, mentre il compito, dell’IP é di gestire l’indirizzamento di ogni pacchetto ad ogni nodo attraversato, in modo che sia selezionato man mano il tratto di linea più conveniente. Ogni computer collegato ad Internet è contraddistinto da un numero, il così detto **indirizzo IP,** composto solitamente da una sequenza di quattro numeri, separati da punti, che individua in modo univoco nella rete quel determinato computer. Ogni computer collegato ad Internet deve avere un indirizzo IP, altrimenti non può comunicare con gli altri. Gli indirizzi definibili sono tuttavia limitati: si cerca quindi di eliminare lo spreco costituito da tutti i computer non collegati in un dato istante. I provider infatti usano un sistema dinamico: ogni volta che un PC si connette ad Internet tramite un modem, gli viene assegnato un indirizzo provvisorio, al termine della comunicazione, lo stesso indirizzo sarà usato da un altro PC che nel frattempo si è collegato. Gli utenti Internet non utilizzano direttamente gli indirizzi numerici IP, ma dei comodi indirizzi sostitutivi di tipo descrittivo. Tali indirizzi descrittivi sono convertiti nei corrispondenti indirizzi IP dal DNS (Domain Name Server), un computer predisposto dal provider che funge da “guida telefonica”.

Internet (INTERconnected NETworks), può logicamente essere vista come una rete di interconnessioni composta da migliaia di nodi e centinaia di migliaia di utenti sparsi in tutto il mondo. Essa permette l’accesso ad una enorme quantità di risorse, servizi ed informazioni. Proprio per questo motivo, unito al fatto che consente la comunicazione fra più tipi di reti, gli è stato attribuito l’appellativo di: *“Rete di tutte le Reti”*.

2 La ricerca di informazioni in Internet.

I servizi di base in Internet sono il trasferimento di file, il login remoto, la posta elettronica ed i servizi di News. I primi due sono basati su applicazioni Client-Server quali FTP (File Trasport Protocol) e telnet consentono di ottenere dei file da un computer remoto e di connettersi ad una macchina remota usando una interfaccia tipo terminale a caratteri. La posta elettronica è forse il servizio più famoso in Internet con il quale due utenti possono scambiarsi messaggi. I servizi di News sono l’equivalente in Internet dei gruppi di discussione e sono organizzati per argomenti.

Risulta difficile dire quanti computer sono connessi alla rete Internet, fino al giugno del 1995 tale numero era stimato in 6,6 milioni con un tasso di crescita del 10-15% al mese o circa del 100% annuo.

Con un così grande numero di risorse a disposizione, cresciute senza alcun organismo di governo o organizzativo, è facile capire che, per l’utente, uno dei problemi principali è il recupero dell’informazione [CCRZ96].

Per soddisfare questa esigenza nascono alcuni servizi quali:

- **Archie,** consiste in un’ampia banca dati che archivia il contenuto dei server FTP di pubblico dominio, associando ad ogni nome di file una serie di informazioni come: locazione, dimensione, data di memorizzazione e cosi via.
- **WAIS,** per facilitare il reperimento di documenti attraverso la ricerca di parole chiave nei documenti stessi.
- **Gopher**, creato nel 1991 da un gruppo di programmatori diretti da Mark P. McCahil nell’Università del Minesota. Il loro obiettivo era quello di sviluppare una interfaccia semplice da usare, che permettesse l’accesso alle risorse della rete, nascondendo all’utente la locazione fisica delle informazioni.

Internet guadagnò fama mondiale quando ai servizi già citati si aggiunse il World Wide Web.

