Web-Usage Mining e Clickstream:

la profiliazione degli utenti nel web

Alberto De Bortoli

giugno 2009

Abstract

Questo paper analizza il Web per quanto riguarda il Web-Usage Mining e affronta le tecniche di profiliazione degli utenti che visitano siti web. Il primo capitolo introdurrà l'ambiente oggetto di studio, ovvero il Web; il secondo affronterà le varie sfaccettature del Data Mining in relazione al Web; il terzo farà luce sulla branca del Web Mining chiamata Web-Usage Mining; il quarto introdurrà il Data Webhouse. Sarà fatta infine una dissertazione sul problema della privacy legato alla profiliazione dell'utente.

Contents

1 Il Web

1.1 Il web per la raccolta di informazioni

Il web è uno straordinario mezzo per la raccolta di informazioni. Innanzitutto è possibile, grazie al web, acquisire dei dati di tipo tradizionale come i dati anagrafici. All'interno di numerosi siti, infatti, sono disponibili delle form di registrazione, in cui vanno inseriti dati personali quali, appunto, informazioni relative al nostro stato civile, al luogo di residenza, indirizzo di posta elettronica e così via.

Ovviamente la compilazione di questi form comporta, per tutti gli utenti, un costo, sia in termini di energia che di tempo. Per questo in cambio di tali registrazioni si ottiene il diritto di accedere a particolari servizi, che possono essere mailing list, sezioni del sito con contenuti riservati, ed altro ancora. Le informazioni raccolte in questo modo sono volontariamente rilasciate dall'utente che decide se e per quali siti desidera dichiarare informazioni private.

Il web consente, per le caratteristiche proprie dello strumento, di rilevare delle informazioni in alcuni casi addirittura indipendentemente dall'utente, in altri con tacito consenso: fa parte del diritto di utilizzare questo strumento il rinunciare ad un po' di privacy.

Vediamo alcune delle informazioni che è possibile ricavare sul navigatore.

- la località ove si trova il provider;
- la lingua che l'utente è in grado di comprendere;
- l'ultima pagina cui l'utente ha avuto accesso;
- l'insieme dei dati identificativi prelevati dal cookie: il riconoscimento di un utente che ha già visitato il sito, il suo comportamento di navigazione all'interno del sito.

  Questi dati sono importanti per definire le caratteristiche del cliente del sito web e offrono molte possibilità per il marketing.

1.2 I dati provenienti dal Web

La potenzialità del Web Mining sta nell'applicazione di algoritmi di Data Mining ai dati relativi all'accesso degli utenti sui server Web. Il vero valore di questa conoscenza è ottenuto però tramite l'integrazione di questi dati con altre informazioni – anche esterne e più convenzionali – su clienti, vendite e prodotti. Tutte queste interazioni tra fornitori di servizi digitali e consumatori possono essere registrate e immagazzinate in database digitali chiamati **Data Webhouse**.

Attraverso il Web Mining le società possono analizzare questa immensa fonte di dati e prevedere il comportamento dei loro utenti, al fine di soddisfare i loro bisogni e distribuire servizi migliori e personalizzati.

Dal punto di vista tipologico, possiamo suddividere i dati nelle seguenti categorie:

- **Contenuto**: i veri dati delle pagine Web. Questi in genere consistono in testi e immagini.
- **Struttura**: i dati che descrivono l'organizzazione del contenuto. Informazioni sulla struttura della pagina includono la sistemazione dei vari tag HTML o XML all'interno di una pagina determinata.
- **Comportamento**: i dati che descrivono il modello di comportamento delle pagine Web, come gli indirizzi IP, gli URL e la data e l'ora degli accessi.
- **Profilo Utente**: i dati che offrono informazioni demografiche sugli utenti del sito web. Questi includono i dati di registrazione e le informazioni sul profilo dei clienti.

Analizzare tali dati può aiutare le organizzazioni a determinare, fra le altre cose, la durata di vita dei clienti e il valore delle strategie di marketing sui prodotti. Può servire a valutare l'efficacia di campagne promozionali. Può offrire anche informazioni su come ristrutturare un sito Web. Analizzare informazioni sull'accesso degli utenti aiuta anche nel designare specifici gruppi di utenti su cui effettuare campagne di marketing mirate.

1.2.1 Le fonti dei dati

I dati raccolti dalle diverse fonti Web aiutano a comprendere i comportamenti di navigazione di uno o più utenti a singoli o multipli siti Web. Possono essere ricavati lato server, lato client o da server proxy.

Dati raccolti a livello server

Un **Web server log** è un'importante fonte per eseguire il Web Mining perché registra automaticamente tutte le richieste fatte dagli utenti attraverso i loro browser e tutte le risposte dei server Web. I dati registrati nei server log riflettono quindi l'accesso ad un sito Web da parte di utenti multipli. Questi file log possono essere immagazzinati in vari formati, come i common log e gli extended log.

Nonostante la loro importanza, i dati sul comportamento nel sito registrati nei server Web non possono essere completamente affidabili a causa della presenza di vari livelli di caching nell'ambiente Web: le pagine memorizzate nella cache, quando vengono visitate, non vengono registrate nei log. Inoltre, alcune importanti informazioni passate attraverso il metodo POST non saranno disponibili nel server log.

La tecnologia **packet sniffing** è un metodo alternativo per il raccoglimento

dei dati di comportamento attraverso i server log: il packet sniffing monitora il traffico di rete che arriva ad un server Web ed estrae i dati sul comportamento direttamente da pacchetti TCP/IP.

I server Web possono anche immagazzinare altri tipi di informazioni sul comportamento, come i **cookie**, ovvero dei metodi di tracciamento generati dal server Web, per i browser dei singoli client, allo scopo di individuare automaticamente i visitatori del sito. Individuare i singoli utenti non è un compito facile, a causa del modello di connessione senza stato del protocollo HTTP. I cookie contano sulla cooperazione implicita degli utenti, per questo hanno elevato preoccupazioni crescenti sulla privacy.

Altre fonti importanti sono i **referrer log**, che contengono informazioni sulle frequenze di visita delle pagine, e le informazioni immesse direttamente dagli utenti tramite i **moduli di registrazione**.

Dati raccolti a livello client

Le raccolte di dati lato client possono essere implementate usando un agente remoto (come JavaScript o applet Java) o cambiando il codice sorgente di un browser esistente per migliorarne le capacità di raccolta. La realizzazione di metodi di raccolta di dati lato client richiede la cooperazione dell'utente, o nell'abilitare il funzionamento di JavaScript e applet Java o nell'usare volontariamente il browser modificato.

Rispetto ai dati raccolti lato server, quelli lato client hanno il vantaggio di diminuire i problemi di caching e di identificazione della sessione. Comunque, le **applet Java** possono sopravvalutare il tempo di visita ad una pagina, specialmente quando vengono caricate per la prima volta. **JavaScript** impiega poco tempo nell'interpretazione di una pagina ma, d'altra parte, non può catturare ogni click dell'utente (come i bottoni "aggiorna" o "indietro"). Questi metodi raccoglieranno solamente il comportamento di visita di singoli utenti a singoli siti.

Un **browser modificato** è molto più versatile e permetterà raccolte di dati su un singolo utente a siti Web multipli. La parte più difficile nell'utilizzo di questo metodo sta nel convincere gli utenti ad usare il browser. Questo può essere fatto proponendogli incentivi.

Dati raccolti a livello proxy

Un **Web proxy** si comporta come un livello di caching intermedio tra il browser del client e il server Web. Il caching del proxy può essere usato per ridurre il tempo di caricamento di una pagina Web richiesta dagli utenti così come il peso del traffico di rete al server.

Le performance delle cache dei proxy dipendono dalla loro abilità di prevedere

correttamente le future pagine richieste. Le tracce proxy possono rivelare le attuali richieste HTTP da clienti multipli a server Web multipli. Questo può servire come sorgente dati per caratterizzare il comportamento di navigazione di un gruppo di utenti anonimi che dividono un server proxy comune.

1.2.2 Astrazione dei dati

Le informazioni fornite dalle fonti di dati descritte sopra possono tutte essere usate per costruire/identificare molte astrazioni di dati, in particolare *utenti*, *sessioni server*, *episodi*, *flussi di click* e *pagine intere*.

Per offrire della consistenza al modo in cui questi termini sono definiti, la W3C Web Characterization Activity (WCA) ha dato delle definizioni per i termini Web attinenti all'analisi del comportamento nel Web.

Un **utente** è definito come un singolo individuo che accede a file di uno o più server Web attraverso un browser. Sebbene questa definizione sembri banale, in pratica è molto difficile identificare gli utenti unicamente e ripetutamente. Un utente può accedere al Web attraverso macchine diverse, o usare più di un browser su una singola macchina.

Una **pagina intera** è formata da ogni file che contribuisce a mostrare ciò che appare sul browser di un utente in un dato momento. Le pagine intere di solito sono associate con una singola azione dell'utente (come un click del mouse) e possono essere formate da molti file, come frame, grafici e script. Quando si analizza il comportamento degli utenti, la vera importanza è data dalla pagina intera globale: l'utente non chiede di caricare nel suo browser esplicitamente "n" frame e "m" grafici, l'utente richiede una "pagina Web". Tutte le informazioni per determinare quali file costituiscono una pagina intera sono accessibili dal server Web.

Un **flusso di click** è una serie sequenziale di richieste di pagine intere. Di nuovo, i dati disponibili dal lato server non sempre forniscono abbastanza informazioni per ricostruire il flusso di click completo per un sito. Alcune pagine intere visitate attraverso un client o cache di livello-proxy non saranno "visibili" dal lato server.

Una **sessione utente** è il flusso di click di pagine intere per un singolo utente attraverso l'intero Web. Da quando le informazioni sugli accessi non sono pubblicamente disponibili dalla vasta maggioranza dei server Web, solamente la porzione di ogni sessione utente che accede ad uno specifico sito può essere utilizzata per l'analisi.

L'insieme delle pagine di un sito Web viste in una sessione utente è interpretato come una sessione server (o, comunemente, visita). Un insieme di sessioni

server è l'input necessario per qualsiasi analisi sul comportamento nel Web o strumento di Data Mining. La fine di una sessione server è definita come il punto in cui la sessione di navigazione dell'utente su quel sito è terminata. Di nuovo, questo è un semplice concetto che è molto difficile individuare in modo affidabile.

