UNIVERSITÀ DEGLI STUDI DI MILANO – BICOCCA

Facoltà di Economia

Corso di laurea in Economia e Commercio

WEB USAGE MINING

Analisi del comportamento di navigazione e classificazione degli utenti.

Applicazione al sito della Biblioteca di Ateneo.

Relatrice: Prof.ssa Silvana STEFANI

Correlatore: Prof. Walter MAFFENINI

Tesi di laurea di

Gianluca TAVELLA

Matricola n. 575829

Anno Accademico 2003-2004

A mia madre e a mio padre,
per l'affetto
e la comprensione
che mi hanno sempre dimostrato.

"La vita è un dono e ogni compleanno è un nuovo inizio:
fa sì che ogni giorno si rivesta di speranza,
perchè le ombre del passato
non offuschino la luce del futuro."

Mary Ann Hathaway Tripp

Si ringraziano per la preziosa collaborazione: lo staff della Biblioteca di Ateneo, in particolare la dottoressa Luisa Berchialla ed il Direttore dottor Maurizio di Girolamo; il prof. Fabio Stella del Dipartimento di Informatica Sistemistica e Comunicazione; lo staff del Laboratorio di Informatica della Facoltà di Economia.

Indice Generale

Introduzione

Il web è diventato un mercato senza confini per acquistare e scambiare prodotti e servizi. Mentre gli utenti ricercano, passano in rassegna e occasionalmente comprano prodotti e servizi dal web, le società competono duramente per raggiungere ogni potenziale cliente. La conoscenza dei bisogni dei clienti potenziali e la capacità di stabilire servizi personalizzati che soddisfino questi bisogni rappresentano la chiave per vincere questa corsa competitiva.

L'unica informazione lasciata dai numerosi utenti che visitano un sito web è la traccia che lasciano attraverso le pagine alle quali hanno avuto accesso. Da questa fonte di dati, il proprietario del sito deve capire che cosa richiedono gli utenti dal sito, che cosa li attrae e che cosa li disturba o li distrae.

E' facile concludere che i prodotti acquistati di rado sono quelli che interessano di meno i potenziali clienti, e che la possibilità di accedere ad una pagina aumenta sistemando un collegamento ad essa in una posizione rilevante. Tuttavia, questa conclusione è valida solo se gli utenti percepiscono il sito e comprendono i suoi servizi, *allo stesso modo in cui gli sviluppatori li hanno concepiti.*

Questa non è sempre la situazione che si verifica. Molti hanno familiarità con il sistema di fare acquisti in un negozio o acquisire un documento o un certificato da un'autorità. Questo non implica che aggiungere e togliere prodotti da un carrello della spesa elettronico sia per loro intuitivo, o che possano formulare domande in modo efficace, ad un sito web di una grande organizzazione governativa.

Per questa ragione, prima di personalizzare i prodotti offerti in un sito web per adattarsi ai bisogni degli utenti, è necessario personalizzare il sito stesso come servizio ai suoi utenti. Altrimenti, potrebbero verificarsi conseguenze negative. Primo, gli utenti che hanno difficoltà nella comprensione di come il sito debba essere esplorato rimangono delusi, ossia si perdono potenziali clienti. Secondo, le loro tracce offuscano le statistiche

riguardo a quali pagine o prodotti sono popolari o correlati. Il risultato potrebbe essere: conclusioni sbagliate e clienti confusi.

Ci sono tre fattori che influiscono sul modo con cui un utente percepisce e valuta un sito: contenuto, design della pagina web, e design complessivo del sito. Il primo fattore riguarda merci, servizi, o dati offerti dal sito. Gli altri fattori riguardano il modo con il quale il sito rende i contenuti accessibili e comprensibili ai suoi utenti. E' necessario distinguere tra il design di una singola pagina e il design complessivo del sito, perché un sito non è semplicemente una raccolta di pagine – è una rete di pagine collegate. Gli utenti non si impegneranno nell'esplorazione, salvo che essi non trovino la sua struttura intuitiva, ed è proprio sul design complessivo del sito che l'attività di Web Usage Mining si concentra.

La facilità e la velocità con le quali le transazioni commerciali possono essere eseguite sul web sono state fattori chiave nella rapida crescita del commercio elettronico. Particolarmente, l'attività di e-commerce che coinvolge il consumatore finale sta affrontando una significativa rivoluzione. La capacità di rintracciare il comportamento di navigazione degli utenti fino ai singoli *click* del mouse ha portato il venditore e l'acquirente vicini come non mai. Oggi è possibile per un venditore personalizzare i propri annunci pubblicitari per i singoli clienti in scala massiccia, un fenomeno al quale ci si riferisce come *customizzazione di massa*.

Lo scenario sopra descritto è uno delle molte possibili applicazioni di *Web Usage Mining*, che è *il processo che impiega tecniche di data mining per la scoperta di modelli di navigazione degli utenti dai dati resi disponibili dal web, rivolto a varie applicazioni*.

La presente trattazione consiste in una parte teorica ed in una parte pratica.

Nel capitolo 1 illustreremo il data mining in generale, i legami con le altre discipline, e l'organizzazione dei dati per realizzare l'attività di data mining.

Si andranno poi a descrivere nel capitolo 2 le principali metodologie statistiche e computazionali proprie del data mining, soffermandosi in particolar modo sulle tecniche impiegate nel Web Usage Mining, oggetto principale di questa ricerca.

Il Web Usage Mining verrà trattato ampiamente dal capitolo 3: in esso si esaminerà la materia dal punto di vista teorico, mostrandone le caratteristiche principali alla luce dell'evoluzione negli ultimi anni di questo nuovo campo di studi interdisciplinare. Si darà ampio spazio anche all'analisi del successo di un sito, fondamentale nel caso del commercio elettronico.

Il Web Usage Mining verrà applicato nel capitolo 4 esaminando il sito della Biblioteca di Ateneo. Si analizzerà il comportamento di visita degli utenti e si classificheranno gli utenti stessi in gruppi omogenei. L'obiettivo è di migliorare la struttura del sito e di consentire quindi una più facile navigazione tra le pagine che risulteranno correlate. Si potrà suggerire anche una personalizzazione dei percorsi di visita in base al diverso gruppo in cui ogni utente andrà a far parte.

In tutta la trattazione verrà dato particolare rilievo ai vantaggi competitivi, dal momento che le analisi effettuate si devono sempre tradurre in termini economici.

1

Data Mining

1.1 Introduzione

Nella società dell'informazione, ogni individuo e organizzazione (azienda, famiglia, istituzione) ha a disposizione grandi quantità di dati e informazioni relative a se stessa e all'ambiente nel quale si trova ad operare.

Questi dati possono costituire un importante fattore di sviluppo ma il loro potenziale - la capacità di prevedere l'evoluzione di variabili di interesse o le tendenze dell'ambiente esterno - fino ad ora è rimasto in larga parte non utilizzato e ciò vale, in particolare, per il contesto aziendale.

Dati male organizzati e una scarsa consapevolezza delle potenzialità degli strumenti statistici di elaborazione delle informazioni hanno impedito il completo impiego della moltitudine di informazioni disponibili.

