UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

TESI DI LAUREA

VALUTAZIONE DI UN PROGETTO DI PRESENZA SUL WEB:IL CASO SNAIDERO

Relatrice: CH.MA PROF.SSA SUSI DULLI

Laureanda: VALENTINA RAMPADO
Matricola: 452607

ANNO ACCADEMICO 2003/2004

INDICE

INTRODUZIONE

Internet ha rivoluzionato le tecniche di comunicazione, l'usufruibilità dei servizi e delle informazioni e più in generale l'approccio al mercato. Sempre un numero più elevato di aziende, infatti, vede il sito web come strumento di relazione con i propri clienti e proprio per questo motivo diventa importante riuscire ad avere una valutazione immediata della produttività e del funzionamento di questo nuovo strumento di comunicazione. Diventa indispensabile ottenere una chiara percezione sul ritorno che il progetto internet sta ottenendo in termini di posizionamento corretto rispetto al target di riferimento, di visibilità sul mercato. Inoltre viene naturale chiedersi come il visitatore si comporta, quante volte lo si può contare in termini di presenza, quali sono le pagine che ha visitato, come è arrivato sul sito, quanto tempo rimane e quali siano i percorsi più seguiti. Da queste misure di sintesi è possibile ricavare informazioni di natura commerciale in quanto si offre la possibilità di definire valide strategie nel campo della pubblicità, di natura tecnica in quanto si valuta se la struttura delle pagine e dell'albero di navigazione invogliano o scoraggiano la permanenza dei visitatori sul sito e di natura editoriale in quanto si capisce quali siano i contenuti sui quali puntare e quali invece quelli da modificare o addirittura eliminare. I dati vengono forniti dai log file, ovvero quei file che ogni server produce con l'obiettivo di tracciare passo per passo tutte le azioni che vengono compiute da un utente per la consultazione delle informazioni disponibili. Con quali strumenti analizzare i dati e come interpretarne il significato in modo corretto è ancora oggi oggetto di scontro in quanto la rilevazione del traffico di un sito web è una scienza giovane che riscontra una serie di oggettivi ostacoli che influiscono in modo sensibile sull'affidabilità dei dati registrati : i file di log e i servizi di analisi valutano solamente l'interazione meccanica tra sito ed utente.

In questo elaborato viene presentata un'applicazione di questa nuova scienza al sito del gruppo Snaidero, leader italiano in Europa nella produzione e commercializzazione di cucine componibili.

La diffusione del web che ha caratterizzato gli ultimi anni ha comportato la disponibilità di un numero sempre maggiore di informazioni dedite a registrare come l'utente si relaziona a questo nuovo strumento di comunicazione. Vista la complessità e la vastità della massa dei dati raccolti, questi vengono elaborati utilizzando tecniche di Data Mining e tale nuovo approccio viene identificato con il nome di Web Mining. Tuttavia i dati raccolti in rete sono di varia natura: dopo una breve panoramica del Web Mining si focalizzerà , quindi, l'attenzione sui log file e su una particolare branca del Web Mining, il Web Usage Mining, sulla descrizione delle fasi in cui essa si articola e su quali tecniche si appoggia per permettere una valutazione dell'interazione dell'utente con il sito. In seguito si procederà a presentare la realtà del gruppo Snaidero ponendo in evidenza il ruolo svolto dal sito: per quali finalità è stato pensato? come è stato strutturato? Si definirà, quindi, la linea su cui indirizzare lo sviluppo dell'analisi in funzione degli obiettivi per i quali questo sito è stato pensato. Si proseguirà nella valutazione di quale software utilizzare per l'elaborazione dei dati raccolti e, operata la scelta sullo strumento più opportuno, si concluderà con l'elaborazione dei log file che permetterà di sviluppare alcune considerazioni sull'efficacia del sito come strumento di informazione e di interazione con l'utente.

CAPITOLO 1

WEB MINING

1.1 PROBLEMI CHE SI RISCONTRANO NELL' INTERAZIONE CON IL WEB

Il Web è un centro di servizi per l'informazione enorme, distribuito e globale, utilizzato nei più svariati ambiti, quali il commercio elettronico, la pubblicità e l'informazione ai consumatori, la finanza, lo sport e i servizi ai cittadini. La quantità, la qualità e la dinamicità dell'informazione su web rendono pertanto indispensabili meccanismi efficaci di gestione di tali informazioni in modo che l'utente che accede ai contenuti e che interagisce con il Web possa trovare quello che cerca.

Alcune osservazioni che danno l'idea dei problemi effettivamente riscontrati nella gestione dell'informazioni sono: [1]

- Ricerca di informazioni utili: gli utenti utilizzano il servizio di ricerca on line quando vogliono trovare delle specifiche informazioni sul Web. Immettono una keyword sul servizio di ricerca e la risposta sarà una lista di pagine in ordine di attinenza con la query fatta. Tuttavia questi servizi di ricerca soffrono di alcuni problemi: una bassa precisione dovuta all'irrilevanza della maggior parte dei risultati ottenuti, una bassa capacità di trovare tutti i documenti attinenti alla query dovuta all'inabilità di indicizzare tutti i documenti presenti nel Web;

- Personalizzazione dei contenuti: questo problema è associato con la tipologia e la presentazione delle informazioni, infatti più la presentazione dei contenuti e i contenuti stessi sono personalizzati più l'utente preferisce interagire con il Web e lo si invoglia a continuare la visita nel proprio sito;

- Conoscenza degli utenti che interagiscono con il Web: questo problema è direttamente collegato al precedente, in quanto l'obiettivo è quello di capire come l'utente interagisce con il sito e quali sono i contenuti che vuole trovare in esso.

1.2 LE TECNICHE DI WEB MINING COME SOLUZIONE

Le tecniche di Web Mining possono essere utilizzate per risolvere i problemi appena presentati. Il Web Mining è la branca del Data Mining che si occupa dell'estrazione di informazioni dai dati provenienti dal Web. Tuttavia in questa definizione generale si possono identificare obiettivi di ricerca diversi in base alla tipologia dei dati che si vogliono analizzare: le pagine web e le informazioni in essa contenute, la struttura degli hyper-link che permettono di connettere un sito al resto del web e i dati che riflettono l'utilizzo delle risorse di un sito. In base, quindi, ai dati di partenza si può parlare rispettivamente di:[2] [3]

Web Content Mining

Il peggior difetto che hanno i tradizionali motori di ricerca risiede nel fatto che la loro ricerca all'interno di un sito web avviene attraverso le tradizionali tecniche che si basano sulla frequenza delle parole o su delle keyword inserite dall'utente in appositi campi. Il principale svantaggio di questo tipo di approccio è la mancanza di informazione strutturata che impedisce una ricerca precisa e mirata. Il Web Content Mining focalizza la sua ricerca sullo sviluppo di tecniche che assistono l'utente nella ricerca di documenti web. Il supporto di queste tecniche sta nella capacità di integrare i risultati

con profili dell'utente che ha posto la query o in una ricerca avanzata dalla keyword di partenza in modo da trovare nella vastità del numero di documenti, quelli che ne soddisfano il bisogno informativo.

Web Structure Mining

In questo caso le informazioni estrapolate non si basano sul contenuto della pagina web, bensì sulla sua struttura. Ovvero la qualità della pagina viene valutata in base ai link in essa contenuti e ai link che ad essa connettono. L'analisi può essere portata avanti intra-page o inter-page: la distinzione avviene in base al fatto che i link connettano, rispettivamente, a risorse interne al sito o contenute in altri siti. L'obiettivo di questa tecnica è, quindi, di catalogare le pagine come authorities o hubs: rispettivamente pagine che sono molto referenziate e alle quali linkano le pagine definite come hubs.

Web Usage Mining

In questo caso le informazioni che interessano riguardano il comportamento dell'utente, ovvero, è lo studio dei cammini che egli compie attraverso un sito web, oppure quali sono i documenti che egli visiona o preleva. Queste informazioni possono essere dedotte principalmente dai log file del web server con l'obiettivo di migliorare il design e l'usabilità di un sito, personalizzare i contenuti web e supportare le decisioni di marketing.

Le tre tipologie di approccio all'analisi dei dati provenienti dal Web non sono nettamente distinte, anzi molto spesso si usa una combinazione di tecniche.

L'attenzione si focalizzerà ora sul Web Usage Mining, processo che consente di verificare come un utente interagisce con il proprio sito.

CAPITOLO 2

WEB USAGE MINING

2.1 EVOLUZIONE DEGLI STRUMENTI DI MISURAZIONE

La presenza nel web è stata pensata per soddisfare ad uno specifico obiettivo: provvedere in modo più veloce e con costi più contenuti a quelli che sono i servizi offerti con le tradizionali tecnologie. Ma per valutare se questo servizio viene erogato secondo gli obiettivi previsti si necessita di strumenti di misurazione. I principali portali e siti informativi e di comunicazione, infatti, hanno da sempre attuato un processo atto a rilevare nel modo più preciso possibile il numero dei propri utenti nel tentativo di misurare il successo del servizio offerto: il primo approccio ha visto l'installazione nella home page di un contatore delle visite, counter, che aumenta di una unità ogni volta che un utente si connette alla pagina. Tuttavia questo strumento era tanto semplice quanto approssimativo essendo commisurato al momento: allora sapere che qualcuno aveva visitato il sito era già una misura di successo. Il counter, infatti, non teneva traccia degli effettivi accessi ma delle semplici visualizzazioni delle home page. Una situazione ricorrente si verificava nel momento in cui il visitatore aveva l'home page in questione impostata come menù di navigazione e quindi ad ogni accesso generava una visita indipendentemente dal fatto che poi andasse a consultare i contenuti del sito.

Parallelamente alla crescita di Internet, per soddisfare le sempre maggiori necessità di conoscenza di ciò che avveniva sul Web e sul

proprio sito, si raffinano le tecniche di misurazione e si passa ad un'analisi statistica descrittiva per determinare in primis i carichi di lavoro dei server e in seguito, anche, per avere un'idea di quella che è l'attività esercitata dagli utenti nel proprio sito. Tuttavia il meccanismo di funzionamento di Internet ed in particolare del web produce una notevole quantità di dati sottoforma di accesi al sito, e-mail inviate e moduli compilati. Questi dati non contengono informazioni interessanti ed utilizzabili direttamente dai manager per gestire l'attività svolta tramite il sito web: per essere utili devono essere rielaborati, aggregati ed organizzati in base alle metriche utili all'azienda. Nasce così l'e-Intelligence, cioè quel processo di gestione di grandi masse di dati web per estrarre informazioni di supporto ai sistemi decisionali per il Web Marketing, che è a tutti gli effetti un'applicazione di Data Warehousing dove i dati da memorizzare, analizzare e sui quali fare Data Mining allo scopo di scoprire relazioni non ovvie e potenzialmente utili, provengono da file di log dei server internet che gestiscono le applicazioni on-line e da altre sorgenti di informazioni a esse correlate, integrati da dati esterni riguardanti essenzialmente prodotti e clienti. Nel caso in cui i dati che si hanno a disposizione si limitino ai file di log derivanti dal traffico sul sito, la sola possibilità è quella di applicazioni di Web Traffic Mining. Nel caso, invece, di integrazione dei dati con altre fonti è possibile ipotizzare applicazioni di e-CRM (Electronic Customer Relationship Management) volte quindi a semi-personalizzare la struttura e la composizione del sito e di definire valide strategie nel campo delle vendite on-line e della pubblicità.

Il Web Usage Mining dovrebbe, quindi, consentire di mettere a punto una vera e propria metodologia che, a partire dall'osservazione, dalla corretta analisi statistica dei dati e dalla definizione di modelli statistico-matematici, consenta di influenzare il comportamento dell'internauta, fornendo una maggiore soddisfazione dell'utente nell'interazione con il sito.

2.2 LE FASI DEL PROCESSO DI WEB USAGE MINING

Il processo di Web Usage Mining si suddivide in cinque fasi identificate come data collection, data pre-processing, data abstraction, pattern discovery e pattern analysis. [3] [4] [5] [6] [7]

2.2.1 DATA COLLECTION

I dati nei processi di Data Mining sono il punto di partenza per un'accurata, quanto significativa analisi. Nel caso più specifico del Web Usage Mining questi sono rappresentati dai log file che registrano la richiesta di connessione alla risorsa di un sito dal lato server. Il Web, infatti, si basa su un meccanismo di funzionamento cosiddetto client-server. Il client è il computer che l'utente utilizza per navigare, mentre il server è la macchina dove sono depositate le informazioni che si cercano, ossia un computer permanentemente connesso ad internet il cui compito è quello di archiviare e distribuire i file richiesti . Il client e il server entrano in contatto tramite internet e i suoi linguaggi, ovvero i protocolli di trasmissione dati. Il funzionamento del meccanismo è abbastanza semplice: il client stabilisce una connessione internet, effettua una richiesta per una ricorsa che risiede su un server e se tutto funziona correttamente nel contatto, il server riceve la richiesta e, se la risorsa è disponibile, procede ad inviarla al client insieme ad alcune informazioni riepilogative dello scambio di file appena avvenuto tra i due computer. Inoltre, il server registra su un opportuno file, il log file, l'operazione appena effettuata: la richiesta e l'eventuale invio della risorsa disponibile. Nel momento in cui il client ricevuta la risorsa, la elabora mediante un browser, programma che permette la visualizzazione dei file in formato html, si parla di resource manifestation. Tuttavia una pagina Web, proprio per le caratteristiche intrinseche della stessa, si presenta come un oggetto di difficile classificazione, in quanto, è concepita come struttura interattiva non gerarchica di banners, immagini, suoni, applets, dati

ed altri oggetti. Il procedimento di registrazione, quindi, si ripete per ogni risorsa che compone una pagina Web a partire dal file html e che viene identificata come Hit. Di conseguenza la visualizzazione di una pagina web completa comporta una serie di richieste ed altrettante linee nel log file. Generalmente i dati di accesso ad un sito Web sono contenuti in file di testo. Esistono diverse configurazioni di file di log dove i più comuni sono Microsoft IIS Log format, NCSA Common Log File, W3C Extended Log File e ODBC Logging Format. In origine, tuttavia, si parlava esclusivamente di Common Log Format dove le informazioni registrate si limitavano ai soli dati di accesso: access log. [8]

Transfer/Access Log

Registra tutte le richieste di trasferimento file pervenute ad un server tramite protocollo http dagli utenti collegati ad Internet. I dati in esso raccolti sono una buona misura del carico di lavoro a cui è sottoposto un server Web.

Host o Indirizzo IP: la sigla IP sta per Internet Protocol. E' un numero di 32 bit che rappresenta univocamente ogni mittente o ricevente di pacchetti dati attraverso Internet. Nella sua forma più comune l'indirizzo IP è espresso come una serie di quattro numeri, separati tra loro da un punto. Ognuno dei quattro numeri può variare tre 0 e 255. Qualsiasi comunicazione che avviene su reti appartenenti ad Internet deve comprendere necessariamente l'indirizzo IP del mittente e quello del destinatario, allo scopo di poter essere rintracciata correttamente. Si fa riferimento, invece, all' host nel momento in cui si può risalire dall'indirizzo IP al nome del dominio tramite un'operazione detta Reverse DNS lookup. Il DNS, domain name system, è un sistema che traduce i nomi di dominio in indirizzi IP. Un nome di dominio è un nome letterale, associato in modo univoco ad un indirizzo IP numerico, per identificare una risorsa su internet. Mentre la trasformazione da un indirizzo letterale al

corrispondente indirizzo numerico è detta forward DNS lookup l'operazione contraria è detta reverse DNS lookup.

Data/ora: si tratta dell'ora e del giorno del luogo dove si trova il server e non dell'utente che ha richiesto la risorsa in questione. Si registra, inoltre, la differenza rispetto all'ora di riferimento del meridiano di Greenwich.

Action: le azioni richieste al server possono essere di vari tipi. Per fare alcuni esempi: GET ,POST e HEAD.

URL: l' Uniform Resource Locator è l'indirizzo, unico e inequivocabile, di una risorsa su Internet. Qualsiasi documento, sia esso un file immagine, un file di testo, una risorsa multimediale, è localizzabile precisamente per mezzo dall'URL che contiene una parte relativa al protocollo di comunicazione invocato, una parte più generale che identifica l'host, ovvero il computer su cui è archiviata la risorsa e una parte di dettaglio, che specifica il percorso e il nome del file da recuperare. Sempre più diffusamente, però, si fa riferimento all'URI (Uniform Resource Identifier) che non si riferisce alla posizione della risorsa ma piuttosto al nome con il quale viene essa identificata tra le altre risorse disponibili in rete.

Dimensione del file in byte

Protocollo Internet usato
I protocolli sono l'insieme delle regole che governano le trasmissioni in rete e nel log file viene registrata anche la versione.

Error Log
Memorizza tutte le richieste http che non hanno prodotto il risultato atteso dall'utente.

