

## **Il controllo degli accessi è di fondamentale importanza per comprendere il successo di un sito Internet.**

La determinazione del successo o del fallimento di un prodotto diffuso tramite il Web, sia esso un sito che offre informazioni, un portale o una soluzione di commercio elettronico è un'impresa assai ardua.

Come è noto, a differenza di un dato di vendita quale può essere il numero di copie vendute di un prodotto, sia esso tradizionale oppure hardware o software, la valutazione numerica oggettiva del "gradimento" di un sito Internet è davvero difficilmente misurabile.

La ragione di questa difficoltà risiede in almeno due aspetti: entrano in gioco molti parametri e diventa assai complicato e soggettivo valutarli e confrontarli con i dati offerti dalla concorrenza. Per meglio dire la rilevazione dei dati è tecnicamente realizzabile, ma l'interpretazione è generalmente opinabile e discutibile, dal momento che i valori possono essere letti con punti di vista diametralmente opposti. Ne deriva che un confronto è spesso assolutamente improponibile.

### **Lo standard di registrazione delle informazioni di log**

In termini strettamente tecnici ogni server Web è in grado di tracciare un log abbastanza preciso degli accessi alle singole pagine. Il protocollo che consente la navigazione e la comunicazione tramite il browser, l'HTTP, permette la memorizzazione dell'indirizzo IP del navigatore dal quale proviene la richiesta.

Sin dagli albori di Internet era stato definito uno standard per la memorizzazione di tali informazioni; i primi server Web accessibili dal pubblico, Ncsa e Cern, adottavano già tale standard e tutti i successivi prodotti della categoria, sia commerciali che di dominio pubblico hanno continuato ad adottare tale standard offrendo file di log compatibili. La proliferazione e la diffusione di moltissimi strumenti di analisi dei file di log è stata tale che oggi sono disponibili un'infinità di prodotti in grado di esaminare e riassumere le statistiche degli accessi utenti alle pagine di un sito.

La fonte più nuova di raccolta di statistiche di rete è rappresentata dai *log files* o *transaction logs*.

Come detto, i *log files* contengono la registrazione delle interazioni che intervengono nel momento in cui un utente accede ad una risorsa in rete, ad esempio una pagina di un sito. L'analisi dei *log files* consente l'ottenimento di informazioni utili sul traffico registrato sul sito, sulle caratteristiche demografiche dell'utente, cioè dall'area geografica da cui avviene l'interazione, e sul comportamento nell'utilizzo delle risorse, sul tempo di permanenza in una pagina o sul percorso che viene seguito per navigare all'interno di un sito.

I software che analizzano i *log files*, però, se da un lato sono estremamente efficaci nel registrare il traffico, gli accessi, il numero di computer connessi, mostrano evidenti limiti nel momento in cui cercano di offrire un quadro completo del comportamento dell'utente.

Per fare un esempio con quanto poco sopra descritto, la valutazione del tempo di permanenza su una pagina non esprime la motivazione per la quale l'utente ha temporeggiato sulla pagina stessa. Chi può dire con certezza se si è attardato nella lettura del testo, se ha letto con rapidità e si è quindi dedicato alla scrittura di appunti e considerazioni personali, oppure se ha proprio abbandonato la postazione perché richiamato da altra attività più urgente.

I *log files* sono comunque un dato da cui è importante partire per successive considerazioni. I dati che un *log file* può mettere a disposizione sono i seguenti:

- il numero di accessi e delle visite; gli accessi si identificano con i download effettuati su una pagina html (ogni volta che si scarica una pagina, si scaricano tutte le immagini contenute in quella pagina), le visite identificano invece il numero di download di un'intera pagina;
- il numero di byte trasferiti;
- quali risorse sono state ricercate;
- da quali workstation;
- quando;
- i numeri di indirizzi IP da cui sono state effettuate le connessioni;
- il tipo di browser con cui sono state effettuate le connessioni.

### **Cosa leggere e interpretare dai log-files**

Occorre capire ora in quale modo si possano interpretare al meglio i dati provenienti dai log-files. Cominciamo con il dire che il concetto di "numero di accessi" ha una validità relativa. Il server per il Web traccia la richiesta di un client per una pagina HTML, per una immagine o per altri oggetti multimediali, ma spesso il rilascio delle pagine non proviene più o non proviene in assoluto dallo stesso server.