Gli sviluppi della tecnologia dell'informazione hanno permesso di correggere questa tendenza. Lo sviluppo di strumenti, sia hardware sia software, più potenti ed economici sta permettendo alle organizzazioni più moderne di raccogliere e organizzare i dati in strutture che facilitano l'accesso e il trasferimento delle informazioni verso tutti i possibili fruitori. Inoltre, la ricerca metodologica nei settori dell'informatica e della statistica ha condotto di recente allo sviluppo di procedure, flessibili e scalabili, per l'analisi di grandi basi di dati.

L'emergere di queste tendenze ha fatto sì che il data mining si stia rapidamente diffondendo in molte organizzazioni aziendali, come importante strumento di *business intelligence*, vale a dire di tecnologia dell'informazione per il supporto alle decisioni.

1.2 Definizione ed evoluzione della disciplina

Risulta difficile fornire una definizione per un qualcosa ancora in evoluzione, ed il **data mining**[1] è sicuramente un processo ancora coinvolto in profondi e continui mutamenti.

Per comprendere il termine data mining, è utile innanzi tutto conoscere qual è la traduzione letterale del termine: *to mine* in inglese significa "scavare per estrarre" ed è un verbo solitamente usato per azioni compiute nelle miniere; l'associazione del verbo alla parola dati rende l'idea di come vi sia una ricerca in profondità per trovare informazioni aggiuntive, non precedentemente note, nella massa dei dati disponibili.

Dalla prospettiva della ricerca scientifica, il data mining rappresenta un'area disciplinare di recente costituzione, che si è sviluppata traendo spunto da altri ambiti disciplinari, quali l'informatica, il marketing e la statistica. In particolare, molte delle metodologie impiegate nel data mining traggono origine principalmente da due campi di ricerca: quello sviluppato dalla comunità scientifica dell'apprendimento automatico (machine learning) e quello sviluppato dagli statistici, specie da coloro i quali si sono occupati di metodi multivariati e computazionali. La novità offerta dal data mining è l'integrazione delle precedenti metodologie con i processi decisionali. Nel particolare contesto aziendale, l'attività di data mining deve produrre risultati interpretabili e usufruibili per il supporto alle decisioni aziendali (DSS: Decision Support System).

L'apprendimento automatico (machine learning) è un campo di studi collegato all'informatica e all'intelligenza artificiale, che si occupa di ricavare dai dati relazioni e regolarità, alle quali fornire, in una seconda fase, valenza generale. Nella seconda fase lo scopo dell'apprendimento automatico è la riproduzione dei processi generatori dei dati, che permette la generalizzazione di quanto osservato, al fine di prevedere

[1] Quest'espressione è stata utilizzata la prima volta da U.Fayaad nei lavori della "First International Conference on KDD", tenutasi a Montreal (Canada) il 20-21 agosto 1995, per indicare un insieme integrato di tecniche di analisi, ripartire in varie fasi procedurali, volte ad estrarre conoscenze non note a priori da grandi insiemi di dati apparentemente non correlati.

l'andamento di certe variabili in corrispondenza di casi non osservati. Dal percettrone[2], il primo modello di macchina per l'apprendimento automatico, si svilupparono le reti neurali, nella seconda metà degli anni '80. Nel medesimo periodo, alcuni ricercatori perfezionarono la teoria degli alberi decisionali, in prevalenza per problemi di classificazione.

La metodologia statistica si è da sempre occupata della realizzazione di metodi e modelli per l'analisi dei dati. Questo ha anche determinato una costante attenzione per gli aspetti computazionali connessi all'applicazione della metodologia. A partire dalla seconda metà degli anni '80, a causa della crescente importanza degli aspetti computazionali per i fondamenti stessi dell'analisi statistica, vi è stato un parallelo sviluppo di metodi statistici computazionali per l'analisi di applicazioni reali, di natura multivariata. Ciò ha condotto, negli anni '90 all'attenzione, anche da parte degli statistici, ai metodi di apprendimento automatico, che ha condotto a importanti risultati di natura metodologica.

Verso la fine degli anni '80 si ebbero le prime applicazioni dei metodi di apprendimento automatico al di fuori dei settori dell'informatica e dell'intelligenza artificiale, particolarmente nelle applicazioni di database marketing, nelle quali i database a disposizione venivano utilizzati per elaborare campagne di marketing mirate. In quel periodo, venne coniato il termine Knowledge Discovery in Databases (KDD), con l'obiettivo di descrivere tutti quei metodi il cui scopo fosse la ricerca di relazioni e regolarità nei dati osservati.

Gradualmente, il termine KDD venne utilizzato per descrivere l'intero processo di estrazione della conoscenza da un database, dall'individuazione degli obiettivi di business iniziali fino all'applicazione delle regole decisionali trovate. In quest'ambito, il termine data mining venne impiegato per descrivere la fase del processo di KDD nel quale gli algoritmi di apprendimento venivano applicati ai dati. A seguito della

[2] Il percettrone venne proposto da F. Rosenblatt nel suo *Principles of neurodynamics: perceptrons and the theory of brain mechanism*, Spartan, Washington, 1962

successiva, graduale ma costante affermazione commerciale, il termine data mining è diventato sinonimo dell'intero processo di estrazione della conoscenza.

Ciò che distingue il processo di data mining da una tradizionale analisi statistica non è tanto la quantità dei dati che vengono analizzati o le particolari tecniche che vengono impiegate, quanto la necessità di operare in una modalità in cui la metodologia di analisi quantitativa e le conoscenze di business devono essere integrate. Fare data mining significa, infatti, seguire una metodologia che va dalla traduzione della problematica di business in una problematica di analisi quantitativa, all'implementazione di regole decisionali economicamente misurabili. Una definizione più completa di data mining è pertanto la seguente.

"Per data mining si intende il processo di selezione, esplorazione e modellazione di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori, e allo scopo di ottenere un risultato chiaro e utile al proprietario del database."[3]

Nel contesto aziendale, l'utilità del risultato si traduce in un risultato di business e, pertanto, ciò che distingue il data mining da una analisi statistica non è tanto la quantità dei dati che vengono analizzati o le particolari tecniche che vengono impiegate, quanto la necessità di operare in una modalità in cui la conoscenza delle caratteristiche del database, la metodologia di analisi e le conoscenze di business devono essere integrate. Fare data mining significa, infatti, seguire un processo metodologico integrato, che va dalla traduzione delle esigenze di business in una problematica da analizzare, al reperimento del database necessario per l'analisi, fino all'applicazione di una tecnica statistica, implementata in un algoritmo informatico, al fine di produrre risultati rilevanti per prendere una decisione strategica. Tale decisione, a sua volta, comporterà nuove esigenze di misurazione e, quindi, nuove esigenze di business, facendo ripartire quello che è stato definito "il circolo virtuoso della conoscenza"[4] indotto dal data mining.

---

[3] P. Giudici, *Data Mining. Metodi statistici per le applicazioni aziendali*, McGraw-Hill, Milano, 2001

[4] M. Berry, G. Linoff, *Data Mining techniques for marketing, sales, and customer support*, Wiley, New York, 1997

In definitiva, il data mining non è il mero utilizzo di un algoritmo informatico, o di una tecnica statistica, ma un processo di business intelligence che, in quanto tale, è volto all'utilizzo di quanto fornito dalla tecnologia dell'informazione per supportare le decisioni aziendali.

Il data mining è pertanto orientato al rilascio di applicazioni integrate nei processi decisionali aziendali piuttosto che al rilascio di "studi" rivolti alla comprensione di determinati fenomeni.

