alessandro.galano@tecnichenuove.com
▷ Alessandro Galano
Ha collaborato Paolo De Nictolis

GLI STRUMENTI PER IL DATA MINING

Dopo avere introdotto i modelli teorici sui quali si basa il Data Mining presentiamo ora alcuni degli strumenti più diffusi per l'analisi dei dati a supporto dell'e-business

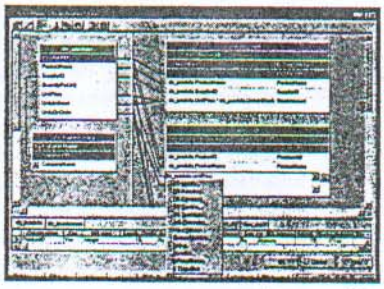
Nell'assunzione di decisioni strategiche per lo sviluppo aziendale il management di una società deve poter contare su strumenti d'analisi business intelligence adeguati alle esigenze. Uno di questi, considerato tra i più efficaci e potenti, è il Data Mining. Soluzione usata oggi in tutti i grandi gruppi, può trovare spazio anche presso Pmi, operatori Web e professionisti di marketing. Nel numero precedente abbiamo introdotto i modelli più comuni utilizzati nel Data Mining. Passiamo ora a illustrare alcuni degli strumenti più diffusi, suddivisi per case di produzione.

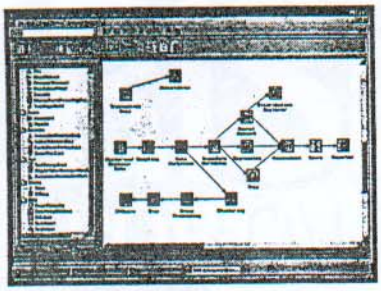
Ascential Software

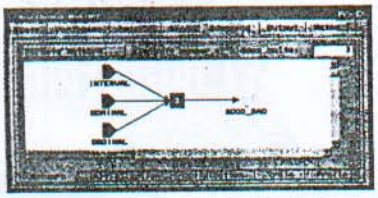
Data Stage XE è la soluzione di Ascential Software

► www.ascential.it

per monitorare, tracciare, analizzare il traffico degli accessi ai siti: tale soluzione permette la raccolta e l'analisi dei dati relativi all'utilizzo del sito e alla navigazione degli utenti, memorizzati nei Web Server e nei Proxy Server. Infatti, tutti i clic degli utenti, corrispondenti a pagine visitate o contenuti specifici richiesti (immagini, audio ecc.), vengono registrati e organizzati nei log file insieme a informazioni descrittive (data e ora, indirizzo IP dell'utente ecc.). L'architettura della soluzione prevede un database relazionale opportunamente organizzato, che permette l'archiviazione e la storicizzazione di tutte le informazioni che saranno oggetto dell'analisi; componenti per estrarre, trasformare e caricare i dati e componenti per la visualizzazione intuitiva dei dati e delle analisi. Avvalendosi di uno strumento Etl (Extract, Transform, Load) Data Stage gestisce il processo di estrazione dei dati contenuti nei log file dei Web Server e dei Proxy Server, di trasformazione degli stessi e di caricamento nella base dati specializzata per l'analisi. Il database garantisce prestazioni adeguate per interrogazioni complesse su grandi volumi di dati e al contempo garantisce agli utilizzatori della soluzione la massima flessibilità di analisi, con la possibilità di approfondire il livello di dettaglio fino al singolo clic di ogni singolo visitatore. La visualizzazione grafica delle informazioni che sintetizzano le modalità di consultazione del sito e i comportamenti dei visitatori è immediata, intuitiva ed espressamente dedicata ai vertici aziendali che possono così disporre di un vero e proprio cruscotto decisionale. Inoltre, le

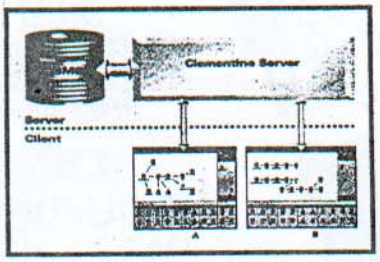






metodologico sono:

