Capitolo 1

Le basi della Web Analytics

Chiunque può sviluppare un sito web, sia per diletto sia per lavoro. Chiunque può facilmente installare un codice di monitoraggio, per capire quante visite riceve un sito e da dove arrivano gli utenti. Farsi un'idea sommaria dello stato di salute di un sito è abbastanza facile. Ma quando si vuole affrontare l'argomento con professionalità e precisione, bisogna fare scelte ragionate, e bisogna sforzarsi di comprendere cosa davvero significano i numeri e la terminologia specifica. È dunque fondamentale, come per qualunque altro argomento che si voglia affrontare con serietà, studiare le basi della Web Analytics, che consentiranno in seguito di valutare e interpretare i dati con una buona dose di certezza e precisione.
In questo primo capitolo dunque parliamo degli elementi fondanti di questo vasto argomento: log file, metriche, terminologia, spider e cookie.

In questo capitolo

- **Comprendere i log file**
- **Log file vs Tag**
- **Terminologia di base**
- **Identificare gli utenti**
- **Gli spider**

Comprendere i log file

Le statistiche di un sito web sono il risultato di un insieme complesso di fattori. Prima di tutto è però necessario concentrarsi su ciò che potremmo definire "livello 0": i file di log dei web server. Ogni qualvolta che un utente, attraverso il suo browser, visita una pagina web, il server che la ospita crea un file di log. In questo file vengono conservate alcune informazioni relative al visitatore della pagina. Questo fa si che, durante il nostro peregrinare in largo e il lungo per il Web, vengano impresse nei vari server le informazioni

che ci identificano, a vari livelli, sulla rete. Ma i log file non vengono creati solo se un utente compie una visita: anche gli spider dei motori di ricerca, che visitano un sito web per includerlo nel proprio indice, vengono registrati. La creazione di un log file è simile concettualmente a quello che avviene quando la polizia o i carabinieri ci fermano per un controllo. Se abbiamo commesso qualche infrazione o qualcosa nei documenti della nostra vettura non va, molto probabilmente verremo multati. Se invece abbiamo tutto in regola, ci lasceranno andare. In entrambi i casi però, la pattuglia annota su un registro i dati principali del controllo: data e ora, vettura, targa, proprietario. Queste informazioni dimostrano dunque, al di là del momento contingente, che il fermo di controllo è davvero avvenuto, e dopo che ce ne saremo andati via, costituiscono un documento che certifica quanto avvenuto. La pattuglia potrebbe dimenticarsi presto della nostra faccia, ma, il nostro passaggio rimane sul registro in maniera permanente. Questo è più o meno ciò che accade quando visitiamo un sito web… anche se apriamo solo una pagina web, e ci rimaniamo una manciata di secondi, la nostra traccia rimane indelebile sul web server. A seconda del tipo di server chiamato in causa, il log file può essere diverso. In ogni caso, possono essere definiti alcuni elementi fondanti del log file, comuni a tutti i web server. Ma quali informazioni lasciamo veramente? È utile dare un'occhiata ad alcuni file di log, per comprendere meglio quali informazioni vengono depositate.

Listato 1.1 Un log file creato dopo la visita di uno spider

```
fcrawler.looksmart.com
- -
[26/Apr/2000:00:00:12 -0400]
"GET /contacts.html HTTP/1.0"
200
4595
"-"
"FAST-WebCrawler/2.1-pre2 (ashen@looksmart.net)"
```

Il Listato 1.1 mostra un log file relativo a uno spider, dunque non a una visita da parte di un utente. Per comodità, ogni elemento del log file è stato posto su una riga, ma normalmente il log file è su un'unica riga. Analizziamolo in dettaglio partendo dalla prima riga.

Indirizzo IP

L'indirizzo IP identifica in maniera univoca la macchina dal quale è partita la richiesta di lettura della pagina sul web server. In alcuni casi l'IP è interamente numerico è dunque è necessario sforzarsi un po' per trovare il vero nome di origine, in altri casi è invece in chiaro, come quello dell'esempio citato nel Listato 1.1. Si tratta in questo caso di uno spider (*fcrawler*) che è partito dal dominio `looksmart.com`.

Cosa accade con la privacy

Questo è un primo elemento sul quale fare una piccola riflessione in materia di privacy. L'indirizzo IP di chi naviga viene elargito automaticamente al server, che provvede a trascriverlo sul log file. Normalmente non c'è nulla che un utente possa fare per evitarlo, a meno che non ricorra a sistemi di anonimizzazione della navigazione. Ovviamente, quanto sia facile associare un indirizzo IP con la persona fisica è un altro paio di maniche. Ma è comunque ovvio che se un utente identificato con un determinato indirizzo IP visitasse più volte le pagine di un determinato sito web, sarebbe già più facile tracciarne l'attività e, attraverso il percorso di navigazione, farsi un'idea basilare degli interessi di questo, almeno limitatamente al sito web che sta visitando.

Nome utente

Il nome utente, che in questo caso è "- -" viene esplicitato solo quando si è in presenza di un utente autenticato. In questo caso il valore è nullo, perché lo spider rappresentato non effettua alcuna autenticazione. Qualora fossimo in presenza di un utente che si è registrato regolarmente a un sito web, e vi accede, al posto dei trattini verrebbe inserito il nome di autenticazione dell'utente. Questo potrebbe avvenire anche in presenza di uno spider che è stato istruito ad autenticarsi come un utente.

Data e ora della richiesta

La data e l'ora della richiesta (chiamata anche Timestamp) definisce in modo inequivocabile il momento esatto in cui è stata richiesta una pagina. Il formato contempla data, mese, anno, ore, minuti e secondi.

Richiesta di accesso

La richiesta di accesso è una stringa che indica l'oggetto della richiesta. Nel caso specifico, l'istruzione `GET` indica al web server che lo spider desidera ottenere una pagina web (contact.html), utilizzando il protocollo di comunicazione HTTP, tipico del browser. Un altro esempio di istruzione è `HEAD`, che invece richiama solo l'intestazione del documento HTML. È in pratica l'equivalente della funzione `PING`, che serve ad avere il riscontro dell'esistenza di una determinata URL, sia essa un dominio o una pagina web specifica.

Codice di stato del risultato

Quando l'istruzione della richiesta di accesso ha un esito positivo, il web server scrive sul log file un codice di stato del risultato. Il valore `200` in questo caso indica che l'operazione ha avuto esito positivo. Se la URL richiesta non dovesse esistere o non fosse disponibile al momento, si otterrebbe un codice errore `404`. Questo errore è noto agli utenti poiché non solo viene trascritto nel log file, ma il browser mostra un messaggio di errore `404`, quando la URL che si tenta di aprire non esiste. Nella Tabella 1.1 sono riportati i principali codici di stato.

Tabella 1.1 Definizione dei codici di stato

Codice di stato	Definizione	Descrizione
200	OK	La richiesta è andata a buon fine. Le istruzioni possono essere: `GET`, `HEAD`, `POST`, `TRACE`.
201	Creata	La richiesta è andata a buon fine, e il risultato è stato memorizzato in una nuova risorsa. Questa è indicata attraverso la URL memorizzata nel log file.
202	Accettata	La richiesta è stata accettata, ma deve essere ancora eseguita. La risposta 202 sostanzialmente indica che il server ha accettato la richiesta in relazione a un particolare processo, e non richiede che la connessione rimanga aperta per effettuarla.
205	Reset del contenuto	La richiesta è stata eseguita, ma per poter essere visualizzata è necessario aggiornare il documento.
301	Spostato in maniera permanente	Il documento richiesto è stato trasferito in modo permanente presso un'altra URL, e qualunque nuovo riferimento deve essere passato a questa. La nuova URL viene trascritta nel campo "Location" del log file.
304	Non modificata	La richiesta è andata buon fine in seguito a un'istruzione GET, ma il documento non appare modificato.
305	Usare il proxy	La risorsa richiesta non è disponibile direttamente, ma il suo accesso è possibile solo tramite un proxy. La URL precisa viene trascritta nel campo "Proxy".
307	Redirect temporaneo	Il documento richiesto è stato trasferito temporaneamente presso un'altra URL. Poiché si tratta di un redirect temporaneo, le richieste devono essere indirizzate sulla URL attuale.
400	Richiesta non corretta	La richiesta non è stata compresa a causa di un errore di sintassi.
401	Non autorizzata	La richiesta non può essere accettata, poiché prevede un'autenticazione utente.
403	Accesso vietato	La richiesta è stata presa in carico, ma non può essere eseguita poiché non è autorizzata. Nel caso di istruzioni HEAD, il motivo del rifiuto vieni trascritto nel campo "Entity".
404	Non trovata	La URL richiesta non è stata trovata, o al momento non è disponibile.
405	Metodo non permesso	Il metodo specificato nella richiesta di accesso non è permesso in relazione alla URL richiesta.
408	Tempo scaduto	Il client non ha effettuato la richiesta entro il tempo di risposta stabilito per il server.
414	URL richiesta troppo lunga	Il server si rifiuta di rispondere poiché la URL richiesta è troppo lunga.
500	Errore interno del server	Il server ha incontrato un errore inaspettato, che impedisce di portare a termine la richiesta.
503	Servizio non disponibile	Il server non è al momento in grado di rispondere alla richiesta, a causa di un sovraccarico temporaneo o di manutenzione.
505	Versione HTTP non supportata	Il server non supporta la versione del protocollo HTTP usata nella richiesta.