Il World Wide Web, conosciuta anche come WWW, W3 o semplicemente Web, è un’architettura Client-Server atta alla diffusione di documenti ipermediali in maniera distribuita ed omogenea sulla rete Internet. Il progetto WWW nasce nel 1992 ad opera di Tim Berners Lee, nei laboratori informatici del CERN di Ginevra ed aveva l’obiettivo di sviluppare un sistema di pubblicazione e reperimento dell’informazione per documenti multimediali, rendendo, globalmente disponibili, in maniera semplice le informazioni distribuite attraverso l’integrazione dei tool esistenti. Un notevole impulso allo sviluppo di WWW venne poco più tardi, 1993, dal National Center for Supercomputing Applications (NCSA) dell’Università dell’Illinois con la nascita di Mosaic, un’interfaccia grafica multipiattaforma per l’accesso ai documenti presenti su WWW, sviluppato da Marc Andressen ed Eric Bina.

World Wide Web, come del resto le applicazioni Internet, come detto prima funziona attraverso una interazione tra un Client ed un server, attraverso il protocollo HyperText Trasfer Protocol (HTTP).

Per essere inserito nel World Wide Web, un documento deve essere memorizzato in un particolare formato, denominato HyperText Markup Language (HTML). HTML è un linguaggio di marcatura, nato per la descrizione di documenti testuali, che si basa sulla sintassi dello Standard Generalized Markup Language (SGML). I linguaggi di marcatura sono costituiti da un insieme di istruzioni, dette Tag (marcatori), che servono a descrivere la struttura, la

composizione e l’impaginazione del documento. I marcatori. I marcatori sono sequenze di normali caratteri ASCII, e sono introdotti, secondo una determinata sintassi, all’interno del documento, accanto alla porzione di testo cui si riferiscono.

L’unico modo a disposizione per avere un accesso immediato alle informazioni è conoscere l’indirizzo del documento, il quale prende il nome di URL (Uniform Resource Locator), una stringa non contenente spazi costituita da tre parti: la prima descrive il protocollo utilizzato per accedere al file (HTTP, FTP, …), la seconda contiene il nome della macchina connessa ad Internet, in cui è memorizzato il documento, (con eventualmente associato il numero della porta se non è utilizzata quella di default), l’ultima parte é il path del documento da recuperare. L’URL offre una forma ragionevolmente intelligibile per identificare o indirizzare in maniera univoca le informazioni su Internet. Gli URL sono ubiquitari, tutti i browser li utilizzano per identificare le informazioni su Internet, ma per chi non è a conoscenza degli URL di un documento può essere impossibile trovare le informazioni cercate. A questo scopo nascono due strumenti fondamentali per la “navigazione” in Internet: le *directory* ed i *motori di ricerca.*

Se l’utente ha una idea chiara delle informazioni che cerca e quindi può esprimerla in una parola, o una combinazione di parole chiave, allora gli strumenti più idonei per assolvere questo compito sono sicuramente i motori di ricerca.

Se non si ha in mente un obiettivo specifico, conviene allora ricorrere a una directory che suddividono le informazioni in categorie e contengono collegamenti ad interi siti anziché a singole pagine sparse. I siti sono catalogati secondo tematiche generali, ognuna suddivisa in modo gerarchico in sotto categorie più specifiche, a loro volta suddivisibili. Diversamente dai motori di ricerca, nelle directory dunque è l’utente stesso, anziché la macchina, ad eseguire la ricerca. Molte directory comunque, contengono recensioni dei siti per aiutare il compito e consentono anche la ricerca per parola chiave. Un inconveniente nell’utilizzo delle directory, sta nel fatto che esse contengono solo i siti segnalati dai rispettivi gestori dei siti (Web Master), per cui può accadere che ciò che si cerca sia presente sul Web senza essere segnalato in una directory. In questo caso conviene provare con un motore di ricerca.