Ogni sottoinsieme semanticamente significativo di sessione utente o sessione server è interpretato come un **episodio** dal W3C WCA.

1.3 Il Database

Un database è spesso realizzato per scopi diversi dal data mining e quindi alcune informazioni importanti possono non essere presenti. Un altro problema è la presenza di dati non corretti, contenenti errori di vario genere riguardanti la misurazione del fenomeno o l'errata classificazione di alcune unità.

Ottenere un valido database è la prima fondamentale operazione da compiere al fine di ottenere informazioni utili nell'attività di business intelligence. Vi sono tre tipi di database: il data warehouse, il data webhouse e il data mart.

1.3.1 Il Data Warehouse

Immon definisce il data warehouse come "una raccolta di dati, orientata al soggetto, integrata, non volatile e variabile nel tempo e volta a supportare le decisioni del management".

- *orientata al soggetto*: in un data warehouse le aree aziendali vengonosuddivise per soggetto piuttosto che per settore;
- *integrata*: il data warehouse deve essere in grado di integrarsi con tutti gli standard utilizzati dalle varie applicazioni che raccolgono i dati; esso deve ricodificare questi standard in modo univoco prima di immagazzinare i dati;
- *non volatile*: l'aggiornamento dei dati, con la perdita di informazioni che ne consegue, non viene effettuato all'interno del data warehouse; in questo modo vengono aggiunte informazioni senza modificare quelle precedenti;
- *variabile nel tempo*: l'orizzonte temporale di un data warehouse è di cinque-dieci anni.

Il data warehouse è un repository, all'interno del quale devono essere raccolti i dati utili per effettuare decisioni di management.

Esistono due tipi di approccio alla creazione di un data warehouse: la creazione di un archivio centralizzato, che raccoglie tutte le informazioni aziendali e le integra con quelle provenienti dall'esterno, e un approccio che crea il data warehouse dall'unione di diversi database tematici (data mart).

Se il primo approccio permette un'alta qualità dei dati, dovuta al controllo da parte degli amministratori, il secondo è relativamente più facile nella fase di

implementazione iniziale. Richiede comunque un notevole sforzo di definizione e pulizia dei dati per ottenere un sufficiente livello di uniformità.

1.3.2 Il Data Webhouse

Il web è un'immensa fonte di dati. Per poter utilizzare questi dati è necessario convogliarli in un particolare tipo di database, il data webhouse, che si distingue dal data warehouse perché deve tener conto delle caratteristiche del web.

Per quanto riguarda la necessità di fornire informazioni in tempo reale, per esempio, il data warehouse risulta avere dei tempi di risposta insufficienti. Anche la raggiungibilità da ogni parte del mondo impone al data webhouse di essere reperibile velocemente e senza interruzioni di disponibilità.

La costruzione di un sistema di Data Warehouse deve ora tenere presente anche gli aspetti nuovi legati al Web. Infatti i dati riguardanti flussi di "click" hanno la capacità di fornire in maniera molto dettagliata informazioni su qualsiasi gesto compiuto da ogni individuo durante la navigazione in Internet.

Questa immensa fonte di dati può essere convogliata all'interno del Data Webhouse per essere analizzata ed eventualmente conformata e combinata con le già esistenti e più convenzionali fonti di dati. Questo tipo di database verrà meglio illustrato nella sezione 4 di questo paper.

1.3.3 Il Data Mart

Data Mart è l'abbreviazione di "database marketing" e può essere considerato un archivio aziendale, contenente tutte le informazioni relative alla clientela acquisita e potenziale.

Solitamente derivato da altre strutture di dati: da un data warehouse è possibile estrarre molteplici data mart, uno per ogni tipo di analisi che si vuol effettuare, anche se è possibile, ma più complesso, costituire data mart in assenza di un sistema integrato di data warehousing.

La costruzione di data mart costituisce il passo fondamentale per qualsiasi operazione di business intelligence.

1.4 Query e reporting

Gli strumenti di query e reporting sono veloci e facili da usare. Permettono di esplorare dati aziendali a vari livelli, recuperando le specifiche informazioni richieste (strumenti di query) e presentandole in modo chiaro e comprensibile (reporting).

Attraverso il loro uso è possibile la costruzione di tabelle e grafici riepilogativi dei dati a due o tre dimensioni, sulla base di interrogazioni normalmente effettuate attraverso un'interfaccia grafica che consente di selezionare le variabili di interesse.

Il reporting viene normalmente utilizzato per la realizzazione di cruscotti per l'alta direzione che evidenziano indicatori grafici dell'andamento di un business (grafici di decomposizione delle vendite sul territorio, alert sul superamento

di determinate soglie, eccetera). Sono analisi molto semplificate che servono soprattutto se integrati con altri tipi di analisi più sofisticate.

1.5 OLAP

OLAP è la sigla di "On-Line Analytical Processing". Un'applicazione OLAP è uno strumento di reportistica multidimensionale: spesso di tipo grafico, permette di visualizzare le relazioni tra le variabili a disposizione seguendo la logica di analisi di un report a due dimensioni.

Utilizzando l'OLAP l'utente forma delle ipotesi sulle possibili relazioni esistenti tra le variabili e cerca delle conferme osservando i dati. Per esempio, l'analista potrebbe ipotizzare inizialmente che le persone con bassi redditi ed un elevato livello di debiti siano persone ad alto rischio di mancato rimborso di un prestito.

Al fine di verificare tale assunzione l'OLAP fornisce una rappresentazione grafica (detta ipercubo dimensionale) della relazione empirica tra le variabili reddito, debito e insolvenza. L'esame del grafico può fornire indicazioni sulla validità dell'ipotesi effettuata.

L'OLAP quindi permette di estrarre informazioni utili dai database aziendali; diversamente dal data mining, tuttavia, le ipotesi di ricerca vengono suggerite dall'utente, e non scoperte nei dati. In definitiva, l'OLAP non è un sostituto del data mining, ma anzi le due tecniche di analisi sono complementari e il loro impiego congiunto può produrre utili sinergie.

L'OLAP può essere impiegato nelle fasi preliminari del data mining (pre- processing) agevolando la comprensione dei dati: ad esempio permettendo di focalizzare l'attenzione sulle variabili più importanti, identificando i casi particolari o trovando le interazioni principali. D'altra parte, i risultati finali dell'attività di data mining, riassunti da opportune variabili di sintesi, possono a loro volta essere convenientemente rappresentati da un ipercubo di tipo OLAP, che permette una comoda visualizzazione.

I sistemi OLAP sono in grado di aggregare i dati a disposizione rispetto a più prospettive (dimensioni) e più livelli di dettaglio contemporaneamente, attraverso l'utilizzo di una semplice interfaccia grafica di tipo drag-and- drop. L'andamento di un business può essere esplorato da varie angolazioni fino a trovare quella più adatta ai propri scopi (di sintesi o di individuazione di fenomeni anomali).

Un sistema OLAP necessita che i dati siano organizzati con modelli multidimensionali detti ipercubi, che rappresentano una concettualizzazione più vicina al modo in cui il manager percepisce la realtà aziendale di quanto non siano le tabelle di un database relazionale. In un modello multidimensionale la dimensione organizza i dati secondo un particolare punto di vista, eventualmente specificabile a diversi livelli di dettaglio secondo una struttura definita gerarchia.

Esempi di dimensioni sono il prodotto, il tempo, l'area geografica, mentre una gerarchia tipica è quella relativa alla dimensione tempo, contenente gli elementi anno, trimestre, mese, settimana e giorno. Le dimensioni costituiscono gli n lati di una struttura ipercubica a n dimensioni.

Il contenuto delle celle dell'ipercubo, individuate dai valori delle dimensioni, sono le grandezze che si intendono analizzare, ad esempio il volume delle vendite o il numero di clienti acquisiti eccetera.

2 Data Mining

2.1 Definizione di Data Mining

Il Data Mining si sta rapidamente diffondendo in molte organizzazioni aziendali come importante strumento per il supporto alle decisioni.

I sistemi di supporto alle decisioni (**DSS**, Decision Support System) sono sistemi sviluppati per fornire le informazioni necessarie per decidere nel modo migliore possibile riguardo a decisioni inerenti la gestione dell'organizzazione. Vengono identificati anche con il termine di BI (Business Intelligence) per sottolinearne l'indipendenza ed autonomia rispetto ai sistemi informativi tradizionali.

In particolare, il Data Mining è una parte del più ampio processo di scoperta di conoscenza a partire dai dati contenuti nei database, conosciuto con il nome di KDD (Knowledge Discovery in Databases).

Il Data Mining, per scoprire modelli e relazioni nascoste nei database (che i metodi ordinari difficilmente scorgono e comunque non sono adatti a tale scopo) sfrutta raffinate analisi statistiche e tecniche di modellazione. Molte delle tecnologie impiegate nel Data Mining traggono origine principalmente da due filoni di ricerca: quello sviluppato dalla comunità scientifica dell'apprendimento automatico (Machine Learning) e quello sviluppato dalla statistica.

Più in generale, il Data Mining si è sviluppato traendo principalmente spunto da quanto sviluppato in altri ambiti disciplinari:

- Ambito epistemologico: complessità, incertezza.
- Ambito scientifico: scienze cognitive, intelligenza artificiale, computer science.
- Ambito statistico: teoria dell'apprendimento statistico (Learning from Data), statistica computazionale, strategie d'analisi.
- Ambito economico: marketing personalizzato, sistemi di supporto al management.

Si può notare che l'elemento che interseca tutti questi ambiti disciplinari è l'informazione vista nel processo

INFORMAZIONE → CONOSCENZA → DECISIONE

Ma la vera novità offerta dal Data Mining è l'integrazione delle precedenti metodologie con i **processi decisionali**.

In termini generali ciò che distingue l'attività di Data Mining dall'analisi statistica comunemente intesa non è, perciò, solamente la mole di dati su cui vengono effettuate le elaborazioni, così come nemmeno la disponibilità di un numero rilevante di tecniche, quanto il fatto che l'attività d'analisi è finalizzata alle esigenze aziendali e viene svolta in un ambiente predisposto per l'interrogazione

di contributi tecnici e conoscenze di business: fare Data Mining significa seguire una metodologia che va dalla definizione della problematica, all'implementazione di regole decisionali economicamente misurabili. Pertanto il Data Mining è un processo, non è il mero utilizzo di un algoritmo o di una tecnica statistica.