1.3 Legami con l'informatica

L'emergere del data mining è strettamente connesso agli sviluppi della tecnologia dell'informazione e, in particolare, all'evoluzione delle forme organizzative dei database, sviluppatesi rapidamente negli ultimi anni.

E' necessario fare chiarezza riguardo alla differenza tra alcuni termini, spesso confusi tra loro:

- **data retrieval**;
- **data mining**;
- **esplorazione multidimensionale dei dati** (OLAP: On Line Analitical

Processing).

Il data retrieval è l'attività consistente semplicemente nell'estrazione da un archivio o da un database di una serie di dati basandosi su criteri definiti a priori, in maniera esogena all'attività di estrazione stessa: un classico esempio è la richiesta fatta dalla Direzione Marketing di un azienda di estrarre i dati anagrafici di tutti i clienti che hanno acquistato almeno una volta il prodotto A e il prodotto B con lo stesso ordine; tale richiesta potrebbe basarsi sulla presunzione dell'esistenza di una relazione (almeno di tipo statistico) tra il fatto di aver acquistato A e B insieme almeno una volta e la propensione

ad acquisire un nuovo prodotto in lancio, ma tale ipotesi non è stata verificata e viene data per scontata basandosi, forse, su esperienze analoghe fatte in passato.

I nominativi così ottenuti potrebbero essere i destinatari di una campagna confidando sul fatto che la percentuale dei successi (ovvero i clienti che acquisiscono effettivamente il prodotto in lancio rispetto al totale dei contattati) sarebbe sicuramente superiore alla percentuale che si otterrebbe qualora i clienti contattati non avessero la caratteristiche indicate dalla Direzione Marketing. Tuttavia non si può dire in questo caso se, specificando meglio le caratteristiche dei clienti da contattare, si sarebbe ottenuto un risultato ancora migliore con uno sforzo uguale o minore.

II data mining va alla ricerca di relazioni e associazioni tra fenomeni non note a priori, e nel contempo fornisce precise misure comparative dei risultati attesi e, quindi, ottenuti.

L'attività di data mining, ancora, non deve essere confusa con l'attività volta alla realizzazione di strumenti di reportistica multidimensionale. Uno strumento OLAP è, essenzialmente, uno strumento, spesso di tipo grafico, che permette di visualizzare le relazioni tra le variabili a disposizione, seguendo la logica di analisi di un report a due dimensioni. Infatti, il data mining permette di andare oltre alla visualizzazione di semplici sintesi presenti nelle applicazioni OLAP, formulando modelli funzionali all'attività di business.

Attraverso l'impiego di metodologie OLAP l'utente forma delle ipotesi sulle possibili relazioni esistenti tra le variabili e cerca delle conferme osservando i dati. Per esempio, l'analista potrebbe voler determinare quali fattori portino al mancato rimborso di un prestito; potrebbe ipotizzare inizialmente che clienti ad alto rischio siano a basso reddito e con un elevato indebitamento. Al fine di verificare tale ipotesi l'OLAP fornisce una rappresentazione grafica (detta ipercubo multidimensionale) della relazione empirica tra le variabili reddito, debito, e insolvenza. L'esame del grafico può fornire indicazioni sulla validità dell'ipotesi fatta.

Pertanto, anche l'OLAP permette di estrarre informazioni utili dai database aziendali; diversamente dal data mining, tuttavia, le ipotesi di ricerca vengono suggerite

dall'utente, e non scoperte nei dati. Inoltre, l'estrazione viene effettuata in modo puramente informatico, senza avvalersi degli strumenti di modellazione e di sintesi forniti dalla metodologia statistica. Pertanto, sebbene l'OLAP possa dare indicazioni utili per database con un numero limitato di variabili, i problemi diventano insormontabili, quando il numero delle variabili da analizzare simultaneamente cresce e raggiunge l'ordine delle decine o delle centinaia. Diventa sempre più difficile e dispendioso in termini di tempo trovare una buona ipotesi e analizzare il database con gli strumenti di OLAP per confermarla o smentirla.

Mentre uno strumento OLAP è capace di incrociare tutte le dimensioni di classificazione per consentire la navigazione multidimensionale tra i dati, seguendo sempre una logica di analisi relativa ad un report a due dimensioni, il data mining permette di estrarre indicazioni sintetiche, combinando in modo multivariato tutte le dimensioni.

In definitiva, l'OLAP non è un sostituto del data mining, ma anzi, le due tecniche di analisi sono complementari e il loro impiego congiunto può produrre utili sinergie. L'OLAP può essere impiegato nelle fasi preliminari del data mining (pre-processing), agevolando la comprensione dei dati: per esempio permettendo di focalizzare l'attenzione sulle variabili più importanti, identificando i casi particolari o trovando le interazioni principali. D'altra parte, i risultati finali dell'attività data mining, riassunti da opportune variabili di sintesi, possono a loro volta essere convenientemente rappresentati in un ipercubo di tipo OLAP, che permette una comoda visualizzazione.

Potremmo concludere che il data mining non si relega in una semplice attività di analisi dei dati, bensì in un processo più complesso, in cui l'analisi dei dati è solo uno degli step caratterizzanti l'intero processo.

Possiamo riassumere quanto detto finora con una semplice relazione che, sia pure in modo schematico, fornisce la linea di evoluzione degli strumenti di business intelligence volti all'estrazione di conoscenza da un database :

DATA RETRIEVAL ⇨ **OLAP** ⇨ **DATA MINING**

La relazione ordina gli strumenti in relazione alla loro capacità informativa e, parallelamente, in relazione alla loro complessità di implementazione. Questa coesistenza suggerisce l'esistenza di un trade-off fra costi e benefici dei diversi strumenti.

Si deve aggiungere, infine, che la scelta fra i diversi strumenti va fatta in relazione alle specifiche esigenze di business e, inoltre, tenendo in debito conto le caratteristiche del sistema informativo aziendale. Per esempio, una delle difficoltà maggiori nell'attuazione di un efficace processo di data mining è la carenza di informazioni; un database è spesso realizzato per scopi diversi dal data mining e quindi delle informazioni importanti possono non essere presenti. Un altro problema che affligge i database è la presenza di dati non corretti, con errori di vario genere, dovuti alla misurazione del fenomeno o all'errata classificazione di alcune unità.

La creazione di un data warehouse[5] può eliminare molti di questi problemi. La combinazione di un'organizzazione efficiente dei dati (data warehouse) con un processo efficace e scalabile di analisi dei dati (data mining) permette di usufruire degli innumerevoli vantaggi che un uso corretto ed efficace delle informazioni disponibili può fornire, a supporto delle decisioni aziendali.

1.4 Legami con la statistica

La costruzione di metodologie per l'analisi dei dati è sempre stata oggetto di studio della statistica. La differenza principale rispetto ai metodi sviluppati nell'apprendimento automatico è che i metodi statistici vengono solitamente sviluppati in relazione ai dati in esame, ma anche secondo un paradigma concettuale di riferimento. Sebbene ciò abbia reso i numerosi metodi statistici coerenti e rigorosi, ne ha limitato la capacità di fronteggiare, in tempi rapidi, le richieste metodologiche avanzate dagli sviluppi della tecnologia dell'informazione e dallo sviluppo delle applicazioni di apprendimento automatico. Recentemente, anche gli statistici hanno volto la loro attenzione al data

[5] Si parlerà più diffusamente di data warehouse nel paragrafo 1.5 relativo all'organizzazione dei dati

mining, e ciò non può che costituire un importante fattore di rigore e sviluppo della disciplina.