I due diversi mezzi per il recupero dell’informazione hanno due criteri-guida diversi. I motori di ricerca hanno come criterio-guida la completezza. Per le directory invece il criterio di scelta è la facilità di reperimento dell’informazione desiderata. Attualmente, non è possibile combinare le due caratteristiche, perché i motori di ricerca sono dei robot telematici che cercano a testa bassa e non sono in grado di capire che cosa veramente interessa all’utente, limitandosi a rintracciare, la parola chiave in qualsiasi documento, anche il meno rilevante. In altre parole la differenza sostanziale tra i motori di ricerca e le directory, è che i primi effettuano una ricerca quantitativa, mentre i secondi optano per una ricerca qualitativa. D’altro canto, le directory sono gestite da risorse umane che non sono in grado di catalogare tutte le centinaia di milioni di pagine esistenti su Internet.

2.1 Le Directory

Gli utenti che utilizzano le directory sono a conoscenza, che esse catalogano solo una porzione del Web, l'utente richiede loro precisione, chiarezza, ordine, semplicità, unite ovviamente a una sufficiente copertura dei principali siti.

Di seguito si darà una descrizione delle principali directory di Internet e i criteri di scelta.

- **Yahoo!**[1]. (Yet Another Hierarchical Officious Oracle) progettata nel 1994 da David Filo e Jerry Yang, studenti di ingegneria elettronica dell'Università di Stanford. Può essere considerata la leader delle directory per semplicità d'uso e numero di siti catalogati. La sua semplicità d'uso dipende da una intelligente divisione delle categorie e da una interfaccia semplice ed efficace. Il grande numero di siti catalogato dipende dal fatto che la politica di Yahoo! è quella di elencare qualsiasi sito segnalato dai suoi creatori, senza operare alcuna scelta qualitativa. Per rintracciare un sito su Yahoo! è anche possibile usare una ricerca per parola chiave, se la ricerca non ha buon fine il testimone è passato ad Altavista (uno dei motori di ricerca in Internet) e la ricerca prosegue su tutta la rete.
- **Lycos.** Lycos Top5%[2] è un servizio, strutturato come Directory ma utilizzabile anche come motore di ricerca, che si limita a indicizzare quello che si sostiene essere "Il miglior cinque percento della Rete". Per ognuno dei circa 100 mila siti catalogati e disponibile una recensione che ne descrive forma e contenuti. Inoltre fornisce un punteggio, con relativa classifica, calcolato tenendo conto della qualità, completezza e chiarezza dell'informazione, oltre che dall'aspetto grafico del sito.
  Altri Directory con prestazioni inferiori sono *Magellan*[3] e la Directory contenuta in *Infoseek.*

2.2 I motori di ricerca

Come esposto in precedenza, esistono vari motori di ricerca, ognuno con particolari qualità che li rendono diversi uno dall'altro. Si tenterà di elencarne alcuni spiegando ciò che li differenzia.

- **Altavista**. Il motore di ricerca Altavista[4], nato nel 1995, è il risultato di un progetto di ricerca della Digital ed è stato recentemente potenziato per tenere a bada gli agguerriti concorrenti. Altavista è considerato il leader del settore, anche se non risulta essere il più usato nel mondo [AV98]. Se si può trovare un difetto in Altavista, questo, è paradossalmente la completezza che spesso genera come risultato di una ricerca un numero di pagine eccessivo e difficile da gestire. Adesso il problema è risolvibile grazie all'uso del comando REFINE, che fornisce un elenco di termini correlati alla parola chiave in modo da aiutare l'utente a selezionare solo le pagine realmente interessanti. Inoltre Altavista è l'unico fra i motori principali che consente di eseguire ricerche a desinenza multipla. Per esempio, digitando "architett*" Altavista cercherà, "architetto", "architettura", "architettonico" e cosi via. Infine Altavista mette a disposizione la possibilità di inserire più parole chiave organizzate secondo le regole dettate dalla logica

[1] Yahoo! URL: http://www.yahoo.com/
[2] Lycos URL http://www.lycos.com/
[3] Magellan URL http://www.masgellan.com/
[4] Altavista URL: http://www.altavista.com/

Booleana[5] (and, or, not). Esiste anche un sito europeo[6] che consente di svincolarsi dal traffico nordamericano e contiene la versione in italiano.