Byte trasferiti

La dimensione totale dei byte trasferiti viene trascritta anch'essa nel file di log. Nel caso specifico, il valore "4595" si riferisce alla pagina "contacts.html" che pesa in pratica 4,6 Kbyte. In alcuni casi, la richiesta di scaricamento può anche essere divisa in più fasi, nel qual caso, il log file riporterà più righe con il numero totale di byte trasferiti per ciascuna richiesta.

NOTA

Quando il file di log riporta un chiaro spezzettamento del download di un file su più righe, è molto probabile che questo sia dovuto a problemi di connessione lato utente. Il client ha cioè cercato di riconnettersi più volte per ottenere la risorsa richiesta.

Origine

L'origine, anche chiamata "referrer", coincide generalmente con la pagina web da cui è partita la richiesta. In effetti, la maggior parte del traffico web ha origine da clic su link in pagine web. È però pur vero che alcune volte ci troviamo a digitare nella barra degli indirizzi del browser un indirizzo specifico a noi noto. In questo caso l'origine non viene trascritta nel log file, e al suo posto viene inserito un trattino. Nel caso specifico, trattandosi del log file di uno spider, che si "muove" in autonomia nella rete, l'origine non è rilevabile. Il referrer, come avremo modo di vedere più avanti nel testo, ricopre un ruolo molto importante nell'analisi delle statistiche. Da un lato perché in questo modo si possono identificare i referrer più rilevanti, e dunque stabilire con questi relazioni utili al sito web. Dall'altro perché offre una panoramica rapida sui siti web che linkano le pagine di un determinato sito, in relazione a termini specifici, cosa questa che può comportare un aumento di link popularity. Tratteremo questi e altri aspetti legati al SEO, nell'Appendice A del libro.

User agent

Lo user agent identifica la piattaforma da cui ha avuto origine la richiesta. Per piattaforma si intende generalmente il browser, la versione di questo e il sistema operativo. La stringa descrittiva dello user agent non è però del tutto standard, perché è specifica per ogni piattaforma. Nel caso specifico, trattandosi di uno spider, la stringa è appunto molto particolare: non viene riportato alcun browser né tanto meno un sistema operativo. È invece interessante notare che alcuni spider, progettati in modo corretto, accodano a fine stringa una casella di posta elettronica che è possibile contattare nel caso in cui lo spider provocasse problemi o errori nel sito web che sta indicizzando.
Diamo ora un'occhiata al log file creato dalla visita di un vero utente (Listato 1.2).

Listato 1.2 Un log file creato dopo la visita di un vero utente

```
111.111.111.111 - - [26/Apr/2000:00:23:47 -0400] "GET /product.html / HTTP/1.0"
200 8130 "http://search.netscape.com/Computers/Data_Formats/Document/Text/RTF"
"Mozilla/4.05 (Macintosh; I; PPC)"
111.111.111.111 - - [26/Apr/2000:00:23:48 -0400] "GET /pics/wpaper.gif HTTP/1.0"
200 6248 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
```

```
111.111.111.111 - - [26/Apr/2000:00:23:48 -0400] "GET /pics/5star2000.gif HTTP/1.0"
200 4005 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
111.111.111.111 - - [26/Apr/2000:00:23:50 -0400] "GET /pics/5star.gif HTTP/1.0"
200 1031 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
111.111.111.111 - - [26/Apr/2000:00:23:51 -0400] "GET /pics/a2hlogo.jpg HTTP/1.0"
200 4282 "http://www.jafsoft.com/asctortf/" "Mozilla/4.05 (Macintosh; I; PPC)"
111.111.111.111 - - [26/Apr/2000:00:23:51 -0400] "GET /cgi-bin/newcount?jafsof3&
width=4&font=digital&noshow HTTP/1.0" 200 36 "http://www.jafsoft.com/asctortf/"
"Mozilla/4.05 (Macintosh; I; PPC)"
```

NOTA

L'indirizzo IP di questo utente è stato volutamente modificato, per proteggerne la privacy.

In questo esempio, siamo di fronte a un log file che descrive una serie di attività di un utente. Con la stessa formula appena descritta, vengono create tante stringhe quante sono state le richieste effettuate dal browser dell'utente. Come si può osservare, ogni singola richiesta genera un log file. L'utente ha richiesto la pagina "product.html". Questa pagina contiene un certo numero di elementi (alcune GIF, e uno script CGI). Ognuno di questi genera dunque una nuova richiesta e una nuova stringa. In questo caso specifico il log file è abbastanza compatto, essendo pochi gli elementi presenti nella pagina. Questo serve però per capire quanto una pagina densa di elementi, possa generare log file molto dettagliati.

NOTA

È molto importante non confondere la richiesta fatta a un web server, con quella fatta a una pagina web. Nei sistemi di statistiche viene usato il termine "HIT" per definire la singola richiesta effettuata a una pagina in un determinato arco temporale. Questo tipo di richieste vengono però conteggiate per la pagina nella sua globalità, e non sono quindi confrontabili con le richieste fatte al web server. In effetti, la richiesta a una singola pagina viene calcolata come unica HIT in un sistema di statistiche, mentre nel file di log per quella stessa pagina, possono essere create innumerevoli stringhe di richiesta, una per ogni elemento che costituisce la pagina.

Lo user agent viene tipicamente identificato come Mozilla; il perché di questo è spiegato nel box che segue. Vediamo poi che nella stringa vengono memorizzati il tipo di computer (Mac) e il processore utilizzato (PPC). Nella stringa non viene invece riportato il browser, poiché non è stato possibile rilevarlo.

Perché tutti gli user agent si chiamano Mozilla?

Nel passato, il browser Netscape generava una stringa user agent composta dal termine Mozilla, seguito da altre informazioni. Per esempio la stringa Mozilla/4.04 (Win95 ; I) stava a indicare il browser Netscape versione 4.04, il sistema operativo Microsoft Windows 95 e il livello di sicurezza (U o I). Alcune volte, era possibile veder apparire anche la localizzazione geografica, tramite l'inserimento dell'abbreviazione di stato, tra il browser e il sistema operativo, come per esempio: Mozilla/4.04 [es] (Win16; I). Quando Internet Explorer iniziò a diffondersi, l'identificativo dello user agent venne mantenuto come "Mozilla". Nacquero diverse discussioni sulla legittimità di questa scelta da parte di Microsoft, che almeno all'apparenza, faceva pensare a un browser che si spacciava per un altro. Oggi tutto questo è storia, e il termine Mozilla resiste immutabile al passaggio del tempo.

SUGGERIMENTO

Il sito `www.useragent.org` permette di verificare istantaneamente il proprio User Agent. È sufficiente collegarsi alla URL con il browser, per ottenere una stringa simile a questa: `Your User Agent is: Mozilla/5.0 (Macintosh; U; PPC Mac OS X 10_5_7; it-it) AppleWebKit/525.28.3 (KHTML, like Gecko) Version/3.2.3 Safari/525.28.3`.