Tuttavia, viene contestato dagli statistici che, nel data mining, non vi è un unico modello teorico di riferimento, ma numerosi modelli in competizione, che vengono selezionati sulla base dei dati in esame. La critica a questo modo di procedere risiede nel fatto che è sempre possibile trovare un modello, sebbene complesso, che si adatti "bene" ai dati. D'altra parte, l'insieme di tecniche statistiche chiamate in generale "analisi esplorative dei dati" si basano sulla stessa logica.

In secondo luogo, si contesta che l'abbondanza di dati a disposizione può indurre a trovare nei dati delle relazioni inesistenti.

Queste critiche sono da tenere in debita considerazione. Ciò nonostante si deve dar rilievo a due caratteristiche del data mining.

1. Le moderne metodologie di data mining prestano particolare attenzione al concetto di generalizzabilità dei risultati: ciò implica che, nella scelta di un modello, si tenga in debito conto la capacità previsiva e, quindi, vengano penalizzati i modelli più complessi;

2. molti risultati di interesse per un'applicazione non sono noti a priori e quindi non sono quantificabili in un'ipotesi di ricerca. Questo accade, particolarmente, in presenza di database di grandi dimensioni.

Questo ultimo aspetto è uno dei tratti distintivi del data mining dall'analisi statistica dei dati: mentre l'analisi statistica si occupa tipicamente di analisi di dati primari, raccolti allo scopo di verificare determinate ipotesi di ricerca, il data mining si può anche occupare di dati secondari, raccolti anche per scopi differenti da quelli dell'analisi. Questa ultima situazione è la regola, per esempio, nell'analisi di dati aziendali provenienti da un data warehouse. Inoltre, mentre in ambito statistico i dati possono avere anche natura sperimentale (possono cioè essere il frutto di un disegno degli

esperimenti, che alloca, per esempio, le unità statistiche in modo casuale a diverse tipologie di trattamenti), nel data mining i dati hanno tipicamente natura osservazionale.

Vi sono almeno altri tre aspetti che distinguono l'analisi statistica dei dati dal data mining.

1. Il data mining si occupa tipicamente dell'analisi di grandi masse di dati. Ciò implica considerazioni nuove per l'analisi statistica. Per esempio, il fatto che, per molte applicazioni, è impossibile elaborare e, perfino, accedere all'intero database, per motivi di efficienza computazionale, ma anche di informatività dei risultati. Si pone pertanto l'esigenza di effettuare un campionamento dei dati dal database in esame. Tale campionamento va effettuato in relazione agli obiettivi del data mining e, pertanto, non può essere analizzato con i tradizionali strumenti della teoria statistica dei campioni.

2. Molti database non sono riconducibili alle forme classiche di organizzazione dei dati della statistica. Ciò vale, come nel nostro caso, per i dati provenienti dall'accesso a Internet, dove è necessario considerare le variabili *data* e *ora* dell'accesso per risalire alle singole sessioni. Questo implica lo sviluppo di metodologie di analisi appropriate, non disponibili in ambito statistico.

3. I risultati del data mining devono essere rilevanti: ciò implica una costante attenzione alla valutazione dei risultati economici ottenuti con i modelli di analisi dei dati.

1.5 Organizzazione dei dati

Soffermiamoci su quello che sta accadendo nel mondo dell'Information Technology (IT). Secondo alcune stime del settore, la quantità di dati disponibili nel mondo in formato digitale raddoppia ogni anno e mezzo. Tale possibilità spinge le imprese verso la realizzazione di sistemi decisionali, per trasformare tali dati in informazioni utili ad acquisire e mantenere un vantaggio competitivo.

Fino a qualche anno fa, istituzioni, organizzazioni ed aziende erano in grado di utilizzarne solo una minima parte: circa il 10%[6].

Il colosso americano della grande distribuzione Sears custodisce attualmente nel suo sistema informativo 4 terabytes di dati sulla sua clientela, e ciò non è niente se confrontato con i 75 terabytes di dati del database dell'U.S. Departement of Energy.

Electricitè de France (EDF) riceve, nell'ambito di un programma di controllo della soddisfazione del consumatore, fino a 130.000 chiamate telefoniche al giorno nei suoi centri di raccolta dati, che devono essere resi prontamente disponibili per le analisi progettate. Sempre Sears raccoglie periodicamente i dati dei suoi 120 milioni di carte fedeltà, che vengono utilizzati per monitorare l'andamento delle preferenze dei consumatori.

Questi sono solo alcuni esempi per dare l'idea di come si stia vertiginosamente espandendo la mole dei dati che aspettano di essere trasformati in informazioni tramite operazioni di business intelligence.

Data warehouse

L'analisi dei dati presuppone indubbiamente che i dati stessi siano organizzati in un database ordinato, ed è influenzata in modo determinante dal modo in cui gli stessi sono organizzati nel database. Nella attuale società dell'informazione vi è abbondanza di dati, ed è crescente la necessità di analizzarli in modo efficace.

Un esempio, che ci riguarda direttamente, è costituito dalla crescente diffusione del commercio elettronico, e dalla conseguente abbondanza di dati sulle visite ai siti web, con le relative transazioni e pagamenti. In questo caso è decisivo, da parte del fornitore di servizi via Internet, comprendere la tipologia di visitatori e clienti, per prevedere e pianificare le modalità di offerta. Le transazioni (corrispondenti a dei *click* sul web) vengono inserite in un database ordinato, denominato data webhouse, che viene successivamente analizzato.

[6] Fonte Gartner Group, 1997

In generale, nei moderni mercati in eccesso di offerta, è divenuto strategico per ogni azienda di medio-grandi dimensioni disporre di un sistema informativo unificato, detto data warehouse.

Il **data warehouse** può essere definito come una raccolta di dati, orientata al soggetto, integrata, non volatile e variabile nel tempo, volta a supportare le decisioni del management.[7]

Andiamo a descriverne in dettaglio le caratteristiche.

- **Orientato al soggetto**: nel data warehouse i dati sono organizzati per soggetti rilevanti – prodotti, clienti, fornitori, periodi di tempo – al fine di offrire tutte le informazioni inerenti una specifica area.

- **Integrato**: il data warehouse deve essere in grado di integrarsi perfettamente con la moltitudine di standard utilizzati nelle diverse applicazioni. I dati devono essere ricodificati, per risultare omogenei dal punto di vista semantico, e devono utilizzare le stesse unità di misura. Per un esempio di ricodifica si osservi la figura 1.1.

- **Variabile nel tempo**: a differenza dei dati operazionali, quelli di un data warehouse hanno un orizzonte temporale molto ampio (anche 5-10 anni), risultando riutilizzabili in diversi istanti temporali.

- **Non volatile**: i dati operazionali sono aggiornati in modo continuo; nel data warehouse i dati sono caricati inizialmente con processi integrali e successivamente aggiornati con caricamenti parziali; i dati, una volta caricati, non vengono modificati e mantengono la loro integrità nel tempo.