- **Excite.** La differenza sostanziale tra Altavista ed Excite[7] sta nel fatto che il primo è un potente mezzo per il completo recupero delle informazioni, mentre il secondo risulta più facile da usare. Le pagine indicizzate sono circa la metà rispetto ad Altavista ma in compenso Excite, sviluppato da studenti di informatica americani [AV98], punta alla semplicità d’uso. La specificazione dei criteri di ricerca viene opzionalmente guidata da una finestra di dialogo (power search) in cui Excite chiede se i risultati devono contenere tutte le parole chiave specificate o solo una, se vi sono parole che non devono apparire e altre combinazioni del genere. Questo è possibile anche con Altavista, che però obbliga a imparare la non immediata sintassi booleana. Oltre a semplificare le ricerche complesse, Excite si assume l’onere di cercare anche concetti correlati alla parola chiave specificata, associando ad ogni URL dato come risposta, un link ad una pagina contenente riferimenti ad argomenti trattati. I concetti correlati sono individuati assumendo che le parole con maggiore frequenza siano molto vicini all’argomento di ricerca. Ad esempio se si ricerca informazioni sul “football”, i concetti correlati saranno relativi alle squadre che giocano nei vari gironi. In questo modo però la ricerca finisce per generare un numero eccessivo di pagine e una soluzione può essere l’opzione che consente di limitare le ricerche a porzioni del Web. Anche Excite suggerisce termini per rifinire la ricerca e consente di visualizzare solo le pagine dal contenuto simile a quella che l’utente individua come più soddisfacente. Molto utile, infine, è la possibilità di riordinare l’elenco delle pagine ottenute raggruppando tutte quelle appartenenti al medesimo sito.

- **Infoseek.** Non ha la chiarezza e la semplicità di Excite, ma Infoseek[8] garantisce prestazioni equivalenti offrendo qualcosa in più, come la segnalazione di eventuali notizie fresche collegate alle parole chiave e l’indicazione di quali siti si occupano dell’argomento ricercato. Questo motore di ricerca è inoltre più solerte di altri motori nel rimuovere link a pagine non più esistenti. Esso si distingue infine perché contiene anche una Web directory le cui categorie appaiono direttamente nella pagine principale.

- **Hotbot.** Anche Hotbot[9] offre peculiarità interessanti, avvicinandosi a Excite quanto a semplicità d’uso. Hotbot consente di limitare la ricerca a pagine modificate in una determinata data. Forse più utile ancora è la capacità maggiore di questo motore di ricerca, rispetto ad altri, di posizionare all’inizio dell’elenco delle pagine trovate quelle più rilevanti. Questo unito al fatto che i riassunti del contenuto delle singole pagine trovate, sono più esaurienti che in altri motori, fa di Hotbot il motore che consente di arrivare prima a capire la rilevanza del documento. Si distingue inoltre per la capacità di trovare anche le pagine inserite di recente (due, tre giorni).

[5] Altavista con notazione Booleana URL: http://altavista.digital.com/
[6] Altavista sito europeo URL: http://altavista.telia.com/
[7] Excite URL: http://www.excite.com/
[8] InfoSeek URL: http://www.infoseek.com/
[9] Hotbot URL : http://www.hotbot.com/

Se la scelta tra questi motori di ricerca dovesse risultare impossibile, si può sempre ricorrere ai meta search engine, vale a dire ai servizi che uniscono automaticamente i risultati di vari motori di ricerca eseguendo per conto dell’utente le interrogazioni. Tre esempi sono: *Metacrawler*[10] *InferenceFind*[11] e *ProFusion*[12].

ProFusion interroga simultaneamente Excite, InfoSeek, Lycos, WebCrawler[13]., Open Text[14] e Altavista, ed ottenute le informazioni automaticamente elimina i documenti duplicati e quelli irrilevanti [GW96].