Log file vs Tag

Quando si intraprende la strada della Web Analytics è necessario scegliere il sistema più adatto, in relazione alle proprie esigenze. La maggior parte degli utenti usano sistemi gratuiti offerti su Internet, e ne vedremo un confronto più avanti. Le aziende e i professionisti si rivolgono invece spesso a servizi di statistiche a pagamento, perché ritenuti più completi e affidabili. Quanto questo sia vero, in un verso o nell'altro, lo valuteremo nel Capitolo 2. Ora è importante guardare con attenzione ai due principali metodi di raccolta dei dati: il log file e il page tag.
Il primo metodo fonda la sua funzionalità interamente sulla lettura dei file di log del server, ed estrapola da questi informazioni strutturate, che vengono poi presentate agli utenti attraverso una specifica interfaccia. È possibile visualizzare i dati divisi per tipologie, come per esempio le visite totali, le pagine visualizzate, i file scaricati e molto altro ancora.
Il secondo metodo utilizza invece il codice Javascript, che viene inserito all'interno delle pagine web da monitorare. Ogni qualvolta una pagina viene richiesta da un utente, le informazioni vengono notificate a un servizio esterno, che raccoglie e memorizza i dati.
Entrambi i metodi dunque offrono la possibilità di raccogliere i dati, e mostrarli in forma strutturata, attraverso vere e proprie web application.

NOTA

Log file e tag sono indubbiamente i metodi basilari per la raccolta e analisi dei dati di accesso ai siti web. È però importante sottolineare che la Web Analytics comprende anche altri strumenti, meno noti e diffusi, ma comunque molto utili. Tra questi, vale la pena citare le analisi delle campagne di e-mail marketing, e i test di eye tracking per i layout delle pagine web.

Analisi dei log file

Come abbiamo visto in precedenza, i web server memorizzano nei file di log gli accessi alle pagine di un sito. Storicamente le unità di misura principali (dette anche metriche) consistevano nel calcolare il numero di pagine viste, e il numero di visite. Per "pagina vista" si intendeva la richiesta di una singola pagina da parte di un utente, mentre per visita si intendeva una sequenza di pagine richieste (e dunque navigate) da parte di un utente unico, seguite da una inattività di circa mezz'ora. Vale a dire che, se un utente esplorava cinque pagine di un sito web, e dopo 40 minuti tornava sullo stesso sito, per esplorarne altre (ma anche le stesse), il sistema di statistiche contava due visite, oltre a un certo ammontare di pagine viste.
Col tempo, questo sistema di analisi ha iniziato a mostrare diversi aspetti negativi, soprattutto perché, con la proliferazione degli spider dei motori di ricerca, è diventato sempre più difficile discernerli dagli umani.

Con il diffondersi dei browser sono inoltre emersi altrettanti problemi, uno tra tutti quello legato alla cache, peraltro ancora attuale. Se una persona richiede la pagina di un sito web una prima volta, e poi lo fa per una seconda volta, il browser mostra all'utente la pagina che ha già memorizzato in cache, impedendo che questa venga riscaricata dal web server. Nella Web Analytics questo provoca problemi di coerenza dei dati, perché il percorso dell'utente viene, a un certo punto, perso.

Page tag

I problemi riscontrati con i sistemi di analisi basati sui log file, hanno spinto, a un certo punto, molte aziende che operano nel campo della Web Analytics a creare un nuovo modello di raccolta e analisi dei dati, basato non più sul file di log, ma su un codice Javascript installato dentro ogni pagina di un sito web. Il codice di monitoraggio viene inserito al termine del sorgente di pagina, quasi sempre prima del tag `</BODY>`, che determina la chiusura del "corpo" della pagina.
Ogni qualvolta una pagina viene visualizzata dal browser, appena la lettura del codice raggiunge il punto in cui è presente il codice di monitoraggio, viene innescato un sistema che notifica la lettura della pagina al servizio esterno di Web Analytics. Vengono dunque trasferite al di fuori del sito, e memorizzate, le informazioni peculiari sulla richiesta appena fatta. Inoltre, viene assegnato un cookie all'utente (sempre che questi non abbia disabilitato i cookie nel browser), che fungerà da identificativo univoco anche per le future visite.

NOTA

L'identificazione di un utente tramite cookie funziona a patto che, nelle visite successive, venga utilizzato lo stesso computer e lo stesso browser. Se una persona si connette a un sito web da luoghi diversi e con computer e browser diversi, la sua unicità non potrà essere memorizzata nel sistema di Web Analytics, e dunque sarà impossibile descriverne i comportamenti, quali le pagine visualizzate, le visite e molto altro ancora.

Indicatori comuni

Indipendentemente dal metodo di raccolta dei dati utilizzato, possono essere identificati alcuni indicatori comuni, che vengono usati nella Web Analytics. Questi includono:

- numero di visite e numero di visitatori unici;
- durata delle visite e ultima visita;
- utenti eventualmente autenticati e ultimo utente autenticato;
- giorni del mese e concentrazione oraria delle visite;
- indirizzo IP degli utenti;
- numero totale di pagine visualizzate;
- pagine maggiormente visitate;
- pagine principali di ingresso e uscita;
- classifica per tipi di file richiesti (GIF, JPG, PDF ecc.);

- sistemi operativi degli utenti;
- browser usati dagli utenti;
- spider che hanno visitato il sito;
- referrer;
- motori di ricerca che hanno originato le visite;
- chiavi di ricerca che hanno originato le visite;
- errori di connessione HTTP.

Vantaggi e svantaggi di entrambi i metodi

Il mercato dei sistemi di Web Analytics è molto vario e offre sia servizi basati su log file sia page tag. In alcuni casi una stessa azienda li offre entrambi. La domanda che nasce spontanea è dunque se questi sistemi possono essere usati in ugual modo, o se uno sia migliore dell'altro. La risposta non è del tutto netta, perché entrambi hanno vantaggi e svantaggi. Si tratta di valutare se, in rapporto a ciò che si vuole ottenere, l'uno sia preferibile all'altro.

Vantaggi dei log file

I vantaggi principali di un sistema di analisi basato su log file possono essere sintetizzati come segue.

- I web server producono "naturalmente" i file di log, senza che l'utente debba intervenire. I sistemi basati su questa metodologia si limitano dunque a leggere il log file, senza richiedere che l'utente debba installare alcun codice nelle pagine web.
- I dati risiedono presso il provider che ha in hosting il sito web, e sono in un formato standard (quello del log file appunto). Questo può essere utile, qualora si decida di cambiare in seguito il software di analisi. I sistemi basati su page tag invece fondano la propria struttura su database proprietari, che legano indubbiamente l'utente al fornitore.
- I file di log contengono informazioni relative alle singole visite degli spider dei motori di ricerca. Sebbene queste non dovrebbero essere riportate nei risultati, come parte dell'attività degli utenti, è utile sapere quali motori di ricerca indicizzano il proprio sito, in un'ottica SEO.
- I file di log non necessitano la chiamata a un server esterno. Su siti di grandi dimensioni, con svariate visite giornaliere, si può notare una netta differenza sia a livello di tempo apertura pagina, sia pure a livello di banda occupata.
- I web server registrano ogni singola richiesta, mentre i sistemi di page tag non sono in grado di assicurare una totale memorizzazione delle richieste. Questo può accadere per vari motivi: gli utenti possono aver disabilitato le funzioni Javascript del browser; le pagine possono venire caricate in modo incompleto dal browser, inibendo l'attivazione del codice Javascript; l'attività su alcuni tipi di documenti (per esempio i PDF) può essere tracciata solo a partire da una pagina che ne contiene il link.

Il falso problema del JavaScript

Una delle questioni su cui spesso si punta l'attenzione, quando si parla di page tag, è che l'utente possa aver disabilitato volontariamente il supporto al codice Javascript nel proprio browser. Questo è sicuramente possibile, a maggior ragione nelle aziende in cui spesso per motivi di sicurezza i responsabili IT bloccano centralmente l'avvio di qualunque codice Javascript. Si può comunque paventare questo timore, ma se non si hanno dati alla mano, non si può asserire con certezza che un certo numero di persone disabilitino il supporto Javascript nel proprio browser. È utile dunque dare un'occhiata alla Tabella 1.2, tratta dal sito web `www.w3schools.com`, che riporta i dati rilevati dal 2000 al 2008.