Infatti tecniche di ottimizzazione sviluppate per migliorare l'occupazione della banda e velocizzare le operazioni di navigazione fanno in modo che, dopo un primo effettivo accesso al sito, le successive consultazioni del browser provengono dalla memoria stessa dell'utente qualora non vi siano stati aggiornamenti sul server; in teoria è possibile che un navigatore visiti più volte e per più giorni lo stesso sito senza che il server Web ne sia a conoscenza.

Se è vero che l'interesse di un utente è stato tracciato almeno una volta, viene perso il conteggio effettivo di visite al sito: l'utilizzo di server proxy, ai quali accedono una grandissima parte di navigatori, maschera la consultazione di un sito e si può in teoria arrivare alla paradossale situazione in cui una certa pagina venga visitata da un numero elevatissimo di utenti senza che sul server venga tracciato un solo accesso. Tecnicamente questo può essere evitato ed il server Web può venire informato degli accessi alla memoria cache del proxy, ma ciò comporta un intervento volontario da parte degli amministratori di sistema su cui è lecito porsi qualche dubbio.

La presenza di macchine quali proxy e firewall, il server che per garantire sicurezza filtra accessi in ingresso e in uscita sui siti, crea anche l'inconveniente di mascherare il client reale. In moltissimi casi i log del Web riportano non l'indirizzo IP vero, ma quello della porta di ingresso alla rete, spesso condiviso da altri utenti dello stesso provider. In questo caso si ha la certezza del dominio di provenienza, derivando così un'informazione utile a livello di aggregazione, ma non valida dal punto di vista analitico. Per questa ragione i numeri relativi agli accessi a un sito Web sono quantomeno discutibili e di conseguenza assai poco comparabili.

Ad esempio se dal tracciamento degli accessi risultasse che una pagina viene consultata dal proxy server di una grande multinazionale, si potrebbe sostenere che il numero di accessi reali alla pagina corrisponde al numero totale di computer dotati di browser appartenenti all'azienda. E' evidente che il ragionamento appare piuttosto forzato ma non lo si può ritenere falso in assoluto.

**Tabella 1 :**

Hit	La singola richiesta a cui risponde il server per generare la singola pagina web. Ogni componente della pagina, immagine, filmato, etc..., rappresenta una hit differente. La pura somma delle hit può fornire un dato sovrastimato rispetto al numero reale dei visitatori.
Page View	Una singola schermata visualizzata sul browser dell'utente. La pagina include tutti i componenti testuali e multimediali contenuti all'interno, quindi il conteggio fornisce un valore più realistico dei visitatori.
Impression	Il numero di richiami pubblicitari, ads, (ad esempio banner) contenuti in una pagina web. Se una pagina contiene 2 ads, vengono calcolate 2 impression. Sebbene sia impossibile dire se l'utente legge veramente la pubblicità, il solo aprire la pagina scatena il calcolo delle impression.
Visit	Una sessione in un sito Web comincia quando l'utente accede alla prima pagina e termina quando l'utente abbandona il sito. Misurare il numero di visite può essere complicato: un utente potrebbe accedere ad un sito, lasciare il browser sulla medesima pagina per un'ora a causa di un impegno, quindi tornare a navigare sullo stesso sito: questa sessione costituisce una visita e due? Molti sosterrrebbero che si tratta di una visita, ma secondo l'Internet Advertising Bureau (IAB) questa verrebbe contata come se fossero due in quanto il periodo di inattività è superiore ai 30 minuti.
Visitor	Un individuo che visita il sito Web. Questa è una delle statistiche più complicate perché sino ad oggi non vi è alcuna evidenza fisica da parte del server che l'utente che sta navigando sia proprio quello che ha inserito ID e password di accesso. Il server viene a conoscenza dell'IP, l'indirizzo numerico per la navigazione in rete.
Cookie	E' un'informazione immagazzinata sul computer del navigatore e che aiuta a riconoscere che un utente è tornato a visitare lo stesso sito. L'utente, adottando una opportuna impostazione, può impedire l'utilizzo dei cookie e perciò evitare che venga violata la privacy durante la sua navigazione.