Alcune sue caratteristiche sono:

1. Usa una singola interfaccia utente per multipli sistemi distribuiti per il recupero dell’informazione.
2. Fornisce una maggiore libertà di scelta rispetto altri meta search, proponendo vari criteri di scelta, da quella manuale dei motori di ricerca a quella della scelta dei migliori o dei più veloci tre.
3. Fornisce una semplice interfaccia per agevolare coloro che non hanno dimestichezza con i tools di ricerca.
4. Visualizza, associato ad ogni documento recuperato, il grado di rilevanza attribuito dal Meta search, espresso con un numero compreso tra [0, 1].
5. Offre una descrizione del documento tramite un breve riassunto.

La rimozione dei documenti duplicati, è basata su poche semplici regole. Il caso più semplice si verifica quando due o più documenti hanno il medesimo URL ottenuti come risultato da più motori di ricerca, mentre regole più complesse sono applicate nel caso in cui due pagine Web identiche hanno un indirizzo di poco diverso. Per esempio l’URL “http://server/” e “http://server/index.html” si riferisce alla stessa pagina Web ed è presentato una volta sola. Infine se più pagine hanno URL diversi ma il medesimo titolo e quindi vi è la possibilità che siano uguali, ProFusion controllerà l’URL suddividendolo in tre parti: protocollo, server e path; se risulta che due delle tre sono uguali allora decide che i documenti sono uguali. Per esempio se due documenti hanno il medesimo titolo, ed URL “http://server/path” e “ftp://server/path”, ProFusion ipotizza che gli URL si riferiscono allo stesso documento. Documenti duplicati rimangono comunque perché la stessa pagina può comparire in diversi siti, con titoli diversi.

Il risultato è una lista ordinata di documenti [FG97], il cui ordine dipende principalmente dal numero di volte che le parole chiave compaio all’interno del documento e dal fattore di confidenza calcolato per ogni motore di ricerca. Infatti studi effettuati dal team che lavora su ProFusion hanno permesso di appurare che i motori di ricerca si differenziano non solo per la diversa quantità di pagine indicizzate, ma anche per la diversità di argomenti trattati. Vale a dire che un motore di ricerca può possedere una quantità maggiore di documenti per uno specifico argomento piuttosto che per un altro. Questa “specializzazione” é espressa dal fattore di confidenza, calcolato scegliendo tredici categorie e sottoponendo quattro query per ogni

[10] MetaCrawler URL: http://www.cs.washington.edu:8080/
[11] InferenceFind URL: http://www.inference.com/find/
[12] ProFusion URL http://www.designlab.ukans.edu/ProFusion.html
[13] WebCrawler URL: http://www.webcrawler.com/
[14] Open Text URL: http://www.opentext.com/omw/f-omw.html/

argomento ad ogni motore usato. Per ogni query sono esaminati i primi dieci documenti prodotti cui é attribuito un voto N_i eguale a zero se il documento é riconosciuto irrilevante, uno altrimenti.

Il fattore di confidenza é quindi calcolato come [GWG96]:

$$\left(\frac{\sum_{i=1}^{10} N_i}{10} \cdot \frac{R}{10} \right) \cdot \frac{1}{0{,}2929} \qquad (3.1)$$

La (3.1) è il prodotto di due fattori, *Rank Order Factor* (ROF) ed il *precision*, normalizzati per produrre un numero tra [0, 1]. Il ROF prende in considerazione la posizione che occupa il documento nella lista, attribuendo maggiore rilevanza ai documenti in cima. Il precision è una misura del numero dei documenti rilevanti nell’insieme recuperato, ed è incurante della posizione che occupano i vari documenti nella lista.

Per tale motivo ProFusion ha classificato i motori di ricerca suddividendoli in *esperti per argomento*, inoltre ha partizionato le informazioni contenute in Internet in tredici categorie. Ogni qualvolta viene selezionata l’opzione “*I migliori tre*” in base alla richiesta vengono interrogati i tre motori di ricerca, scelti sulla base della loro specializzazione.