Status (Codice di Ritorno): risposta del server alla richiesta di risorse da parte del client. “200” indica che la richiesta ha avuto successo. Tuttavia non tutte le richieste producono il risultato atteso: può succedere che il server non è riuscito a trovare il file richiesto, che ci siano degli errori di time out (scadenza del tempo di attesa), connessioni rifiutate o interrotte, messaggi di “server too busy”. Si può procedere ad un’analisi degli errori per correggere possibili squilibri nella struttura di un sito, per verificare se la potenza di elaborazioni della CPU o la banda di connessione ad Internet siano sufficienti rispetto al volume di traffico generato.

Questi rappresentano il set minimo di campi che possono essere registrati in un CLF. Tuttavia, in seguito è stata possibile una caratterizzazione degli access log volta a registrare altre variabili utili alla gestione del sito: referrer log e agent log.

Referrer Log

Questa tipologia di log file permette di conoscere i collegamenti attraverso i quali i visitatori hanno raggiunto il sito. E’ evidente l’importanza di conoscere la provenienza dell’utente: per valutare l’efficacia di una campagna di posizionamento, per conoscere e valutare di che tipo è la diffusione sul Web, per sapere se il sito è stato segnalato da qualcuno. I dati di referrer sono anche utili per capire quali chiavi di ricerca abbiano consentito all’utente di raggiungere le pagine del sito preso in considerazione.

Motori di Ricerca

Particolari siti web che aiutano l’utente nella ricerca di risorse in internet attinenti ad un termine inserito nei campi di ricerca.

Keyword

Termine che si inserisce per avviare il processo di ricerca.

Agent Log

Registra il tipo di software che invia una richiesta al server. le informazioni fornite da questo file riguardano il tipo e la versione di browser utilizzato dall'utente, il sistema operativo e la risoluzione video.

Browser e Sistema Operativo

Spider e Robot: programmi che automaticamente effettuano una serie di richieste di file ad un server Web allo scopo di indicizzare i contenuti di quel sito per conto di un motore di ricerca.

Inizialmente queste nuove tipologie di log file erano registrare in file diversi, ora si può parlare di Extended Log File dove tutti i dati sono registrati in una singola riga.

Informazioni ausiliarie

Nel momento in cui gli utenti si registrano al sito per accedere a particolari sezioni o servizi questi forniscono oltre ai propri dati personali , dati demografici ed altre utili informazioni riguardanti le proprie preferenze e i propri interessi. Spesso si discute a proposito dell'affidabilità di questi dati, tuttavia questo argomento non rientra tra gli obiettivi di questa approfondimento.

2.2.2 DATA PREPROCESSING

Uno degli step chiave nel processo di Usage Mining è quello di creare un data set di partenza che possa essere informativo ai fini dell'analisi e che sia anche contenuto in termini di dimensioni, in

modo da perseguire obiettivi di efficienza. Si parla rispettivamente di data cleaning e data compression.

DATA CLEANING

Per quanto riguarda la prima fase è necessario tornare a fare riferimento al concetto di Hit. L'hit è costituito da qualsiasi richiesta di file pervenuta ad un server Web. Così, se una pagina Web è costituita da un file html e da sei immagini, la visualizzazione completa all'interno della finestra di un browser sia della pagina sia delle immagini in essa contenute corrisponderà alla registrazione di sette hit nell'apposito file di log del server. Ma non solo i file HTML hanno diritto ad essere considerati "pagine": anche i file cosiddetti "dinamici", cioè quelli che contengono elementi di programmazione in grado di generare contenuti differenti a seconda dei casi, sono da considerarsi "pagine". Rientrano in questa categoria i file con estensione ASP, PHP, PHP3, PL e simili. Ma possono rientrarvi - ed è una scelta dell'amministratore di rete impostare di conseguenza opportuni filtri sul server - anche i file TXT, i file RTF, i file DOC, i PDF ed altri ancora. Insomma: non è per niente semplice creare una categoria astratta chiamata "pagina", che comprenda alcuni tipi di file e ne escluda altri e che fornisca, allo stesso tempo, un parametro attendibile per la valutazione del numero di pagine viste da utenti umani. Un equivoco comune, soprattutto in passato, consisteva nel confondere le richieste di accesso con le pagine realmente caricate: un numero, quest'ultimo, che è in realtà quasi sempre nettamente inferiore al numero di hit registrato. Si fa, quindi, riferimento ad un nuovo concetto: hit qualificato. Questo non è molto diverso da un hit normale. L'unica differenza è che esclude le informazioni non pertinenti registrate dai log file. Il numero degli hit qualificati registrati da un sito è utile per definire grossolanamente il volume di traffico di un sito, ma non il numero di persone che lo hanno visitato. Un'altra fase di data cleaning è quella che tratta i log file dove gli accessi sono riconducibili a spiders, crawlers o robots. Una percentuale rilevante di traffico Web mondiale è generata non da

persone fisiche, bensì da automatismi che navigano la rete alla ricerca di informazioni. Gli spider (chiamati anche crawlers o robots) sono software che attraversano automaticamente la struttura ipertestuale del Web seguendo i link che uniscono i vari documenti presenti nella rete. Dunque i robots non fanno altro che chiedere ai server le pagine Web, al pari di comuni browser installati sui client, per poi seguirne i link interni con specifici obiettivi quali l'indicizzazione dei siti visitati (Indexing), la link e l'html validation ovvero un servizio che si occupa di controllare la giustezza formale delle pagine Web e dell'esistenza di link errati e il monitoraggio di novità per controllare lo stato di aggiornamento di alcuni siti specifici. La navigazione dello spider lasciando traccia nei log file del sito tende a falsare in modo significativo le informazioni che si possono dedurre da una normale analisi del traffico. In termini di hit e pagine viste la percentuale di incidenza degli spider può apparire bassa, questa diventa più rilevante nel momento in cui si stilano i profili medi di permanenza, visto che uno spider potrebbe trattenersi nel sito per diversi giorni facendo una richiesta ogni 15 minuti e quando si va ad analizzare i percorsi di navigazione degli utenti nel sito in quanto non corrispondono a preferenze dei visitatori. Tuttavia anche l'analisi degli accessi effettuati da spider può dare risultati interessanti sul posizionamento corretto del sito nei motori di ricerca. Un'altra tipologia di accessi che potrebbero essere cancellati sono quelli in cui il trasferimento delle risorse non è andato a buon fine, quindi, le righe riconducibili agli error log.

DATA COMPRESSION

La fase di data compression si occupa di eliminare quei campi contenuti nei log file che non sono informativi al fine di un'analisi statistica con il solo obiettivo di ridurre le dimensioni del data set di partenza.

2.2.3 DATA ABSTRACTION

La fase di data abstraction è necessaria per standardizzare i dati registrati nei log file secondo le disposizioni del W3C Web Characterization Activity (WCA) che rendono necessari procedimenti di user identification, session identification, path completion e transaction identification.

User Identification

La misura derivante dall'identificazione di un visitatore unico è considerata indicativa del traffico generato da un sito, ma anche del livello di fidelizzazione degli utenti. Tuttavia le aspettative di conoscenza legate ai valori numerici rilevati si scontrano purtroppo con difficoltà oggettive: l'impossibilità di identificare univocamente un utente guardando esclusivamente all'indirizzo IP. Su ogni server, infatti, è installato un protocollo di comunicazione DHCP (Dynamic Host Configuration Protocol) che consente di governare automaticamente l'assegnazione degli indirizzi IP a ciascuna macchina connessa ad Internet nella rete. Il DHCP è in grado di assegnare IP statici, uguali nel tempo, o IP dinamici, cioè indirizzi con scadenza a breve termine (generalmente la durata di una connessione). Proprio questo ultimo caso conferma come è impossibile sapere se visite successive eseguite da uno stesso IP provengano dalla stessa persona e se visite eseguite da IP diversi possano essere state eseguite dal medesimo utente. Per ovviare a questo problema sono state pensate due possibili soluzioni che però richiedono la collaborazione degli utenti. La prima è l'utilizzo di cookies. I cookies sono dei file di testo che vengono generati dal browser dell'utente in seguito ad un messaggio inviato dal server Web in risposta alla richiesta di collegamento ricevuta. Il cookies viene memorizzato sul client e contiene delle informazioni che identificano univocamente quell'utente rispetto al sito Internet che lo ha generato. Ad ogni successiva connessione, il server richiederà al browser il cookie precedentemente memorizzato. Tute le rilevazioni statistiche di traffico Web originate dall'uso di cookies sono però

soggette principalmente a due variabili: che l'utente collegato abbia abilitato nel proprio browser il supporto cookies e che questo effettui i successivi collegamenti per mezzo dello stresso browser. Una seconda possibilità è quella di richiedere all'utente una fase di login per accedere ad un servizio offerto dal sito. Questo consiste nell'inserire in appositi campi userID e password: codici che identificano univocamente l'utente rispetto al sito Internet e che sono rilasciati dopo che l'utente ha effettuato una registrazione sul sito, registrazione in cui ha fornito dai personali e demografici oltre ad altre informazioni circa le sue preferenze e i suoi interessi. Tuttavia, proprio perché queste operazioni richiedono la cooperazione dell'utente molto spesso possono non essere attuabili, resta, però, l'eventualità di identificare l'utente attraverso gli indirizzi IP. Una possibilità è data dall'operazione di tipo reverse DNS look up che consiste nel risalire da un indirizzo IP noto al corrispondente nome di dominio. Altrimenti risultano necessarie delle operazioni che combinano gli indirizzi IP con le informazioni derivanti dagli agent o dai referrer log: differenti valori nelle variabili dell'agent log sullo stesso indirizzo IP rappresentano necessariamente utenti diversi o diversi utenti che hanno lo stesso indirizzo IP e lo stesso browser possono essere considerati come distinti guardando il percorso che hanno utilizzato per accedere alla risorsa. Anche la combinazione di più informazioni può comunque risultare inattendibile in quanto ci sono delle particolari situazioni che generano incertezza e che non possono essere risolte:

- Single IP address/ Multiple server session: gli internet service providers (ISP) sono composti da un insieme di proxy server attraverso i quali gli utenti possono accedere indistintamente alle risorse Web. L'indirizzo IP in questo caso identifica non l'utente ma ISP: visite fatte con lo stesso server da più utenti nello stesso periodo saranno tutte identificate con lo stesso IP.

- Multiple IP address/ Single server session: gli ISP possono assegnare ad ogni richiesta fatta da un utente IP diversi anche all'interno della stessa sessione: in questo caso ai diversi indirizzi IP verranno ricondotte a visite diverse pur essendo state eseguite dalla stessa persona.

- Multiple IP address/ Single user: un utente che accede ad Internet da diversi client avrà IP diversi da sessione a sessione.

- Multiple Agent/ Single User: un utente che utilizza sullo stesso client più browser per navigare in Internet verrà identificato come utenti diversi.

Session Identification

La fase di session identification si rifà al concetto di visita: tutte le richieste ricevute in successione ininterrotta da un server web, provenienti da un medesimo indirizzo IP. Il problema sta nell'impossibilità di sapere quando un utente lascia il sito e, quindi, quando la visita finisce. L'unico criterio per identificare una sessione è quello di considerare terminata una visita da parte di un utente identificato se tra due successive richieste di pagina intercorre un tempo superiore al timeout di sessione impostato che è la durata massima predefinita di una visita ad un sito da parte di un utente unico. Non esiste uno standard per questa durata e neppure un consistente accordo in proposito. La lunghezza di una sessione può variare da un minimo di 10-15 minuti ad un massimo di un'ora. Nella maggior parte dei casi essa è impostata su 20 o 30 minuti. Se dura 20 minuti, ciò significa che ad un utente unico - riconosciuto come tale perché ha il medesimo indirizzo IP - vengono attribuite due visite al sito, nel caso in cui una sua richiesta di pagina giunga oltre 20 minuti dopo la precedente richiesta registrata. Viceversa, se l'intervallo

trascorso tra questi due eventi è inferiore a 20 minuti, allora viene conteggiata per quell'utente un'unica visita. Come è facile comprendere, la durata di sessione è un parametro del tutto arbitrario, che nulla ha a che vedere con l'effettivo comportamento degli utenti collegati ad un sito e che può tuttavia influenzare le valutazioni del settore commerciale di un'azienda, circa la misura della fedeltà degli utenti ai siti presi in considerazione. Il totale delle visite che si ricava in tal modo dall'analisi dei file di log per un dato periodo di tempo rappresenta evidentemente un'approssimazione statistica, il cui indice di affidabilità rimane imprecisato.

Path completion

Ci sono degli accessi che non vengono registrati nei file di log a causa dell'interazione del proxy server, si possono, quindi, trovare delle richieste di pagine che non sono direttamente collegate all'ultima pagina visitata. Il proxy server, infatti, agisce da filtro tra le richieste di connessione del client ai siti. La richiesta di accedere ad una risorsa su Internet viene intercettata dal proxy di rete in modo del tutto trasparente per l'utente. Se la pagina richiesta non è presente nella cache, memoria buffer del proxy, la richiesta viene inoltrata al sito che ospita la ricorsa, così da recuperare la pagina ed inviarla all'utente. Se, invece, la pagina è già presente nella cache del proxy, questa viene inoltrata direttamente all'utente, senza che occorra inviare alcuna richiesta al sito Internet che ospita la risorsa. L'uso di un proxy server fornisce essenzialmente la possibilità di aumentare notevolmente le prestazioni, risparmiando tempo e banda di connessione: ciò avviene perché il proxy memorizza una copia locale nella cache delle risorse più richieste, con lo scopo di servire all'utente quella copia locale, il luogo del documento originale presente nel sito Internet. È possibile prevenire questo inconveniente, inserendo nel codice delle pagine dl sito sottoposto a rilevazione statistica un comando che, definendo la scadenza immediata della validità di ogni pagina, costringa l'utente interessato

a collegarsi effettivamente ala risorsa richiesta, non potendola recuperare dalla cache. Questa soluzione ha però degli svantaggi: in primo luogo una maggiore occupazione di banda, in secondo luogo un'attesa più lunga per il caricamento delle pagine da parte dell'utente, il quale potrebbe essere negativamente influenzato per quanto riguarda le future visite. Il processo di path completion cerca di risolvere questo problema identificando e aggiungendo gli accessi mancanti.

Transaction Identification

La transaction identification ha come obiettivo quello di creare particolari sottoinsiemi di visite: per questo motivo ci sono più tecniche applicabili in relazione all'analisi che si vuole portare avanti. La prima, time window, assume che le transazioni significative siano quelle a cui è associato un tempo di visita non inferiore ad un determinato parametro. Le altre due, reference length e maximal forward reference, permettono di suddividere le pagine delle transazioni in due tipologie: auxiliary e content. Una risorse ausiliaria rappresenta uno step intermediario per l'utente nella ricerca della pagina di contenuto, che è quella che contiene le informazioni a lui utili. Gli algoritmi creati per questo obiettivo guardano l'uno, reference length, il tempo speso da un utente nella consultazione della pagina, l'altro, maximal forward reference, il percorso seguito dall'utente. Mentre nel primo caso basta fissare un parametro di tempo per suddividere le pagine in due categorie, nel secondo nessun parametro viene coinvolto. Infatti, si definiscono le backward reference che sono la visualizzazione di un pagina precedentemente già consultata e le forward reference che è la visita di una nuova risorsa da parte dell'utente. Analizzando il percorso di ogni visita, nel momento in cui si incontra una backward reference la visita delle pagine seguenti viene identificata come una nuova visita. Quindi la pagina precedente alla backward viene definita maximal forward reference, cioè quella pagina che è la meta della visita di un utente e che quindi può essere considerata una risorsa content.

2.2.4 PATTERN DISCOVERY E PATTERN ANALYSIS

Nella fase di pattern discovery, implementando tecniche ed algoritmi derivanti dalla statistica, Data Mining e Machine Learning, si evidenziano relazioni prima implicite. Tuttavia le relazioni scoperte possono non essere significative: si necessità, quindi, di una fase di pattern analysis in cui si evidenziano le regole che possono essere di supporto al miglioramento delle performance del sito e alle decisioni di marketing.

Statistical Analysis

L'analisi statistica è stata una delle prime tecniche applicate ai dati provenienti dai file di log per determinare i carichi di lavoro dei server prima e in seguito implementata per misurare l'audience nel web in quanto internet si era proposto come una strumento proficuo per l'advertising. Viene implementata da diversi software che comunque focalizzano la propria attività sull'interpretazione di alcune categorie standard di informazioni che danno la dimensione dei punti di forza e di debolezza del sito: quantificazione degli utenti, comprensione della tipologia degli accessi, comprensione del comportamento degli utenti e profilazione delle caratteristiche tecniche degli utenti. L'obiettivo dell'analisi è puramente di natura tecnica, in quanto, si valuta se la struttura delle pagine e dell'albero di navigazione invogliano oppure scoraggiano la permanenza dei visitatori nel sito e di natura editoriale per capire quali siano i contenuti sui quali puntare e quelli invece da eliminare o modificare. [9]

<u>Ad impression</u>

Nel linguaggio della pubblicità in Rete, si conta una impression ogni volta che un banner viene caricato in una pagina web. Poiché in una

singola pagina può essere contenuto più di un banner, il numero di impression - registrato in un apposito file - è in genere superiore al numero delle pagine servite. Tuttavia è erroneo associare alla registrazione di una impression l'idea che il banner corrispondente sia stato effettivamente visto dall'utente che ha richiesto la pagina: non esiste infatti un modo per sapere se, ad esempio, l'utente ha attivato la visualizzazione delle immagini nel proprio browser oppure se ha scorso la pagina fino alla fine (nel caso che il banner si trovi al di sotto di ciò che lo schermo gli mostra inizialmente).