## **Analisi traffico Web con Data mining**

Diversi prodotti di analisi dei log-files presenti sul mercato hanno cercato di sofisticare la loro funzionalità per cercare di fornire informazioni più complete sui visitatori e sui loro modelli di comportamento.

Lo sforzo fornisce un buon risultato quando si riescono ad abbinare informazioni quantitative, immediatamente ricavabili da log, a quelle qualitative.

Il passo finale è la fusione e l'analisi di tutti i dati attraverso algoritmi di data-mining. Tale tecnica permette di derivare e prevedere risultati sulla base di dati raccolti dall'esperienza. Quando i visitatori interagiscono con un sito forniscono informazioni su se stessi e su come rispondono ai contenuti presentati. Ad esempio forniscono nomi, indirizzi, utilizzano termini quando cercano informazioni, danno giudizi attraverso sondaggi. Questo insieme di informazioni, naturalmente desumibili non solo dai log-files, possono essere strutturate in un database al punto da ottenere un mole di dati tale da alimentare gli algoritmi di data-mining.

I sistemi di data-mining sono chiamati con termine tecnico online analytical processing (OLAP). I sistemi OLAP generano report sulla base di informazioni osservate direttamente e messe in relazione in modo semplice, ottenendo così modelli riutilizzabili.

Per fare un esempio, i sistemi OLAP non spiegano la ragione per la quale un navigatore che visita un portale di cinema sia poi attratto da un banner che gli consente di cercare l'anima gemella. In compenso i sistemi OLAP evidenziano questi fatti sottolineandone le ricorrenze, anche quelle più nascoste.

Il buon funzionamento di un algoritmo di data-mining ha luogo solo attraverso la conoscenza e la memorizzazione delle caratteristiche dei visitatori. Queste comprendono informazioni demografiche, psicologiche e tecnologiche.

Le informazioni demografiche sono attributi reali quali l'indirizzo, il reddito, la responsabilità di acquisto o la proprietà di apparecchiature per il tempo libero.

Le informazioni psicologiche sono i profili psicologici che possono essere rilevati in un'analisi, come il fatto di essere eccessivamente protettivi nei confronti dei figli, la tendenza ad effettuare acquisti impulsivamente, l'interesse per la tecnologia e così via.

Le informazioni tecnologiche si riferiscono al sistema del visitatore, ad esempio, il sistema operativo, il browser, il dominio e la velocità del modem utilizzato. Se si dispone di un numero telefonico o di un indirizzo spesso è possibile ottenere informazioni di tipo demografico o psicologico tramite provider di servizi di marketing diretto.

Per un sito di commercio elettronico è importante abbinare la caratteristiche del visitatore con il prodotto che viene acquistato. In questo caso le informazioni sugli articoli comprendono il tipo, la categoria, la dimensione, il prezzo, il margine, la disponibilità. Le interazioni tra il visitatore e l'articolo acquistato (o semplicemente consultato) comprendono la cronologia dell'acquisto o dell'interessamento, la cronologia della pubblicità e le informazioni sulle referenze. La cronologia dell'acquisto è costituita da un elenco di prodotti e dalle date di accesso ad essi. Mentre la cronologia della pubblicità indica quali articoli sono stati presentati a un visitatore.

Le informazioni sul percorso dei clic è una cronologia dei collegamenti ipertestuali selezionati da un visitatore. Le opportunità di collegamento sono collegamenti ipertestuali che sono stati presentati a un visitatore.

Le statistiche visitatore-sito sono generalmente tipiche di una sessione, ad esempio il tempo totale, le pagine visualizzate, il ricavo e il guadagno per sessione con un visitatore. Le informazioni visitatore-società possono comprendere il numero totale di clienti presentati da un visitatore, il guadagno totale, il numero totale di pagine visitate, il numero di visite al mese, la data dell'ultima visita e così via. Le informazioni visitatore-società possono comprendere valutazioni di marche. Le associazioni con le marche, ad esempio, sono elenchi di concetti positivi o negativi che un visitatore associa alla marca stessa, che possono essere misurati considerando periodicamente i visitatori.