- *sample*, lo step in cui viene estratta una porzione di dati abbastanza grande per contenere ancora informazioni significative, e abbastanza piccola per analizzarla velocemente;
- *explore*, l'esplorazione dei dati e serve per scoprire in anticipo relazioni e anomalie nei dati e per capire quali possono essere d'interesse;
- *modify*, la fase per creare, selezionare e trasformare le variabili, al fine di mettere a punto il processo di costruzione del modello;
- *model*, la ricerca automatica delle variabili significative e dei modelli che forniscono le informazioni contenute nei dati;
- *assess*, infine, è la parte di metodo che permette la valutazione dell'utilità e affidabilità delle informazioni scoperte nel processo di Data Mining. In questa fase vengono portate nell'ambiente di produzione le regole estratte dai modelli. Il processo SEMMA è di per sé un ciclo le cui fasi possono essere sviluppate interattivamente come desiderato. I progetti che seguono questa metodologia possono analizzare milioni di record e rivelare relazioni che permettono agli analisti di raggiungere gli obiettivi di Data Mining.



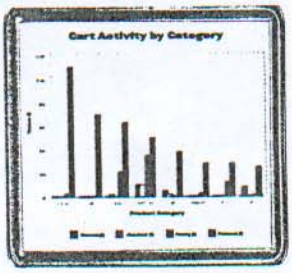
La struttura di Clementine di Spss

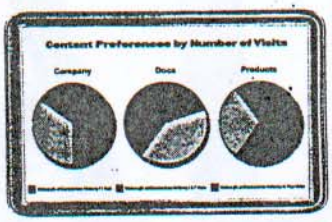
SPSS

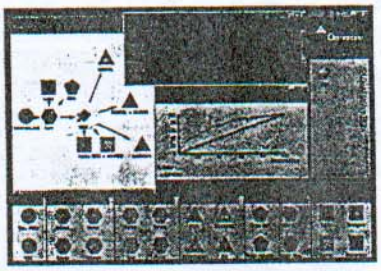
SPSS

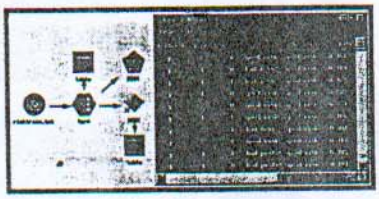
► www.spss.it

è il produttore di *Clementine*, software di Data Mining che consente una comprensione profonda delle realtà aziendali trovando soluzioni efficaci a numerosi problemi di business. I processi interattivi di Data Mining implementati in Clementine sfruttano le competenze aziendali a ogni step per strutturare modelli predittivi che tengono conto di esigenze reali. Più precisamente, il software di Data Mining visuale si chiama *Clementine Server* e ottiene buone prestazioni anche su una grande mole di dati poiché scarica l'elaborazione del database su server più potenti, limitando il traffico di rete e gestendo più efficacemente il flusso dei dati. Questi vantaggi derivano dall'architettura a tre livelli di Clementine Server, che integra client, server e database. Durante l'esecuzione, il client manda una descrizione dello stream – codificata in *Stream description language* (Sdl) – alla componente server di Clementine. Utilizzando tabelle di ottimizzazione, Clementine Server elabora le operazioni traducibili in Sql direttamente sul database, creando le interrogazioni appropriate. Il database esegue le interrogazioni Sql e passa i dati a Clementine Server, che procede a elaborare le operazioni non traducibili in Sql. Al termine dell'elaborazione, solo i risultati vengono restituiti al client, ottimizzando così sia i tempi di elaborazione sia il traffico di rete. Clementine Server può anche eseguire tutte le operazioni fuori dal database, se necessario, bilanciando automaticamente l'utilizzo di RAM e disco per il trattamento dei dati. In questo modo le prestazioni sono ottimizzate anche quando non si dispone di un database relazionale, ma soltanto di dati piatti come, per esempio, fogli elettronici o file Ascii.









Alcune funzioni del software Clementine Server

Accrue Software

L'applicativo per il Data Mining della Accrue Software

► www.accrue.com

si chiama *Accrue Insight 5*. Questo software offre soluzioni adattabili a ogni categoria aziendale, funzionali e flessibili grazie a quattro moduli specifici per analisi differenti:

- *Accrue Insight Campaign*: modulo per l'analisi delle campagne di marketing e pubblicitarie che razionalizza l'uso di Internet, determina il Roi e identifica i fattori critici ecc.
- *Accrue Insight Content*: analizza il responso dei visitatori in base ai contenuti del sito, permettendo una migliore segmentazione dei consumatori;
- *Accrue Insight Commerce*: aiuta le aziende ad andare incontro alle esigenze degli acquirenti, personalizzando gli acquisti;
- *Accrue Insight Affiliate*: permette alle aziende di valutare quali partnership sul Web danno i maggiori risultati e il maggior numero di visite che si trasformano in vendite.