Average page view duration

È il tempo medio speso da un utente unico su una singola pagina del sito. Può essere calcolato in due modi:

1. dividendo il tempo complessivo speso da un utente sul sito per il numero di page view registrate per quello stesso utente nel periodo considerato (ad es. un giorno)
2. facendo la stessa operazione, ma con la differenza di considerare solo le page view la cui durata sia inferiore al timeout di sessione impostato. In questo secondo caso, se il timeout è ad esempio di 30 minuti, una richiesta di pagina a cui non ne seguano altre da parte dello stesso utente per oltre 30 minuti viene scartata: si considera cioè come una pagina che l'utente non sta più guardando.

Average pages view per visit

Il valore si ottiene dividendo il numero complessivo di pagine richieste da un utente unico per il numero di visite effettuate da quell'utente nell'arco di tempo considerato. Incrociando i dati ottenuti per questo parametro con quelli relativi al tempo medio per visita, è possibile ipotizzare il comportamento-tipo degli utenti del sito. Ad esempio, una media di poche pagine viste per utente, accoppiata ad una lunga durata media delle visite registrate, potrebbe indicare che i visitatori del sito trovano con relativa facilità ciò che stanno cercando e leggono a fondo i contenuti reperiti. Viceversa,

una media di molte pagine viste in rapida successione nel corso di poche e brevi visite potrebbe indicare che la struttura del sito è caotica, che gli utenti non riescono a trovare ciò che stanno cercando e che perciò non sono invogliati a ritornare. Naturalmente queste supposizioni devono essere avanzate considerando tutti i possibili fattori di incertezza dei dati statistici rilevati.

Average time per visit

È il tempo medio speso da un utente unico per una visita al sito. Il valore si ottiene dividendo il tempo complessivo speso dall'utente sulle pagine del sito nel periodo considerato per il numero di visite che ha effettuato nello stesso periodo di tempo. Poiché il numero di visite effettato da un utente unico in un certo arco di tempo dipende dal parametro arbitrario della durata di sessione, è evidente che anche la durata media di una visita risulta influenzata dal valore assunto da questo parametro.

Browser used

È la classifica espressa in valori percentuali dei browser utilizzati dagli utenti che si collegano ad un sito. Questa informazione è utile soprattutto ai responsabili tecnici, per tarare al meglio la struttura delle pagine e la presentazione dei contenuti, in modo che siano navigabili per mezzo di ciascuno dei vari tipi di browser che risultano presenti in questa classifica. Se, ad esempio, analizzando l'elenco dei browser utilizzati, si scopre che una discreta percentuale di visitatori utilizza un browser non compatibile con alcune soluzioni tecniche implementate sul sito, sarebbe opportuno ricalibrare le pagine in modo da renderle accessibili anche alla parte di utenza penalizzata dalle precedenti scelte tecniche. Va comunque precisato che la verifica pratica di quali tipi di browser si colleghino alle pagine di un sito non dovrebbe aver alcuna importanza, se quelle pagine sono state codificate fin dall'inizio nel rispetto dei linguaggi standard per il Web definiti dal W3C.

Click rate

È il rapporto percentuale tra il numero di volte che un utente ha fatto clic su un banner presente su una pagina web ed il numero di volte che quel banner è stato caricato. È cioè il rapporto percentuale tra click-through ed impression. Un click rate del 5% significa, ad esempio, che un banner ha ricevuto 5 clic per ogni 100 richieste di caricamento registrate.

Click-through

È il numero di volte in cui un utente fa clic su un banner pubblicitario presente su una pagina web, collegandosi in tal modo al sito dell'azienda che vende il prodotto o il servizio reclamizzato dal banner.

Hourly (daily, weekly, monthly, yearly) pages count

Un grafico o una tabella che mostra la distribuzione oraria delle page view registrate nell'arco di una giornata (o la distribuzione giornaliera nell'arco di una settimana oppure di un mese, o la distribuzione mensile nell'arco di un anno solare).

Least requested, o popular, pages

È il complemento del parametro most requested pages, è cioè la classifica delle pagine meno richieste di un sito in un certo arco di tempo. Lo studio di questa classifica è utile per cercare di capire se i pochi accessi registrati per alcune pagine dipendono da contenuti non interessanti o dalla scarsa visibilità di quelle risorse nella struttura generale del sito.

Most common countries

È l'elenco in ordine decrescente delle nazioni da cui proviene il maggior numero di accessi ad un sito.

Most common operating systems

È la classifica espressa in valori percentuali dei sistemi operativi più utilizzati. Scorrendo i risultati forniti da questi resoconti si ha di solito la riprova di quanto sia schiacciante il monopolio ormai raggiunto dalla Microsoft.

Most requested pages

È la classifica, in ordine decrescente, delle pagine che hanno ricevuto più contatti in un determinato arco di tempo. È utile, per i tecnici e per i responsabili editoriali di un sito, considerare attentamente la classifica delle pagine più richieste, sia per correggere eventuali problemi di natura tecnica - come un sovraccarico del server web dovuto ad errori di programmazione - sia per correggere problemi di struttura logica del sito: alcune pagine, ad esempio, potrebbero essere al vertice della classifica delle più richieste non per i loro contenuti, ma perché sono delle strettoie obbligate da cui passare per raggiungere determinati altri contenuti.

Pages view

Il numero di pagine viste su un sito è forse l'informazione più importante che le statistiche web possano fornire, ma è anche l'informazione di gran lunga più ambigua e difficile da determinare, sia per la difficoltà di definire univocamente cosa sia una pagina sia per l'impossibilità oggettiva di conoscere il rapporto preciso tra pagine servite e pagine caricate da un utente umano (a causa dell'interferenza di numerose variabili quali proxy, NAT, cache locali, spider, ecc.). Ciò dovrebbe far comprendere quanto poco siano attendibili le valutazioni effettuate sulla base del numero di page view riportato dai programmi che analizzano i file di log. Soprattutto va tenuto presente che la comparazione delle page view registrate per due o più siti differenti può essere un'operazione dai risultati molto poco informativi: infatti, pur ponendo come uguali gli strumenti di rilevazione del dato e i filtri impostati, la struttura dei siti - in termini di composizione delle pagine e di oggetti in esse

presenti o da esse richiamati - può essere motivo sufficiente per generare, nel numero di page view rilevato, uno scarto nettamente superiore (o nettamente inferiore) alla reale differenza nella quantità di pagine viste da visitatori umani su ciascuno di essi. In definitiva, per dare sostanza ai dati numerici dei log relativi alle page view, occorre, più che in altre circostanze, lo studio approfondito di ogni singolo caso.

Request By Organization Type

È la classifica in valori percentuali delle richieste di accesso ad un sito, ordinate in base al tipo di dominio da cui parte la richiesta (COM, NET, ORG, MIL, EDU, GOV, identificativi nazionali).

Single access pages

È la classifica delle pagine uniche più richieste, visitate in un certo intervallo di tempo. Si tratta cioè di quelle pagine che, per motivi che i responsabili di un sito dovrebbero studiare a fondo, suscitano l'interesse degli utenti, ma allo stesso tempo non li invogliano a proseguire la visita appena iniziata. Potrebbe trattarsi di pagine con contenuti chiusi in se stessi (ad esempio una serie di collegamenti o una recensione), referenziate da altri siti. In questo caso andrebbe studiato il modo per indurre il visitatore a continuare la navigazione all'interno del sito, ad esempio inserendo nelle pagine individuate dei collegamenti ad altre sezioni con contenuti affini.

Top directories

È l'elenco in ordine decrescente delle directory (in genere solo quelle di primo livello) che hanno ricevuto complessivamente più richieste di accesso dagli utenti collegati. Questa classifica tende a dare un'idea dell'importanza reciproca delle sezioni in cui è suddiviso un sito. Perché questo resoconto abbia un qualche valore conoscitivo, occorre che la struttura logica del sito sia stata progettata in modo razionale, raggruppando i vari contenuti, in base alla loro omogeneità, sotto apposite directory.

Top entry pages

È la classifica in ordine decrescente delle pagine iniziali più richieste per ciascuna visita al sito registrata in un certo arco di tempo. Normalmente al vertice di questa classifica c'è la home page. Se così non è, diventa importante identificare i motivi per cui altre pagine funzionano meglio della home page come ingressi al sito. Ciò può essere fatto, ad esempio, analizzando i referrer log, cioè i dati sulla provenienza delle visite, per capire se e da quali altri siti sono referenziate le pagine che si trovano al vertice della classifica delle top entry.

Top exit pages

È la classifica in ordine decrescente delle pagine più richieste in un certo arco di tempo come pagine finali di una visita ad un sito. È, in altre parole, l'elenco delle pagine che sembrano più di tutte invogliare l'utente ad interrompere una visita in corso. Anche qui è importante uno studio approfondito, allo scopo di capire cosa c'è in quelle pagine che spinge i visitatori a lasciare il sito. Molto spesso la causa è da ricercarsi in una serie di collegamenti diretti ad altri siti; altre volte può trattarsi di un cattivo sviluppo dell'albero di navigazione, che finisce con il condurre gli utenti verso pagine prive sia di informazioni utili sia di collegamenti verso altre sezioni del sito.

Top paths

È la classifica dei più comuni percorsi di navigazione seguiti dagli utenti nel corso delle loro visite ad un sito. Per ogni elemento della classifica vengono forniti di solito:

- la sequenza delle pagine visitate, che costituisce il path (= percorso);
- la percentuale delle visite sviluppatesi seguendo quel path, rispetto alle visite totali registrate;
- il numero delle visite per quel path nel periodo considerato.

Top referring pages (o URLs)

È la classifica in ordine decrescente delle singole pagine che hanno reindirizzato degli utenti verso un sito. È utile che il servizio di rilevazione statistica adoperato permetta di aggregare i reindirizzamenti, in modo tale da separare quelli provenienti dall'interno del dominio di appartenenza del sito da quelli provenienti dall'esterno.

Top referring sites

È la classifica in ordine decrescente dei siti che hanno reindirizzato il maggior numero di contatti a file presenti su un sito. Spesso un'elevata percentuale di contatti è accoppiata in questa classifica all'etichetta "no referrer": ciò significa che un utente si è collegato direttamente ad una risorsa su un sito, senza esserci arrivato per via di collegamenti. Questo caso si verifica, ad esempio, quando un visitatore, conoscendo l'indirizzo della pagina richiesta sul sito di destinazione, inserisce manualmente la URL nella barra degli indirizzi del proprio browser.

Top Search Engines

È l'elenco in ordine decrescente dei motori di ricerca che hanno generato più contatti al sito. Se il numero complessivo di contatti generato da motori di ricerca è basso rispetto al numero complessivo di contatti registrato per un sito, allora se ne può dedurre che le pagine e i contenuti di questo sito non sono sufficientemente indicizzati dai motori di ricerca. Occorrerebbe in questo caso effettuare le apposite procedure per migliorare l'indicizzazione dei contenuti messi in linea. Essere ai vertici delle classifiche generate dai principali motori di ricerca può essere, infatti, un formidabile strumento per incrementare il numero di visite ricevute.

Top Search Keywords

È l'elenco in ordine decrescente delle parole chiave con più frequenza utilizzate dai visitatori di un sito nell'interrogare i motori di ricerca; parole chiave che hanno prodotto, come risultato dell'interrogazione, dei collegamenti e delle conseguenti visite al sito. È importante che un sistema di rilevazione del traffico sia in grado di fornire la classifica delle parole chiave più utilizzate dagli utenti. Studiare con attenzione questa classifica è infatti molto utile al fine di comprendere che tipo di contenuti gli utenti riescono a trovare sul proprio sito grazie ai motori di ricerca. Per via di esclusione si può poi cercare di definire quali altri contenuti, pur presenti sul sito, non generano contatti tramite i motori di ricerca, e perché.

Visitors Gained since Previous Period

Si tratta di visitatori mai registrati nei precedenti periodi di rilevazione statistica, ovvero di nuovi visitatori.

Visitors Lost since Previous Period

È l'elenco dei visitatori registrati nel corso di precedenti periodi di osservazione e mancanti, invece, dalle rilevazioni per il periodo corrente: si tratta cioè di visitatori perduti.

Visitors Returning from Previous Period

Si tratta di visitatori che hanno già visitato un sito in un precedente intervallo di tempo. Per la significatività di questa classifica è importante tarare con intelligenza i periodi presi in considerazione (non si può considerare, ad esempio, come un visitatore abituale di un sito un utente la cui precedente visita è stata registrata tre anni prima).

Association rules discovery

La scoperta di regole di associazione si inserisce inizialmente nel contesto della market basket analysis dove l'obiettivo è di trovare un sottoinsieme di prodotti che sono frequentemente acquistati insieme dal cliente. Nel contesto del Web Usage Mining l'algoritmo è implementato al fine di evidenziare correlazioni tra pagine a cui l'utente accede nella medesima sessione, senza il vincolo che queste pagine siano direttamente connesse e, quindi, visitate in sequenza.

Tuttavia le relazioni individuate non sono tutte statisticamente significative, si necessità dell'introduzione di due concetti: l'indice di support e l'indice di confidence. Infatti, la ricerca di una regola di associazione si può ricondurre alla ricerca, fra tutte le possibile regole costruibili, di quelle che soddisfano un minimo di supporto e un minimo di confidenza. Una regola di associazione è rappresentata dai simboli A→B dove A è chiamato antecedente e B conseguente. Si indichi con A→B il numero di sessioni utente in cui tale sequenza compare, almeno una volta e sia N il numero complessivo delle sessioni utente. Il supporto per la regola A→B si ottiene dividendo il numero di sessioni utente che soddisfano tale regola per il numero totale di sessioni utente. Si tratta di una frequenza relativa che indica la percentuale degli utenti che hanno visitato in successione le due pagine. In presenza di un numero elevato di sessioni utente, si può affermare che il supporto per la regola A→B esprime la probabilità che una sessione utente contenga le due pagine, in sequenza.

La confidenza per la regola A→B si ottiene invece dividendo il numero di sessioni utente che soddisfano la regola per il numero di sessioni utente che compongono la pagina A. L'indice di confidence esprime la frequenza che in una sessione utente in cui è stata visitata la pagina A possa essere successivamente visitata la pagina B. E' possibile individuare anche delle sequenze costituite da un numero di pagine maggiori di due. Facendo riferimento alla regola ABCD l'interpretazione diventa se ABC allora D. Si ottiene che le sequenze di pagine che attraggono di più i visitatori sono quelle a cui sono associati gli indici di support più elevati. Inoltre, gli indici di

confidence permettono di valutare, subordinatamente ad un certo corpo di partenza, quali siano i percorsi successivi preferiti. Il limite principale di tali indici è che, in quanto indici descrittivi, permettono di trarre conclusioni valide solo per il data set osservato e non delle previsioni di comportamento affidabili, per nuovi utenti. La presenza o l'assenza di alcune regole, tuttavia, possono aiutare il Web Designer in un restyling del sito, per esempio aggiungendo link che connettano pagine frequentemente viste insieme. [10]

Sequential Pattern discovery

La tecnica della scoperta di pattern sequenziali riprende i concetti dell'association rules discovery con la differenza che in questo caso si prende in considerazione anche la successione in sequenza delle pagine visitate. I risultati ottenuti possono essere un valido supporto per decisione di marketing all'interno del sito.

Cluster analysis

La segmentazione o profilazione ha come obiettivo quello di segmentare l'utenza in gruppi omogenei di comportamento. I dati che possono essere presi in considerazione per la profilazione sono molteplici: la serie delle scelte di navigazione effettuate sul sito in esame dagli utenti unici identificati, la dichiarazione esplicita di preferenze e interessi ottenuta tramite procedure di registrazione o sondaggi, la raccolta di dati demografici, la risposta degli utenti identificati a promozioni o a contenuti particolari. Il valore aggiunto di questa tecnica sta nella scoperta di correlazioni che possono essere utili a fini commerciali: Content affinities (affinità di contenuto) - gli insiemi di contenuti che tendono ad essere visti insieme dagli utenti del sito esaminato, Content effectiveness (efficacia dei contenuti) - per i siti di commercio in Rete, i contenuti che tendono ad essere visti in sessioni-utente che si concludono con un acquisto, Product affinities (affinità di prodotto) - sempre per i siti di commercio in Rete, l'elenco dei prodotti che sono più spesso acquistati insieme. In

base ai dati di partenza del processo di profilazione si può parlare di profilazione implicita, dove il tracciamento del comportamento di utenti anonimi nel corso delle loro visite ad un sito può avvenire tramite i cookies, utente rintracciato, e profilazione esplicita dove la segmentazione dell'utenza avviene in base ai dati forniti dall'utente durante la procedura di registrazione al sito, utente identificato. La tecnica di segmentazione può essere spiegata come segue: dato un insieme di punti in uno spazio p-dimensionale, dove p è il numero della variabili prese in considerazione, si cerca di suddividere questi punti in sottoinsiemi. Il numero dei sottoinsiemi può essere definito a priori o essere il risultato dell'applicazione della tecnica a seconda del metodo utilizzato. Si sta applicando, rispettivamente, il metodo di partizione e il metodo gerarchico.