Alla luce di ciò, una definizione completa di Data Mining è la seguente:

> *Per Data Mining si intende il processo di selezione, esplorazione e modellazione di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori, e allo scopo di ottenere un vantaggio di business.*

2.2 Il supporto informatico ai dati

Una caratteristica fondamentale del Data Mining è quella della disponibilità di grandi masse informative:

- Giacimenti informativi: archivi di istituzioni pubbliche, dati amministrativi di enti e imprese.
- Basi di dati gestionali: gestori di telecomunicazioni, grandi banche e imprese finanziarie e assicurative, catene di supermercati.
- Basi di dati di ricerca: basi di dati genetiche e di grandi esperimenti (onde gravitazionali, acceleratori di particelle).
- Flussi informativi su Rete (Web Mining): testi, suoni, immagini, dati, comportamenti d'utenza.

L'ottenimento di un valido database è la prima e fondamentale operazione da compiere al fine di ottenere informazioni utili nell'attività di Data Mining.

2.3 Il Data Mining come processo

Per data mining si intende "il processo di selezione, esplorazione e modellazione di grandi masse di dati al fine di scoprire regolarità o relazioni non note a priori, e allo scopo di ottenere un risultato chiaro e utile al proprietario del database".

Diversamente dall'OLAP, il data mining combina in modo multivariato tutte le variabili a disposizione. Permette di andare oltre la visualizzazione dei riassunti presenti nelle applicazioni OLAP, formulando modelli funzionali all'attività di business.

Il data mining non si esaurisce nell'attività di analisi dei dati, bensì in un processo più complesso, in cui l'analisi dei dati è solo uno degli aspetti. Dal punto di vista più strettamente operativo, il data mining è un processo di analisi dei dati, consistente in una serie di attività.

Le fasi di tale processo possono essere così schematizzate:

- Definizione degli obiettivi dell'analisi;

- Selezione, organizzazione e pre-trattamento dei dati;
- Analisi esplorativa dei dati e loro eventuale trasformazione;
- Specificazione dei metodi statistici da impiegare nella fase di elaborazione;
- Elaborazione dei dati sulla base dei metodi scelti;
- Valutazione e confronto dei metodi impiegati e scelta del modello finale di analisi.

2.3.1 Definizione degli obiettivi dell'analisi

Questa fase del processo è sicuramente una delle più critiche, perché secondo quanto stabilito in essa, verrà organizzata tutta la metodologia successiva.

Si tratta anzitutto di definire gli obiettivi dell'analisi. Non sempre il fenomeno che si vuole analizzare è di facile definizione: infatti, mentre gli obiettivi aziendali cui si vuole mirare sono generalmente ben chiari, le problematiche sottostanti possono risultare complesse da tradursi in obiettivi dettagliati di analisi.

Una chiara esplicitazione del problema e degli obiettivi che si vogliono raggiungere è il presupposto per impostare correttamente l'analisi.

Gli obiettivi devono quindi essere formulati chiaramente e non lasciare spazio a dubbi e incertezze.

2.3.2 Selezione, organizzazione e pre-trattamento dei dati

Identificati gli obiettivi di analisi, è necessario selezionare i dati necessari per l'analisi. Anzitutto è necessario individuare le fonti dei dati. Solitamente si scelgono dati da fonti interne: più economici e affidabili, hanno inoltre il vantaggio di essere il risultato di esperienze e processi diretti dell'azienda stessa.

La fonte ideale dei dati è rappresentata dal data warehouse aziendale, un "magazzino" di dati storici non più soggetti a mutamenti nel tempo, dal quale è semplice estrarre dei database tematici (data mart) di interesse. In assenza di un sistema di data warehouse, i data mart devono essere costruiti incrociando le diverse basi di dati aziendali. In generale, la creazione dei data mart di ana-lisi fornisce l'input fondamentale alla successiva analisi dei dati. Conduce alla rappresentazione dei dati, spesso in forma di tabella detta matrice dei dati, dise-gnata sulla base delle esigenze di analisi e degli obiettivi preposti.

Ottenuta la matrice dei dati, è spesso necessario effettuare operazioni di pulizia preliminare dei dati. In altre parole, bisogna attuare un controllo di qualità dei dati disponibili (data cleansing). Si tratta sia di un controllo formale per l'individuazione di variabili non utilizzabili, esistenti ma inadatte all'analisi, ed anche di un controllo sostanziale per la verifica del contenuto delle variabili e della eventuale presenza di dati mancanti o errati

Nel caso emergesse la mancanza di elementi informativi essenziali sarà necessario rivedere la fase di individuazione delle fonti. Da ultimo, nell'attività di data mining è spesso opportuno impostare l'attività di analisi su un sottoinsieme dei dati a disposizione (campione).

Nelle applicazioni di data mining infatti le dimensioni del database analizzato sono spesso notevoli e, pertanto, l'utilizzo di un campione, ovviamente rappresentativo, permette di ridurre notevolmente i tempi di analisi ed elaborazione.

Lavorando su campioni si ha l'importante vantaggio di poter validare il modello costruito sulla rimanente parte dei dati, ottenendo così un importante strumento diagnostico.

Infine, il vantaggio di lavorare su base campionaria consiste nel tenere sotto controllo il rischio che il metodo statistico, adattandosi anche alle irregolarità e alla variabilità propria dei dati sui quali è stimata, perda capacità di generalizzazione e previsione.

2.3.3 Analisi esplorativa dei dati e loro eventuale trasformazione

L'analisi statistica vera e propria inizia con l'attività di analisi preliminare, o esplorativa, dei dati. Si tratta di una prima valutazione della rilevanza dei dati raccolti che può condurre, eventualmente, a una trasformazione delle variabili originarie, per una maggiore comprensione del fenomeno o per la sua riconducibilità a metodi statistici che poggiano sul soddisfacimento di determinate ipotesi iniziali.

L'analisi esplorativa può suggerire inoltre l'esistenza di dati anomali, differenti rispetto agli altri. Questi dati anomali non vanno necessariamente eliminati, perché potrebbero contenere delle informazioni preziose al raggiungimento degli obiettivi dell'analisi.

L'analisi esplorativa dei dati è una fase indispensabile e necessaria per permettere all'analista di pervenire, nella fase successiva, alla formulazione dei metodi statistici più opportuni per il raggiungimento degli obiettivi dell'analisi. Ciò naturalmente deve tener conto della qualità dei dati a disposizione, ottenuti nella fase precedente.

L'analisi esplorativa potrebbe, eventualmente, suggerire una nuova estrazione di dati, essendo quella considerata insufficiente per gli scopi preposti.

2.3.4 Specificazione dei metodi statistici da impiegare nella fase di elaborazione

I metodi statistici che possono essere utilizzati sono numerosi, e di conseguenza anche gli algoritmi che li implementano. Il data mining è un processo guidato dalle applicazioni, quindi i metodi utilizzati possono essere classificati in base

allo scopo immediato per il quale l'analisi viene effettuata.

In conformità a tale criterio si possono distinguere, essenzialmente, quattro grandi classi di metodologie, che possono essere esclusive, oppure corrispondere a distinte fasi del processo di data mining: esplorativi, descrittivi, previsivi, locali.

La scelta di quale metodo utilizzare nella fase di analisi dipende essenzialmente dal tipo di problema oggetto di studio e dal tipo di dati disponibili per l'analisi.

2.3.5 Elaborazione dei dati sulla base dei metodi scelti

Specificati i modelli statistici, si tratta di tradurli in opportuni algoritmi di calcolo informatico che permettano di ottenere, per il database a disposizione, i risultati sintetici desiderati.

L'ampia disponibilità di software, anche specialistico, per l'attività di data mining, fa sì che, nella maggioranza delle applicazioni standard, non sia necessario sviluppare un algoritmo di calcolo ad hoc, ma semplicemente impiegare quello implementato nel software a disposizione.

I gestori del processo di data mining devono avere comunque un'adeguata conoscenza delle differenti metodologie, e non solo delle soluzioni software, per adattare il processo alle specifiche esigenze aziendali e saper interpretare correttamente i risultati delle elaborazioni in termini decisionali.

2.3.6 Valutazione e confronto dei metodi impiegati e scelta del modello finale di analisi

Per poter produrre una regola decisionale finale è necessario scegliere, fra i vari metodi statistici considerati, il "modello" migliore di analisi dei dati. La scelta del modello e, quindi della regola decisionale finale, si basa su considerazioni che riguardano il confronto tra i risultati ottenuti con i diversi metodi.

Questa fase costituisce un importante controllo diagnostico della validità dei metodi statistici specificati, successivamente applicati ai dati a disposizione: potrebbe darsi che nessuno, fra i metodi impiegati, permetta un soddisfacente raggiungimento degli obiettivi di analisi; in tale caso, si tratterà di "tornare indietro" e specificare una nuova metodologia, più opportuna per l'analisi in oggetto.

Nella valutazione delle performance di uno specifico metodo concorrono, oltre a misure diagnostiche di tipo statistico, anche considerazioni relative ai vincoli di business, sia in termini di risorse sia in termini di tempo, oltre alla qualità e disponibilità dei dati. Nel contesto del data mining, risulta spesso poco utile utilizzare un solo metodo statistico per l'analisi dei dati.

Ogni metodo è potenzialmente in grado di far luce su aspetti particolari, possibilmente trascurati da altri. Per la scelta ottimale del modello finale è necessario applicare e confrontare una pluralità di tecniche in modo semplice e rapido, confrontare i risultati prodotti e dare una quantificazione economica delle diverse regole costruite.

2.3.7 Utilità del Data Mining

La forza del data mining consiste nell'essere in grado di fornire modelli predittivi a supporto delle decisioni riguardanti diverse aree di interesse:

- *customer profiling*: predice gli interessi del cliente, l'affinità ed il comportamento;
- *web marketing*: predice i click-throught rate;
- *gestione dell'abbandono*: predice il tasso di abbandono, l'intervallo e l'attrito;
- *gestione della frode*: predice transazioni fraudolente;
- *assegnazione del credito*: predice la bancarotta;
- *manutenzione*: predice proattivamente guasti alle apparecchiature;
- *gestione d'inventario*: predice le richieste.

Le analisi dei comportamenti d'acquisto possibili grazie all'uso di data mining sono molteplici. Le principali sono l'analisi delle associazioni di dati e delle forme di comportamento sequenziali.