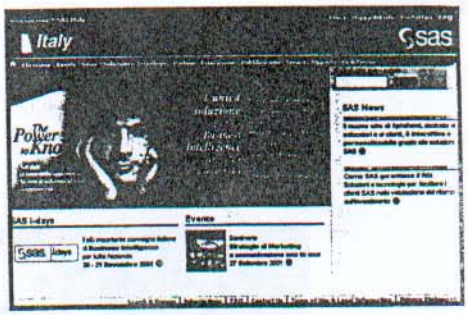
protette attraverso sistemi di crittografia che mostrano solo le transazioni lasciando anonimi i consumatori. In seguito Falcon utilizza un particolare tipo di rete neurale brevettato per scoprire le relazioni che intercorrono tra centinaia di variabili «rischiose» e che non sarebbero individuabili senza l'ausilio del computer. Lo scopo è abbassare al minimo il numero dei falsi allarmi e, quindi, aumentare il grado di efficienza, precisione e velocità del sistema. Un modulo aggiuntivo permette, inoltre, ai responsabili di aggiustare e adattare Falcon alle condizioni e caratteristiche specifiche della società che lo utilizza, per esempio creando o riaprendo casi sospetti basandosi sul punteggio o «score» del sistema o altri parametri chiave. Falcon viene utilizzato da 16 dei 25 principali fornitori di carte di credito e/o debito al mondo e controlla oggi un volume di circa 300 milioni di carte (il 65% del traffico mondiale). Falcon è disponibile per server Sun Microsystems e sistemi Unix.

eFalcon è un potente servizio anti frodi e di controllo del rischio per negozi online e Isp. Utlizza la stessa tecnologia di Falcon che produce punteggi (*score*) e, successivamente, accetta o rifiuta un determinato acquisto. In pratica, funziona in questo modo: per ogni transazioni valuta le probabilità di frode in base al valore di diverse variabili matematiche associate ai clienti, al prodotto che intendono comperare e al venditore. Per esempio, due transazioni relative a un medesimo articolo possono essere una accettata e l'altra rifiutata sulla base dell'indirizzo del destinatario. Se, infatti, un cliente cambia spesso indirizzo di spedizione in tempi ravvicinati, oppure proviene da regioni o Paesi noti per frodi relative a quella tipologia di prodotto acquistato in quel modo, a quell'ora e con quell'indirizzo di spedizione, viene bloccata la transazione. Come si può notare, è un sistema probabilistico raffinato che cerca di limitare al massimo i problemi senza eliminare grossolanamente quegli acquisti inizialmente sospetti, ma alla fine legittimi. Il sistema è in grado di evitare oltre il 50% delle frodi online rifiutando meno del 5% delle transazioni buone.

ProfitMax è, infine, il prodotto di punta di HNC Software per il settore delle telecomunicazioni. Il suo obbiettivo non è solo quello di evitare le frodi telefoniche, ma anche quello di individuare e prevedere con precisione quei clienti che sottoscrivono un abbonamento senza l'intenzione di pagare, oppure quelli morosi. In questo caso, la tecnologia è di nuovo identica a quella utilizzata negli altri prodotti, ma qui i parametri cambiano essendo le transazioni ora rappresentate dalle chiamate telefoniche. Il sistema funziona anche per la telefonia via Internet (VoIP) così come per la nuova tecnologia 3G (trasmissione di dati).