Tabella 1.2 Statistica di funzionalità del codice JavaScript nei browser dal 2000 al 2008

Anno	Percentuale di utenti con JavaScript attivo	Percentuale di utenti con JavaScript non attivo
2008	95%	5%
2007	94%	6%
2006	90%	10%
2005	89%	11%
2004	92%	8%
2003	89%	11%
2002	88%	12%
2001	81%	19%
2000	80%	20%

Vantaggi dei page tag

I vantaggi principali del page tag sono invece i seguenti.

- Il codice di monitoraggio Javascript viene avviato nuovamente ogni volta che la pagina viene ricaricata. In questo modo, si eliminano le problematiche legate alla cache del browser.
- È molto facile modificare il codice di monitoraggio affinché includa informazioni aggiuntive, come per esempio la risoluzione video dell'utente, piuttosto che altri dati relativi alla sessione in corso (per esempio il prezzo di acquisto di un determinato prodotto). Con il sistema dei log file, le informazioni che non vengono raccolte dal server, devono essere necessariamente incluse nella URL, rendendola esageratamente lunga.
- Attraverso il codice di monitoraggio Javascript è possibile rilevare attività che non sono strettamente collegate al web server, come per esempio: interazioni varie all'interno di documenti di Adobe Flash; eventi del mouse (tramite le funzioni *onClick*, *onMouseOver*, *onBlur*, *onFocus* ecc.), e il completamento dei moduli da parte degli utenti.
- Il codice di monitoraggio si occupa direttamente dell'associazione dei cookie agli utenti. Nel caso di utilizzo del log file, è necessario istruire il server in tal senso.
- Il page tag è disponibile per tutti quei soggetti che non hanno il controllo del web server su cui risiede un sito web. In effetti, per poter usare efficacemente i sistemi di rilevamento tramite log file, bisognerebbe avere il controllo totale del server su cui risiedono i siti web da controllare.

Alcuni aspetti economici

Preso atto delle differenze tecniche dei due sistemi, può essere utile fare anche alcune riflessioni di natura economica. L'approccio "log file" prevede solitamente l'acquisto di una licenza client, che si può utilizzare su uno o più computer, senza limiti di utilizzo. È però pur vero che alcuni fornitori di queste soluzioni, hanno iniziato a introdurre un massimale di pagine viste in un determinato periodo (per esempio un anno solare), per evitare che l'analisi dei dati, soprattutto in presenza di siti web con moltissimi accessi, incida pesantemente sulla banda utilizzata. Oltre alle versioni commerciali si trovano in rete molti software Open Source, che svolgono il lavoro di raccolta dati in maniera egregia. Anche in questa ultima condizione, il problema principale che si incontra con l'approccio log file è che i dati, una volta letti, devono essere memorizzati in maniera permanente da qualche parte. Sebbene il costo delle periferiche di archiviazione sia ormai irrisorio, può diventare oneroso, in termini di tempo e di gestione, dover mantenere un archivio che cresce rapidamente giorno dopo giorno. Inoltre, è necessario aggiornare costantemente il software con tutti gli update e le patch di sicurezza. Per quanto riguarda i sistemi di page tagging, molti fornitori di questo servizio offrono soluzioni a pagamento basate su un numero massimo di accessi/visualizzazioni di pagine mensili. Se dal punto di vista economico questo può essere un limite, dal punto di vista della gestione nulla è richiesto all'utente finale, che si limita a utilizzare la web application messa a disposizione senza curarsi di spazio occupato sul disco, aggiornamenti di sicurezza e altre questioni. Anche in questo caso, le soluzioni free non mancano. Le più usate sono: Google Analytics, e Yahoo Web Analytics.

Le metriche di Audiweb

Le aziende che operano nel campo della Web Analytics hanno spesso parametri di valutazione proprietari, le cosiddette *metriche*. Le metriche non sono dunque le stesse per tutti i sistemi di Web Analytics e, per questo motivo, il raffronto dei risultati provenienti da sistemi diversi può essere difficile. Le metriche sono dunque un fattore chiave del rilevamento statistico, ed è quindi importante conoscerne le regole principali. In Italia, la società Audiweb è un soggetto realizzatore e distributore di dati sulla audience online (Figura 1.1).
Il sistema Audiweb integra diverse fonti.

- *Ricerca di base*. Ricerca quantitativa costituita da una serie (waves) di interviste, su un campione della popolazione italiana residente (individui di 11-74 anni), effettuata mediante questionari strutturati con metodologia CAPI (*Computer Aided Personal Interview*) e finalizzata alla definizione dell'universo degli utenti Internet e alla descrizione delle loro caratteristiche in termini di profilo socio-demografico e attitudinale.
- *Catalogo*. Informazioni su tutta l'offerta editoriale disponibile su Internet, strutturata per gerarchia di navigazione (Property, Brand, Domain, Channel…), organizzata per categorie di contenuti editoriali ed eventualmente per macro-aggregazioni.
- *Sistema censuario*. Rilevazione oggettiva e completa (censuaria) di tutta l'attività di Internet sui siti del catalogo, tramite feedback tecnici verso i server di raccolta dati o web server attivati dall'apertura di pagine web e altre azioni.

Figura 1.1 La home page del sito di Audiweb (www.audiweb.it).

- *Panel*. Rilevazione oggettiva (automatizzata attraverso un software meter) della fruizione di Internet sui siti del catalogo da parte di campioni statisticamente rappresentativi di popolazione. L'operazione avviene nel pieno rispetto delle norme sulla privacy.

L'obiettivo di Audiweb è quello di fornire al mercato informazioni oggettive e imparziali, sia dal punto di vista qualitativo sia quantitativo, in merito alla fruizione del mezzo Internet e su altre reti online, utilizzando opportuni strumenti di rilevazione. Per poter espletare questo compito, Audiweb ha definito dei parametri standard per le metriche, che sono riassunti nella Tabella 1.3.

Tabella 1.3 Metriche di Audiweb

Parametro	Descrizione
Visita	Una o più richieste consecutive fatte dallo stesso visitatore all'interno di un sito con un tempo limite di inattività di 30 minuti. La ripresa dell'attività dopo 30 minuti sarà conteggiata come una seconda visita. Non devono essere considerate le attività di robot e spider.
Browser unici	Il numero di browser diversi che, in un determinato arco temporale, effettuano una o più visite a un sito. Sono convenzionalmente identificati da un cookie unico nel caso dei sistemi browser based, o nel caso dei sistemi server based da una combinazione unica di IP address + un altro indicatore che può essere uno user agent, un cookie, un registration ID. Deve essere comunque sempre disponibile, per eventuali controlli, il dato relativo al numero degli IP univoci nel medesimo arco temporale.
Pagine viste	Si intende per pagina un documento che contiene testo, immagini, suoni o altri oggetti. Si intende per "pagina vista" una richiesta esplicita, fatta da un utente, depurata dalle attività di robot, spider ecc., da eventuali codici di errore. Nel caso il sistema lo consenta, il tag di misurazione deve essere posizionato in fondo alla pagina e quindi sarà considerata pagina vista quella interamente scaricata e visualizzata sullo schermo dell'utente. Si considera "pagina vista" anche quella proveniente da cache e da proxy, se puntualmente misurabile. • Pop-up, pop-under e splash pages = misurate solo se richieste dall'utente; • Interstitial e jump pages = non misurate; • Sondaggi (surveys) = misurate se richiesti esplicitamente dall'utente; • Autorefreshed pages = sono misurate tutte. Quelle con un tempo di refresh inferiore ai 15' dovranno essere esposte in modo separato e identificabile rispetto alle altre. Dovrà in ogni caso essere dichiarato dall'editore il tempo di refresh applicato; • Pagine a frame = per ogni pagina non può essere conteggiato più di un frame. Anche in questo caso sarà importante la corretta dichiarazione dell'editore su quale è il frame considerato più rilevante e che quindi deve essere conteggiato; Le pagine generate da sistemi chat, forum, instant message, personal page sono misurate ma devono poter essere identificabili anche separatamente, se necessario.
Tempo medio per pagina	Il tempo in minuti e secondi trascorso sulla pagina.
Tempo totale	Il tempo complessivo in minuti e secondi trascorso sulle pagine prese in esame.
Durata della visita	Il tempo in minuti e secondi della visita.
Utenti unici	Sono i singoli individui che si sono collegati in un determinato arco temporale, a un sito e/o a pagine di esso effettuando una o più visite. Differiscono dai browser unici in quanto rappresentano persone fisiche.
Penetrazione utenti attivi (*Active Reach*)	Percentuale di utenti, rispetto a un universo di riferimento (per esempio la popolazione italiana attiva nel periodo di osservazione), che in un determinato arco temporale ha visitato un sito o parte di esso. Esempio: popolazione Internet italiana attiva: 20 milioni di persone; utenti di un sito: 10 milioni => Active Reach = 50%.
Penetrazione universo (*Universe Reach*)	Percentuale di utenti, rispetto all'universo totale degli utenti potenziali (cioè coloro che hanno la possibilità di collegarsi anche se non lo hanno necessariamente fatto nel periodo di osservazione), che in un determinato arco temporale ha visitato un sito o parte di esso.