Associazioni di dati

Dato un database di transazioni (dove spesso ogni transazione è un insieme di voci), si possono scoprire tutte le associazioni ove la presenza di una specifica voce comporta la presenza di altre (informazioni ottenute attraverso l'analisi degli scontrini registrati alla cassa).

Ciò rappresenta un ausilio per esempio nella disposizione dei prodotti sugli scaffali, nella preparazione di campagne promozionali, nel processo degli ordini, nella gestione delle scorte.

Forme di comportamento sequenziali

Dato un insieme di sequenze di transazioni, il Data mining aiuta a trovare le sottosequenze contenute in una specifica parte di queste sequenze. Una sequenza di transazioni può essere la sequenza di transazioni di acquisti di un gruppo di clienti (informazioni ottenute attraverso analisi di carte fedeltà o di credito).

Si vanno ad individuare comportamenti di acquisto relativi non ad una sola visita ma a più visite consecutive.

2.4 Web Mining

Il Web Mining è l'applicazione di tecniche di Data Mining al World Wide Web. È stato al centro di molti recenti progetti di ricerca e articoli. Ciononostante c'è molta confusione intorno a questo termine, perché comprende concetti diversi tra loro. Il Web Mining può, infatti, essere suddiviso in Web-content Mining, Web-structure Mining e Web-usage Mining.

- Il **Web-content Mining** è il processo di scoperta di informazioni da pagine Internet, frequente nella prossima generazione di motori di ricerca basati su XML/RKF.
- Il **Web-structure Mining** è l'applicazione di tecniche di Data Mining per ricostruire la struttura di uno o più siti web.
- Il **Web-usage Mining** è l'applicazione di tecniche di Data Mining per scoprire modelli di comportamento degli utenti a partire dai dati del Web, in modo di capire, e meglio soddisfare, le necessità delle applicazioni basate sul Web.

In questo paper, con il termine Web Mining si farà sempre riferimento al **Web-usage Mining**.

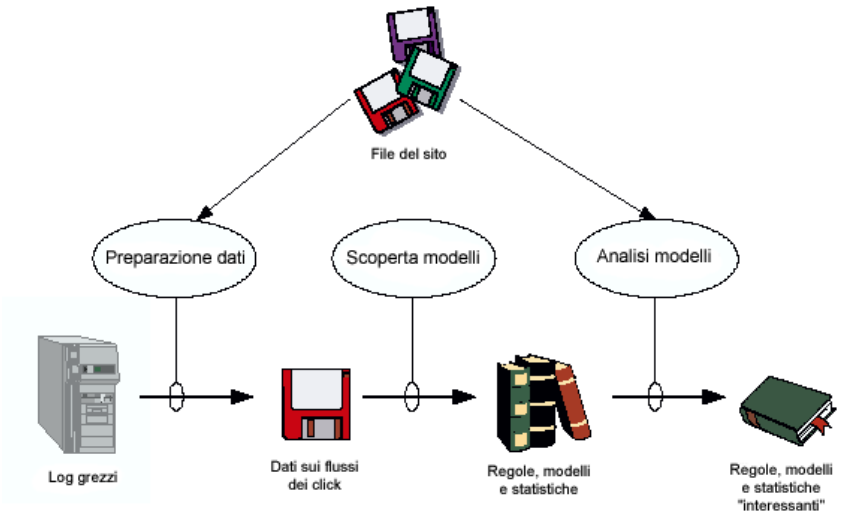
Rispetto alle transazioni tradizionali, le transazioni online permettono di utilizzare un marketing in tempo reale poiché virtualmente basato su "tutte" le informazioni disponibili rispetto a un particolare cliente e i suoi pari, appartenenti allo stesso segmento, in qualsiasi momento.

Una tecnica è quella del collaborative filtering: se due individui condividono la propria preferenza per un certo paniere di prodotti e il primo individuo acquista una nuova offerta, vi è molta probabilità che anche l'altro individuo apprezzi il nuovo item.

Fatta eccezione per questa applicazione particolare per il web, con i dati acquisiti è possibile svolgere tutte le altre funzioni del data mining più tradizionale.

3 Il Web-usage Mining

Come si può vedere in figura, il Web-usage mining è formato da tre passi principali: la preparazione dei dati, la scoperta dei modelli e l'analisi dei modelli.



3.1 Preparazione dei dati

La preparazione dei dati consiste nella conversione delle informazioni sul comportamento, il contenuto e la struttura, contenute nelle varie fonti di dati disponibili, in astrazioni di dati necessarie per la scoperta dei modelli. Questo è, indiscutibilmente, il compito più difficile nel processo di Web Mining a causa dell'incompletezza dei dati disponibili.

3.1.1 Preparazione dei dati comportamentali

La prima operazione da effettuare per la preparazione dei dati è la **pulizia** degli stessi. Le tecniche per pulire i server log ed eliminare gli elementi irrilevanti sono importanti per ogni tipo di analisi dei log, non soltanto per il Web Mining.

La scoperta di associazioni o statistiche è utile, infatti, solo se i dati rappresentati nei server log danno un'accurata rappresentazione degli accessi degli utenti al sito Web. L'eliminazione degli elementi non rilevanti può essere ragionevolmente completata controllando il suffisso dell'URL. Per esempio, tutti i log con un suffisso come gif, jpeg, jpg e map possono essere eliminati.

Un problema analogo, ma più complicato, è determinare se vi sono importanti accessi che non sono stati archiviati nel log. Abbiamo visto, infatti, che meccanismi come le cache locali e i server proxy possono distorcere fortemente il quadro complessivo degli utenti che visitano un sito Web.

L'unica soluzione per localizzare le visite alle pagine in cache è monitorare il comportamento dal lato client. I metodi attualmente utilizzati per cercare di superare questi problemi comprendono l'uso dei cookie, il cache busting e la registrazione esplicita degli utenti.

Nessuno di questi metodi è, però, senza seri svantaggi: i cookie possono essere cancellati dall'utente; il cache busting elimina il vantaggio della velocità (per cui la cache è stata creata) e può essere disabilitato; gli utenti forniscono spesso false informazioni durante la registrazione.

Metodi di comportamento in presenza di problemi di cache comprendono l'uso della topologia dei siti o dei refferrer log, insieme a informazioni temporanee per desumere i riferimenti mancanti.

Un altro problema associato con i server proxy è l'**identificazione degli utenti**: l'uso del nome della macchina come identificativo dell'utente può portare molte volte a trattare erroneamente gruppi di utenti come singoli individui. A meno che non sia usato un meccanismo di localizzazione lato client; per identificare univocamente gli utenti si possono utilizzare solamente indirizzo IP, browser agent e flussi di click.

La seconda maggiore operazione di preparazione dei dati è l'**identificazione delle sessioni server**. Prima di effettuare qualsiasi operazione di Web Mining occorre, infatti, raggruppare le sequenze di pagine visitate in unità logiche rappresentanti le operazioni Web o le transazioni.

Una sessione comprende tutte le pagine viste da un utente durante una singola visita a un sito. Identificare le sessioni è simile al problema di identificare gli utenti singoli, come discusso sopra. Una transazione può variare dalla visita di una singola pagina alla visita di tutte le pagine della sessione, a seconda del criterio usato per identificarle.

A parte nei casi tradizionali del Data Mining, non ci sono metodi adeguati per classificare le visite delle pagine in transazioni più piccole di un'intera sessione.

Nella pratica, alcuni dei problemi tipicamente incontrati nell'identificazione di utenti e sessioni sono i seguenti:

- **Singolo indirizzo IP/multiple sessioni**: gli internet service provider (ISP) tipicamente utilizzano dei server proxy attraverso cui gli utenti accedono al Web. Un solo server proxy può avere molti utenti che visitano un sito Web, potenzialmente nello stesso periodo di tempo.
- **Multipli indirizzi IP/singola sessione**: alcuni provider assegnano casualmente un indirizzo IP differente ad ogni richiesta dell'utente. In questo caso una sola sessione server può avere indirizzi IP multipli.
- **Multipli indirizzi IP/singolo utente**: un utente che accede al Web da macchine diverse avrà un indirizzo IP diverso da sessione a sessione. Ciò rende difficile il rilevamento delle visite ripetute dallo stesso utente.

- **Multipli browser agent/singolo utente**: un utente che usa più di un browser sempre sulla stessa macchina sarà scambiato per utenti multipli.

Per individuare unicamente gli utenti sono stati realizzati vari algoritmi. Uno di essi, ad esempio, controlla se ogni richiesta pervenuta è raggiungibile da pagine già visitate; se è stata richiesta una pagina che non è direttamente collegata alle precedenti, si assume che esistano utenti multipli sulla stessa macchina.

In un altro algoritmo la durata delle sessioni, determinata automaticamente da modelli di navigazione, viene usata per identificare gli utenti. Altre euristiche utilizzano l'uso combinato di indirizzo IP, nome della macchina e informazioni temporanee per l'identificazione.

Assumendo di aver identificato gli utenti, il flusso di click per ognuno di essi deve essere diviso in sessioni. Se non sono state effettuate richieste a pagine su altri server, è difficile sapere quando l'utente ha lasciato un sito Web.

Spesso viene utilizzato un limite di tempo in cui, se un utente non ha eseguito operazioni, la sessione viene interrotta. Quando un identificativo di sessione è inserito in ogni URL, la definizione della sessione è impostata dal server dei contenuti.

Sebbene l'esatto contenuto trattato è spesso ricavabile dai server log, è a volte necessario avere accesso alle informazioni del server dei contenuti. Da quando questi possono utilizzare variabili di stato per indicare se una sessione è attiva, le informazioni necessarie per determinare esattamente quale contenuto è trattato da una richiesta utente non è sempre disponibile nell'URL.

La figura seguente mostra un log di esempio che illustra molti dei problemi discussi sopra. L'indirizzo IP 79.50.12.199 è responsabile di tre sessioni server, e gli indirizzi IP 66.249.68.179 e 66.249.68.180 sono responsabili di una quarta sessione.

Usando una combinazione delle informazioni dei campi referrer (indirizzo di provenienza) e agent (il browser), le righe che vanno dalla 1 alla 11 possono essere divise in tre sessioni:

- *Music - Biography - Live - Phone - System*
- *Links - Paper*
- *Music - Biography - Contacts - Band*

Il percorso completo dovrebbe aggiungere due visite alla prima sessione *"Music - Biography - Live - Phone - Live - Biography - System"*, ed uno alla terza sessione *"Music - Biography - Music - Contacts - Band"*.