## **Gli obiettivi del data-mining**

Rispetto ad aree commerciali di vecchio stampo o al direct mailing, il marketing su Internet gode del grande vantaggio di riuscire a misurare le interazioni dei visitatori in modo decisamente più efficace. Quando si vogliono applicare tecniche di data-mining alla mole di dati raccolta è necessario avere molto chiari gli obiettivi. Fra gli altri, possono essere presi in considerazione questi obiettivi:

- aumentare il numero medio di pagine visualizzate per ogni sessione,
- aumentare il guadagno medio per visita,
- ridurre la quantità di prodotti restituiti,
- aumentare il numero dei clienti,

- aumentare la conoscenza del prodotto,
- aumentare il numero di visitatori che sono ritornati entro un lasso di tempo (velocità di ritorno),
- ridurre il numero di clic per uscire (numero medio di pagine visualizzate per concludere un acquisto o per ottenere le informazioni desiderate),
- aumentare la velocità di conversione (risultati per visita).

Inoltre, se il sito ha particolare struttura per raccogliere le caratteristiche anagrafiche dei visitatori e il tipo di interazione, ci si trova in una condizione qualitativamente ancora migliore. Non resta che sfruttare ed estrarre le informazioni contenute nei dati raccolti.

### **Comprensione a approccio alla raccolta di informazioni per il data mining**

La raccolta dei dati per un corretto approccio ad una analisi dei visitatori di un sito può essere realizzata analizzando attentamente il problema che si intende risolvere. I problemi più comuni che devono essere risolti dal marketing riguardano come indirizzare gli annunci pubblicitari, personalizzare le pagine Web, creare pagine Web che mostrino prodotti comunemente acquistati contemporaneamente, classificare automaticamente gli articoli, suddividere in gruppi di visitatori simili, calcolare i dati mancanti e prevedere i comportamenti futuri. Tutte queste attività comportano la scoperta e lo sfruttamento di diversi tipi di modelli nascosti.

### TARGETING

Gli esperti di marketing utilizzano questa tecnica per selezionare le persone che ricevono un annuncio pubblicitario fisso. Il risultato della selezione consente di ottenere una categoria selezionata di visitatori grazie ai quali si addebitano cifre superiori per la vendita di spazi pubblicitari. Sui siti in cui i visitatori si registrano, è possibile indirizzare la pubblicità in base alle informazioni demografiche. Persone che vivono in parti diverse di un paese o che visitano diversi siti Web possono avere diverse propensioni agli acquisti di abbigliamento sportivo, di viaggi o di accessori per auto. Altri siti consentono di destinare gli annunci pubblicitari in base all'indirizzo IP, in base cioè all'ubicazione fisica dell'Internet Service Provider, ma tale metodo non è sufficientemente affidabile.

Il data mining consente la selezione dei criteri di indirizzamento di una campagna pubblicitaria. Una tecnica consiste nel lanciare un annuncio pubblicitario di test senza alcun target specifico. Il risultato del lancio è l'ottenimento di una "conversione", quindi di una serie di variabili demografiche di quegli utenti che hanno reagito con un clic, con una registrazione o con un acquisto, fornendo così un profitto. Il data mining aiuta nell'identificazione della combinazione di criteri che danno un profitto.

## PERSONALIZZAZIONE

La personalizzazione è un metodo per selezionare gli annunci pubblicitari da inviare ad una determinata persona. Si può dire che la personalizzazione è il contrario del targeting: quest'ultimo ottimizza i tipi di persone che vedranno una certa pubblicità. Di contro, la personalizzazione ottimizza gli annunci che una persona riceve, incrementando i ricavi poiché la persona vede più articoli di possibile interesse.

Alcuni sistemi di personalizzazione si basano su regole scritte dall'esperto di marketing per personalizzare gli annunci verso i visitatori. Questi sono sistemi di personalizzazione basati su regole. Se si dispone di informazioni storiche, è possibile acquistare strumenti di data mining di terze parti per generare le regole.

I sistemi di personalizzazione basati su regole vengono generalmente utilizzati in situazioni in cui sono presenti limitate quantità di prodotti o di servizi offerti, ad esempio compagnie di assicurazioni e società finanziarie, in cui i responsabili del marketing possono scrivere un piccolo numero di regole.