NOTA

I provider "censuari" accreditati per effettuare la rilevazione oggettiva dei dati di audience dei siti aderenti ad Audiweb sono (al momento in cui questo libro va in stampa): Nedstat, Nielsen SiteCensus, ShinyStat, Trackset, WebTrends. Questi provider hanno accettato di effettuare le proprie rilevazioni, seguendo le metriche di Audiweb, al fine di garantire risultati oggettivi e imparziali, sia dal punto di vista qualitativo, sia da quello quantitativo. Sono invece molte le aziende che usufruiscono dei servizi di Audiweb, tra queste figurano: concessionarie di pubblicità, testate nazionali, società telefoniche e molte altre. Una lista completa di tutti gli aderenti di Audiweb si trova all'indirizzo: `http://www.audiweb.it/links/index.php`.

Terminologia di base

Gli elementi della Web Analytics vengono identificati attraverso una terminologia globale, che gli addetti ai lavori ben conoscono. Spesso però i termini usati vengono mal interpretati, portando a risultati ambigui, soprattutto dal punto di vista espressivo. È quindi consigliabile avere sempre in mente i termini di base e soprattutto il loro significato. La Tabella 1.4 risponde efficacemente a questa necessità.

Tabella 1.4 Terminologia di base della Web Analytics

Termine	Descrizione
Hit (richiesta)	La richiesta di un file a un server web. Questo parametro è disponibile solo in relazione all'analisi dei file di log. Il numero di hit ricevute da un sito web è frequentemente (ed erroneamente) utilizzato come indice di popolarità del sito. Ciò accade perché ogni pagina è composta da numerosi elementi, che vengono richiamati durante la sua apertura, generando un numero molto elevato di hit. In questo caso dunque un numero di hit molto elevato indica più che altro la complessità di una pagina, e non come si pensa, la popolarità del sito web. La somma totale dei visitatori e delle pagine viste indica invece con maggiore precisione il grado di popolarità di un sito web.
Page view (pagina visualizzata)	La richiesta di un file che viene identificato come "pagina" nel caso di log file, oppure l'avvio di uno script di monitoraggio, nel caso di utilizzo di page tag. Nel caso di analisi dei file di log, una singola pagina visualizzata può generare numerose hit, poiché tutte le risorse necessarie per comporla (immagini, codice Javascript, file CSS) vengono richieste allo stesso modo al web server.
Visita/sessione	Una serie di richieste provenienti da un medesimo e unico client all'interno di un arco temporale che solitamente è di 30 minuti. Una visita contempla normalmente una o più visualizzazioni di pagina.
Prima visita/ prima sessione	La prima visita in assoluto di un visitatore che non è mai stato in precedenza su un determinato sito web.
Visitatore unico	Un visitatore identificato in maniera univoca, sia tramite log file, sia tramite page tag, all'interno di un arco temporale ben definito (giorno, settimana, mese ecc.). Un visitatore unico viene conteggiato una sola volta all'interno dell'arco temporale definito, sebbene questi possa poi tornare nuovamente a visitare un determinato sito web. Poiché l'identificazione di un visitatore avviene generalmente attraverso l'attribuzione di un cookie al suo computer/browser, qualora questi si connettesse da due postazioni diverse, non sarebbe più possibile identificarlo come visitatore unico e dunque verrebbe conteggiato come se si trattasse di due utenti unici.

Visitatore di ritorno	Un visitatore che ha compiuto in precedenza almeno una visita. Il periodo che intercorre tra la prima e l'ultima visita viene tecnicamente chiamato "recency", e si misura in giorni.
Nuovo visitarore	Un visitarore che non ha compiuto in precedenza alcuna visita.
Impression	Impression è un termine legato alla pubblicità online. Si parla di impression ogni qualvolta una pubblicità, come per esempio un banner, viene visualizzata su una pagina web.
Singleton	Con il termine "singleton" ci si riferisce a un particolare tipo di visita in cui viene visualizzata una sola pagina. Il singleton viene erroneamente considerato poco utile nelle metriche della Web Analytics, sebbene ricopre una grande importanza nei casi in cui si cerca di identificare truffe ai danni delle campagne di pay-per-clic. Risulta inoltre utile per identificare la *frequenza di rimbalzo* degli utenti in relazione a una determinata pagina web.
Frequenza di rimbalzo	La frequenza di rimbalzo corrisponde alla percentuale di visite in cui un visitatore approda sulla pagina web di un sito e abbandona subito dopo il sito senza visitare altre pagine. Si tratta di un comportamento ricorrente, che fonda spesso le sue ragioni in una cattiva strutturazione del layout di pagina e di navigazione. È dunque possibile, analizzando le pagine che ottengono un elevato numero di rimbalzi, identificare i probabili motivi di questo fenomeno e modificarle di conseguenza.
Tempo di visualizzazione	Il tempo dedicato dall'utente alla visualizzazione di una singola pagina.
Durata della sessione	La durata media della visita di un utente su un determinato sito web.
Permanenza media sulle pagine	La durata media dedicata alla consultazione delle pagine di un sito web.
Pagine viste per visita	Il numero medio di pagine che vengono visualizzate da un utente durante una visita. Questo valore viene calcolato dividendo il numero totale di visualizzazioni di pagine per il numero totale di sessioni.
Frequenza di visita	La frequenza di visita definisce la ricorrenza con cui gli utenti visitano un determinato sito web. Questo valore viene calcolato dividendo il numero totale di sessioni (o visite) per il numero totale di utenti unici. Questo valore ricopre un ruolo interessante nella Web Analytics perché indica il grado di "affezione" degli utenti rispetto a un determinato sito web.
Percorsi	I percorsi definiscono la sequenza di pagine che un utente visualizza durante una visita. L'analisi dei percorsi può risultare molto utile per identificare i "nodi" principali di un sito web, su cui concentrare azioni di web marketing.

Identificare gli utenti

Per poter funzionare correttamente i sistemi di Web Analytics devono poter identificare in maniera univoca i visitatori. Il modo più semplice per identificare un utente, sarebbe quello di associarlo all'indirizzo IP della sua connessione. Questo, sebbene tecnicamente possibile, condurrebbe presto a problemi di analisi, dato che la maggior parte degli utenti che si collega alla rete utilizza connessioni ADSL con indirizzo IP dinamico. In pratica, ogni qualvolta un utente si collega a Internet, il provider associa a questi uno degli indirizzi IP disponibili in quel momento.

Dammi la zampetta che ti do un biscottino: i cookie

Visto che con l'indirizzo IP risulta del tutto impossibile identificare un utente in maniera stabile, i sistemi di Web Analytics utilizzano normalmente i cookie. Per comprendere cosa realmente sia un cookie, è sufficiente pensare al rapporto tra un cane ammaestrato e il suo padrone. Il padrone esorta il proprio cagnolino a dargli la zampetta, attraverso un gesto o una frase. Il cane ubbidisce e alza la zampetta, perché sa che a quel punto il padrone lo premierà con un biscottino. Ciò che abbiamo appena descritto è sostanzialmente ciò che avviene ogni qualvolta il browser di un utente tenta di collegarsi a un sito web. Quando il browser si collega a un sito web, riceve in "regalo" un biscottino: il cookie. Il biscottino viene dunque associato a un utente specifico, quello che si è appena collegato. La prossima volta che quell'utente si collega al medesimo sito web, il cookie viene letto dal browser, e dunque l'utente viene nuovamente identificato.