Ricordiamo che il data warehouse deve essere orientato a produrre informazioni rilevanti per le decisioni del management. La vera differenza fra il data warehouse e

[7] W.H. Immon, *Building the Data Warehouse*, Wiley, New York, 1996

qualsiasi database aziendale è quella di configurarsi come un contenitore in cui sono collezionati dati utili per effettuare operazioni di business intelligence.

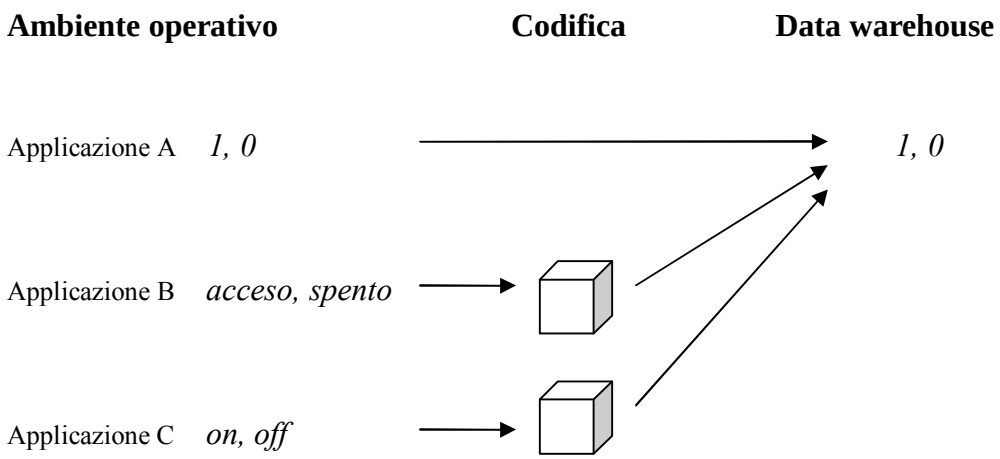

Figura 1.1 - Esempio di ricodifica

Data webhouse

La nascita del web, ed il suo impatto rivoluzionario nel mondo dell'Infomation Technology, ha incrementato ulteriormente il forte sviluppo avuto dal data warehouse durante gli anni '90. Il web rappresenta il nuovo ambiente nel quale operare ed impone al data warehouse di dotarsi di nuovi requisiti; la natura del data warehouse deve cambiare rispetto a quella dello scorso decennio, andando a configurarsi come web data warehouse o in breve *data webhouse*.

Il web è una ricchissima fonte di dati sul comportamento di coloro che interagiscono attraverso i propri browser con i siti Internet. Nonostante i dati riguardanti flussi di "click" siano in molti casi grezzi ed estremamente semplici, hanno la capacità di fornire

in maniera molto dettagliata informazioni su qualsiasi gesto compiuto da ogni individuo durante la navigazione in Internet.

Questa immensa e indisciplinata fonte di dati può essere convogliata all'interno del data webhouse per essere analizzata ed eventualmente conformata e combinata con le già esistenti e più convenzionali fonti di dati.

Come data webhouse può anche essere inteso il data warehouse convenzionale fruibile attraverso il web, con interfacce utilizzabili da semplici browser, attraverso le quali è possibile effettuare diverse operazioni: dalle più semplici, quali l'inserimento e l'aggiornamento dati, alla visualizzazione di reporting, fino alle più complesse come la gestione del sistema stesso.

Data mart

Il data mart (database di marketing) è un database tematico, orientato all'attività di marketing che contiene dati di tipo descrittivo e di tipo comportamentale, utili per valutare attentamente i propri clienti, identificare esigenze e stili di comportamento, stabilire strategie commerciali differenziate.

Il data mart può essere estratto da un data warehouse, ed è possibile estrarre tanti data mart quante sono le finalità che si vogliono perseguire con le successive analisi.

La costruzione di data mart rappresenta il primo e fondamentale passo nella predisposizione di un ambiente informativo per l'attività di data mining.

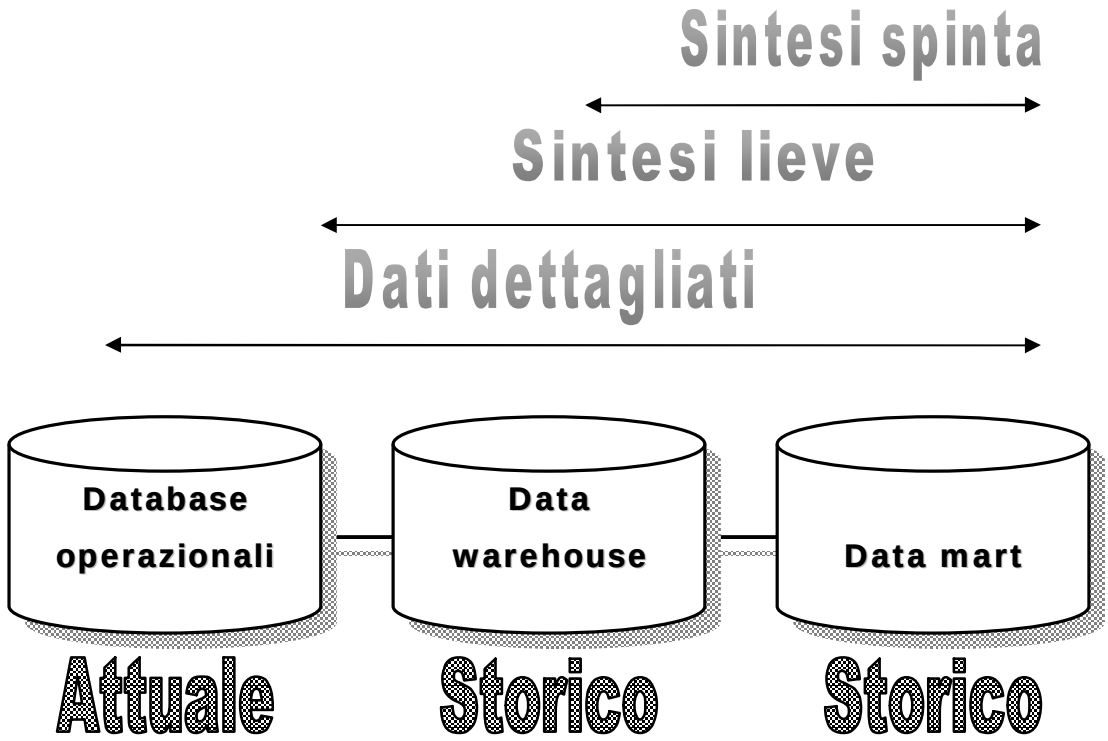

Figura 1.2 – Livelli di sintesi ed attualità dei dati

I dati contenuti nel data warehouse possono essere sintetizzati su più livelli rispetto ai dati operazionali, in modo da fornire una sequenza di immagini che parta da una visione di alto livello di business, caratterizzata da informazioni aggregate, la quale diventa sempre più dettagliata, in modo da spiegare aspetti particolarmente critici dell'analisi in oggetto.

Classificazione dei dati

Un data mart può essere organizzato secondo due dimensioni principali:

1. le **unità statistiche**, vale a dire gli elementi del collettivo di interesse ai fini dell'analisi;

2. le **variabili statistiche**, ovvero l'insieme delle caratteristiche di interesse per l'analisi, misurate su ciascuna unità statistica.