3. Agenti di supporto alla navigazione e alla ricerca nel Web.

Agenti Intelligenti capaci di assistere l’utente nella navigazione in Internet sono stati già implementati sia per assolvere servizi off-line ad esempio per alleggerire il carico di rete sia per assistere on-line l’utente nella navigazione.

Esistono vari agenti che operano in maniera off-line, tipico esempio gli spider [DEEWA94] i quali ricercano dei documenti in Internet simulando il comportamento umano, ponendo ai vari motori di ricerca una richiesta. Ottenuta la risposta lo spider archivia gli URL al fine di utilizzarli solo quando il carico di rete è sceso sotto una soglia. Oltre al recupero dei documenti (Information Retrieval IR) altri agenti offrono servizi come il filtraggio (Information Filtering IF) e salvataggio [JWR97], [LS96]. I documenti devono essere catalogati e indicizzati prima di archiviarli al fine di un rapido recupero. L’indicizzazione, che consiste nell’estrazione delle Meta Informazione, include il riassunto del documento con l’aggiunta di altre informazioni utili quali: Uniform Resource Locator (URL), insieme di parole chiavi, data e orario di accesso, annotazioni da parte dell’utente.

Per assistere l’utente nella ricerca nel Web, sono stati ideati degli agenti on-line che anticipano l’esplorazione dell’utente consigliando i link da seguire in modo tale da proporre documenti ritenuti di interesse; la scelta finale, comunque, è sempre a carico dell’utente. L’unico modo che hanno questi agenti per elidere o consigliare una pagina Web è basarsi sul personale profilo utente, alcuni agenti lo creano dinamicamente come in [HL95], oppure come l’agente citato in [AFJM95] che opera su specifiche raccomandazioni da parte dell’utente più l’esperienza accumulata in precedenti esplorazioni.

L’ultimo compito dell’agente è di visualizzare, in un’opportuna interfaccia, i risultati così ottenuti [ABOTH96].

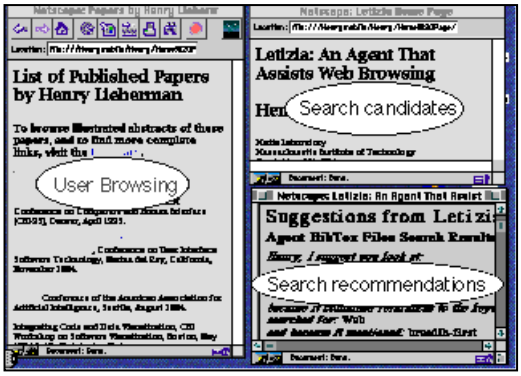
Esistono molti agenti che affiancano l’utente nella ricerca o nel salvataggio di informazioni due esempi sono Letizia e JASPER.

Letizia è un esempio di agente che assiste l’utente nella ricerca di documenti nel Web. Questo agente, creato da Henry Lieberman e consultabile in Internet all’URL “http://www.media.mit.edu/~lieber/Lieberary/Letizia/Letizia-Intro.html”, coadiuva l’utente nell’individuazione di documenti rilevanti [HL95].

La particolarità di Letizia è di non ricercare nel Web documenti rilevanti da sottoporre all’utente, ma navigare in Internet assieme lui. Ogni qual volta l’utente si sofferma a leggere una pagina Web, Letizia esplora i link associati ad essa e successivamente consiglia i link da seguire. Questo agente non possiede un modello statico degli interessi dell’utente, ma lo crea dinamicamente durante la ricerca, semplicemente memorizzando una lista di parole chiave delle pagine che l’utente si sofferma a leggere. Sull’analisi di questa lista di parole chiave si basa il consiglio della successiva pagina da esplorare.