NOTA

Un utilizzo tipico dei cookie consiste nel "personalizzare" la visita degli utenti. Spesso infatti, dopo essersi registrati su un sito web, è facile vedere a una visita successiva, anche lontana nel tempo, una frase che recita: "Bentornato…" seguita dal proprio nome. Questo accade grazie al riconoscimento del cookie, depositato in precedenza dentro al browser dell'utente, da parte del sito web.

L'ambigua storia dei cookie

Qualche anno fa, i sistemi di Web Analytics utilizzavano dei cookie erogati dai propri server. Il problema era che questo tipo di cookie, non solo erano in grado di consentire l'identificazione di un utente su un determinato sito web, ma consentiva la tracciatura dello stesso anche su siti web tra loro diversi. In questo modo le società di Web Analytics erano in grado di monitorare l'attività degli utenti anche al di fuori di uno specifico dominio, infrangendo indubbiamente aspetti legati alla privacy dell'utente. Quando gli utenti hanno iniziato a prendere coscienza di quanto i cookie violassero potenzialmente la proprio privacy, in molti iniziarono a bloccare l'uso dei cookie da parte dei browser, o addirittura presero l'abitudine a rimuoverli del tutto. La cancellazione o il blocco dei cookie è un problema non da poco per le società che si occupano di Web Analytics, poiché questo corrisponde a perdere totalmente traccia dell'attività di un utente, oltre che non riuscire più a identificarlo in maniera univoca. Oggigiorno, la maggior parte dei sistemi di Web Analytics usa un approccio diverso alla gestione dei cookie, limitandosi a utilizzarli solo per l'identificazione e la tracciature degli utenti in relazione al dominio per cui i sistemi vengono attivati. Ciò comunque non risolve il problema della cancellazione dei cookie: quando un utente, deliberatamente, li rimuove dal proprio browser, la raccolta dei suoi dati di accesso presso un sito web viene totalmente falsata. Senza un identificazione persistente dell'utente dunque, le conversioni, l'analisi dei clic, i percorsi di navigazione e molto altro, diventano impossibili da monitorare con precisione, e dunque minano l'affidabilità dei dati prodotti dal sistema.

Funzionamento dei cookie

I cookie permettono di identificare gli utenti in base ad alcuni elementi: browser, computer, ma non solo. Rispetto all'utilizzo degli indirizzi IP infatti, i cookie sono in grado di fornire anche l'ID utente. Un indirizzo IP identifica infatti il singolo computer. Questo significa che, qualora più persone usino la stessa postazione per accedere alla rete, anche con account diversi, l'indirizzo IP rimarrà invariato. Sia che si tratti di una connessione telefonica (PSTN o ISND) sia ADSL, non sarà comunque possibile tracciare i singoli utenti collegati. D'altra parte però, se un medesimo utente esplora un sito web con browser diversi, il sito non collegherà gli accessi alla stessa persona. Ogni browser ha infatti una propria cartella di cookie. Con una verifica congiunta dell'indirizzo IP e dei cookie, il server è dunque in grado di collegare gli accessi alla stessa persona.
Un utilizzo combinato di cookie e indirizzo IP permette quindi di ottenere la tracciatura della navigazione: collegandola alla stessa persona, con l'indirizzo IP, quando un cambiamento di browser porta a usare cookie differenti per entrare nello stesso sito; a persone diverse, con i cookie, quando più persone con un proprio account utente condividono lo stesso computer e sono tracciabili con un comune indirizzo IP.

Tabella 1.5 Struttura di un cookie

Attributo	Descrizione
Nome	È una variabile a cui viene assegnato il nome del cookie. Si tratta di un campo obbligatorio.
Dominio	Specifica il dominio di provenienza del cookie.
Scadenza	È un attributo opzionale. Permette di impostare la data di scadenza del cookie, ovvero quando questi dovrà cessare di funzionare. Si può impostare una data specifica, oppure la decorrenza di un numero preciso di giorni. Altre possibilità sono "now", che elimina il cookie subito dopo che ha assolto al suo compito, e "never", che rende la durata del cookie illimitata.
Percorso	Specifica il percorso preciso da cui ha origine il cookie che viene inviato all'utente.
Modalità di accesso	Attraverso il parametro "HttpOnly" fa sì che il cookie diventi invisibile ai linguaggi di scripting lato client presenti nella pagina.
Sicuro	Istruisce il sistema a trasmettere il cookie in modalità criptata, tramite protocollo HTTPS.

NOTA

Il Web odierno è ormai costellato di malfattori che, tramite vari sistemi (come per esempio il Phishing), cercano di truffare gli utenti. Anche i cookie possono essere usati per tali scopi. Questi possono essere manipolati tramite la tecnica del "cookie poisoning", che prevede la modifica dei contenuti di un cookie con il fine ultimo di eludere i meccanismi di sicurezza e immedesimarsi nei panni di un utente durante l'accesso a un sito web con sistema di autenticazione. Basti pensare che oggigiorno gli hacker che sferrano attacchi contro i sistemi di e-commerce utilizzano primariamente la manipolazione dei cookie come metodo di accesso.

Scadenza dei cookie

La durata e validità dei cookie non è ovviamente infinita. Abbiamo visto che basta un'azione da parte dell'utente per rimuoverli o bloccarli. Se questo non dovesse accadere è probabile che i cookie conservati nel nostro browser scadano, dopo un certo periodo. Le modalità e i tempi di scadenza possono essere diversi, e vengono impostati specificatamente in base al tipo di cookie. Talvolta, è persino possibile che i cookie non scadano mai. Per esempio, il sistema di monitoraggio Google Analytics imposta una durata massima di 2 anni per i cookie relativi a un visitarore unico. La visita è invece definita tramite il metodo `_setSessionTimeout()` in modo che il cookie relativo scada dopo 30 minuti. Il cookie destinato al monitoraggio dell'origine delle visite ha invece una durata di 6 mesi. Questo è considerato un tempo sufficiente per poter mettere in relazione più visite di un medesimo utente all'interno di una campagna pay-per-clic. Nella Tabella 1.6 sono riportati, a titolo di esempio, alcuni metodi utilizzati nei cookie di Google Analytics e la relativa descrizione.

Tabella 1.6 Principali metodi dei cookie di Google Analytics

Metodo	Descrizione	Scadenza
`__utma`	Questo cookie identifica un browser nel momento in cui avviene la prima visita a un sito web.	2 anni
`__utmb`	Questo cookie viene usato per stabilire l'avvio di una sessione utente, quando questi visita un sito per la prima volta.	30 minuti
`__utmz`	Questo cookie memorizza l'origine (referral) di una visita.	6 mesi

I cookie di Flash

I cookie tradizionali sono in grado di memorizzare circa 4 Kbyte di informazioni, e sono facilmente rimovibili dal browser attraverso le impostazioni di questi. Esiste però un altro tipo di cookie, che viene generato durante la navigazione su un sito costruito con Adobe Flash (Figura 1.2).

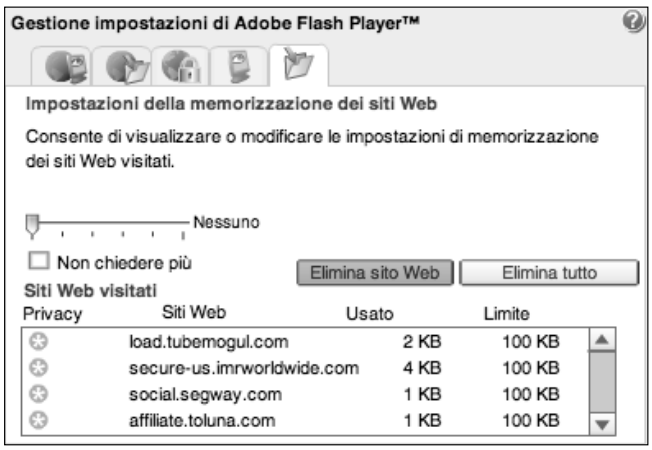

Figura 1.2 Il pannello di gestione delle impostazioni di Flash.