Senza usare cookie, un identificativo di sessione incorporato o un metodo di raccolta dei dati lato client, non c'è un metodo per determinare che le righe 12 e 13 sono una sola sessione server.

#	IP address	User ID	Time	Method/URL/Protocol	Status	Size	Referrer	Agent
1	79.50.12.199	-	[25/Apr/2007:03:04:41]	GET Music.html HTTP/1.0	200	3290	-	Mozilla/3.04 (WinXP)
2	79.50.12.199	-	[25/Apr/2007:03:04:49]	GET Biography.html HTTP/1.0	200	2050	Music.html	Mozilla/3.04 (WinXP)
3	79.50.12.199	-	[25/Apr/2007:03:05:01]	GET Links.html HTTP/1.0	200	4130	-	Mozilla/3.04 (WinXP)
4	79.50.12.199	-	[25/Apr/2007:03:05:52]	GET Live.html HTTP/1.0	200	5096	Biography.html	Mozilla/3.04 (WinXP)
5	79.50.12.199	-	[25/Apr/2007:03:05:59]	GET Music.html HTTP/1.0	200	3290	-	Mozilla/3.01 (WinVista)
6	79.50.12.199	-	[25/Apr/2007:03:06:41]	GET Biography.html HTTP/1.0	200	2050	Music.html	Mozilla/3.01 (WinVista)
7	79.50.12.199	-	[25/Apr/2007:03:07:46]	GET Paper.html HTTP/1.0	200	8140	Links.html	Mozilla/3.04 (WinXP)
8	79.50.12.199	-	[25/Apr/2007:03:07:59]	GET Contacts.html HTTP/1.0	200	1820	Music.html	Mozilla/3.01 (WinVista)
9	79.50.12.199	-	[25/Apr/2007:03:09:01]	GET Phone.html HTTP/1.0	200	2270	Live.html	Mozilla/3.04 (WinXP)
10	79.50.12.199	-	[25/Apr/2007:03:09:32]	GET Band.html HTTP/1.0	200	9430	Contacts.html	Mozilla/3.01 (WinVista)
11	79.50.12.199	-	[25/Apr/2007:03:09:50]	GET System.html HTTP/1.0	200	7220	Biography.html	Mozilla/3.04 (WinXP)
12	66.249.68.179	-	[25/Apr/2007:08:19:02]	GET Music.html HTTP/1.0	200	3290	-	Mozilla/3.04 (WinXP)
13	66.249.68.180	-	[25/Apr/2007:08:19:32]	GET Time.html HTTP/1.0	200	1680	Music.html	Mozilla/3.04 (WinXP)

3.1.2 Preparazione dei dati di contenuto

La preparazione dei dati sul contenuto consiste nella conversione di testo, immagini, script e altri file, come quelli multimediali, in forme che sono utili per il processo del Web Mining. Spesso questo consiste in applicazioni del Web-content Mining per ottenere classificazioni o raggruppamenti.

Sebbene l'applicazione del Data Mining al contenuto dei siti Web sia un'interessante area di ricerca, nel contesto del Web-usage Mining il contenuto di un sito può essere usato per applicare filtri agli algoritmi di scoperta di modelli. Per esempio, i risultati di un algoritmo di classificazione potrebbero essere usati per limitare i modelli scoperti a quelli che contengono pagine intere su un certo soggetto o categoria di prodotti.

Per classificare o raggruppare pagine intere in base agli argomenti, si può utilizzare anche il loro uso intenzionale. Infatti le pagine intere sono state create per comunicare informazioni (attraverso testi, grafici o contenuti multimediali), raccogliere informazioni dagli utenti, permettere la navigazione (attraverso una lista di link) o per combinazioni di questi usi.

L'uso intenzionale di una pagina intera può anche filtrare le sessioni prima o dopo la scoperta di modelli.

Per poter eseguire algoritmi di Web-content Mining sulle pagine intere, le informazioni devono prima essere convertite in un formato quantificabile.

Per fare questo sono tipicamente usati modelli di spazio vettoriale: i file di testo vengono scomposti in vettori di parole. I grafici e gli elementi multimediali possono essere sostituiti da parole chiave o testi descrittivi.

Il contenuto di una pagina intera statica può facilmente essere preparato analizzando l'HTML e ristrutturando le informazioni, oppure eseguendo algoritmi addizionali. Pagine intere dinamiche presentano più difficoltà.

I server sui contenuti che impiegano tecniche di personalizzazione e/o utilizzano database per costruire le pagine intere possono creare più pagine intere che vengono pre-elaborate. Un insieme di sessioni server può accedere solo a una frazione delle pagine intere possibili per un grande sito dinamico.

Anche il contenuto può essere riveduto: per essere pre-elaborato deve essere "assemblato" anche da una richiesta HTTP o una combinazione di template, script e accessi a database. Se viene pre-elaborata solo una porzione di pagina

intera, l'output di ogni algoritmo di classificazione o raggruppamento può essere distorto.

3.1.3 Preparazione dei dati strutturali

La struttura dei dati è creata dai link tra le pagine intere. La struttura può essere ottenuta e pre-elaborata nello stesso modo del contenuto di un sito.

Inoltre i contenuti dinamici (e perciò i link) incontrano più problemi delle pagine intere statiche. Per ogni sessione server si può costruire una struttura del sito differente.

3.2 Scoperta dei modelli

La scoperta dei modelli è nata su metodi e algoritmi sviluppati in diversi campi, come la statistica, il Data Mining, l'apprendimento automatico e il riconoscimento dei modelli. Nel Web Mining, però, questi metodi devono tener conto dei differenti tipi di astrazioni dei dati e della conoscenza disponibile a priori.

Per quanto riguarda la scoperta di regole associative, ad esempio, la nozione di transazione nella *market-basket analysis* non prende in considerazione l'ordine con cui gli articoli sono stati selezionati. Nel Webusage Mining, però, una sessione server è una sequenza ordinata di pagine richieste da un utente. Per di più, a causa delle difficoltà nell'identificazione di sessioni univoche, è richiesta un'addizionale conoscenza a priori (per esempio imporre un timeout predefinito).

3.2.1 Analisi statistiche

Le tecniche statistiche sono il metodo più comune per estrarre conoscenza dai visitatori di un sito Web. Analizzando il file di sessione si possono eseguire differenti tipi di analisi statistiche descrittive (frequenza, media, mediana, ecc.) su variabili come pagine intere, tempo di visita e lunghezza di un percorso di navigazione.

Molti strumenti di analisi del traffico Web producono report periodici contenenti informazioni statistiche come le pagine più visitate, il tempo di visita medio ad una pagina o la lunghezza media di un percorso nel sito.

Questi report possono includere limitate analisi di errori di basso livello, per esempio possono scoprire punti di entrata non autorizzata o trovare l'URL errato più frequente.

Questi pacchetti software, però, sono stati creati per produrre report sull'attività del server e non sul comportamento dei suoi visitatori. Di conseguenza sono poco utili nell'analisi di relazioni tra i dati riguardanti gli accessi a file e directory di uno spazio web.

Ciononostante, questo tipo di conoscenza può essere potenzialmente utile per migliorare le performance del sistema, aumentane la sicurezza, facilitare le

operazioni di modifica del sito e fornire supporto per le decisioni di marketing.

3.2.2 Analisi delle sequenze di visita

Per eseguire analisi sulle sequenze di visita si possono rappresentare le relazioni definite sulle pagine Web (o su altri oggetti) attraverso vari tipi di grafici. Il grafico più banale è quello che rappresenta il layout fisico di un sito Web con le pagine come nodi e i link tra le pagine come archi.

Altri grafici di questo tipo possono rappresentare sugli archi la similarità tra le pagine o il numero di utenti che vanno da una pagina a un'altra. Da questi grafici si possono determinare i modelli di attraversamento frequenti o le grandi sequenze di pagine visitate. Altri esempi di informazioni che possono essere scoperte attraverso l'analisi dei percorsi di visita sono:

- Il 70% degli utenti che ha avuto accesso alla pagina */company/product2* l'ha fatto iniziando la sua visita da /company e procedendo attraverso */company/new*, */company/products* e */company/product1*;
- L'80% degli utenti che ha visitato il sito ha iniziato da */company/product1*;
- Il 65% degli utenti ha lasciato il sito dopo aver visitato al massimo quattro pagine.

La prima regola suggerisce che c'è un'informazione utile in */company/product2* ma, siccome gli utenti tendono a fare un giro tortuoso per arrivare alla pagina, questa non è chiaramente evidenziata. La seconda regola semplicemente afferma che la maggior parte degli utenti accede al sito da una pagina diversa dalla home page (/company in questo esempio) e potrebbe essere una buona idea includere un menu su questa pagina, se non è presente.

L'ultima regola indica il tasso di abbandono del sito. Siccome molti utenti non visitano il sito per più di quattro pagine, bisognerebbe assicurarsi che le informazioni importanti siano contenute nelle pagine che gli utenti maggiormente visualizzano.

3.2.3 Regole associative

Le tecniche di scoperta di regole associative generalmente vengono applicate quando ogni sessione è formata da un insieme di elementi. In queste strutture il problema è scoprire tutte le associazioni e le correlazioni tra i dati quando la presenza di un insieme di elementi in una sessione implica (con un certo grado di confidenza) la presenza di altri elementi.

Nel contesto del web mining il problema consiste nella scoperta di correlazioni tra le pagine di un sito visitate da un dato utente. Queste pagine possono anche non essere direttamente connesse tramite link. Per esempio, usando tecniche di scoperta di regole associative possiamo trovare correlazioni come le seguenti:

- Il 40% degli utenti che ha visitato la pagina */company/product1* ha visitato anche la pagina */company/product2*;
- Il 30% degli utenti che ha visitato la pagina */company/special* ha eseguito un ordine on-line su */company/product1*.

Siccome solitamente i database contengono un ammontare estremamente grande di dati, le attuali tecniche di scoperta di regole associative provano a sfoltire la ricerca secondo un *valore-soglia* per gli elementi sotto esame. Il valore-soglia è calcolato in base al numero di sessioni utenti all'interno dei log.

La scoperta di queste regole organizzative nel commercio elettronico può aiutare lo sviluppo di effettive strategie di mercato. Ma, inoltre, la scoperta di regole associative dai log di accesso può dare un'indicazione su come meglio organizzare lo spazio Web.