2.3 METODI DI MISURAZIONE ALTERNATIVI

La presenza nel web è stata pensata per soddisfare ad uno specifico obiettivo: provvedere in modo più veloce e con costi più contenuti a quelli che sono i servizi offerti con le tradizionali tecnologie. Per valutare se questo servizio viene erogato secondo gli obiettivi previsti si necessita di strumenti di misurazione. Tuttavia, la misurazione del traffico di un sito Web analizzando i file di log, misurazione site centric, è soggetta a fattori che generano incertezza: l'impossibilità di definire univocamente un utente esclusivamente mediante l'indirizzo IP, il processo di caching generato dai proxy server, la difficoltà di definire quali tipi di file sono da considerare pagine e quali no. Possibili alternative possono essere una misurazione basata sul browser, tecnologia BBM (browser based measurement), o una misurazione centrata sull'utente, user centric measurement . La tecnologia BBM si basa ancora sui dati registrati in file di log, ma la generazione dei log non è più determinata dall'attività del server web bensì dal caricamento nel browser dell'utente di un apposito frammento di codice, denominato page tag. Si tratta in genere di alcune righe contenenti un javascript, il cui compito è di inviare una richiesta HTTP al server addetto alla

registrazione del traffico-web, per informarlo che è stata generata una page impression. Questo metodo di rilevazione offre alcuni indubbi vantaggi:

- Consente ad esempio di superare il problema - tipico delle misurazioni site centric - di definire dei filtri omogenei per separare i tipi di file associabili ad una page view (HTML, ASP, ecc.) dai tipi di file non associabili (JPG, GIF, PNG, ecc.). Con il sistema dei page tag, infatti, solo i file che contengono l'apposito frammento di codice javascript sono in grado di generare una page impression.
- Consente di eliminare l'incertezza legata alla non quantificabile interferenza di strumenti di caching interposti tra il server web ed i visitatori del sito. Infatti anche le pagine recuperate da cache locali, se dotate dell'opportuno page tag, invieranno al server deputato della registrazione dei log le chiamate necessarie a generare una page impression.

Per contro, la misurazione basata sul browser presenta anche degli svantaggi:

- In primo luogo, va tenuto presente che le statistiche di traffico generate con questo sistema riguardano solo ed esclusivamente le pagine in cui è stato inserito, e nel modo corretto, l'apposito page tag. Ciò significa che, laddove vi sia una realtà aziendale molto complessa - con molti siti da monitorare, moltissime pagine pubblicate e numerose persone addette al processo produttivo -, sarà molto difficile (per non dire impossibile) avere la certezza che la totalità delle pagine da sottoporre ad analisi statistica sia stata effettivamente modificata con l'inserimento dell'opportuno codice javascript. Si rischia cioè, con l'andar del tempo, di trovarsi di fronte ad un nuovo tipo di incertezza: di non sapere, cioè, se le rilevazioni del traffico browser-based di cui si è in possesso coprano la totalità degli accessi effettuati ai propri siti e siano perciò davvero attendibili.

- La presenza - per quanto minima e tendenzialmente non avvertibile - di un ritardo nel caricamento della pagina, dovuto alle chiamate HTTP aggiuntive presenti nel codice del page tag, indirizzate - nel caso che il servizio sia fornito da terzi - ad un server differente da quello che ospita il sito.

La seconda alternativa, conosciuta anche come Web audit, è completamente differente dalle altre due. Non si tratta, infatti, di un'elaborazione statistica effettuata a partire dai dati registrati nei file di log; si tratta piuttosto di una vera e propria indagine di mercato, basata su elementi tipici del settore:

- un campione di popolazione significativo del tipo di utenza che si vuole misurare;
- uno strumento per la rilevazione del comportamento degli individui che compongono il campione, applicato ai loro computer ed in grado di monitorare attimo per attimo qualsiasi tentativo, riuscito o non riuscito, di navigazione su Internet;
- strumenti statistici studiati per effettuare proiezioni più o meno attendibili, che estendono alla totalità della popolazione i dati ricavati dall'analisi del campione.

La validità delle proiezioni di traffico ricavate per mezzo di indagini di mercato è strettamente dipendente dalla significatività del campione selezionato - che è evidentemente una variabile difficilmente quantificabile - e dalla raffinatezza ed affidabilità degli strumenti statistici adoperati.

CAPITOLO 3

APPLICAZIONE AD UN CASO REALE

3.1 DEFINIZIONE DEL PROBLEMA

Ogni realizzazione di un progetto di presenza sul web richiede competitività e successo al fine di confermare, ma anche, migliorare la percezione dell'immagine dell'azienda e offrire un servizio migliore alla propria clientela. A questo proposito, soprattutto, negli ultimi anni si è verificato un sensibile aumento negli investimenti rivolti ad assicurare la propria visibilità anche in rete. Tuttavia gli obiettivi per i quali si sono spesi soldi e tempo spesso non vengono raggiunti, si necessità così di soluzioni che possano valutare in modo effettivo ed immediato la produttività e il funzionamento del progetto. La metodologia qui adottata per valutare il successo di un sito si basa sui percorsi di navigazione effettuati dal visitatore durante la sua permanenza e su alcune statistiche descrittive che sintetizzano l'attività esercitata sul sito dagli utenti. Questo perchè si richiede un approccio che deve essere capace di tener conto di tutti gli utenti, deve risultare appropriato per essere eseguito molto frequentemente e deve, eventualmente, condurre a concreti indicatori dei difetti del sito e ai rimedi per attenuarli. Il sito preso in considerazione è quello del gruppo Snaidero.

3.2 IL GRUPPO SNAIDERO

Il gruppo Snaidero è oggi leader italiano in Europa nella produzione e commercializzazione di cucine componibili. La mission dell'azienda, però, denota la volontà di distinguersi nei principali mercati europei con una posizione di leadership incontrastata. I due pilastri della filosofia del gruppo, fattori importanti per una politica di creazione di valore di lungo periodo, sono l'innovazione e la passione per l'eccellenza declinati nell'orientamento al consumatore, nell'internazionalità e nel modello organizzativo. Questi fattori di successo che da tempo garantiscono al gruppo Snaidero una certa continuità sono il risultato di tappe che hanno segnato negli anni l'identità di questa azienda.

La storia del gruppo Snaidero s'identifica con la storia del suo fondatore, il cavalier Rino Snaidero che nel 1946 inaugurò il suo primo laboratorio per la produzione di mobili, sicuro della crescente domanda alimentata dalla ricostruzione del secondo dopoguerra. L'evoluzione storica del gruppo s'identifica nella trasformazione da impresa artigiana a leader europeo data dall'evoluzione della piccola azienda artigianale in una struttura aziendale che punta sulla specializzazione nella produzione di cucine componibili sullo stampo di quelle americane.

Gli anni 70 e 80 sono quelli in cui cresce l'azione verso l'internazionalizzazione sia dal punto di vista commerciale che produttivo e gli anni 90 vedono l'inizio di un processo di crescita del gruppo attraverso acquisizioni di aziende europee leader nei loro mercati. Si sente, tuttavia, la necessità di una riorganizzazione strategica. Dagli anni 60 Snaidero, infatti, ha saputo imporsi grazie ad un'attenta politica di valorizzazione del prodotto in quanto il problema era produrre e non vendere ma verso la metà degli anni 90 il gruppo è invaso da una generale insoddisfazione rispetto alla propria posizione competitiva. L'azienda comprende che il punto di svolta sta nello spostare l'attenzione dal prodotto al cliente: questo deve diventare il punto di partenza per lo sviluppo di politiche d'offerta e per la riorganizzazione di tutte le funzioni aziendali

lasciando che il prodotto sia esclusivamente la realizzazione di quello che il cliente desidera.
Snaidero commissiona, quindi, un'indagine che avrà come obiettivo quello di definire l'identità e la vocazione del gruppo, visualizzare il target con l'obiettivo di sviluppare un nuovo prodotto ed elaborare un nuovo modello comunicazionale.
L'identità Snaidero, fino a quel momento, era quella di un'azienda caratterizzata da valori di prodotto consistenti e credibili, una cultura d'impresa coesa ed omogenea nella quale però mancavano i contenuti emotivi. Per rivitalizzare il proprio marchio Snaidero doveva interpretare le tendenze socio-culturali più in sintonia con la sua identità, la sua vocazione e le caratteristiche del suo target strategico.
I risultati dell'indagine rivelano che il target Snaidero si identifica in un segmento che vive il recupero della tradizione come espressione di modernità in quanto è alla ricerca di affettività, autenticità e valori duraturi: chi sceglie Snaidero non sceglie una cucina ma uno stile di vita. L'idea è quella, quindi, di segmentare la produzione per stili di vita ed anche il sito viene rinnovato con l'obiettivo di poter cambiare la percezione del marchio. Diventa uno strumento di supporto al consumatore sia nella fase di selezione ed acquisto del bene che in quella d'utilizzo, consente un rapporto con il consumatore più personale ed interattivo, è perfettamente integrato con le politiche comunicative e la rinnovata immagine aziendale e diventa uno strumento per comunicare il nuovo mondo Snaidero.

3.3 IL SITO SNAIDERO

Le motivazioni che spingono un'azienda a creare il proprio sito web sono molteplici: pubblicità, visibilità, conoscenza dell'azienda, facilitazione del contatto, individuazione dei punti vendita e maggiore assistenza al cliente. Internet è, infatti, sia un'efficace vetrina in cui l'azienda può presentare e far conoscere i propri prodotti, lo stile e la mission dell'impresa, sia un canale attraverso il quale questa si rende più facilmente individuabile, raggiungendo

l'obiettivo di poter interagire direttamente, più efficacemente con il cliente e con costi più contenuti rispetto ad una buona campagna pubblicitaria attuata attraverso i mezzi di comunicazione tradizionali. Gli obiettivi del restyling del sito Snaidero, quindi, reinterpretano quelle che solitamente sono le motivazioni che spingono un'azienda a creare un proprio sito web in funzione della percezione della nuova immagine del gruppo. Inoltre, essendo il consumatore il punto di partenza per lo sviluppo di una nuova concezione Snaidero, si punta ad un sito che coinvolga all'interno della piattaforma anche questa nuova rilevante figura, favorendo una più forte interazione tra consumatore, rivenditore ed azienda e che si distingua tra gli altri mezzi di comunicazione come un ottimo canale per portare potenziali acquirenti nei punti vendita.

Individuate le motivazioni del restyling del sito, risulta ora necessaria una panoramica sulla sua struttura in modo da avere un'idea degli strumenti utilizzati e delle innovazioni apportate per conseguire gli obiettivi prefissati. Sarà, infatti, su questi strumenti che si concentrerà in seguito l'analisi utile a valutare la loro efficacia.

3.3.1 I CONTENUTI

Il sito, www.snaidero.it, è dotato di domain name corrispondente al nome dell'azienda che ne assicura immediatezza e facilità nella reperibilità anche se questa avviene attraverso i motori di ricerca.

Con l'accesso all'home page si nota come il sito, caratterizzato da colori caldi e da un tratto tradizionale, sia strutturato in più sezioni facilmente consultabili cliccando sui diversi collegamenti: news, catalogo veloce, e-mail-me, la scelta del cuore, la scelta della mente, essere snaidero e snaidero.

La sezione essere snaidero è la presentazione, particolarmente chiara e ricca di contenuti, dell'azienda. Si trovano informazioni sulla mission, sulla storia che indica le tappe fondamentali dello sviluppo dell'azienda, sulle news che la riguardano e il link al sito www.snaiderogroup.it in cui si presentano i partners dell'azienda.

Un'altra sezione dedicata all'immagine aziendale è quella delle news in cui si fa riferimento a elementi di comunicazione integrata quali la segnalazione di eventi o fiere. Si ritrova, infatti, un elenco in ordine cronologico non solo d'occasioni nelle quali l'azienda ha partecipato con l'esposizione dei propri prodotti, ma anche una sorta di rassegna stampa con l'indicazione delle riviste sulle quali viene pubblicizzata.
L'ultimo collegamento che vuole ribadire la credibilità del gruppo Snaidero è quello che linka al sito della società sportiva sponsorizzata dall'azienda: "Se lo sport rappresenta da sempre uno straordinario veicolo di comunicazione, la pallacanestro in particolare si rivolge ad un pubblico che corrisponde perfettamente al target Snaidero: quasi automaticamente si è pensato quindi di comunicare attraverso questa disciplina sportiva che, oltretutto, aveva già solide radici all'interno dell'azienda".
Il collegamento e-mail-me offre la possibilità di interagire con l'impresa e, a seconda delle informazioni che si vogliono ottenere, si opera una scelta tra i vari indirizzi messi a disposizione:

- per saperne di più, in merito a rivenditori e promozioni;
- per un approfondimento sul servizio Snaidero dove si fanno richieste circa i ricambi per i modelli fuori produzione, l'identificazione di un punto vendita presso il quale trovare uno specifico modello e l'eventualità di un credito al consumo;
- per usufruire dell'assistenza, che da la possibilità di segnalare disservizi sulla consegna e dei punti vendita;
- per diventare partner;
- per prendersi cura della propria Snaidero.

Sempre nell'ottica di instaurare una relazione di feedback con il consumatore e, quindi, comprendere le sue esigenze e fornirgli un servizio migliore, nel modulo da compilare per le eventuali segnalazioni sono anche richieste dati del cliente.
Si trovano informazioni commerciali cliccando sul collegamento catalogo veloce, sezione che si presta ad essere una sorta di vetrina dell'azienda: i prodotti vengono presentati suddivisi a seconda dello

stile e del modello per consentire all'utente di indirizzarsi fin da subito verso un prodotto il più possibile conforme alle sue esigenze. Si può scegliere tra il design contemporaneo, il moderno emozionale e il tradizionale rivisitato. La scheda prodotto si articola per ogni cucina in diverse parti: versioni, colori e finiture, particolari, accessori e sedie&tavoli. In fine, per facilitare il contatto e l'individuazione dei punti vendita si trova anche un collegamento con informazioni chiare e complete sui rivenditori, i modelli che hanno a disposizione in mostra e i servizi da loro offerti.

La vera innovazione del sito, però, sta nelle sezioni la scelta del cuore e la scelta della mente: si offre la possibilità al potenziale cliente di vivere la scelta del modello della propria cucina in modo, rispettivamente, emozionale o razionale. Questo nell'ottica di una fruizione di internet che deve sì rispondere ad un'esigenza informativa ma deve anche riscontrare gradimento e piacevolezza, a conferma anche della nuova immagine dell'azienda che Snaidero che vuole trasmettere.

Nella sezione la scelta del cuore troviamo i seguenti links : life style, di che cucina sei, progetta la tua cucina, feng shui e guardati attorno, che, nonostante l'impossibilità di acquistare on line, offrono un ampio supporto alla decisione di acquisto. Life style è un catalogo basato sulla personalità dell'utente che può scegliere tra sei diversi profili caratteriali: dinamico, curioso, creativo, sereno, raffinato e schietto. Nel caso non riuscisse a riconoscersi in nessuna descrizione, il cliente ha la possibilità di completare il test di che cucina sei? che lo porta ad identificare una serie di cucine affini alle scelte evidenziate. Un'altra iniziativa che arricchisce i servizi offerti al visitatore del sito nella fase della scelta è progetta la tua cucina: il cliente, rispondendo ad alcune domande può farsi una prima idea della soluzione compositiva che maggiormente si adatta alle sue necessità arredative. Al termine di questa progettazione semplificata si deve anche indicare un rivenditore a cui rivolgersi per poter avere una un'elaborazione più dettagliata e, quindi, operativa. Allo stesso tempo, il rivenditore viene automaticamente informato del

nominativo del potenziale cliente e della scelta che lui ha fatto in modo da avere già un'idea delle sue esigenze e delle preferenze. La sezione feng shui dà l'opportunità di arredare la propria cucina nel rispetto della natura ed infine entra nella tua cucina che permette di elaborare un'immagine più completa della cucina in quanto la si vede secondo diverse prospettive ed angolazioni, in modo da rendersi maggiormente conto delle dimensioni e del rapporto tra le varie componenti e di capire l'esatto sviluppo spaziale della cucina nel suo insieme.

La sezione la scelta della mente risponde invece all'ideale informativo richiesto ad internet in quanto strumento di comunicazione, infatti, presenta dei collegamenti a pagine che in modo chiaro ed esaustivo danno informazioni circa la qualità dei prodotti e la loro garanzia, sui consigli pratici di come pulirli e mantenerli in buono stato, sul rispetto dell'ambiente che Snaidero segue nella sua produzione ed infine sulla possibilità per un acquirente di ottenere un credito al consumo per il pagamento personalizzato della sua cucina.