## ASSOCIAZIONE

La tecnica di marketing detta "associazione" identifica gli articoli che verranno probabilmente acquistati o visti nella stessa sessione. Se si inseriscono dei riferimenti a tali articoli sulla stessa pagina in un catalogo Web, si potrebbe indurre il visitatore ad acquistare o visualizzare qualche altro prodotto dimenticato.

Se si definisce una promozione su un prodotto in un gruppo omogeneo, si potranno probabilmente incrementare gli acquisti di altri prodotti dello stesso gruppo.

L'associazione è quella soluzione di data mining utilizzata da molti siti commerciali quali Amazon.com.

### **Gli algoritmi utilizzati**

Per il Web, gli algoritmi di data mining utilizzati ricadono in categorie utilizzate per altri settori. L'algoritmo a "reti neurali", le quali quando viene presentato un modello sono in grado di derivarne delle regole comportamentali, richiede un elevato grado di addestramento. Inoltre per comprendere la previsione di un comportamento di singoli navigatori devono essere in qualche misura personalizzate.

Un altro metodo di analisi è il "clustering", talvolta detto segmentazione, che identifica le persone raggruppandole al di sotto di caratteristiche comuni derivando una media di tali caratteristiche. Il clustering viene utilizzato direttamente da alcuni fornitori per generare report su caratteristiche generali di diversi gruppi di visitatori. Anche questa tecnica richiede un addestramento e non sono particolarmente adatte per siti Web con pagine dinamiche.

L'algoritmo "stima e previsione" tenta di indovinare un valore sconosciuto, ad esempio il reddito, quando si conoscono altre informazioni su una persona. In particolare la previsione cerca di indovinare un valore futuro, ad esempio la probabilità di acquistare un'automobile entro l'anno successivo, quando una persona non l'ha ancora fatto oppure il numero previsto di azioni che verranno acquistate da una persona nel corso dell'anno successivo. Gli stessi algoritmi possono generare una

stima e una previsione. La stima viene spesso utilizzata nelle analisi demografiche per calcolare valori mancanti.

Gli esperti di marketing utilizzano il metodo dell' "aggregazione" delle informazioni per comprendere il comportamento di gruppi di clienti. Anche l'aggregazione o la valutazione di eventi passati in base a diverse dimensioni, ad esempio la categoria dei visitatori, la categoria del contenuto, il referente e la data, possono fornire utili informazioni.

Questa semplice aggregazione viene chiamata OLAP; online analytical processing: online perché l'esperto di marketing utilizza un motore di reportistica per muoversi in maniera interattiva all'interno della base dei dati; analytical perché l'esperto di marketing effettua analisi in modo passivo sui dati del passato, senza modificarli.

Infine il metodo ad "albero decisionale" è costituito da un flusso di domande o punti di dati che in definitiva riportano a una decisione. I sistemi ad alberi decisionali cercano di creare percorsi ottimizzati, ordinando le domande in modo da poter prendere una decisione nel minor numero di passi.

## **Conclusioni**

Le tecniche di data mining non sempre valgono anche per il Web. Non esistono soluzioni pronte all'uso e quelle che esistono sconfinano nella mera statistica. Occorre prestare attenzione ad adottare per il Web tecniche di data mining applicate solitamente in aree tradizionali.

Infatti in rete tutto funziona ad una velocità superiore e si hanno a disposizione maggiori quantità di dati. Gli esperti più tradizionalisti sono abituati a operare in un mondo meno stressante, in cui il data mining viene utilizzato una volta al mese, invece che, almeno in teoria, una volta ad ogni clic del mouse. In più sul Web l'ammontare dei dati viene misurato in gigabyte al mese, un ordine di grandezza superiore rispetto alle aree più tradizionali.

In generale gli algoritmi di data mining per la rete devono essere appropriati per l'attività che si intende portare a termine. Memorizzando i dati associati ai visitatori, ai contenuti e all'interazione per lo meno si è certi di poter utilizzare tali informazioni in un momento successivo. Tuttavia, prima si inizia a imparare dai dati e prima si potranno distanziare i proprio concorrenti.