3.2.4 Analisi dei gruppi

L'analisi dei gruppi è una tecnica per raggruppare un insieme di elementi che hanno caratteristiche simili. Nel Web Mining, si possono scoprire due interessanti tipi di gruppi: gruppi di comportamento e gruppi di pagine.

Il raggruppamento degli utenti mira a stabilire gruppi di visitatori che mostrano modelli di navigazione simili. Tale conoscenza è utile soprattutto per eseguire strategie di mercato mirate, o contenuti Web personalizzati, agli utenti che cadono nello stesso gruppo.

Il raggruppamento delle pagine scoprirà gruppi di pagine che hanno contenuti collegati. Questa informazione è utile per motori di ricerca e provider di assistenza Web. In entrambe le applicazioni possono essere create pagine HTML, statiche o dinamiche, che suggeriscono link personalizzati in base alle parole chiave immesse o alla cronologia dei bisogni informativi.

3.2.5 Regole di classificazione

La scoperta di regole di classificazione permette di creare dei profili di elementi in base ad una suddivisione dei dati in una o più classi predefinite. Ciò richiede estrazioni e selezioni di caratteristiche che meglio descrivono le proprietà della classe data. I profili ricavati possono essere poi utilizzati per classificare i nuovi elementi che vengono aggiunti al database.

Nel Web Mining le tecniche di classificazione consentono di sviluppare dei profili di utenti che accedono a particolari file, basandosi su informazioni demografiche o modelli di accesso. La classificazione può essere fatta usando algoritmi di apprendimento induttivi supervisionati. Ad esempio, classificazioni sui server log possono condurre alla scoperta di interessanti regole come:

- Gli utenti che lavorano in enti o agenzie statali quando visitano il sito sono interessati alla pagina *company/product1*;

- Il 50% degli utenti che ha eseguito un ordine on-line su *company/product2*, ha 20-25 anni e vive nel Nord Italia.

3.2.6 Modelli sequenziali

Le tecniche di scoperta dei modelli sequenziali tentano di trovare dei modelli tra le sessioni tali che la presenza di un insieme di elementi sia seguita da un altro elemento nel periodo di tempo dato da un insieme di sessioni. Nei log sulle transazioni dei server Web la visita di un utente è registrata lungo un periodo di tempo. L'ora associata ad una sessione sarà, in questo caso, un intervallo temporale determinato, e collegato alla sessione, durante le fasi di pulizia dei dati o di identificazione delle stesse.

Analizzando queste informazioni il Data Mining può determinare relazioni temporali tra gli elementi, come i seguenti:

- Il 30% degli utenti che ha visitato */company/products/* ha eseguito una ricerca su Yahoo nella settimana precedente;
- Il 60% degli utenti che ha eseguito un ordine on-line sulla pagina */company/product1* ha eseguito un ordine anche su */company/product4* entro 15 giorni.

La scoperta di modelli sequenziali nei log degli accessi consente alle organizzazioni Web di prevedere i modelli di visita degli utenti e di effettuare campagne di mercato mirate a gruppi di utenti determinati in base a questi modelli. Altri tipi di analisi temporali che possono essere eseguiti sui modelli sequenziali includono analisi sulle tendenze, rilevazioni dei punti di cambiamento o analisi simili.

Un altro importante tipo di dipendenza tra i dati, che può essere scoperto usando le caratteristiche temporali dei dati, sono le sequenze di tempo simili. Per esempio, potremmo essere interessati a trovare le caratteristiche comuni di tutti i clienti che hanno visitato un file particolare in un periodo di tempo 44 [t1, t2]. O, al contrario, potremmo essere interessati a un intervallo temporale (un giorno, una settimana, ecc.) in cui un particolare file ha avuto più accessi.

3.2.7 Modellazione della dipendenza

La modellazione della dipendenza è un'altra utile operazione di scoperta di modelli nel Web Mining. Il punto qui è sviluppare un modello capace di rappresentare le dipendenze significative tra le varie variabili del contesto Web. Per esempio, si può essere interessati a costruire un modello, basato sulle azioni degli utenti, che rappresenta le differenti fasi in cui un visitatore passa facendo compere in un negozio on-line (es. da visitatore casuale a serio compratore potenziale). Ci sono diverse tecniche di apprendimento probabilistiche che possono essere impiegate per modellare il comportamento di navigazione degli utenti.

La modellazione dei modelli di comportamento nel Web non solo fornisce una struttura teorica per analizzare il comportamento degli utenti, ma è potenzialmente utile per prevedere futuri consumi di risorse web. Queste informazioni possono aiutare a sviluppare strategie per aumentare le vendite dei prodotti offerti tramite sito Web o migliorare l'interesse di navigazione degli utenti.

3.3 Analisi dei modelli scoperti

L'analisi dei modelli è l'ultimo passo del processo globale di Web-usage Mining, come descritto nella Figura 2.1. Viene eseguita per filtrare le regole e i modelli scoperti allo scopo di comprenderli meglio. Infatti, i modelli sul comportamento nel Web, scoperti con le tecniche precedentemente descritte, non sono molto chiari senza meccanismi e strumenti che aiutino gli analisti. Questi sono stati implementati basandosi su diversi campi e comprendono statistiche, grafici e report, analisi di utilizzabilità e query sui database. L'esatta metodologia da applicare dipende comunemente dall'applicazione per cui il Web Mining è stato eseguito.

Tecniche di visualizzazione, come modelli grafici o l'assegnazione di co-lori ai differenti valori, possono spesso evidenziare modelli globali o tendenze nei dati. Oggi sono una scelta naturale per comprendere il comportamento degli utenti Web.

Secondo il *paradigma dei percorsi Web*, un insieme di log viene usato per estrarre sotto-sequenze di modelli di attraversamento del Web chiamati *percorsi Web*. Sul principio di questo paradigma sono sorti sistemi per analizzare in modo selettivo la porzione di Web interessata, filtrando le porzioni irrilevanti. Il Web viene qui visualizzato come un grafo diretto ciclico dove i nodi sono pagine e gli archi sono link.

Informazioni sul contenuto e la struttura possono essere usate per filtrare i modelli che contengono pagine di un certo tipo di utilizzo o contenuto, o pagine che hanno una certa struttura di collegamenti.

L'**OLAP** sta risultando un potente strumento di analisi strategiche sui database.

È stato recentemente dimostrato che le necessità funzionali e di efficacia dell'OLAP richiedono che siano disegnate nuove strutture informative. Questo ha portato allo sviluppo di modelli di informazione basati su ipercubi di dati e di tecniche per la loro efficiente implementazione. Recenti studi hanno mostrato che i bisogni di analisi dei dati sul comportamento nel web hanno molto in comune con quelli di un Data Warehouse, e qui le tecniche di OLAP sono piuttosto applicabili.

Le informazioni sull'accesso nei server log sono modellate come un archivio di dati che cresce nel tempo. Finché la dimensione dei server log cresce troppo rapidamente, non è possibile effettuare analisi on-line su tutti quanti i dati. C'è

perciò bisogno di riassumere i dati dei log per rendere fattibile la sua analisi online. Per ragioni di sicurezza si possono rendere porzioni dei log selettivamente (in)visibili ai vari analisti.

La più comune forma di analisi dei modelli è data da **meccanismi di query**, simili all'SQL. Una delle ragioni che viene attribuita al grande successo della tecnologia dei database relazionali è proprio l'esistenza di un query language, di alto livello e dichiarativo, che consente a un'applicazione di esprimere quali condizioni devono essere soddisfatte dai dati di cui ha bisogno, piuttosto che dover specificare come ricavare i dati richiesti.

Dato il grande numero di modelli che possono essere scoperti, sorge il bisogno di definire il fulcro dell'analisi. Questo si può trovare in almeno due modi. Primo, si possono porre delle restrizioni nel database (presumibilmente in un linguaggio dichiarativo) per restringere la porzione di dati su cui eseguire il processo di estrazione. Secondo, le query possono essere eseguite sulla conoscenza che è stata estratta dal processo di data mining. In questo caso c'è bisogno di un linguaggio per interrogare la conoscenza piuttosto che i dati. Per esempio, la pseudo-query:

```
SELECT association-rules(A*B*C)
FROM log.data
WHERE date >= 970101
AND domain = ‘‘edu’’ AND support = 1.0 AND confidance = 90.0
```

estrae le regole che coinvolgono il dominio ".edu" dopo il 1° gennaio 1997, che iniziano con l'URL A e contengono B e C in quest'ordine, e che hanno una soglia minima dell'1% e una confidenza minima del 90%.

4 Il Data Webhouse

4.1 Cenni storici

Fino a poco tempo fa, il concetto e le tecnologie di Data Warehousing, diffusesi nel corso degli anni '90, venivano regolarmente associate alla Grande Impresa: per costi, mole di dati implicati, complessità progettuale, i Data Warehouse, che per essere messi in esercizio richiedono da 6 a 18 mesi di lavoro e un investimento che spesso superava i 500 milioni di lire, venivano utilizzati per esaminare gli andamenti nel tempo di numerose grandezze, cercando di trovarne ogni tipo di correlazione possibile.

Problema tipico delle aziende di Grande Distribuzione, alle prese con le difficoltà connesse, per esempio, a individuare i migliori criteri di segmentazione per capire le dinamiche dei propri mercati, analizzandoli per prodotti, canali, aree geografiche e via dicendo.

Ancora una volta, Internet e il Web hanno dato un tale scossone al tradizionale modo di operare da stravolgere questi concetti, rendendo quasi impellente l'installazione di un Data Warehouse presso qualsiasi azienda che intenda rimanere sul mercato anche in un futuro sempre più proiettato verso l'e-Business.

Ormai non si parla più di segmentazione di mercato, ma di Marketing One-to-one, sfruttando il canale bidirezionale con ogni singolo cliente in grado di accedere al Web dell'azienda attraverso il proprio Browser.

A ogni richiesta, a ogni visita al sito, l'utente parla di sé, sia compilando i moduli che di volta in volta gli vengono proposti, sia visitando, in modo più o meno approfondito, determinate pagine, sia effettuando ordini, sia proveniendo da un posto o l'altro, a fronte delle parole chiave utilizzate per la ricerca, per non citare elementi quali l'immissione di ordini, l'effettuazione di pagamenti attraverso carta di credito e di tutte le azioni nelle quali l'utente esplicita le motivazioni che lo hanno portato a interagire con l'azienda.