3.3.2 LA STRUTTURA

Oltre all'organizzazione dei contenuti bisogna anche soffermarsi sulla particolare struttura del sito: gli autori hanno utilizzato come linguaggio di descrizione delle pagine l'html e hanno presentato i documenti in esso contenuti suddividendo la stessa pagina in vari riquadri differenti, i frame. Questa soluzione offre ai progettisti dei siti un modo per mantenere determinate informazioni visibili, facendo scorrere o addirittura sostituirne delle altre. Un documento HTML organizzato in frame è chiamato documento frameset. Questa soluzione porta di sicuro a dei vantaggi tangibili: non costringe a ricaricare tutta la pagina, accelerando così la navigazione dell'utente all'interno del sito e facendo risparmiare banda anche dal lato server, consente ai webmaster di non ripetere le parti comuni nelle varie pagine del sito dal momento che il contenuto delle pagine è organizzato in riquadri e inoltre, mantenendo fisso su un lato del

monitor il menù di navigazione e facendo scorrere sull'altro il contenuto, la visualizzazione della pagina risulta più gestibile anche a basse risoluzioni. La struttura di un frameset differisce da sito a sito, ma anche all'interno di uno stesso sito da pagina a pagina. Tuttavia, prendiamo un caso standard per valutare il problema che si insinua nella valutazione dell'attività esercitata dagli utenti su un sito costruendo le pagine seguendo l'espediente dei frame.

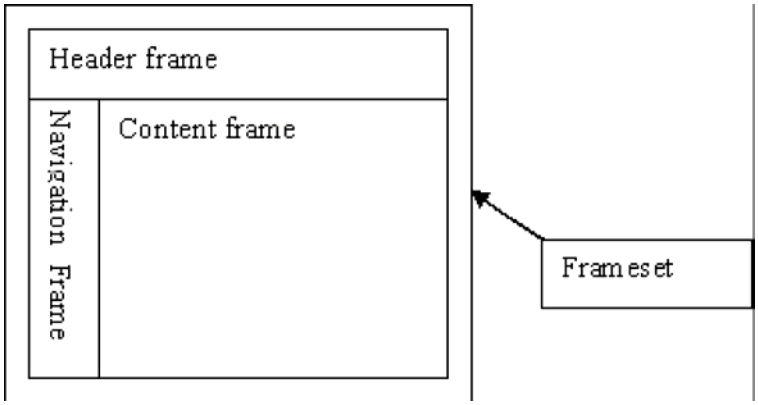


Figura 1: tipica struttura di una pagina costruita con frames

Nell'esempio particolare il frameset è suddiviso in tre riquadri; se dal punto di vista dell'utente la pagina viene visualizzata come una, dal lato server risulta che il visitatore si muova all'interno di essa spostandosi da frame a frame nel seguente modo: frameset, header frame, navigation frame e content frame. Questa sequenza di pagine, quindi, viene registrata all'interno dei log file portando a dei risultati falsati nella loro elaborazione. Nel caso specifico del sito Snaidero, le pagine sono costruite seguendo 3 diverse strutture del frameset:

1. header frame-content frame;
2. header frame-navigation frame-content frame;
3. header frame-navigation frame-content1 frame-content2 frame-content3.

La descrizione dell'organizzazione del sito sia nei contenuti sia nella costruzione delle pagine pone in evidenza i punti sui quali si dovrà prestare maggiore attenzione nell'analisi utile a valutare il successo e la qualità del sito stesso e a, eventualmente, suggerire azioni correttive.

3.4 OBIETTIVI DELL'ANANLISI

Il punto di partenza per valutare il successo di un sito è quello di specificare il problema, relazionando, quindi, il concetto di successo agli scopi oggettivi per i quali lo stesso sito è stato pensato e ai dati che si hanno a disposizione per l'analisi. Come introdotto in precedenza il nuovo sito Snaidero è stato progettato con la prospettiva di riqualificare l'immagine dell'azienda e di mettere a punto un efficace strumento di interazione che fungesse, anche, da canale per guidare i visitatori nei punti vendita come potenziali acquirenti. Per quanto riguarda il primo obiettivo si necessita di un campione rappresentativo di utenti al quale somministrare un questionario che valuti in modo appropriato l'eventuale cambio di percezione della realtà Snaidero. Questo approccio ha degli indubbi svantaggi, tra i quali i costi molto elevati per la costruzione di un ambiente sperimentale adeguato, la difficoltà nella selezione di un gruppo d'utenti rappresentativo da inserire nel campione per la particolarità dell'utenza di internet ed infine per il fatto che l'analisi condotta deve anche dare dei concreti suggerimenti per un possibile miglioramento.

La nozione di successo per il sito Snaidero in questo specifico caso viene,quindi, relazionata esclusivamente al grado di interazione che questo media riesce a stabilire con gli utenti che lo visitano. L'approccio, in questo caso, è più accessibile in quanto si tratta di concentrare l'attenzione sull'attività che gli utenti esercitano sul sito quantificandola in una serie di indicatori e guardando ai percorsi di visita che seguono per ottenere le informazioni a loro necessarie. I dati di cui si ha bisogno sono i log file e la metodologia per elaborarli guarda alle tecniche di data mining. Queste, tuttavia, non sono sufficienti in quanto si necessita di un modello ideale del comportamento di navigazione cosicché i risultati ottenuti dalle elaborazioni siano ad esso confrontabili e utili a fornire indicazioni sul miglioramento della struttura del sito.

Quando si fa riferimento ad un modello di navigazione ideale, si hanno in mente valori e metriche che i risultati ottenuti dall'elaborazione dei log file dovrebbero confermare. Facendo riferimento, in particolare, ai percorsi di visita si potrebbe definire una sorta di modello di consultazione delle risorse disponibili: l'ideale sarebbe che in seguito alla consultazione del catalogo l'utente avesse come target le pagine dedicate alle informazioni sui rivenditori, in quanto pensate come un reale strumento per acquisire nuovi contatti. La consultazione del catalogo può avvenire cliccando nell'home page sia sul collegamento catalogo veloce, a conferma che il sito viene esclusivamente visto come una vetrina, sia cliccando sulla scelta della mente o del cuore che da l'opportunità di scegliere la propria cucina in modo razionale od emozionale. Di conseguenza ci si aspetta che nell'elaborazione dei percorsi di visita, ci siano valori molto significativi degli indici che relazionano le pagine del catalogo a quella in cui vengono fornite indicazioni sui rivenditori. Inoltre, siccome il sito è stato costruito con l'intenzione di avere maggior feedback con la clientela ci si aspetta un valore molto alto nella consultazione della sezione e-mail-me dove l'utente ha a disposizione vari indirizzi mail ai quali rifarsi per avere informazioni di ogni tipo.

Bisogna fare un discorso a parte per quanto riguarda la sezione la scelta del cuore: al suo interno sono presenti, come detto prima, dei servizi aggiuntivi che aiutano il navigatore nella scelta della propria cucina e che sono del tutto innovativi in quanto frutto di un progetto in cui si pensa ad internet non solo come strumento di comunicazione ma anche di interazione e di svago. La possibilità che questa sezione risulti dall'analisi poco consultata pone in essere due tipologie di problemi: l'uno è la poco funzionale presentazione dei contenuti e l'altro è l'impossibilità di comunicare, anche se solo in parte, l'immagine di Snaidero come azienda che è in sintonia con le emozioni della propria clientela. Tuttavia, se dall'analisi si dovesse riscontrare questa bassa consultazione non si potrebbero fornire utili suggerimenti per migliorare l'usufruibilità dei servizi: non si

riuscirebbe a capire con i dati a disposizione se la causa è attribuibile alla poco chiara presentazione dei contenuti o al fatto che la maggior parte degli utenti che ha consultato il sito non faccia parte di quella schiera di clienti che vive la scelta della propria cucina in modo emozionale. Tuttavia, si cercherà la migliore interpretazione solo avendo i dati sottomano.

Lo studio dei percorsi di visita deve essere, però, spiegato anche in relazione ad alcune statistiche descrittive come in particolare alla media delle pagine per visita, al tempo medio di visita e alle single access pages e alla lista delle keyword che portano al sito. La media delle pagine per visita è un dato importante ma è opportuno non considerarlo come valore assoluto: va pesato in funzione della quantità delle pagine disponibili sul sito, della profondità con cui viene trattato un argomento e del valore di session time out impostato. Per quanto riguarda il valore di session time out questo è stato fissato a 30 minuti, uno standard per le elaborazioni di log file che hanno come obiettivo la valutazione della visibilità di un sito. Considerando, invece, il numero delle pagine del sito e della profondità con cui un argomento viene trattato ci si aspetterebbe un valore minimo di 4 pagine. Questa considerazione viene fatta assumendo una navigazione base del tipo: home page, catalogo veloce, primo modello di cucina, secondo modello di cucina, rivenditore. In realtà, visto e considerato che per ogni modello di cucina ci sono altre cinque sezioni consultabili riguardanti le finiture, gli accessori e che il sito offre all'utente svariati servizi per supportarlo nella scelta della cucina a lui più congeniale ci si aspetterebbe un valore più alto. Anche per il valore del tempo medio di visita vanno fatte le stesse considerazioni: più risulta essere elevato più si dimostra con ragionevolezza un discreto livello di interesse, tuttavia bisogna distinguere la tipologia del sito preso in analisi. Nel caso del sito Snaidero, questo può essere classificato come un sito di presentazione che ha al suo interno sia una informazione testuale, con immagini ma anche un accenno di interattività. Ci si aspetta, quindi, un valore della statistica che si

aggira intorno ai quattro minuti. Per quanto riguarda le single access pages, questa statistica indica quanti visitatori, raggiunto il sito Snaidero, non hanno continuato la loro visita perché non hanno trovato quello che vogliono. Verosimilmente il valore dovrebbe essere molto basso a conferma che il sito ha raggiunto un posizionamento corretto rispetto al target che il progetto aveva prefissato. L'analisi si fa ancora più elaborata se si tiene conto del fatto che il sito in questione può essere visitato da utenti che effettivamente volevano consultarne i contenuti o che ci sono capitati per caso. A questo proposito si può portare un esempio tanto banale quanto esplicativo: un indice di permanenza sul sito che raggiunge una media di 1 minuto. Di primo acchito il risultato sembrerebbe piuttosto scoraggiante, ma si vada a distinguere i visitatori in base alle keyword digitate nei motori di ricerca. Le situazioni riscontrabili potrebbero essere di due tipi: l'una che vede i visitatori a tema essere pochi ma con un tempo medio di visita molto elevato e per contro numerosi visitatori fuori tema che restano giusto il tempo di capire che il motore di ricerca li ha indirizzati verso un sito che non ha contenuti attinenti a quello che loro stavano cercando, l'altra in cui non si nota una distinzione tra visitatori a tema e non: indistintamente le tue tipologie di visitatori rimangono lo stesso tempo sul sito. Indubbiamente la seconda situazione è molto più sconfortante della prima in quanto si viene a conoscenza che gli utenti che rientrano nel target non hanno trovato nel sito proposto quello che cercavano e scoraggiati nella visita lo hanno lasciato. Stessa identica considerazione può essere fatta per il numero di pagine visitate. I parametri di valutazione in precedenza presentati, quindi, se ben ponderati sono interessanti e utili per comprendere se un progetto di presenza sul web ha raggiunto la competitività e gli obiettivi per i quali era stato pensato. Anche sull'accuratezza dei dati registrati nei log file non si può dubitare in quanto questi considerano la totalità degli utenti che hanno visitato il sito: a differenza di altre indagini statistiche qui si ha l'opportunità di guardare alla popolazione di riferimento e non ad un campione

selezionato. L'aspetto a cui, invece , si deve prestare maggiore attenzione è l'affidabilità e la precisione con cui i vari strumenti di misura utilizzati implementano il processo di analisi.

3.5 STRUMENTI

Per l'elaborazione delle statistiche di traffico si è operata una scelta tra i numerosi software che si sono trovati in rete. Questo segmento, infatti, negli ultimi periodi ha avuto uno sviluppo sensibile dato dalla sempre crescente richiesta di valutare l'attività che gli utenti esercitano nel proprio sito al fine di migliorare il servizio offerto ma soprattutto quantificare i vantaggi in relazione agli investimenti sostenuti. Le specifiche iniziali erano quelle di trovare un software che fosse un log file analyzer, che non necessitasse di un'installazione lato server e che fosse free o che almeno desse la possibilità di testarlo gratuitamente per un periodo di tempo. Esiste, infatti, una distinzione primaria tra software per l'analisi del traffico di un sito in cui si distinguono due tipologie: l'una è quella dei log file analyzer che analizzano i log file in cui è registrata l'attività del server web e l'altra è quella dei tracker in cui i dati registrati nei file di log sono dati dal caricamento nelle pagine del sito che si ritengono più interessanti di un tag, ossia di un frammento di codice nella sintassi html. Per le differenze tra i due metodi di misura si faccia riferimento alla pagina… Inoltre, si necessitava di un software che non dovesse essere installato su lato server in quanto i log file a disposizione erano salvati in una cartella del computer utilizzato per l'elaborazione. Solo marginalmente si è fatto riferimento alle tipologie di log file e server web supportate in quanto di questi tempi i software sono molto adattabili sotto questo punto di vista. I software in questione sono ClickTracks Analyzer version 5.1.7 e NetTracker Enterprise version 7.5. Entrambi si sono rivelati essere soluzioni molto sofisticate per l'analisi del traffico del sito, infatti, oltre alle statistiche base forniscono elaborazioni che guardano ai percorsi di navigazione, ai risultati delle campagne pubblicitarie

sulle quali si è investito per far conoscere il sito e ad altre utili elaborazioni per siti più propriamente di e-commerce.
Da questa prima operazione di selezione sono, quindi, stati presi in considerazione due software sui quali operare una più accurata valutazione delle prestazioni e funzionalità, in base alla quale scegliere in seguito quello più congeniale all'elaborazione da condurre.

3.5.1 CONFRONTO TRA SOFTWARE COMMERCIABILI

L'approccio seguito per il confronto si è basato su un'analisi-campione dei log file di uno dei mesi a disposizione, in questa prova si sono testati i programmi prestando particolare attenzione ai punti qui di seguito:

- facilità nell'utilizzo: questa è una specifica molto richiesta da tutti gli utenti e per una corretta valutazione bisogna guardare contemporaneamente alla difficoltà nell'installazione e nell'utilizzo del software, all'intuitività dell'utilizzo e alla veloce comprensibilità dei reports generati;
- performance: la cosa più ovvia per misurare le performance di un software è quella di misurare il tempo di elaborazione necessario per generare un report. Nonostante questo sia un ottimo punto di partenza, tuttavia bisogna anche sottolineare che non è così realistico in quanto ogni software genera reports tra loro molto diversi che, quindi, fanno la differenza sul tempo di elaborazione necessario;
- accuratezza: questo è uno degli obiettivi sui quali si sta ancora lavorando. Un esempio pratico può essere fatto riconducendosi al concetto di visita che per ogni software può essere diverso: c'è quello che la definisce solo in base al session time out e quindi in base ad una semplice temporizzazione, quello invece che mette in relazione gli indirizzi IP con il browser, il referer o la localizzazione geografica, resa possibile utilizzando un programma di reverse DNS lookup;

- documentazione e help on-line: visto e considerato che per questo confronto si era ad un primo approccio con i software analizzati questo punto è stato essenziale. L'aspettativa era di trovare una guida all'utilizzo del software che fosse intuitiva, che permettesse una consultazione veloce, esaustiva ma non troppo tecnica;
- report pre-definiti: i report pre-definiti devono soddisfare in modo più ampio possibile le esigenze espresse dall'analisi che si deve condurre, in questo caso si necessita delle statistiche inerenti al tempo media di visita, alla media delle pagine per visita, alle single access pages;
- filtri da impostare nelle elaborazioni: di solito l'attenzione, dopo un'analisi generale, si sposta su un sottoinsieme di dati per valutare se ci siano delle differenze che prima rimanevano implicite. Nei report si possono applicare filtri? quali informazioni possono essere filtrate?;
- trend dei report: utile per la comparazione delle statistiche nel tempo e identificare i cambiamenti per trovare delle spiegazioni ragionevoli;
- report riguardanti i referer e le keyword: utili per verificare il posizionamento del sito, la sua visibilità e reperibilità in rete;
- grouping: opzione che permette di raggruppare i dati dando così la possibilità di scoprire e valutare diversi comportamenti nell'attività esercitata sul sito.

Prima di passare alla valutazione dei software secondo i punti presentati è d'obbligo, però, fare una premessa: i programmi sono stati valutati in relazione agli obiettivi dell'analisi che si stava portando avanti per cui è possibile, se non certo, che molte loro potenzialità siano state sottovalutate se non addirittura nemmeno considerate. Questo in particolare riguarda tutte le elaborazioni riguardanti le campagne di marketing attuate e più specificatamente per NetTracker tutti i report che sono molto più appropriati per valutare il successo e la produttività esclusivamente di un sito di e-commerce.

ClickTracks, in quanto a facilità nell'utilizzo, si è rivelato essere il più intuitivo: l'interfaccia è molto semplice, accattivante e, molto più importante, dotata di una toolbar che permette una navigazione tra le sue molteplici funzionalità in modo semplice e veloce in quanto ognuna di queste è rappresentata da un'icona. Anche i report sono organizzati in modo da essere leggibili a colpo d'occhio e, quindi, poter trovare facilmente quello che si sta cercando. NetTracker al contrario è molto più articolato e complesso nella sua struttura e ciò comporta in un primo approccio una sensibile difficoltà in una visione d'insieme delle sue prestazioni. Questa particolarità, tuttavia, deve essere relazionata al grado di dettaglio con cui i report vengono generati e alla possibilità di interagire con il software in profondità anche per quanto riguarda le opzioni di configurazione.