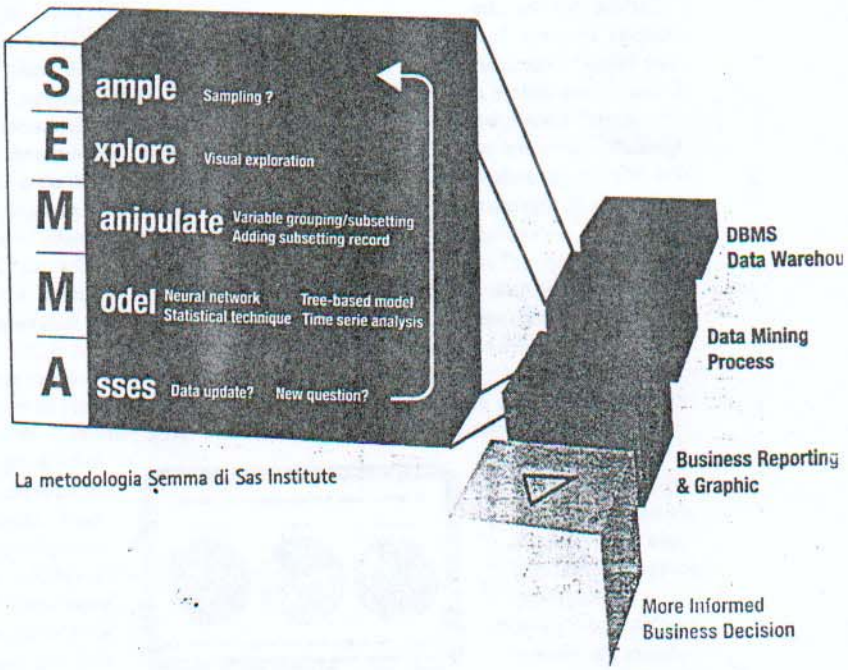
SAS Institute

SAS Institute
► www.sas.com/italy
offre una soluzione di business completa e integrata per il Data Mining: il software di punta è *Enterprise Miner* che si è classificato positivamente in una checklist predisposta dal Gartner Group per la valutazione delle funzionalità dei prodotti di Data Mining. Enterprise Miner comprende, infatti, un vasto repertorio che copre tutto il processo di Data Mining secondo la metodologia *Semma*, un acronimo che la multinazionale americana utilizza per indicare un approccio pratico al processo di Data Mining suddiviso in cinque fasi: *sample, explore, modify, model e assess.*

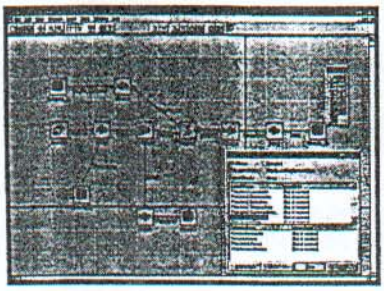


La metodologia Semma rende facili all'analista di business l'applicazione di tecniche d'esplorazione statistica e di visualizzazione, la selezione e la trasformazione delle variabili più importanti, la loro modellazione e la conferma della validità del modello scelto.

Nel dettaglio le fasi di questo approccio



La metodologia Semma di Sas Institute



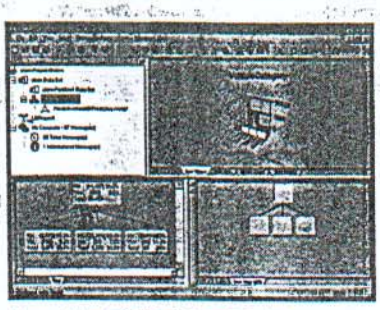
Data Stage XE di Ascential Software

funzionalità di visualizzazione possono essere ulteriormente arricchite grazie all'integrabilità del software con i più diffusi strumenti di visualizzazione analitica. Data Stage si integra sia con siti e portali già esistenti sia con le soluzioni orientate al commercio elettronico, con i sistemi di gestione dei contenuti multimediali e con qualsiasi sistema orientato al rapporto con i clienti (Crm) e al marketing personalizzato.

CrossZ Solution

CrossZ Solution
► www.crossz.it
propone tre diverse tipologie di soluzioni di Data Mining integrabili tra loro.

La prima è basata sulla tecnologia di Angoss Software Corporation
► www.angoss.com
e si chiama *Knowledge Web Miner;* consente di «*trovare soluzioni di Data Mining al confine tra il Data Warehouse e il Web*», cioè di reperire le informazioni tipiche di un approccio di Data Mining (trend, pattern, relazioni tra i dati) su Web, anche a partire da dati presenti in Data Warehouse aziendali. Il focus di questa soluzione sta nella necessità di monitorare correttamente la rapida evoluzione del contesto Web mediante la risoluzione di alcune problematiche come, per esempio, l'identificazione del profilo dei visitatori, i segmenti di mercato raggiunti e quelli mancanti, l'analisi delle opportunità più redditizie, dei prodotti e servizi oggetto di cross selling ecc.