Le unità statistiche possono esaurire l'intera popolazione di riferimento oppure costituire un campione.

Le variabili statistiche costituiscono la fonte principale di informazione su cui lavorare per estrarre conclusioni sulle unità considerate ed, eventualmente, estendere tali conclusioni a una popolazione più ampia. E opportuno che tali variabili siano presenti in un numero sufficientemente elevato, per gli scopi che ci si prepone.

Una volta stabilite le unità e le variabili di interesse nell'analisi statistica dei dati, si da inizio al processo di classificazione delle variabili, assegnando ogni unità statistica ad una classe di misura, o modalità attraverso una regola di assegnazione. Le modalità si devono escludere mutuamente e devono essere esaustive (cioè ogni singola unità non può appartenere a due diverse classi e tutte le unità d'osservazione devono essere classificate). Il procedimento di classificazione è fondamentale per la successiva analisi e dipende dalla natura logica delle variabili a disposizione. Ogni dato relativo a una variabile costituisce una "misurazione" del fenomeno o carattere descritto dalla variabile stessa e, pertanto, la classificazione è funzione della scala di misurazione adottata per ciascuna variabile.

In generale, ciò porta a due tipologie differenti di variabili: qualitative e quantitative. I dati relativi a variabili qualitative sono tipicamente espressi in forma di aggettivo verbale, e danno origine a classificazioni in categorie. Le variabili quantitative sono invece legate a quantità intrinsecamente numeriche. A sua volta i dati qualitativi si distinguono in nominali ed ordinali. Nel caso di dati qualitativi nominali, non si può stabilire un ordine tra le varie categorie e la misurazione consente esclusivamente di stabilire una relazione di eguaglianza o diversità fra le singole modalità. Quando si trattano dati qualitativi ordinali, le categorie distinte presentano un ordine, esplicito o implicito. La relativa misurazione consente di stabilire una relazione d'ordine tra le diverse categorie, ma non consente alcuna asserzione numerica significativa sulle differenze tra le categorie. Più specificamente possiamo affermare quale categoria è più grande, o migliore, ma non di quanto.

Infine, le variabili quantitative, per le quali si possono anche stabilire relazioni e rapporti numerici fra le modalità, si distinguono in quantitative discrete, quando assumono un numero finito di valori, e quantitative continue, se assumono un'infinità numerabile di valori.

Tipo di variabile	Tipologie	
Quantitativa	*Discreta* (n. finito di valori)	*Continua* (infinità numerabile di valori)
Qualitativa	*Ordinale* (=, >, <)	*Nominale* (=, ≠)

Tabella 1.1 – Tipi di variabili

Da ultimo sottolineiamo che, spesso, le modalità ordinali di variabili vengono "etichettate" con dei numeri. Tuttavia, questa operazione non trasforma le variabili in quantitative e, pertanto, non permette di stabilire rapporti e relazioni fra le modalità stesse, che rimangono qualitative, sia pure ordinali.

1.6 L'attività di data mining

Indipendentemente dal tipo di applicazione specifica, l'attività di data mining si caratterizza in una serie di fasi che vanno dalla definizione degli obiettivi dell'analisi, fino alla valutazione dei risultati, all'interno di un processo che si autoalimenta in termini di ridefinizione degli obiettivi e di conseguimento di risultati. Le fasi di questo processo possono essere schematizzate nel seguente modo:

A. Definizione degli obiettivi dell'analisi

B. Predisposizione e pre-trattamento dei dati

C. Analisi preliminare dei dati ed eventuale trasformazione

D. Determinazione dei metodi statistici e computazionali

E. Elaborazione dei dati in base ai metodi scelti

F. Scelta del modello finale di analisi sulla base della valutazione dei metodi adottati

G. Implementazione del modello nei processi decisionali

Andiamo ora a descrivere le singole fasi.

Definizione degli obiettivi dell'analisi

Il primo passo consiste nella definizione degli obiettivi a cui l'attività di analisi è preposta. Questa fase è sicuramente quella più delicata dell'intero processo, perché a seconda di quanto stabilito in essa, verrà organizzata tutta la metodologia successiva. Infatti, l'intero flusso di lavoro, la scelta dei dati, l'utilizzo delle tecniche, il rilascio in produzione dei risultati conseguiti, dipende dagli obiettivi di business che s'intendono raggiungere. E' dunque necessario tradurre tali obiettivi in obiettivi di analisi, senza lasciare spazio a dubbi o incertezze su cosa s'intende perseguire e sulle modalità che verranno impiegate.

Predisposizione e pre-trattamento dei dati

Una volta individuati gli obiettivi di analisi, bisogna selezionare i dati necessari per l'analisi. Per prima cosa è necessario individuare le fonti dei dati. La fonte ideale dei dati è rappresentata dal data warehouse aziendale dal quale è semplice estrarre dei data mart di interesse.

In generale, la creazione dei data mart di analisi fornisce l'input fondamentale alla successiva analisi dei dati. Conduce alla rappresentazione dei dati, spesso in una forma tabellare, detta matrice dei dati, disegnata sulla base delle esigenze di analisi e degli obiettivi preposti.

Ottenuta la matrice dei dati, è spesso necessario effettuare operazioni di pulizia preliminare dei dati. In altre parole, si tratta di effettuare un controllo di qualità dei dati disponibili (data cleansing).

Bisogna sottolineare che, nell'attività di data mining, è spesso conveniente impostare l'attività di analisi su un campione dei dati a disposizione. La convenienza ad operare su base campionaria si può riassumere in tre motivi:

1. la qualità delle informazioni estratte da analisi complete, sull'intero data mart a disposizione, non è sempre superiore di quella ottenibile mediante indagini campionarie. Nelle applicazioni di data mining infatti, le dimensioni del database analizzato sono spesso considerevoli e, pertanto, l'utilizzo di un campione, ovviamente rappresentativo, permette di ridurre notevolmente i tempi di analisi ed elaborazione;

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

3. si riesce a tenere sotto controllo il rischio che il metodo statistico, adattandosi anche alle irregolarità e alla variabilità propria dei dati sui quali è stimata, perda capacità di generalizzazione e previsione.

Analisi preliminare dei dati

L'analisi vera e propria inizia con un'attività di analisi preliminare, o esplorativa, dei dati. Si tratta di una prima valutazione della rilevanza dei dati raccolti che può portare ad una ulteriore selezione e trasformazione delle variabili originarie. In particolare la trasformazione potrebbe essere dettata da esigenze di miglior comprensione del

fenomeno, da esigenze puramente matematico-statistico o ancora da esigenze di sintesi: l'applicazione del metodo delle componenti principali, ad esempio, riduce le dimensioni del problema individuando un numero limitato di variabili, capaci di spiegare la gran parte della variabilità del fenomeno studiato.