In quanto a performance bisogna notare la notevole lentezza di NetTracker nel caricare i log file: per il mese campione che sono 1,80GB di dati si aggira sui 45 minuti quando per ClickTracks la fase di inserimento dati non prende più di 20 minuti. Si rovescia però la situazione quando si va a valutare i tempi delle elaborazioni che con NetTracker risultano essere inferiori, tuttavia sappiamo che questa considerazione è poco attendibile in quanto i report generati tra i due software sono molto diversi l'uno dall'altro.

In quanto ad accuratezza i software identificano in modo diverso la sessione di visita: NetTracker combinando le sole informazioni raccolte nei campi IP e useragent e ClickTracks guardando anche al campo del referrer. Entrambi, quindi, vanno oltre alla semplice temporizzazione delle visite: non basta che per lo stesso IP la richiesta successiva avvenga dopo il tempo del session time out per considerare l'inizio di una nuova visita ma si combinano le informazioni raccolte nel campo dello user agent e anche del referrer in modo che l'identificazione di una sessione diventi più accurata soprattutto per quelle visite che sono supportate da provider che cambiano l'IP dell'utente ad ogni sua richiesta.

Per quanto riguarda guide ed help on-line per entrambi i programmi la valutazione è positiva: i supporti sono esaustivi e comprensibili.

Oltre a statistiche comuni quali numero medio di pagine visitate in una sessione, tempo medio di permanenza sul sito, pagine più visitate, lista delle entry ed exit page, classifica delle keyword e dei referrer, i report generati per default sono molto diversi sia per tipologia che per grado di dettaglio. ClickTracks offre la particolarità di valutare come gli utenti navigano all'interno del sito sovrapponendo i risultati dell'analisi alle pagine del sito stesso: il programma diventa così una sorta di browser e per la pagina visualizzata si hanno report che indicano le pagine di provenienza, le pagine in seguito visitate e il tempo in media speso su quella pagina. Un'altra particolarità di ClickTracks è che permette di raggruppare le visite a seconda di alcune variabili tipo il tempo di permanenza sul sito, le keyword utilizzate per la ricerca e la pagina di entrata di uscita. Questa particolare elaborazione da la possibilità di concentrare l'attenzione solo su una parte delle visite e di confrontare queste ultime con la totalità in modo da rendere espliciti particolari comportamenti di visita. NetTracker, invece, punta più sull'aspetto quantitativo dell'utilizzo del sito piuttosto che su quello qualitativo, infatti, genera report riguardanti le directory più visualizzate, gli errori più frequenti, la distribuzione delle visite tra le settimane e nella stessa settimana tra i vari giorni, opera una traduzione, dove possibile, dei singoli IP ed infine visualizza i percorsi di visita più seguiti. Inoltre, per ogni report generato permette di scendere nel dettaglio della singola visita al fine di identificare particolari comportamenti.

Per quanto riguarda i filtri che si possono impostare ClickTracks risulta molto meno personalizzabile: navigando tra le pagine del sito riconosce automaticamente i frame in cui viene suddivisa la pagina e consiglia di considerare il content frame. L'unica nota negativa in questa fase è che alcuni frame delle pagine del sito preso in considerazione sono pagine esterne per le quali, quindi, non è possibile avere i risultati dalle elaborazioni. Per gli spider, inoltre, non si può apportare nessuna modifica, se non dal lato server, in quanto il programma per default elimina tutte quelle righe in cui il

campo della risorsa richiesta comincia per robot.txt. NetTracker, invece, non riconosce automaticamente né i frame né gli spider per cui è dotato tra le sue opzioni di configurazione anche di appositi filtri di esclusione per entrambe le variabili. Questo perché con NetTracker è possibile anche analizzare il comportamento di queste visite per valutare la propria visibilità nei motori di ricerca.

Una considerazione finale valida per entrambi i programmi presi in considerazione, ma in linea di massima per tutti i software commerciabili di questo tipo, è che questi sono poco configurabili da parte dell'utente: non c'è la possibilità di costruire altri report oltre a quelli generati per default dal programma.

Il confronto tra i due software ha portato alla scelta di ClickTracks e la variabile che ha inciso maggiormente è stata la tipologia dei report che il programma riesce ad implementare: la possibilità di visualizzare in modo sintetico come gli utenti navigano all'interno del sito e di raggruppare le visite al fine di scovare particolari comportamenti risulta più in linea con gli obiettivi dell'analisi che si sta portando avanti.

3.5.2 CONFRONTO TRA SAS E SOFTWARE COMMERCIABILI

Il maggior svantaggio derivante dall'utilizzo di questi programmi commerciabili sta nel fatto che sono poco configurabili dall'utente e hanno come prerogativa solo quella di dare un aspetto più leggibile a masse di log file che risulterebbero, altrimenti, poco interpretabili. Si è pensato, quindi, di prendere in considerazione anche un software di data mining completamente configurabile dall'utente quale SAS. SAS Enterprise Miner combina un sistema di analisi statistica e reporting con un'interfaccia utente grafica (GUI) di facile utilizzo. Il SAS è stato pensato, infatti, per implementare con successo progetti di Data Mining che richiedono una soluzione che offra analisi statistiche e tecniche di reporting avanzate ma di facile utilizzo. L' Enterprise Miner offre una vasta gamma di funzionalità di Data Mining che soddisfano le esigenze di utenti diversi in quanto è in

grado di risolvere problemi di identificazione di clienti maggiormente redditizi e i motivi della loro fedeltà, determinare il motivo per cui i clienti passano alla concorrenza, determinare quale combinazione di prodotti è più probabile che i clienti acquistino e tra le tante applicazioni anche di analizzare i dati di un sito Web per migliorare le strategie di e-commerce. Esso è dotato di un'interfaccia grafica user-friendly in grado di facilitare lo svolgimento delle analisi e di tool di reportistica per condividere i risultati ottenuti all'interno dell'azienda. L'interfaccia, di tipo point-and-click, è progettata, tenendo presenti due tipologie di utenti: gli analisti aziendali, che pur avendo nozioni minime di statistica possono comunque navigare rapidamente e facilmente nel processo di Data Mining e gli esperti statistici, che spesso vogliono esplorare i dettagli ed accedere ai sottostanti processi analitici per ottimizzarli. La GUI utilizza oggetti comuni del desktop come barre degli strumenti, menù, finestre e pagine di dialogo per fornire ad entrambi i gruppi una gamma completa di strumenti di Data Mining. Gli elementi della GUI possono essere utilizzati per implementare la metodologia SEMMA: SAS suggerisce come approccio pratico al processo di Data Mining una metodologia suddivisa in cinque fasi rappresentata, appunto, dall'acronimo SEMMA. In sintesi, le fasi della metodologia SEMMA sono:

- Sample: prevede l'estrazione di una porzione di dati abbastanza grande da contenere ancora informazioni significative, ma sufficientemente piccola da risultare velocemente analizzabile;
- Explore: serve a scoprire in anticipo potenziali relazioni ed anomalie nei dati e per capire quali possono essere quelli d'interesse;
- Modify: include la creazione, selezione e trasformazione delle variabili al fine di mettere a punto il processo di costruzione del modello;

- Model: in questa fase vengono ricercate automaticamente le variabili significative e i modelli che forniscono le informazioni contenute nei dati;
- Assess: è la fase finale nella quale viene valutata l'utilità e l'affidabilità delle informazioni scoperte nel processo di Data Mining, portando nell'ambiente di produzione le regole estratte.

I principali componenti della GUI comprendono:

- Diagram Workspace: utilizzato per costruire, modificare ed eseguire diagrammi di flusso del processo di analisi (PFD);
- Tools Palette: contiene un sottoinsieme degli strumenti di Enterprise Miner, utilizzati per costruire i PFD nel Diagram Workspace;
- Nodes: una serie di icone, organizzate secondo la metodologia SEMMA, che consentono all'utente di eseguire i diversi passi di un progetto di Data Mining, come accedere ai dati, analisi e reporting.

Affinché sia possibile un confronto significativo tra SAS e gli altri software fino a qui considerati l'elaborazione campione sarà fatta esclusivamente sul mese di marzo.

Il processo di analisi dei log file con SAS risulta molto più articolato in quanto sia l'operazione di pulizia sia quella di identificazione delle sessioni di visita devono essere implementate dettagliatamente dall'utente.

Per quanto riguarda il processo di pulizia, la prima fase vede l'identificazione degli spider: è necessario a questo proposito avere a disposizione una lista aggiornata degli automatismi che navigano il web. Nel caso specifico si è consultata la sezione del sito www.submission.it in cui è presente un data base aggiornato degli spider utilizzati dai motori di ricerca. Tuttavia sarà piuttosto improbabile identificare tutti i robots in quanto alcuni riescono a non farsi riconoscere come tali nel momento in cui visitano il sito. Si procede, quindi, con l'identificazione dei frame da inserire nell'analisi: in continuità con l'esigenza che i risultati siano

confrontabili con i software precedentemente considerati si ritiene opportuno considerare solo i contente frames delle pagine visitate. L'ultima fase del processo di pulizia vede l'eliminazione di tutti quei campi registrati nel file di partenza che però non sono utili al fine dell'analisi: il date set risultante sarà così composto dalle variabili IP dell'utente, ora e data della visita che sono sintetizzate in un'unica variabile time, pagina visitata e user agent.

A questo punto si deve procedere con l'identificazione delle sessioni di visita: fase elementare per ottenere un data set di partenza che possa fornire delle analisi attendibili e significative. Come già visto in precedenza ci sono vari approcci da seguire per implementare questa fase ma, ancora, affinché i risultati siano il più possibile paragonabili a quelli ottenuti con i software precedentemente considerati si sceglie di utilizzare l'algoritmo che vede la combinazione delle informazioni ricavate dal campo time, user agent e IP. Questo è così costruito:

1. lettura della prima riga del data set tenendo in memoria i dati dei tre campi;
2. lettura della seconda riga e confronto. A questo punto si possono verificare tre situazioni:
    a. il campo dell'IP è diverso da quello precedente → inizia una nuova visita
    b. il campo dell'IP è uguale a quello precedente ma differisce il campo dello user agent → inizia una nuova visita
    c. sia il campo dell'IP sia quello dello user agent sono uguali. Bisogna, quindi, guardare al campo time:
        i. il tempo passato dalla richiesta precedente è superiore ai 30 minuti → inizia una nuova visita
        ii. il tempo passato dalla richiesta precedente è inferiore ai 30 minuti → questa è la seconda pagina visitata nella sessione di visita appena identificata.

La fase dell'identificazione delle sessioni di visita comporta una nuova ridefinizione delle variabili contenute nel data set: la variabile dell'IP viene sostituita da una nuova variabile costituita da un codice numerico che identifica univocamente la visita mentre nessun cambiamento per le variabili che segnalano la pagina visitata e il giorno e l'ora della richiesta. Costruito il data set di partenza si può passare all'analisi delle associazioni e delle sequenze tra le pagine visitate per individuare percorsi di navigazione più frequentemente seguiti dal visitatore al fine di ottimizzare la creazione del sito e facilitare quanto più possibile la navigazione. Un'ulteriore tipologia di analisi possibile può esser quella che permette di definire i profili dei visitatori: il data set ottenuto deve essere integrato con nuove variabili come il numero di clicks, la durata della connessione, il giorno, una variabile binaria che indica il fatto che la visita al suo interno includa o meno una certa pagina, il giorno della visita ecc… Tuttavia prima di proseguire nell'analisi si è ritenuto opportuno calcolare alcune statistiche descrittive tipo il numero medio di pagine visitate in una sessione e il tempo medio di permanenza sul sito. I risultati ottenuti con SAS per i valori di queste statistiche sono molto diversi rispetto a quelli ottenuti con ClickTracks e NetTracker:

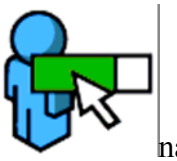
	Tempo medio di permanenza	Numero medio di pagine visitate
SAS	1,5 min	4
ClickTracks	6 min	17
NetTracker	8 min	20

Una possibile spiegazione della differenza tra questi risultati dopo che si è cercato di rendere il più possibile simili le impostazioni delle operazioni di pulizia potrebbe essere data dalla stima o meno della durata di visita dell'ultima pagina di ogni sessione: l'ipotesi talvolta semplificativa è che l'ultima pagina vista in una visita abbia durata zero (soltanto perché l'ultima interazione con il server da parte del client è quella di richiesta dell'ultima pagina) ma alcuni software possono essere configurati per stimare anche la durata di permanenza sull'ultima pagina. Questa possibile soluzione

è però scartata visto che nessuno dei software è configurato per stimare la permanenza sull'ultima pagina di una sessione di visita ma soprattutto per il fatto che i risultati non differiscono solo per tempo di permanenza in media ma anche per il numero di pagine visitate. Da testimonianze trovate nel web, questo è uno dei problemi che maggiormente incide sulle valutazioni dei siti: si è infatti stimato che elaborazioni fatte con software diversi possono differire tra loro anche per un 50%. Nel caso particolare la percentuale è superata ampiamente soprattutto se si paragona SAS a NetTracker. Questa notevole discordanza nei risultati ha portato un'indecisione su che strumento utilizzare per proseguire nella valutazione del sito Snaidero. La scelta è ricaduta ancora una volta su ClickTracks per due motivi: le potenzialità offerte da questo software, in quanto a tipologia di elaborazioni disponibili, sono più in linea con gli obiettivi dell'analisi utili alla valutazione del successo del sito Snaidero ed inoltre in termini di accuratezza dei risultati ottenuti per il mese-campione questo si è dimostrato più affidabile.

3.5.3 PRESENTAZIONE DI CLICKTRACKS

Operata la scelta su quale software utilizzare per l'elaborazione dei dati è d'obbligo una panoramica per presentare nel dettaglio le sue funzionalità. Innanzitutto bisogna sottolineare che i report sono suddivisi in varie sezioni, tutti identificati da un'icona:

-  navigation report: l'interfaccia è un browser, dotata di toolbar con controlli standard tipo l'icona di back, forward, stop, reload, e home. La particolarità è data dal fatto che si può navigare all'interno del sito, sovrapponendo alle sue pagine, dati di sintesi per valutare come gli utenti navigano al suo interno e quali risorse visualizzano. Le statistiche che si possono consultare per la pagina visualizzata sono:la percentuale di utenti che la visita, il tempo medio trascorso dall'utente prima di consultare le

risorse in esso contenuta, il tempo medio di permanenza (questa non include le visite che terminano con questa pagina in quanto i log file per come sono registrati non lo permettono), la percentuale di visite che ha la pagina in questione come pagina d'entrata e quante come pagina di uscita. Sempre nello stesso report vengono indicate anche in percentuale le pagine visualizzate prima e quelle consultate dopo;

-  search report: combina le statistiche delle keyword con i motori di ricerca per valutare per ogni motore quale sia la parola più utilizzata. Questo tipo di report è utile per valutare l'efficacia di campagne PPC, ossia di quei meccanismi di on line advertising in cui chi promuove il proprio sito paga per ogni click che porta l'utente a visitarlo.
-  campaign report: questo report è strettamente collegato a quello precedente, infatti, per ogni campagna pubblicitaria attuata calcola i costi e il ritorno sull'investimento, il ROI;
-  site overview: fornisce alcune statistiche tipo il numero totale di visitatori, il tempo medio di permanenza, il numero medio di pagine visitate, la lista dei referer e delle keyword utilizzate per raggiungere il sito, le pagine più visitate, la lista delle pagine di entrata e di quelle di uscita;
-  options: sezione in cui si possono cambiare le impostazioni delle elaborazioni. In particolare si possono configurare il tempo di session timi out, la lista delle pagine da escludere e degli IP;

-  label visit: tutte le sezioni fino ad ora descritte danno la possibilità di valutare l'attività degli utenti esercitata sul sito nella loro totalità. Tuttavia risulta interessante anche distinguere queste visite in gruppi a seconda di diverse variabili per valutare se ci siano dei comportamenti particolari che prima rimanevano impliciti. Le variabili in questione possono essere la durata della visita, la pagina di entrata, quella di uscita, presenza di referer o meno, keyword Definite, quindi, le variabili con cui si vogliono suddividere le visite il programma implementerà nuovamente l'elaborazione dei dati e mostrerà i risultati, opportunamente distinti da colori diversi, per ognuna delle sezioni di report precedentemente presentate.

3.6 DEFINIZIONE DEI DATI DI PARTENZA E PULIZIA

I log file a disposizione riguardano un periodo che va da gennaio 2004 a maggio 2004 e sono registrati secondo il formato standard W3C Extended Log File che prevede i seguenti campi:

- date;
- time;
- c-ip;
- cs-username;
- s-sitename
- s-computername;
- s-ip;
- s-port;
- cs-method;
- cs-uri-stem;
- cs-uri-query;
- sc-status;
- sc-win32-status;
- sc-bytes;
- cs-bytes;
- time-taken;

- cs-version;
- cs-host;
- cs(User-Agent);
- cs(Cookie);
- cs(Referer).