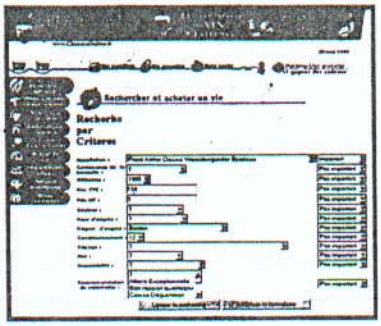


KnowledgeSTUDIO di CrossZ Solutions

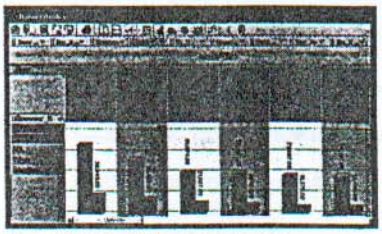
La seconda soluzione è basata su tecnologia Kaidara
► www.kaidara.com
si fonda su una metodologia nota come *Case Based Reasoning (CBR)* e consente l'individuazione, per siti di e-commerce, delle necessità di un potenziale acquirente e di come il catalogo in linea può rispondere a queste necessità: in presenza di piena rispondenza di un oggetto a catalogo alle esigenze del cliente, questo oggetto viene proposto per l'acquisto; in assenza di tale oggetto, viene proposta una lista di oggetti potenzialmente d'interesse per l'acquirente, secondo il principio della minima distanza pesata tra la richiesta e gli oggetti presenti nel catalogo.

L'ultima utilizza la tecnologia di Query Objects System Corporation
► www.queryobject.com
consente di venire incontro a una necessità impellente delle aziende che operano in uno scenario di e-business: la visione accurata e rapida di tutti i dati rilevanti per l'analisi del business dell'azienda. Tradizionalmente, tali sistemi hanno richiesto investimenti per molti milioni di dollari, con significativi potenziamenti dei sistemi hardware e software. Con la soluzione CrossZ è possibile sviluppare e rilasciare informazioni a migliaia di business manager a costi inferiori, e con l'utilizzo dei sistemi hardware già in opera, garantendo un accesso direttamente via browser.

Per tutte le soluzione disponibili CrossZ propone ai suoi clienti uno «*sviluppo pilota per validare l'adeguatezza delle soluzioni alle effetti-*



Kaidara di CrossZ Solutions



Query Objects di CrossZ Solutions

ve necessità, e quindi la fornitura dei prodotti e degli eventuali servizi a supporto richiesti».

WebTrends

Le soluzioni di Data Mining di WebTrends
► www.webtrends.com
sono pensate principalmente per l'analisi e il monitoraggio delle attività legate a Internet. Rivolte al mondo enterprise e agli Internet service provider, offrono specifici strumenti anche per le piccole e medie imprese. I prodotti di punta sono *Commerce Trends* e *Analysis Suite*, che permettono, attraverso funzioni di Path e Proxy Server Analysis, una visione approfondita delle preferenze negli acquisti online e nelle visite dei siti. Analizzando e segmentando l'utenza e i consumi in Rete, servono come supporto alle attività di marketing. Molto veloci da installare (da 1 a 3 ore) funzionano su piattaforme Microsoft.

HNC Software

HNC Software
► www.hnc.com
ha sviluppato software di Data Mining che permettono di operare in tempo reale. La maggior parte dei produttori crea software che utilizzano a posteriori tecniche statistiche rivolte ai dati registrati nel passato, mentre HNC Software offre soluzioni di Data Mining istantaneo, che analizzano dati presenti, fornendo in questo modo previsioni future più efficaci.

HNC Software è specializzata nel controllo e nella prevenzione in tempo reale delle frodi e nelle telecomunicazioni. I tre prodotti principali sono *Falcon, eFalcon, ProfitMax.*

Falcon è il prodotto leader mondiale per il controllo e prevenzione in tempo reale di frodi con carte di credito e/o debito. Questo software aiuta le imprese a monitorare tre tipologie di dati: le informazioni relative alle transazioni, i dati sul proprietario e quelli relativi alla spesa. Tutte le notizie raccolte sono transformate e