Questo patrimonio di informazioni va conservato, analizzato e gestito dall'impresa, così da poter divenire efficaci nel definire nuovi prodotti, nel realizzare campagne di vendita, nell'indirizzare messaggi pubblicitari.

Occorre tener presente che, al di là della nuova tipologia di dati da trattare, ciò a cui occorre far fronte è l'improvvisa espansione del numero di interlocutori con i quali occorre confrontarsi: non più un mercato con i suoi segmenti in un territorio geograficamente definito, ma un universo di individui e di imprese dislocati ovunque sul pianeta.

Questo problema è indipendente dalle dimensioni dell'azienda che si affaccia su Internet con la propria offerta, quindi ci si può facilmente render conto che il problema di gestire e analizzare i dati, così da poter ottimizzare le politiche industriali e commerciali, diventa universale.

A parziale soccorso arriva la tecnologia dei Data Mart che, di fatto, corrispondono a un Data Warehouse focalizzato su un problema specifico, risultando più flessibile, rapido e meno complesso da realizzare rispetto al classico Data Warehouse.

4.2 Direzione Web

L'accrescersi di importanza del Web comporta tuttavia la revisione di alcuni dei criteri sui quali si fondavano i Data Warehouse del passato.

Un Data Warehouse è costituito essenzialmente di tre componenti funzionali, ciascuno dei quali va opportunamente configurato e personalizzato, così da risultare idoneo a soddisfare le esigenze dell'azienda:

- Acquisizione: si tratta di popolare il Data Warehouse con i dati catturati dalle applicazioni Legacy dell'azienda e integrate da quelli provenienti da qualsiasi altra fonte, così da definire il quadro completo della situazione relativamente alla propria azienda e a come questa si confronta con il mercato (concorrenti, esigenze degli utenti, sistema produttivo in generale). Per svolgere tali attività esistono numerosi strumenti Software che semplificano, ma soprattutto automatizzano, le varie operazioni.

- Memorizzazione: una volta individuati, i dati vanno archiviati in motori in grado di conservarli in modo affidabile, permettendo la condivisione e l'analisi secondo i diversi criteri scelti dagli utenti. In quest'ambito vengono comunemente impiegati Database (relazionali, multidimensionali, a oggetti) capaci di trattare in breve tempo considerevoli moli di dati, eventualmente impiegando sistemi Multiprocessor di vario genere. Stando ai dati pubblicati da Meta Group, la maggior parte degli attuali Data Warehouse si basano su piattaforme UNIX: Oracle, Sybase, IBM e Informix controllano il 65% dell'intero mercato.

- Accesso: i dati debbono essere facilmente accedibili ed elaborabili dagli utenti attraverso comuni Workstation, possibilmente da semplici Web Browser, con però la disponibilità di effettuarvi sofisticate analisi multidimensionali, utilizzando sistemi neurali, di Data Mining e più in generale, di individuazione di correlazioni nascoste. La qualità del Data Warehouse si misura soprattutto con i risultati ottenibili con i Tool che appartengono a questa categoria di funzioni, dal momento che sono proprio questi a qualificare e giustificare l'intero progetto. A questo segmento funzionale fanno riferimento ben sei categorie di prodotti: Intelligent Agent, che rilevano i valori scelti degli utenti come elementi critici di controllo e segnalano il raggiungemento delle soglie di attenzione o di intervento; Query Tool che permettono di fare delle interrogazioni al sistema in modo semplificato, senza l'impiego di alcun linguaggio di programmazione; strumenti di Analisi Statistica con i quali è possibile tracciare i Trend e anticipare l'evoluzione delle grandezze da tener sotto controllo; Data Discovery e

Data Mining che, sfruttando i principi dell'Intelligenza Artificiale, la logica Fuzzy, le reti neurali, gli alberi decisionali e altri metodi innovativi, esaminano i dati cercandovi correlazioni o significati altrimenti non rilevabili; OLAP (OnLine Analytical Processing) e Tool Multidimensionali attraverso i quali è possibile rappresentare i dati seguendo criteri a n-dimensioni; Tool di visualizzazione che permettono di creare grafici, Report e documenti con i quali illustrare in modo accattivante le conclusioni raggiunte nel corso delle attività di analisi.

Per trasformare un Data Warehouse in un Data Webhouse occorre:

1. Impostare il sistema sin dall'inizio in modo che risulti totalmente distribuito sia in termini di collocazione dei dati, sia per ciò che attiene alla loro immissione, fruizione, elaborazione.

2. Adottare sistemi con architetture basate su Web e non derivati dagli ambienti Client/Server con l'aggiunta di interfacce per Web Bowser. Occorre cioè che siano utilizzabili con Client privi di Software a bordo, garantendo la necessaria sicurezza e riservatezza nell'accesso ai dati. Nello stesso tempo, non è necessario che i dati risiedano su una stessa macchina, ma devono essere previste tutte le funzioni che servono ad acquisire ed elaborare dati provenienti da fonti esterne che vengono forniti come servizi via Internet e costantemente aggiornati (ad esempio, i dati di borsa o le oscillazioni delle valute). Nella vista complessiva, siamo dunque in presenza di sistemi con un Browser sul lato Client e, per lo meno sul piano logico, un Server con le funzioni di elaborazione, uno per la gestione della sicurezza, uno per la gestione del sito Web ed n con i dati.

3. Prevedere la possibilità di trattare dati alfanumerici, ma nello stesso tempo, grafici, immagini, suoni, filmati e ogni altro genere di File multimediali così come ormai standard su Internet.

4. Essere ampiamente scalabile sia in termini di quantità di utenti da supportare, sia per quanto riguarda la mole di dati da immagazzinare. Ciò può essere ottenuto creando una serie di Data Mart specializzati, ma facilmente correlabili l'uno all'altro. L'importante è che questi Data Mart siano progettati sin dall'inizio come componenti di un'unica infrastruttura integrata e non per risolvere problemi specifici così come spesso viene fatto.

5. Assicurare buoni tempi di risposta, per lo meno dello stesso livello di quelli normalmente rispettati per gli altri servizi erogati via Web. Tali valori corrispondono infatti alle usuali aspettative degli utenti che non accettano facilmente scadimenti delle prestazioni passando da un'applicazione all'altra.

6. Creare delle interfacce utente di uso assolutamente intuitivo, in quanto diventa sempre più difficile, se non impossibile, controllare meticolosamente il proprio parco utenti, formandoli, addestrandoli e assistendoli come nel passato.

4.3 Costruire il Clickstream Data Mart

Una parte molto importante del Data Webhouse è il Data Mart adibito alla memorizzazione dell'attività sul Web. Una frase molto in voga negli anni '90 era la seguente: "*Soon, we will have databases recording every customer interaction, no matter how seemingly insignificant...*". Tale frase racchiudeva in sé la volontà di ottenere una tecnologia che permettesse di analizzare l'immensa mole di click effettuati su un sito; tale realizzazione verrà chiamata "clickstream data mart" all'interno del Data Webhouse.

Il clickstream data mart ci può fornire una quantità sorprendente di informazioni sui visitatori di un sito, ed essendo a conoscenza di ogni singola azione per quanto essa possa sembrare insignificante si può rispondere alle seguenti domande:

- Che parte del sito ottiene più visite?
- Quale parte del sito associamo più frequentemente alle vendite efffettive?
- Quali parti del sito sono superflue o raramente visitate?
- Quali parti del sito sono "killer sessions", ovvero punti in cui l'utente rimane fermo e decide di abbandonare il sito?
- Quale è il profilo di click un utente appena registrato?
- Quale è il profilo di click di un cliente abituale?
- Quante visite effettua un utente tipicamente prima di registrarsi per acquistare un prodotto?

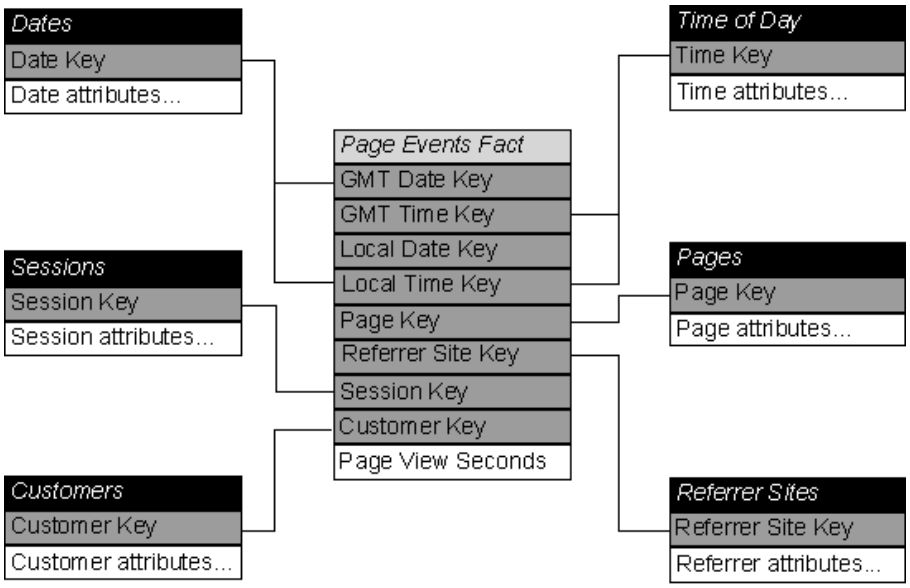
Date queste informazioni sarebbe utile costruire il clickstream data mart usando il modello dimensionale e poter effettuare operazioni di slice & dice per poter rispondere a tutte queste domande.

Per costruire tale Data Mart si può seguire la seguente metodologia:

- Definire la sorgente dei dati
- Scegliere la granularità della tabella dei fatti
- Scegliere le dimensioni appropriete per la granularità
- Scegliere i fatti appropriati

Ogni web server è capace di immagazzinare informazioni molto dettagliate; al più basso livello si può avere: istante temporale preciso del click; indirizzo IP dell'utente che esegue il click, pagine richieste e loro ordine, informazioni sui cookie, etc...

Un problema importante è che le richieste di pagine sono stateless, non ci si riesce a ricordare cosa l'utente abbia fatto nella pagina precedente. Senza un



Clickstream star schema

contesto appropriato, una richiesta di pagina potrebbe essere un singolo evento isolato, difficilmente interpretabile come parte di una sessione.