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

Determinazione dei metodi statistici e computazionali

La scelta di quale metodo utilizzare nella fase di analisi dipende essenzialmente dal tipo di problema oggetto di studio e dal tipo di dati disponibili per l'analisi. I metodi utilizzati possono essere classificati in base allo scopo immediato per il quale l'analisi viene effettuata. In base a questo criterio si possono distinguere, essenzialmente, quattro grandi classi di metodologie, che possono essere esclusive, oppure corrispondere a distinte fasi del processo di data mining.

a. **Metodi esplorativi**. Si tratta di metodologie interattive e, generalmente, visuali, che servono per trarre le prime conclusioni ipotetiche dalla massa di dati disponibili, oltre che per fornire indicazioni su eventuali trasformazioni della matrice dei dati, ovvero sulla necessità di integrare o sostituire il database disponibile.

b. **Metodi descrittivi**. Questo gruppo di metodologie (chiamate anche simmetriche o non supervisionate o indirette) hanno lo scopo di descrivere l'insieme dei dati in un modo più "parsimonioso". Questo può riguardare sia la sintesi delle osservazioni, che vengono pertanto classificate in gruppi non noti a priori (distanze, analisi di raggruppamento, mappe di Kohonen[8]) sia la sintesi delle variabili, che vengono fra loro relazionate, secondo legami non noti a priori

[8] Le mappe di Kohonen sono dei particolari tipi di reti neurali che permettono di classificare oggetti senza alcun tipo di supervisione e nascono dallo studio della topologia della corteccia del cervello umano. Verranno trattate in relazione alle reti neurali nel paragrafo 2.5.

(metodi associativi, modelli log-lineari, modelli grafici). In questo tipo di metodologie tutte le variabili a disposizione sono trattate allo stesso livello e non si fanno ipotesi di causalità.

c. **Metodi previsivi**. In questo gruppo di metodologie (chiamate anche asimmetriche o supervisionate o dirette) l'obiettivo è spiegare una o più variabili in funzione di tutte le altre, ricercando delle regole di classificazione o previsione. Tali regole permettono di prevedere o classificare il risultato futuro di una o più variabili risposta o target, in funzione delle variabili esplicative o input. Le principali metodologie di questo tipo sono sia quelle sviluppate nell'ambito dell'apprendimento automatico, quali le reti neurali supervisionate (percettroni multistrato) e gli alberi decisionali, ma anche classici modelli statistici, quali i modelli di regressione lineare e di regressione logistica.

d. **Metodi locali**. In questo caso l'obiettivo dell'analisi non è, come in tutti i casi precedenti, la descrizione delle caratteristiche del database nel suo complesso (analisi globale), ma l'individuazione di caratteristiche specifiche, relative a sottoinsiemi di interesse del database (analisi locali). Un esempio di quest'ultima tipologia di analisi è rappresentato dalle regole associative per l'analisi di dati transazionali.

Elaborazione dei dati in base ai metodi scelti

Una volta determinati i modelli statistici, si tratta di tradurli in opportuni algoritmi di calcolo informatico, che permettano di ottenere i risultati di sintesi desiderati. In commercio, si trova una grande disponibilità di software, anche specialistico, per l'attività di data mining. Nella maggioranza dei casi non è necessario sviluppare un algoritmo di calcolo su misura, ma è sufficiente impiegare quello implementato nel software a disposizione. In ogni caso, va evidenziata l'importanza che gli analisti abbiano un’adeguata conoscenza delle differenti metodologie, e non solo delle soluzioni software, al fine di poter adattare il processo alle specifiche esigenze aziendali. Gli analisti devono anche saper interpretare le elaborazioni in termini decisionali.

Scelta del modello finale di analisi

Al fine di produrre una regola decisionale finale è necessario scegliere, fra i vari metodi statistici considerati, il "modello" migliore di analisi dei dati. La scelta del modello e, quindi della regola decisionale finale, si basa su considerazioni che riguardano il confronto dei risultati ottenuti con i diversi metodi. Questa fase costituisce un importante controllo diagnostico della validità dei metodi statistici specificati, successivamente applicati ai dati a disposizione. Potrebbe darsi che nessuno, fra i metodi impiegati, permetta un soddisfacente raggiungimento degli obiettivi di analisi; in tale caso, si tratterà di "tornare indietro" e specificare una nuova metodologia, più opportuna per l'analisi in oggetto.

Nella valutazione della performance di uno specifico metodo concorrono, oltre a misure diagnostiche di tipo statistico, la considerazione dei vincoli di business, sia in termini di risorse che di tempo, oltre alla qualità e disponibilità dei dati.

Nel contesto del data mining, risulta spesso irragionevole utilizzare un solo metodo statistico per l'analisi dei dati. Ogni metodo è potenzialmente in grado di fare luce su aspetti particolari, magari trascurati da altri. Disporre di una tecnologia altamente performante e ricca di tecniche costituisce l'elemento caratterizzante l'attività di data mining: produrre una grande quantità di modelli in modo semplice e rapido, confrontare, in termini di robustezza, i risultati da essi prodotti su diversi campioni test, dare una quantificazione economica della regola costruita, sono gli elementi necessari per la scelta ottimale del modello finale.

Implementazione del modello nei processi decisionali

L'attività di data mining non è semplice analisi dei dati ma integrazione dei risultati nei processi decisionali aziendali.

La conoscenza del business, l'estrazione delle regole e il loro inserimento nel processo decisionale, permettono di passare dalla fase di analisi alla produzione di un motore decisionale. Scelto il modello e verificata la correttezza su di un eventuale data set di

validazione, si può applicare la regola di classificazione sull'intera popolazione di riferimento. Si potrà, per esempio, distinguere a priori quali clienti saranno redditizi, oppure calibrare politiche commerciali differenziate rispetto al target di consumatori, così aumentando la redditività aziendale.

Preso atto dei benefici che il data mining può apportare, diventa cruciale, al fine dell'adeguato sfruttamento delle sue potenzialità, riuscire a implementare correttamente il data mining nei processi aziendali.

Il progetto di inserimento del data mining nell'organizzazione aziendale deve essere affrontato in modo graduale, ponendosi obiettivi realistici e misurando i risultati lungo il percorso. L'obiettivo finale è il raggiungimento della piena integrazione del data mining con le altre attività di supporto alle decisioni, all'interno delle procedure operative dell'impresa.

2

Metodi di Data Mining

2.1 Introduzione

In questo capitolo si andranno a descrivere le metodologie statistiche e computazionali proprie del data mining, partendo dall'analisi esplorativa dei dati. Poiché le metodologie sono molte ed in molti casi notevolmente complesse e dal momento che questa esposizione non vuole essere esaustiva in questo senso, si darà maggior rilievo ai metodi principali e particolarmente a quelli utilizzati dal Web Usage Mining.

L'attività di data mining consiste essenzialmente nella ricerca di relazioni e risultati non noti a priori. Con l'analisi preliminare, o esplorativa, dei dati si elaborano le informazioni a disposizione al fine di descrivere in modo sintetico l'insieme dei dati a disposizione. Si utilizzano solitamente le rappresentazioni grafiche oppure indicatori statistici. L'analisi esplorativa può essere univariata, bivariata o multivariata, nel seguito si mostreranno i diversi metodi e i vari indici per ogni tipo di analisi esplorativa. Andremo quindi ad esaminare i principali metodi computazionali, dove non si richiede necessariamente una formulazione in termini di modello probabilistico. Queste metodologie, spesso ideate e sviluppate in ambito informatico, si contrappongono ai metodi statistici elencati successivamente, dove si assume invece un modello probabilistico che descrive il meccanismo generatore dei dati osservati.

Questo capitolo è stato svolto basandosi principalmente sul lavoro di Giudici (2001) e per quanto riguarda le reti neurali sul lavoro di Del Ciello *et. al* (2000).