Questa tipologia di cookie è in grado di memorizzare informazioni con un ordine di grandezza ben più grande dei cookie tradizionali. Sono inoltre archiviati in maniera diversa rispetto a quelli tradizionali e dunque più difficili da trovare e rimuovere. Per poterli gestire è necessario avviare il pannello di gestione delle impostazioni di Flash. Poiché non si tratta di un client installato presso il computer dell'utente, è necessario richiamarlo via Web, dal sito Adobe: `http://www.macromedia.com/support/documentation/it/flashplayer/help/settings_manager02.html`. Il pannello permette di accedere a cinque diverse categorie (*Privacy generale*, *Memorizzazione*, *Sicurezza*, *Notifica*, *Privacy siti web*) dalle quali si possono definire le impostazioni più adeguate per l'utente.

Applicazioni dei cookie

I cookie hanno un campo di applicazione molto esteso. Gli utilizzi più comuni comprendono:

- gestione dei carrelli di spesa nei siti di e-commerce.Vengono usati per dare all'utente la possibilità di aggiungere e rimuovere dal carrello i prodotti che si intendono acquistare, in qualunque momento;
- riconoscimento automatico degli utenti che devono accedere a un'area riservata;
- personalizzazione di layout e aspetto di una pagina da parte di un utente autenticato. Un esempio tipico è Google, che permette di modificare le impostazioni della home page;
- tracciatura dei percorsi degli utenti sul Web, con il fine ultimo di inviare banner pubblicitari coerenti con gli interessi di questi;
- tracciatura dei visitatori di un sito web per scopi di Web Analytics.

Gli spider

I motori di ricerca recuperano le informazioni sulla rete utilizzando dei piccoli programmi, chiamati comunemente spider, che leggono le pagine web e ne estraggono informazioni peculiari per costituire o aggiornare gli indici dei motori di ricerca. Questi piccoli software possono assumere vari nomi (crawler, spider, robot ecc.) ma tutti, comunque, vengono accomunati da un medesimo scopo: esplorare le pagine web, alla ricerca di informazioni di rilievo da riportare a "casa". Uno dei problemi che la circolazione libera degli spider comporta, a chi si occupa di Web Analytics, è proprio che questi vengono tracciati al pari degli utenti e dunque, se non vengono "trattati" in modo adeguato, possono falsare i risultati delle analisi.

NOTA

Il sito web `www.robotstxt.org` riporta in maniera precisa la definizione di uno spider (Figura 1.3): "un programma che viaggia in modo autonomo attraverso la struttura ipertestuale del Web, leggendo documenti già indicizzati in precedenza, e trovandone di nuovi, tramite i collegamenti ipertestuali. In questo peregrinare, gli spider applicano una sorta di selezione euristica, che spesso li può far assomigliare agli utenti, ma si tratta esclusivamente di scelte automatiche basate su regole prefissate."

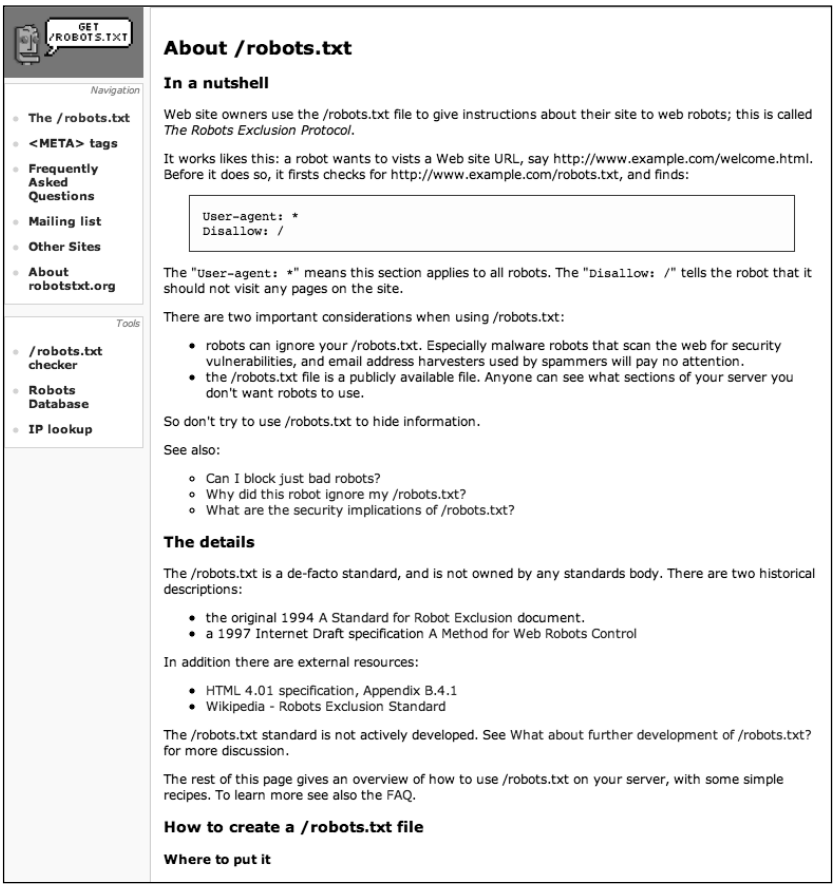

Figura 1.3 La home page del sito di Robotstxt.org.

Tabella 1.7 Nomenclatura dei principali spider

Nome spider	Motore di ricerca
Googlebot	Google
MSNbot	MSN
Ask Jeeves/Teoma	Ask.com
Architext Spider	Excite
FAST-WebCrawler / crawler@fast	AllTheWeb
Slurp	Inktomi
Yahoo Slurp	Yahoo
Ia_archiver	Alexa
Scooter	Altavista
InfoSeek sidewinder	InfoSeek
Lycos_Spider_(T-Rex)	Lycos

NOTA

Esistono molti tipi di spider, ma è bene sapere che non tutti sono davvero "buoni". Alcuni software che vengono usati per duplicare interi siti web (come per esempio Teleport Pro) generano un proprio spider in maniera dinamica, con lo scopo di analizzare rapidamente le pagine dei siti. Questi spider non hanno niente a che vedere con quelli dei motori di ricerca e dunque, se identificati, sarebbe meglio bloccarne l'accesso a un sito. Nell'esempio citato (Teleport Pro), sarebbe sufficiente creare un file `robots.txt` nella cartella principale del sito e inserire dentro questo la sequenza di comandi: 1) `User-agent: Teleport Pro` e 2) `Disallow: /`. La prima riga identifica lo spider attraverso il suo nome, la seconda ne impedisce l'accesso al sito.

Discernere il traffico generato dagli spider da quello generato dai normali utenti, è fondamentale ai fini di una corretta analisi statistica. Se non si è in grado di filtrare le visite provenienti da questi software automatici, i risultati delle analisi appariranno indubbiamente "gonfiati" rispetto alle reali connessioni umane e qualunque strategia che si basi su quei dati verrebbe a essere compromessa. In genere gli spider non interferiscono con il codice di monitoraggio dei sistemi di page tag, ma questo non è sempre vero. Se pensiamo all'approccio log file, mettendo in collegamento l'indirizzo IP degli spider con il motore di provenienza si può, nel tempo, identificarli e isolarli dall'analisi, ma anche in questo modo, a lungo andare, le cose si possono complicare. Bisogna dunque sapere analizzare i dati di accesso, provando a discernere le visite automatiche da quelle umane. In genere gli spider hanno taluni comportamenti che possono sin da subito renderli riconoscibili.

- Hanno una frequenza di accesso ai siti web che spesso è ripetitiva. Alcune volte, possono leggere una pagina web ripetutamente, ogni 10 minuti, ogni ora, ogni settimana, ogni mese, e così via.
- Seguono dei percorsi di navigazione che non sono affatto comuni. Partendo per esempio da una pagina web, esplorano rapidamente tutti i link a partire da quella pagina, compresi anche eventuali documenti collegati, come i PDF. Questo tipo di comportamento è facilmente rilevabile poiché un vero utente sceglierà in genere un percorso ben preciso a partire da una determinata pagina, e non perlustrerà tutti i link contenuti in questa.
- Il numero medio di pagine per visita è molto elevato rispetto agli utenti. In genere infatti, questo valore attribuito agli utenti è moderatamente basso. Gli spider invece si "cibano" di tutto quel che trovano durante una sessione di esplorazione, dimostrando chiaramente di non essere degli utenti.
- Richiedono spesso alcuni documenti a cui i normali utenti non possono avere accesso diretto, come per esempio i file XML.