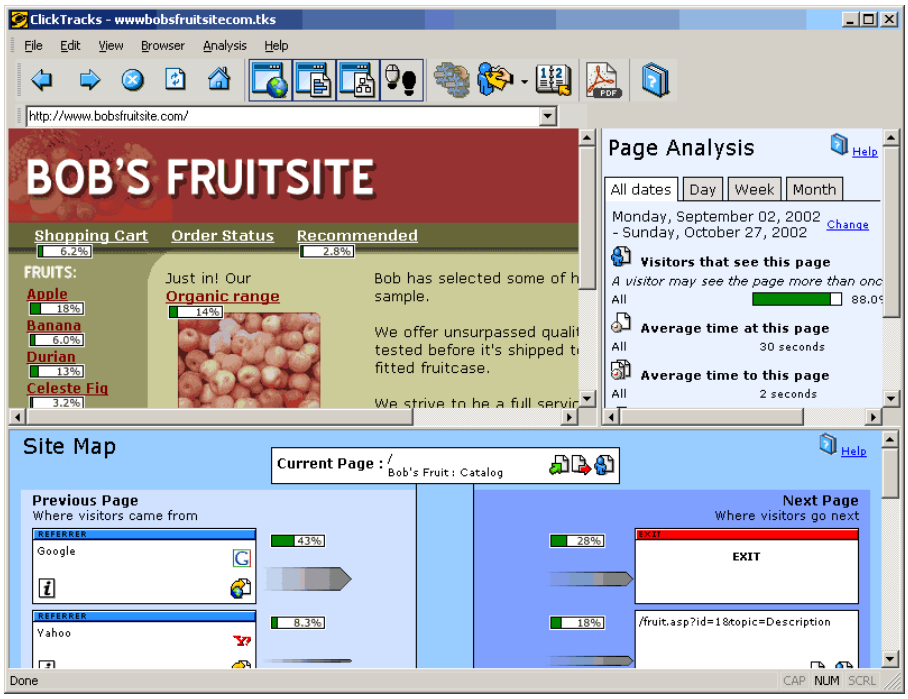
Un altro problema è l'individuazione dell'utente tramite l'indirizzo IP, gli ISP attribuiscono indirizzi IP dinamici a ogni connessione, e ciò non permette di sapere se un utente ritorna sul sito.

Questi problemi sono risolvibili se il sito in questione crea dei cookie sul browser degli utenti. Un cookie, se non cancellato permette di indentificare precisamente un utente.

La granularità di maggior dettaglio è data da ogni singolo evento che un utente esegue all'interno di una pagina dentro una sessione.

Nella figura seguente si può vedere un tipico modello dimensiona per il clickstream data mart. Le dimensioni sono *dates*, *time of day*, *sessions, customers*, *referrer sites*, e *sessions*.

Questa tabella dei fatti contiene una sola misura "page views seconds" che corrisponde al tempo trascorso. Tale stima è ottenuta partendo dal primo click dell'utente fino all'abbandono della pagina, ma poiché il browsing delle pagine è stateless, non si può essere sicuri sulle vere azioni dell'utente, che potrebbe aver semplicemente ridotto la finestra a icona lasciando aperta la sessione. Si possono fare stime precise solo quando è pervenuto un evento successivo facente parte della stessa sessione.



Esistono alcuni, non molti, software in commercio per le analisi clickstream, e tutti sono a pagamento. Questi software dovrebbero mostrare all'utente le informazioni estratte dall'analisi in maniere il più intellegibile possibile.

Come si nota dall'immagine, che mostra una schermata di esempio del software "ClickTracks", tutte le informazioni di interesse si cercano di esprimere in maniera chiara, integrata al sito in oggetto, mostrando le percentuali di click vicino ai link in modo da permetterne una comprensione quasi immediata.

4.4 Il problema della privacy

La privacy è un argomento sensibile che negli ultimi anni ha avuto molta attenzione a causa della rapida crescita del commercio telematico. La natura globale e auto-regolata del Web hanno accentuato il problema, infatti molti utenti vogliono mantenere un rigoroso anonimato nel Web e si rafforza la sensazione che qualcuno monitori i siti che stanno visitando e il tempo che spendono su quei siti.

D'altra parte, gli amministratori dei siti sono interessati a scoprire le caratteristiche degli utenti, realizzare statistiche sul comportamento nelle differenti sezioni dei loro siti Web per permettere loro di migliorare il disegno del sito Web e assicurarsi che il contenuto si rivolga alla maggior parte possibile di utenti. Gli amministratori del sito vogliono anche poter identificare un utente univocamente, per personalizzare il sito Web e migliorare l'esperienza di navigazione conb le tecniche e le modalità che sono state discusse in questo documento

La principale difficoltà è trovare regole e linee guida tali che gli amministratori dei siti possano eseguire varie analisi sui dati sul comportamento senza compromettere l'identità di un singolo utente. Dovrebbero esserci regolamentazioni severe per impedire ai dati sul comportamento di essere scambiati/venduti tra i siti, cosa che spesso in passato è successa e stata segnalata al Garante della Privacy. Gli utenti dovrebbero essere consapevoli delle polizze sulla privacy seguite da ogni sito, in modo che possano prendere una decisione informata sul rilevare i loro dati personali. Una linea guida può essere garantita solamente se gli utenti sono salvaguardati da una struttura legale.

La complessa normativa sulla privacy rende difficile questa attività e l'intera procedura sopra descritta esula quasi sempre dalle conoscenze dell'utente, che viene raramente informato sull'utilizzo dei dati raccolti su di esso, cosa che appare pertanto in contrasto con il D.lgs 196/03.

Il tracciato dei cookie utilizzati con questi scopi non può ogni volta materialmente essere sottoposto ad un vaglio di autorizzazione da richiedere all'utente, e quindi non si può sollevare una questione di illecito utilizzo dei dati personali. Ma nel caso della profilazione, e cioè dell'utilizzo di quei dati da parte di un titolare di un sito, allora l'illecito si realizza, in quanto la privacy del navigatore è violata.

Infatti, tra i principi generali cui si ispira la normativa a tutela della riservatezza, vige il cosiddetto principio della limitazione degli scopi: i dati possono essere raccolti ed impiegati solo per scopi precisamente individuati e con l'assenso e la consapevolezza di tale utilizzo da parte del titolare degli stessi. Ne consegue che anche la profilazione operata su internet deve necessariamente essere sottoposta al consenso dell'utente, potendo altrimenti incorrere il soggetto che la pone in essere in sanzioni previste dal D.lgs 196/03 ed in specie di quelle prevista dagli

artt. 141 e seguenti, che sanzionano il trattamento illecito dei dati personali avvenuto senza consenso del titolare e per trarne profitto.

Quest'ultimo infatti è conseguenza naturale della profilazione: "*il titolare di un sito commerciale capace di individuare e classificare i gusti dei propri navigatori, può incrementare pubblicità e distribuzione del prodotto che risulti, dalla profilazione, come maggiormente ambito dagli utenti*".

Inoltre, sempre secondo quanto stabilito dalla legge a tutela della privacy, coloro che raccolgono ed impiegano dati personali, sono tenuti a garantirne la sicurezza dalla conoscibilità degli stessi da parte di terzi: nel processo di profilazione, non adottando misure di sicurezza idonee, il titolare che raccoglie illecitamente i dati rischia di esporre gli stessi all'accesso da parte di terzi, commettendo ennesimo illecito ai sensi dell'art. 130 del D.lgs 196/03..

La legge italiana non si è espressa sul caso della privacy online in maniera chiara e specifica, cosa che richiederebbe norme speciali soprattutto negli ultimi anni. Eppure anche fenomeni quali lo spam possono essere conseguenza della profilazione, potendo un estraneo durante la nostra navigazione arrivare persino a raccogliere dati ed indirizzo e-mail, associarli con relativi gusti commerciali come emersi dal click-stream, al fine di farci pervenire in modo diretto una "pubblicità personalizzata". È facile concludere che la profilazione così utilizzata rappresenta un fenomeno tutt'oggi impunito o comunque ignorato, a discapito del navigatore poco esperto che può non sospettare di essere spesso "spiato" ed "elaborato".

Il W3C ha in corso una iniziativa chiamata Platform for Privacy Preferences (P3P), il cui ideatore principale è un italiano, il Prof. Massimo Marchiori. Il P3P fornisce un protocollo che permette agli amministratori dei siti di pubblicare le polizze sulla privacy, seguite dal sito, in un formato leggibile dalla macchina. Quando gli utenti visitano il sito per la prima volta il browser legge le polizze sulla privacy e le confronta con le sue impostazioni sulla sicurezza, configurate dall'utente stesso. Se le polizze sono soddisfacenti il browser continua a richiedere le pagine dal sito, altrimenti viene usato un protocollo di negoziazione per giungere a un'impostazione che è accettata dall'utente. Un altro scopo del P3P è preparare delle linee guida per le organizzazioni indipendenti per assicurare che i siti si attengono alle dichiarazioni della polizza che hanno pubblicato. L'Unione Europea ha organizzato una struttura regolatoria per la privacy su Internet e ha emesso una direttiva che pone regole per il trattamento e trasferimento dei dati personali.

References

[1] COOLEY R, MOBASHER B, SRIVASTAVA J. Web Mining: Information and Pattern Discovery on the Wordl Wide Web. Department of Computer

Science and Engineering University of Minnesota.

[2] SRIVASTAVA J, COOLEY R, DESHPANDE M, TAN P. Web Usage Mining : Discovery and Application on Usage Patterns from Web Data. Department of Computer science and Engineering University of Minnesota.

[3] COPPI R. Data Mining. Un quadro teorico-metodologico. Università degli Studi di Roma "La Sapienza": Slides del corso di Data Mining.

[4] GIUDICI P. Data Mining. Metodi statistici per le applicazioni aziendali. Milano: McGraw-Hill, 2001.

[5] FREDIANI V. Click-stream e spamming: la violazione della privacy del navigatore. Articolo legale.

[6] SWEIGER M, MADSEN M, LANGSTON J, LOMBARD H. Clickstream Data Warehousing. John Wiley publications.

[7] FORTUNATO L. Aspetti e Problemi del "Customer Relationship Management" (CRM). Tesi universitaria, Università degli studi di Roma.

[8] CONVERTINO G, DI PACE L, LEO P, MAFFIONE A, MALERBA D, VESPUCCI G. Tecniche di Web Mining per supportare la navigazione in rete.

[9] AA.VV. Database, Data Warehouse e Data Webhouse. fonti: http://www.itware.com/, http://www.consulentelegaleinformatico.it/