Quelli utili all'elaborazione sono date, time, c-ip, cs-method, cs-uri-stem, sc-status, cs(user-agent) e cs(referer) che rispettivamente indicano: data, ora, indirizzo IP, metodo, URI della risorsa richiesta, codice di stato, informazioni sull'utente e sul referer. Al fine di rendere più precise le identificazioni delle sessioni di visita sarebbe stato utile avere a disposizione anche le informazione registrate nei campi cs(cookie) e cs-username ma il sito Snaidero non è stato configurato per inviare agli utenti che lo visitano i cookies ed, inoltre, al suo interno non esistono sezioni per accedere alle quali si necessiti di login. I rimanenti campi possono essere presi in considerazione se si vuole avere un'idea esclusivamente del carico di lavoro del server che ospita il sito.

L'elaborazione dei dati comincia con l'identificazione delle sessioni di visita, fase essenziale affinché i risultati elaborati siano accurati e quindi significativi. Prima, tuttavia, sono necessarie alcune operazioni di pulizia:

- eliminazione di tutte le righe che hanno nel campo cs-uri-stem un file con un'estensione diversa da html;
- eliminazione delle richieste con cs-method diverse da GET;
- eliminazione delle richieste con sc-status diverse da 304 in quanto solo queste ultime identificano che la pagina è stata correttamente visualizzata;
- eliminazione delle visite effettuate dagli automatismi del web. A questo proposito il software guarda al campo cs(useragent): se questo riporta all'inizio la dicitura robot.txt vuol dire che la visita in questione deve essere attribuita ad uno spider.

La fase dell'identificazione delle sessioni di visita comincia guardando al campo cs(referer): se questo si riferisce ad una risorsa esterna al sito senza dubbio identifica una nuova sessione. Di seguito l'algoritmo prende in considerazione prima il campo IP e poi quello cs(useragent): se ad uno stesso IP corrisponde un diverso useragent vuol dire che la sessione presente è terminata e ne è cominciata una nuova altrimenti la sessione in questione sarà considerata tale finché la richiesta successiva non avverrà oltre il tempo di session time out. Questo parametro può essere impostato a piacere dall'utente, nel caso specifico è stato fissato a 30 minuti che è uno standard utilizzato nelle analisi di questo genere. L'ultima fase di pulizia necessaria prima dell'elaborazione dei dati è la configurazione del filtro sui frame. Questo perché, se da un lato i frame velocizzano la visualizzazione della pagina, in sede di valutazione dell'attività esercitata sul sito dagli utenti comportano non pochi problemi. ClickTracks facilita il processo di identificazione dei frame automatizzando in parte questa fase: il programma riconosce la struttura di ogni singola pagina del sito e richiede all'utente di scegliere uno tra i frame che la compone preferendo tra questi il content frame. La scelta del content frame non è casuale. La motivazione va ricercata nel fatto che questo frame è l'unico che non può essere comune a due o più pagine. Proprio la particolarità dell'automatizzazione di questo processo, però, crea problemi nell'identificazione dei frames su alcune pagine:

- pagine come quella dei rivenditori e del gioco progetta la tua cucina hanno i content frame che appartengono al sito dei rivenditori Snaidero e che, quindi, risultano essere esterni. Per queste particolari pagine, quindi, non si può visualizzare il navigation report. Queste, però, sono molto importanti dal punto di vista degli obiettivi prefissati dall'analisi perché, guardando alla modalità con cui sono consultate, si può valutare il grado di interazione tra azienda e cliente che il

sito, come strumento di comunicazione, riesce a sviluppare. Per ovviare si fa riferimento al frameset e non al content frame: in questo modo saranno usufruibili solo ed esclusivamente i risultati generati del report di site overview;

- un altro problema si riscontra in particolare per la home page: i frame di questa pagina, infatti, sono ripetuti anche nelle pagine dedicate alla presentazione dei vari modelli di cucina. In questo modo, quando si visualizza le pagine del catalogo e si escludono i frame della home page non è più possibile consultare alcun indice calcolato nel navigation report per la pagina in questione. Anche per la home page però sono di rilevante importanza le informazioni che si possono trarre dal navigation report perché permettono di valutare come gli utenti, una volta entrati nel sito, si muovono e verso quale sezione dirigono la propria attenzione. La soluzione del problema sta nella consultazione del report di navigazione in due fasi distinte: l'una dedicata esclusivamente alla home page dove, quindi, nei risultati sarà compreso anche il relativo content frame e l'altra volta alla valutazione delle pagine del catalogo in cui sarà escluso il frame per la home page.

Per la significatività dei dati si è ritenuto necessario eliminare anche le visite effettuate da indirizzi IP uguali a 62.101.98.214. Questo IP, infatti, corrisponde a visite effettuate dal server che ospita il sito Snaidero. Si assume, a ragione, che queste non siano così significative per valutare il comportamento degli utenti esterni perché corrispondenti a persone che conosco bene il sito e che sanno dove trovare i contenuti che a loro interessano.

3.7 ELABORAZIONE DEI DATI

Fatte queste considerazioni si può partire con l'elaborazione dei dati. I log file dei 5 mesi non saranno trattati singolarmente in quanto l'interesse non è focalizzato su un confronto tra i mesi o sullo studio

di un trend ma più sul comportamento degli utenti. Per rendere i dati più affidabili si è pensato, quindi, di considerarli nella loro totalità.
Una prima valutazione sui risultati ottenuti è quella da farsi considerando le statistiche descrittive risultanti nel report di site overview. In particolare, si guardi alla classifica dei referer. Innanzitutto bisogna notare come il 60% dei contatti non sia mediato da altri siti o motori di ricerca: chi visita il sito Snaidero digita correttamente nella barra l'indirizzo www.snaidero.it. Questo grazie, soprattutto, ad un domain name uguale a quello dell'azienda che ne permette una facile e veloce reperibilità in rete ma anche ad un'immagine aziendale ormai consolidata e già fissata nel target di riferimento anche attraverso i più comuni mezzi di comunicazione quali riviste, televisione e radio. I primi posti della lista sono occupati da motori di ricerca che vedono svettare su tutti Google e Virgilio, il secondo con un sensibile distacco dal primo. Nei posti successivi si trovano MSN, Yahoo, Arianna, Altavista e Tiscali. Indubbiamente si può affermare che il sito Snaidero è ben piazzato all'interno dei motori di ricerca ma su questo punto si farà un approfondimento in seguito guardando anche alla classifica delle keyword. Sempre nella stessa classifica si guarda, quindi, ai referer, ossia ai siti che al loro interno contengono un collegamento al sito Snaidero. Guardando a questa tipologia di link si possono ottenere significative informazioni sulla visibilità che si riesce ottenere sui siti partner e sulla tipologia di siti che pubblicizzano di loro iniziativa il sito Snaidero. Primo fra tutti www.webmobili.it che si distingue come il primo portale in Italia dedicato esclusivamente all'arredamento e sul quale Snaidero è presente costantemente con un banner. Segue il sito www.bravacasa.it rivista nota per le idee e i consigli sull'arredamento della casa. Le prime posizioni della classifica portano un numero considerevole di contatti, dai 24831 di Google ai 526 del sito del Salone del Mobile che si tiene ogni anno a Firenze. Ma scendendo nella lista si può notare come la maggior parte dei referer porti dai 100 a solo un contatto, si pensi solo al fatto che la classifica è composta nella sua totalità da più di 900 voci. È su

queste che bisogna concentrare maggiormente l'attenzione per capire se è possibile delineare una nuova categoria di siti sui quali pubblicizzare il sito dell'azienda. Tuttavia, questa analisi non porta alcuna nuova informazione: la maggior parte sono siti di negozi di arredamento o di soluzioni on line che aiutano l'utente ad arredare la propria casa. Da sottolineare, invece, come i contatti generati arrivino sì dall'Italia ma anche da numerosi altri paesi europei quali Olanda, Germania e Francia a conferma della posizione con la quale il gruppo Snaidero si sta distinguendo in tutta Europa dopo il processo di internazionalizzazione che ha investito oltre che il settore commerciale anche quello produttivo.

Strettamente correlata alla classifica dei referer è quella della lista delle keyword. In testa alla classifica con una percentuale rispettivamente del 7,8% e del 3,5% troviamo le parole snaidero e cucine. A scendere, ma con un notevole distacco da queste, altre keywords in linea con la tipologia del sito: cucine componibili, Snaidero cucine, arredamento cucine, tavoli, mobili cucina, cucine design, cucine tradizionali, moderne e classiche. Si può senza dubbio affermare che il sito riesce a comunicare in modo corretto i contenuti in esso presentati. Per contro generano bassi contatti le parole chiave come Pininfarina, Iosa Ghini che sono alcuni tra i designers che hanno lavorato per l'azienda nella realizzazione di nuovi modelli di cucina e mobili ecologici. Soprattutto questa ultima keyword dimostra la scarsa efficacia nella comunicazione del gruppo Snaidero come azienda attenta nella sua produzione all'ambiente. Da notare che generano contatti anche i nomi dei singoli modelli di cucina e degli accessori proposti sul sito, segno che l'utente ha già raccolto in precedenza informazioni riguardo a quello che ora sta cercando in rete. Sarebbe utile indagare su questo punto per capire, quindi, la rete che funzione ha all'interno del processo di ricerca del target Snaidero: si preferisce trovare informazioni prima presso i punti vendita, riviste specializzate o prima in rete? Quale è la tipologia di utenti che compie una ricerca in rete rispetto al target Snaidero? Il traffico generato da referer e motori di ricerca è, quindi, nel

complesso consistente, consolidato, in linea con i contenuti del sito, positivamente correlato alle campagne pubblicitarie su stampa e alla presenza di banner e di link su pagine di siti partner.

Passando a valutare le statistiche descrittive di base vediamo che il tempo medio di permanenza sul sito si aggira intorno ai 5 minuti con una media di 15 pagine visitate. Per quanto riguarda il tempo medio di permanenza il dato è un po' quello che ci si aspettava mentre per le pagine visitate la media è abbastanza buona, anche se bisogna sottolineare che negli obiettivi fissati all'inizio dell'analisi questa era stata di gran lunga sottostimata. Avendo a disposizione questi dati si può anche calcolare la permanenza media su ogni singola pagina che viene ad essere pari a 20 secondi. Il dato risultante è piuttosto confortante anche in relazione al fatto che le pagine costituenti il sito hanno per contenuti immagini e file multimediali. La considerazione è prematura e alquanto rischiosa ma limitatamente ai dati analizzati sembra che gli utenti dimostrino interesse per i contenuti presentati e siano invogliati a continuare la loro visita. Ma quali contenuti attirano maggiormente l'attenzione dei navigatori? Su quali questi si soffermano più a lungo tempo nella consultazione? A queste domande si può rispondere combinando le informazioni ottenute sia dalla lista delle pagine più visitate sia dal navigation report con il quale si può indagare più nel dettaglio circa il comportamento degli utenti una volta raggiunto il sito. Per quanto riguarda la classifica delle pagine più visitate in testa si trova la home page seguita dalle pagine che presentano i modelli di cucina. Di rilevante importanza è andare a guardare quale sia la posizione occupata nella classifica da tutte quelle pagine che compongono la sezione la scelta del cuore e che offrono al navigatore un servizio di supporto nella scelta della propria cucina. Questo perché questi particolari servizi sono stati pensati appositamente per questo sito e ne rappresentano l'innovazione. Nella primissima parte della classifica spicca progetta la tua cucina, e solamente dopo il trentesimo posto della classifica si trovano il test di che cucina sei, il catalogo basato sulla personalità e il servizio entra nella tua cucina. L'alta incidenza nella

visualizzazione della pagina progetta la tua cucina sembra alquanto strana visto e considerato che supera ampiamente anche il numero di contatti avuti dalla pagina la scelta del cuore dove questa è proposta. Da una breve ricerca in rete però si trova una spiegazione: questo supporto è stato ampiamente pubblicizzato e in molti siti c'è un collegamento diretto a questa pagina, fatto confermato anche da una seconda posizione occupata nella classifica delle pagine d'entrata. Anche la pagina dei rivenditori occupa una posizione di nota all'interno della classifica: con una percentuale del 6,8% spicca tra tutti gli altri servizi offerti dal sito come tuttavia si sperava: il sito è stato pensato per far convogliare possibili acquirenti nei punti vendita e il fatto che la pagine dei rivenditori si distingua con una buona percentuale di consultazione non fa che confermare il fatto che l'obiettivo del sito è stato ampiamente conseguito.

Per quanto riguarda l'incidenza delle sezioni la scelta del cuore, la scelta della mente queste sono visualizzate più o meno nella stessa percentuale. Tuttavia, i navigatori dopo aver guardato all'offerta di ciascuna preferiscono consultare i contenuti della sezione la scelta del cuore che sono più accattivanti e consentono un approccio nella raccolta di informazioni più divertente ed interattivo. Complessivamente, quindi, la scelta di strutturare il sito secondo un modello di fruizione di internet che relazioni la necessità informativa alla piacevolezza ha centrato in pieno l'obiettivo. Quello che smorza un po' l'entusiasmo è che la pagina della sezione e-mail-me occupa la parte bassa della classifica. Andando a ricercare le cause di questa bassa incidenza una plausibile motivazione potrebbe essere data dal fatto che chi visita il sito Snaidero trova tutte le informazioni necessarie o preferisce chiedere assistenza direttamente nei punti vendita o ai call center.

Un primo accenno sui percorsi di visita seguiti può essere fatto guardando ai risultati delle classifiche riguardanti le pagine di entrata e di uscita. Prese singolarmente, tuttavia, sono poco informative: l'ideale è considerarle insieme per capire se ci siano particolari pagine che attirano i visitatori sul sito Snaidero ma che non li

incentivano a proseguire nella loro visita. Per quanto riguarda la classifica delle pagine di entrata ai primi posti si trovano la home page, la pagina di progetta la tua cucina e a scalare quelle dei vari modelli. Per la classifica delle pagine di uscita troviamo tra i primi posti ancora il servizio di progetta la tua cucina, la home page e quella dei rivenditori. Le pagine che hanno in comune le due classifiche sono quella del servizio progetta la tua cucina e la home page. Se si guardano alle percentuali vediamo che la home page domina entrambe le classifiche: in quella di entrata con una percentuale del 50,6% e quella di uscita con un 3,1% mentre la pagine progetta la tua cucina genera in entrata il 17,8% dei contatti mentre in uscita il 2%. Di acchito questo dato potrebbe essere preoccupante in quanto vediamo che un numero consistente di contatti generati non sono invogliati nella loro visita e lasciano il sito solo dopo aver visitato una pagina. Per valutare più approfonditamente questa particolare situazione l'ideale è evidenziare il gruppo delle visite che entrano ed escono con le due rispettive pagine, guardare ai termini di ricerca utilizzati e nel particolare ai loro comportamenti nella visita. Mentre per quanto riguarda la pagina dei rivenditori il dato può non essere così preoccupante se si pensa che un percorso tipo all'interno del sito vede questa pagina come l'ultima: quando l'utente arriva a consultare i contenuti della pagina dei rivenditori si assume abbia già raccolto tutte le informazioni necessarie per l'acquisto della cucina e guardi al punto vendita più vicino al quale affidarsi.

Si considerino ora i navigation report dai quali si possono trarre informazioni sui percorsi seguiti in media da chi visita il sito. Entrati dalla home page, assunzione confermata ampiamente dalla posizione di questa pagina nella classifica delle pagine di entrata, i visitatori si dirigono verso la sezione catalogo veloce con una percentuale quasi del 30%. Ben più distanziati sono i collegamenti con la sezione la scelta del cuore 6,5%, essere Snaidero 5,6%, la scelta della mente 3,6% e news con un 2,7%. Arrivati alle pagine dedicate al catalogo, non si notano particolari comportamenti,anzi. In media i visitatori

consultano con ordine le sezioni dedicate ai colori con i quali i vari modelli sono disponibili, alla visione di particolari finiture delle cucine e agli accessori complementari all'arredamento della cucina. Anche nella scelta dei modelli da visionare proseguono in media con ordine: dal design contemporaneo vanno al tradizionale rivisitato passando per il moderno emozionale. Da evidenziare su queste pagine è il tempo di permanenza in media su ognuna al fine di valutare quanto approfonditamente questi consultano i contenuti presentati: i visitatori scorrono in modo veloce tutte le pagine dedicate ai complementi d'arredo della cucina ma sui modelli e sulle loro versioni si soffermano con medie che vanno dai 20 ai 25 secondi per pagina. Una buona media considerato il fatto che le pagine presentano soprattutto immagini e file multimediali sui quali non ci si deve fermare in una lettura approfondita.

Guardando,invece, alla statistica tempo trascorso in media prima di arrivare sulla pagina in questione possiamo notare un percorso medio di consultazione dei contenuti delle varie sezioni così articolato:

- catalogo veloce;
- la scelta del cuore;
- la scelta della mente;
- essere Snaidero;
- e-mail-me.