NOTA

Gli spider si sono evoluti molto negli ultimi anni e, da quel che potevano fare alle origini, hanno oggi funzionalità molto più complesse. Gli spider odierni sono in grado di eseguire codice Javascript contenuto nelle pagine web; riescono ad accedere a un sito web tramite i referrer con diversi metodi; non hanno necessariamente un medesimo indirizzo IP; richiamano ripetutamente una medesima pagina web creando spesso problemi di utilizzo della banda; sono in grado di camuffare il proprio user agent per evitare di essere identificati; possono accettare i cookie, al pari di quanto farebbe un comune browser.

Figura 1.4 La home page del sito di IAB.

IAB: le pagine gialle degli spider

Analizzare gli accessi a un sito web con lo scopo di identificare gli spider può essere un'attività lunga e difficoltosa. Sapendo sin da subito a quali indirizzi IP corrispondono alcuni spider, il tutto risulta sicuramente molto più facile. Lo IAB (*Interactive Advertising Bureau*) pubblica costantemente un elenco aggiornato di tutti gli spider delle aziende che ne confermano l'esistenza. Queste informazioni possono essere attinte dal sito web `www.iab.net` dopo essersi registrati sulla pagina `http://www.iab.net/sites/login.php`. Come si può notare una volta aperta la pagina, le aziende che hanno aderito a questa iniziativa sono molteplici e di spicco.

Il file robots.txt

Nel tentativo di isolare le visite "robotiche" da quelle degli utenti, si ricorre spesso all'utilizzo del file `robots.txt`. Si tratta di un comune file testuale che viene depositato nella cartella principale (root) di un sito web e contiene alcune istruzioni specifiche per gli spider. Sostanzialmente un file di questo tipo è in grado di evitare che gli spider abbiano

accesso ad alcuni file e cartelle, che devono rimanere privati. Se da un lato l'utilizzo del file `robots.txt` potrebbe sembrare una strada utile per fare in modo che le statistiche non vengano inficiate dalle loro visite, dal punto di vista dell'indicizzazione dei contenuti presso i motori di ricerca, questi possono arrecare qualche problema. Impedendo che uno spider acceda a una pagina, si fa in modo che questi non venga conteggiato tra le visite di quella pagina, ma contemporaneamente, si impedisce che il motore di ricerca da cui proviene lo spider possa indicizzare il contenuto e dunque posizionare correttamente la pagina nelle SERP (le pagine dei risultati dei motori di ricerca appunto), in relazione alle keyword principali contenute su quella pagina. La normale sintassi del file `robots.txt` contempla due comandi: `User-agent`, che viene usato per identificare lo spider, e `Disallow`, che serve appunto a inibire l'accesso a determinate risorse, siano esse cartelle o file. La Tabella 1.8 riporta alcuni esempi di utilizzo.

Tabella 1.8 Esempio di utilizzo dei comandi nel file robots.txt

Comandi	Descrizione
`User-agent: *` `Disallow: /`	Il carattere asterisco nella riga "User-agent" viene usato come wildcard, intendendo qualunque spider. Il carattere `/` nel comando `Disallow` inibisce l'accesso a tutti i file del sito web.
`User-agent: googlebot` `Disallow: /`	Il nome "googlebot" nella riga "User-agent" viene usato come identificativo dello spider di Google. Il carattere `/` nel comando `Disallow` inibisce l'accesso a tutti i file del sito web, unicamente da parte di Google. Qualunque altro spider potrà invece accedere ai file del sito.
`User-agent: *` `Disallow: /cartella/prodotti.html`	Il carattere asterisco nella riga "User-agent" viene usato come wildcard, intendendo qualunque spider. Il percorso `/cartella/prodotti.html` nel comando `Disallow` inibisce l'accesso a quello specifico percorso, per tutti gli spider.
`User-agent: msnbot` `Disallow: /cartella/`	Il nome "msnbot" nella riga "User-agent" viene usato come identificativo dello spider di Microsoft. Il termine `/cartella/` nel comando `Disallow` inibisce l'accesso a quella specifica cartella, unicamente da parte di Microsoft. Qualunque altro spider proveniente da motori di ricerca diversi potrà invece accedere a quella cartella.
`User-agent: *` `Disallow: /cartella`	Il carattere asterisco nella riga "User-agent" viene usato come wildcard, intendendo qualunque spider. Il termine `/cartella` nel comando `Disallow` inibisce l'accesso a quella specifica cartella, e a qualunque file che inizi con il nome "cartella". In questo caso la differenza, rispetto all'esempio precedente, è data dall'assenza del carattere `/` dopo il nome della cartella.

`User-agent: *`
`Disallow: /forums/post-`
`Disallow: /forums/posting`
`Disallow: /forums/search`
`Disallow: /forums/updates-topic`
`Disallow: /forums/stop-updates-topic`
`Disallow: /forums/ptopic`
`Disallow: /forums/profile`
`Disallow: /forums/login`
`Disallow: /forums/privmsg`
`Disallow: /forums/memberlist`
`User-agent: Slurp`
`Disallow: /`
	In questo caso viene inibito l'accesso a qualunque spider (*) alle pagine di un forum. In fondo all'elenco delle cartelle inibite, viene istruito poi uno spider specifico (Slurp) impedendogli totalmente l'accesso al sito web.

NOTA

Se si desiderano inserire delle note all'interno del file `robots.txt`, senza che queste vengano identificate come errori di sintassi, è sufficiente iniziare la riga di testo con il carattere "#", seguito da uno spazio e dalla nota stessa. Es: `#Robots.txt creato per il sito web: www.davidevasta.biz`.

Il meta tag "robots"

Esiste un metodo alternativo all'utilizzo del file `robots.txt`, ma sicuramente molto meno efficace. Nelle pagine che non devono essere indicizzate dagli spider, si inserisce un meta tag specifico per i robot, che li istruisce su cosa fare in presenza di quella pagina. La stringa deve essere inserita tra i tag `<head>` e `</head>`, come si può osservare nel listato che segue:

Listato 1.3 Esempio di inserimento del meta tag "robots"

```
<html>
<head>
<title>Davide Vasta</title>
<meta name="keywords" content="esperto seo, adobe guru, consulente web design ">
<meta name="description" content="La home page del sito web di Davide Vasta">
<meta name="robots" content="index,follow">
</head>
<body>
Contenuto della pagina web
</body>
</html>
```

Le varianti delle istruzioni sono invece rappresentate nella Tabella 1.9.

Tabella 1.9 Sintassi del meta tag "robots"

Meta tag	Istruzione	Descrizione
`meta name="robots"`	`Content="index,follow"`	Indica allo spider di indicizzare la pagina in questione e di seguire i link inseriti in questa.
`meta name="robots"`	`Content="noindex,follow"`	Indica allo spider di non indicizzare la pagina ma di seguire i link inseriti in questa.
`meta name="robots"`	`Content="index,nofollow"`	Indica allo spider di indicizzare la pagina in questione ma di non seguire i link inseriti in questa.
`meta name="robots"`	`Content="noindex,nofollow"`	Indica allo spider di non indicizzare la pagina né tanto meno seguire i link inseriti in questa.

L'utilizzo di questa metodologia resta poco utile, soprattutto se nell'analisi statistica si comprende anche il log file. In effetti, intercettando lo spider al momento in cui legge la pagina, viene comunque già registrata una richiesta al web server, che può tranquillamente apparire come quella effettuata da un utente. Anche nel caso in cui si utilizzi il page tag il problema persiste, poiché al termine della lettura della pagina, lo script di monitoraggio verrà comunque avviato, dando luogo al conteggio di una visita, che proviene in realtà da uno spider.

SUGGERIMENTO

Questa metodologia è scarsamente utile nel caso in cui si voglia inibire l'accesso degli spider a una pagina web, poiché le istruzioni contenute nel meta tag (`index`, `noindex`, `follow`, `nofollow`) non indicano allo spider di fermarsi, bensì lo istruiscono su come gestire i contenuti (`index`) e i link (`follow`) della pagina. In effetti, questo metodo trova utile applicazione in tutti quei casi in cui si vuole rendere pubblica una pagina per gli utenti, ma non si vuole che questa venga indicizzata dai motori di ricerca, come per esempio un modulo per l'inserimento dei dati utente. Oltretutto, usando il tag "nofollow", si fa in modo che i link contenuti nella pagina siano navigabili da parte degli utenti, ma non lo siano da parte dei motori di ricerca.