L’interfaccia utente consiste in tre separate finestre all’interno del browser (cfr. Fig. 3.1). La prima finestra contiene la pagina selezionata dall’utente, la seconda quella candidata ed infine quella consigliata.

Figura 3.1. Un esempio di Interfaccia utente di Letizia.



Letizia non ha il controllo dell’interfaccia ma semplicemente fornisce dei suggerimenti, che possono essere accettati o rifiutati dall’utente.

Il sistema JASPER (Jasper Access to Stored Pages with Easy Retrieval) è stato progettato allo scopo di permettere il salvataggio in un database e la circolazione, a più utenti con interessi simili, di informazioni prelevate dal Web. Per ogni documento ritenuto di interesse da uno degli utenti di JASPER vengono salvate nel database locale alcune informazioni quali un sunto ed il titolo, una rappresentazione del documento originale espressa da una lista di parole chiave, l’URL, la data, l’ora del salvataggio e le eventuali annotazioni. In questo modo e’ possibile fornire agli utenti un’idea del contenuto del documento originale e permetterne la indicizzazione per un più rapido recupero.

Ogni utente di JASPER ha associato un agente ed un profilo aggiornato dinamicamente e modificabile attraverso l’aggiunta o la eliminazione di parole chiave. Quando un utente memorizza una pagina, JASPER si fa carico di controllare automaticamente tutti i profili utente e se gli interessi dell’utente sono simili agli argomenti trattati dal nuovo documento memorizzato, lo comunica inviando una e-mail.

BIBLIOGRAFIA

[ABOTH96] Michele Angelaccio: "BOTH: Cooperative Automatic Web Navigation and Hierarchical Filtering"; URL: http://russell.ce.utovrm.it/~angelac/both/final/paper.html. 1996.

[AFJM95] Robert Armstrong, Dayne Freitag, Thorsten Joachims, Tom Mitchell; "WebWatcher: A Learning Apprentice for the World Wide Web"; Proceeding of the AAAI Spring Symposium of Information Gathering from Heterogenous, Distributed Resources, Stanford, CA, March 1995.

[AV98] Alessandro Venturi: "Motori di ricerca, Internet con un clic"; PC Inter@ctive, pp 99-103, Gennaio 1998.

[CCRZ96] M. Calvo, F.Ciotti, G. Roncaglia, M.A. Zela: "Internet 96 Manuale per l’uso della rete"; 1996.

[DEEWA94] David Eichman: "Ethical Web Agent"; American Scientist, v.82, September-October, 1994 URL: http://www.tamu.edu/wha/what2482/accademics/researchs/ethics.html

[FG97] Yizhong Fan, Susan Gaunch: "An adaptive Multi-Agent Architecture for the ProFusion Meta Search System"; To appear in Proc. of WebNet ‘97: The First World Conference of the Web Society, Toronto, october 1997.

[GW96] Susan Gauch, Guijun Wang: "Information Fusion with ProFusion"; To appear in Proc. of WebNet ‘96: the First Conference of the Web Society, San Francisco, CA, October 1996

[GWG96] Susan Gauch, Guijun Wang, Mario Gomez: "ProFusion: Inteligent Fusion from Multiple, Distributed Search Engines"; Journal of Universal Computer Scienze, 1996. URL: http://www.designlab.ukans.edu/ProFusion.html.

[HL95] Henry Lieberman: "Letizia: An Agent That Assists Web Browsing"; Proceedings of the International Joint conference an AI, Montreal August 1995; URL: http://lcs.www.media.mit.edu/people/lieber/Lieberary/Letizia.html

[JWR97] John Davies, Richard Weeks, Mike Revett: "Jasper: Communicating Information Agents for WWW"; URL: http://www.labs.bt.com/jasper/html/Jasper.html; January 1996.

[LS96] Luca Savelli: "La ricerca di documenti in INTERNET con l’approccio monotematico"; pp 287-298, 1996.