Per quanto riguarda le pagine della sezione la scelta del cuore quella che ha la più alta percentuale di visita, 22%, è quella dedicata al test che aiuta a scegliere un modello di cucina in base alle risposte date, a seguire con percentuale che si aggirano tutte sul 12% il servizio multimediale che permette di entrare nella cucina scelta, il catalogo basato sulla personalità e progetta la tua cucina. La sezione feng shui riscontra solo un basso 3%. Nella sezione la scelta della mente la percentuale più alta va al catalogo veloce, 38%, con un netto distacco si trovano le pagine dedicate ai rivenditori con un 8,5%, subito a seguire la sezione dedicata ai consigli pratici e con percentuali che vanno dal 3,5% al 2,3% a seguire le pagine con le informazioni sul credito a consumo, sull'attenzione che Snaidero

riserva all'ambiente nella produzione delle sue cucine, sulla qualità e sulla garanzia che distinguono Snaidero. Quando, invece, il visitatore capita sulla sezione essere Snaidero o conclude la visita oppure da un'occhiata veloce ai suoi contenuti: i tempi di permanenza su questa pagine sono, infatti, tra i più bassi pur avendo contenuti scritti: siamo intorno ai 4-5 secondi. La sezione e-mail-me si trova ancora tra i posti più bassi della classifica: forse il titolo della sezione non fa onore alla vasta tipologia di informazioni che si possono richiedere mediante i form presenti in questa pagina. Personalmente, e-mail-me da l'idea di un unico indirizzo in cui si può chiedere tutto o niente o proprio questa ambiguità forse scoraggia i navigatori a chiedere assistenza usando il web.

Si scende ancora più nel dettaglio se si evidenziano particolari gruppi di visita. Questo espediente può essere utile per rendere espliciti comportamenti di visita non resi noti dall'analisi fino a qui condotta. Le variabili scelte per discriminare questi nuovi gruppi sono:

- tempo di vista;
- pagine di uscita;
- keywords/referer.

La variabile tempo di visita definisce un gruppo che raccoglie al suo interno tutte quelle visite che hanno una durata inferiore ai 5 secondi. La scelta di una durata pari o inferiore ai 5 secondi è ricaduta in quanto il programma definisce per default queste visite come short visits. Evidenziando questo primo raggruppamento si può avere un'idea della percentuale di queste visite rispetto alla totalità: 25,6%. L'ideale è che il valore di questa percentuale sia il più possibile basso, ma anche in questo caso la situazione non è poi così negativa. Se si guarda, però, anche ai primi posti della classifica delle keyword che generano questi contatti si trovano parole chiave a tema e ben il 45% delle visite non ha referer. Spicca su tutte la parola chiave Snaidero; una possibile spiegazione può essere data dal fatto che il gruppo sponsorizza anche una nota squadra di basket e chi cerca il sito della squadra inserendo la keyword Snaidero trova come

primo collegamento quello al sito che invece presenta il catalogo cucine. Nel caso in cui la parola chiave sia cucine, si può pensare che chi ha condotto questa ricerca fosse intenzionato a cercare informazioni sull'arredamento delle cucine ma non del marchio Snaidero. Inoltre, tra le pagine di entrata e di uscita di questa tipologia di visite troviamo ai primi posti progetta la tua cucina: per valutare a fondo il problema bisognerebbe studiare la tipologia di siti sui quali è stato è pubblicizzato questo servizio innovativo per capire se da questi siti si possono avere dei contatti migliori. Per mezzo di una personale ricerca in rete si è scoperto che questo servizio è pubblicizzato su articoli riguardanti il web marketing, argomento non del tutto pertinente con i contenuti presentati sul sito Snaidero.

La variabile pagine di uscita ha evidenziato le visite che terminavano con la pagina dei rivenditori. Come già spiegato in precedenza, si assume che le persone che arrivano a tale pagina abbiano ultimato il loro percorso di visita perché hanno raccolto tutte le informazioni necessarie. In questo particolare caso, quindi, si vuole valutare se nei percorsi di visita ci siano dei particolari percorsi di consultazione. Innanzitutto, si nota che il tempo medio di permanenza sul sito è superiore a quello riscontrato considerando le visite nella loro totalità: si passa da 294 a 463 secondi. Anche il tempo di permanenza su ogni singola pagina del catalogo è maggiore rispetto alla media: questa tipologia di visitatori è ancora più interessato ai contenuti proposti sul sito. L'ideale sarebbe guardare al tempo che questi passano sul sito prima di raggiungere questa pagina in modo da confrontarla con la totalità, ma purtroppo questa pagina ha un frame esterno e non è possibile calcolare le statistiche del navigation report.

I gruppi di visite risultanti da particolari keyword e referers sono stati evidenziati con l'intento esclusivo di valutare se questi contatti siano di qualità, ossia generino delle visite che vanno oltre i 5 secondi. Le statistiche risultanti da questi gruppi non verranno, quindi, confrontati con la totalità delle visite ma con quelle che sono superiori a 5 secondi. Questo essenzialmente perché abbiamo visto

che le short visit sono in percentuale un numero abbastanza elevato da falsare in modo sensibile i valori del tempo medio di permanenza sul sito e delle pagine visitate. Il gruppo not short visit ha una media di 27 pagine visitate e di 444 secondi di permanenza sul sito. I gruppi che hanno come discriminante le keyword Snaidero e cucine hanno delle medie di queste statistiche sensibilmente inferiori mentre per la parola chiave cucine Snaidero queste sono superiori. Questo cosa comporta? Una conferma del fatto che chi mette come termine di ricerca Snaidero forse non sta cercando il sito delle cucine ma quello della squadra di basket e chi invece inserisce cucine non sta cercando in particolare il marchio Snaidero. Niente di negativo se si considera che Snaidero si è sviluppata anche nel settore sportivo, quanto al fatto che alcuni utenti non cercano il marchio Snaidero.. forse l'azienda deve ancora lavorare sulla propria immagine.

3.8 RISULTATI

Di seguito si evidenziano alcuni punti: questi riassumono brevemente lo stato di sviluppo e di successo del sito.

- i contenuti presentati sul sito riscontrano un buon livello di gradimento, in particolare si sta facendo riferimento ad innovazioni quali il test, il catalogo basato sulla personalità, il servizio progetta la tua cucina;
- relativamente al punto precedente il sito Snaidero va oltre la tipologia di sito-vetrina in cui l'azienda pubblicizza esclusivamente il suo catalogo di prodotti: internet si presenta come mezzo di comunicazione che svolge la sua funzione informativa riuscendo ad essere anche accattivante e coinvolgente;
- nella consultazione si riscontra un'alta incidenza delle pagine dedicate ai rivenditori segno che l'obiettivo di far convogliare più utenti nei punti vendita è raggiunto.

Nel complesso, quindi, il progetto di presenza sul web del gruppo Snaidero ha riscontrato un buon successo. Tuttavia, internet è un canale informativo che evolve in tempi brevissimi. Di seguito si

presentano delle linee guida che vogliono essere degli spunti per una possibile ristrutturazione del sito:

- puntare un po' più sul fatto che il web può essere un ottimo canale per mettersi in contatto con l'azienda e per avere assistenza. La sezione e-mail-me così com'è presentata riscontra una bassa consultazione. Sulla struttura del sito deve occupare uno spazio diverso, molto più centrale e deve essere presentata con un titolo diverso, che dia l'idea della ampia tipologia di informazioni che si possono richiedere;
- per quanto riguarda la presentazione dei contenuti si potrebbe pensare di togliere le sezioni la scelta del cuore e della mente, lasciare quella dedicata al catalogo ma ristrutturare le sue pagine in modo da trovare spazio anche per i contenuti presentati nelle sezioni eliminate;
- collaborazione con nuovi siti partner dedicati ai complementi d'arredo per la cucina per cercare di generare nuovi contatti di qualità;
- valutazione di una soluzione e-CRM abbinando i dati del traffico del sito ai dati anagrafici dei visitatori che l'hanno generato. Il punto è alquanto complesso in quanto l'unico modo sarebbe quello di proporre sul sito delle sezioni per accedere alle quali si necessita di login. Tuttavia, questo espediente è alquanto discusso perché l'utente si sente spesso violato nella sua privacy. Questa soluzione comporterebbe numerosi vantaggi: in primis la possibilità di profilare gli utenti che cercano informazione in rete. Da qui si potrebbe muoversi per capire se questi rientrano nel target già individuato da Snaidero oppure sono un nuovo segmento al quale bisogna proporre nuove e mirate strategie di marketing.

CONCLUSIONI

Gli obiettivi per i quali il sito del gruppo Snaidero è stato pensato sono stati identificati nella volontà fare convogliare i visitatori nei punti vendita e di sviluppare un nuovo processo di raccolta di informazioni in funzione anche dell'espresso desiderio degli utenti internet di interazione e svago. Per valutare il successo del sito web in relazione agli obiettivi definiti si è guardato ai percorsi di consultazione delle risorse disponibili, focalizzando l'attenzione su particolari pagine, e ad alcuni parametri di sintesi quali la media del numero delle pagine visitate e del tempo di permanenza sul sito. In seguito al confronto dei software a disposizione per l'elaborazione dei dati la scelta è ricaduta su ClickTracks essenzialmente per due motivi: il primo in riferimento alle potenzialità espresse in termini di tipologia di analisi possibili e il secondo in merito all'affidabilità delle elaborazioni. Si è passato, quindi, all'elaborazione dei dati che ha portato ai seguenti risultati: i contenuti presentati sul sito riscontrano un buon livello di gradimento, il sito Snaidero va oltre la tipologia di sito-vetrina in quanto riesce ad interagire con il proprio visitatore attraverso i servizi di supporto alla scelta della propria cucina ed inoltre le pagine dei rivenditori hanno un'alta incidenza nella classifica delle pagine più visitate segno che il sito è un ottimo strumento per far confluire i visitatori verso i punti vendita. Inoltre sono stai indicati alcuni spunti per migliorare i servizi offerti dal sito: valorizzare la sezione e-mail-me che rappresenta un valido strumento di comunicazione tra l'azienda e il proprio cliente, collaborazione con nuovi siti partner dedicati ai complementi d'arredo per la cucina per cercare di generare nuovi contatti di qualità e valutazione di una soluzione e-CRM abbinando i dati del traffico del sito ai dati anagrafici dei visitatori che l'hanno generato.

BIBLIOGRAFIA

[1] R. Kosala, H. Blockeel (2000). *Web mining research: a survey*

[2] G. Chang (2001). *Mining the world wide web: a search approach*

[3] N. Zhong (2003). *Web intelligence*

[4] G. Bartolini (2001). *Web usage mining and discovery of association rules from HTTP servers logs*

[5] M. Eirinaki, M. Vazirgiannis (2003). *Web mining for web personalization*

[6] J. Srivastava, R. Cooley, M. Deshpande, P. Tan (2000). *Web usage mining: discovery and applications of usage patterns from web data*

[7] O. Zaiane, M. Xin, J. Han (2003). *Discovering web access patterns and trends by applying OLAP and data mining technology on web logs*

[8] *www.w3.org*

[9] M. Diodati (2001). *www.diodati.org*

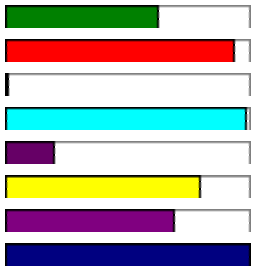
[10] P. Giudici (2001). *Data mining: metodi statistici per le applicazioni aziendali*

APPENDICE

Per problemi di spazio si riportano solo i risultati più significativi ottenuti dall'elaborazione del data set di partenza.

Site Overview
March 01, 2004 - May 31, 2004

Average Time On Site



All visitors	295 seconds
not short visit	440 seconds
short visit	1 second
rivenditori (pg ...	464 seconds
progettacucina (...	96 seconds
snaidero	376 seconds
cucine	324 seconds
snaidero cucine	470 seconds

Page Views per Visitor

All visitors	15,5
not short visit	28
short visit	2,2
rivenditori (pg ...	31,4
progettacucina (...	4,2
snaidero	27,6
cucine	26,1
snaidero cucine	34

Top Search Keywords

All visitors	
snaidero	7,80%
cucine	3,50%
cucine componibili	1,10%
snaidero cucine	0,90%
cucine snaidero	0,90%
www.snaidero.it	0,70%
arredamento cucine	0,40%
tavoli	0,20%
cucine arredamento	0,20%
mobili cucina	0,20%
cucine moderne	0,20%
snaidero.it	0,20%
arredamento	0,10%
cucine classiche	0,10%
.....	
None	60,60%

Top Referrers

All visitors	
Google	12,60%
Virgilio	5,60%
www.webmobili.it	5,00%
MSN	3,10%
Yahoo!	1,20%
ARIANNARICERCA	0,90%
www.bravacasa.it	0,70%
AltaVista	0,60%
search-dyn.tiscali.it	0,60%
www.soloarquitectura.com	0,50%
www.lietti.ch	0,50%
www.snaideropartners.com	0,40%
www.salonedelmobile.it	0,30%
www.arredamento.it	0,20%
www.rational.de	0,20%
www.frattali2.it	0,20%
www.proning.hr	0,20%
www.zortziko.com	0,20%
www.rosada-arredamenti.it	0,20%
www.enex.co.kr	0,20%
www.snaidero.com	0,20%
www.snaiderogroup.com	0,20%
www.salonedelmobile.com	0,10%
Voila.fr	0,10%
www.shopping.it	0,10%
Netscape	0,10%
www.pagesjaunes.fr	0,10%
.....	

Pages with Most Visitors

All visitors	
/html/homlay.html	60,20%
/html/olavers2.html	43,00%
/html/catveloce.html	41,40%
/html/progettacucina.html	24,60%
/html/ideavers1.html	21,90%
/html/vivavers1.html	21,70%
/html/esvers1.html	20,10%
/html/smeraldovers1.html	19,30%
/html/acropolisvers1.html	18,70%
/html/miticavers1new.html	16,70%
/html/sintesivers1.html	16,60%
/html/giocondavers1.html	15,80%
/html/opalevers14.html	15,70%
/html/timevers1.html	15,10%
/html/sistemazetavers1.html	14,80%
/html/topaziovers1.html	14,60%
/html/ginestravers13.html	14,50%
/html/temavers1.html	14,20%
/html/amicavers1.html	14,00%
/html/lineavers1.html	13,70%

/html/certosavers1.html	13,00%
/html/kentvers1.html	12,40%
/html/decorvers1.html	11,90%
/html/basecuoref.html	10,70%
/html/olavers14.html	8,80%
/html/olavers13.html	8,60%
/html/basementef.html	8,50%
/html/baseesseref.html	8,30%
/html/olavers12.html	8,20%
/html/olavers16.html	8,20%
/html/olavers15.html	7,90%
/html/catvelocehf.html	7,50%
/html/rivenditori.html	6,80%
/html/schemanews.html	6,10%
/html/epsolaversioni1.html	6,00%
/html/ipixgin11.html	5,40%
/html/ideavers15.html	5,10%
/html/homegioco.html	4,90%
/html/stepgioco.html	4,90%
/html/ideavers12.html	4,80%
/html/ideavers5_new.html	4,80%
…….	

short visit	
/html/progettacucina.html	46,10%
/html/homlay.html	21,90%
/html/catveloce.html	0,40%
/html/olavers2.html	0,40%
/html/schemanews.html	0,30%
……..	

Top Entry Pages

All visitors	
/html/homlay.html	50,60%
/html/progettacucina.html	17,70%
/html/catveloce.html	0,40%
……..	

short visit	
/html/progettacucina.html	41,50%
/html/homlay.html	3,10%
……..	

Top Exit Pages

All visitors	
/html/homlay.html	3,10%
/html/rivenditori.html	3,00%
/html/olavers2.html	2,50%
/html/acropolisvers1.html	2,10%
/html/progettacucina.html	1,80%
/html/schemanews.html	1,70%
/html/ipixgin11.html	1,60%
/html/certosavers1.html	1,50%

/html/baseesseref.html	1,50%
/html/timevers1.html	1,30%
/html/basecuoref.html	0,80%
/html/basementef.html	0,60%
/html/rivenditoriola.html	0,60%
/html/ideavers1.html	0,60%
/html/stepgioco.html	0,60%
/html/smeraldovers1.html	0,50%
……..	

Page Analysis /html/homlay.html
Next Page
Where visitors go next

/html/catveloce.html		
	All visitors	26,20%
exit		
	All visitors	18,10%
/html/basecuoref.html		
	All visitors	6,30%
/html/baseesseref.html		
	All visitors	5,50%
/html/basementef.html		
	All visitors	3,70%
/html/schemanews.html		
	All visitors	2,50%
/html/progettacucina.html		
	All visitors	1,60%
/html/homegioco.html		
	All visitors	1,50%
/html/rivenditorif.html		
	All visitors	1,10%

RINGRAZIAMENTI

Ringrazio innanzitutto la Prof.ssa Susi Dulli per avermi permesso di sviluppare in questo mio lavoro un argomento di mio interesse

Ringrazio la Dott.ssa Michela Giacomini per avermi supportato nell'elaborazione dei dati con il software SAS