2.2 Analisi esplorativa univariata

Come prima importante fase di analisi preliminare abbiamo l'analisi delle singole variabili a disposizione. I principali strumenti di analisi univariata sono le rappresentazioni grafiche ed una serie di indici statistici. I principali indici statistici unidimensionali si possono dividere in:

- indici di posizione;
- indici di variabilità
- indici di eterogeneità;
- indici di asimmetria;
- indici di curtosi.

Rappresentazioni grafiche

A seconda della tipologia di dati esaminati si utilizzeranno diverse rappresentazioni che possiamo riassumere nella seguente tabella.

Tipologia di dati	Rappresentazione grafica
Qualitativi nominali	Diagrammi a barre / Diagrammi a torta
Qualitativi ordinali ***Quantitativi discreti***	Diagramma a barre (delle frequenze)
Quantitativi continui	Istogramma (dopo la riclassificazione in classi intervallari)

Tabella 2.1 – Tipi di rappresentazioni grafiche

Nel caso di dati qualitativi nominali, l'ordine in cui le variabili vengono inserite sull'asse orizzontale del diagramma a barre è ininfluente e non ha un significato preciso. Se si tratta invece di dati qualitativi ordinali o quantitativi discreti l'ordine in cui le variabili vengono inserite sull'asse delle ascisse deve necessariamente corrispondere all'ordine numerico delle modalità.

Indici di posizione

I più utilizzati indici di posizione sono le medie e possono essere calcolate solamente per caratteri quantitativi.

La **media aritmetica** μ per un insieme $x_1, x_2, ..., x_N$ di N osservazioni è data da:

$$m = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Nel caso di una distribuzione di frequenze la media aritmetica ponderata risulta pari a:

$$m = \sum_{i=1}^{N} x_i \cdot p_i$$

dove p_i è la frequenza relativa ($\sum p_i = 1$).

La **moda** è una media che può essere determinata per tutti i tipi di caratteri. La moda è la modalità cui è associata la massima frequenza e sintetizza tanto meglio la distribuzione quanto più elevata è la sua frequenza relativa.

La **mediana** è la modalità del carattere che in una distribuzione ordinata occupa la posizione centrale, ripartendo la distribuzione in due parti uguali. E' calcolabile solo se tra le modalità del carattere è possibile istituire un ordinamento, perciò il carattere deve essere almeno ordinale.

Dati N dati ordinati in senso non decrescente si dice mediana:

$$Me = x_{\left(\frac{N+1}{2}\right)} \quad \text{per } N \text{ dispari}$$

$$Me = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} \quad \text{per } N \text{ pari}$$

Come generalizzazione della mediana si possono considerare i valori che suddividono la distribuzione di frequenza in parti, con quote percentuali prefissate. Questi valori si dicono **quantili**. In particolare risultano interessanti i quartili, che corrispondono ai valori che dividono la distribuzione in quattro parti uguali. In formula:

$$Q_1 = x_{\left(\frac{N+1}{4}\right)} \; ; \; Q_2 = x_{\left(\frac{N+1}{2}\right)} \text{ (per } n \text{ dispari)}; \; Q_3 = x_{\left(\frac{3}{4}(N+1)\right)}$$

Si noti che Q_2 coincide con la mediana.

Indici di variabilità

Oltre alle misure che forniscono informazioni intorno alla posizione, è interessante studiare anche la dispersione o variabilità di una distribuzione. La misura di variabilità più comunemente usata, per dati quantitativi, è la **varianza**.

Dato un insieme $x_1, x_2, ..., x_N$ di N osservazioni relative ad una variabile X, e indicata con μ la loro media aritmetica, la varianza è definita da:

$$s^2(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2$$

Volendo mantenere l'unità di misura originaria, si fa spesso riferimento allo scarto quadratico medio $s(X) = \sqrt{s^2(X)}$. Mentre per agevolare il confronto fra diverse distribuzioni si può utilizzare il coefficiente di variazione $CV = \frac{s}{|m|}$.

Indici di eterogeneità

La varianza e le misure ad essa riconducibili non sono calcolabili per i caratteri qualitativi. In questo caso si utilizzano indici che ricorrono al concetto di eterogeneità di una distribuzione di frequenze. Per illustrare il concetto di eterogeneità si consideri la rappresentazione in distribuzione di frequenze di una variabile qualitativa, con riferimento alla tabella 2.2.

Modalità	**Frequenze relative**
x_1	p_1
x_2	p_2
…	…
x_K	p_K

Tabella 2.2 – Distribuzione di frequenza teorica per una variabile qualitativa

Nella realtà potremmo avere due situazioni estreme, entro le quali si collocherà la distribuzione osservata. Tali situazioni sono:

- eterogeneità nulla, quando tutte le unità presentano la medesima modalità del fenomeno in oggetto, ovvero se $p_i = 1$ per un certo i; $p_i = 0$ per ogni altro i;
- eterogeneità massima, quando le unità sono ripartite uniformemente tra le K modalità del carattere, ovvero se: $p_i = 1/K$ per $i = 1,\ldots, K$.

Un indice di eterogeneità dovrà essere pertanto minimo nella prima situazione e massimo nella seconda. Due indici di eterogeneità che soddisfano tali condizioni sono l'indice di eterogeneità di Gini e l'indice entropico.

L'indice di **eterogeneità di Gini** è definito da:

$$G = 1 - \sum_{i=1}^{K} p_i^{\,2} \qquad\qquad 0 \le G \le (K-1)/K$$

Per ottenere un indice normalizzato, che assume valori nell'intervallo [0,1], si può utilizzare l'indice relativo $G' = \dfrac{G}{(K-1)/K}$.

L'indice entropico o **entropia** è definito da:

$$E = -\sum_{i=1}^{K} p_i \log p_i \qquad\qquad 0 \le E \le \log K$$

Volendo ottenere un indice normalizzato, che assume valori nell'intervallo [0,1], si può utilizzare l'indice relativo $E' = \dfrac{E}{\log K}$.

Indici di asimmetria

La rappresentazione grafica dei dati considerati, mediante istogrammi o diagrammi a barre, risulta utile per indagare sulla forma della distribuzione considerata, tuttavia esistono anche indici statistici sintetici per informare sul grado di simmetria di una distribuzione.

Un **indice di asimmetria** della distribuzione è definito da:

$$g = \frac{m_3}{s^3}$$

dove $m_3 = \dfrac{\sum_{i=1}^{N} (x_i - m)^3}{N}$ è il momento terzo della distribuzione, e s^3 è il cubo dello scarto quadratico medio.

L'indice di asimmetria γ è calcolabile solo per le variabili quantitative e può assumere ogni valore reale. In particolare:

- $g = 0$ se la distribuzione è simmetrica;
- $g < 0$ se la distribuzione è asimmetrica a sinistra;
- $g > 0$ se la distribuzione è asimmetrica a destra.

Indici di curtosi

Nel caso in cui si lavori con dati di tipo quantitativo e si possa approssimare la distribuzione dei dati con una distribuzione normale, può essere opportuno costruire un indice statistico che misuri la "distanza" della distribuzione osservata dalla distribuzione teorica corrispondente alla perfetta normalità.

L'indice che permette di controllare se i dati seguono una distribuzione normale è l'**indice di curtosi**, definito da:

$$b = \frac{m_4}{m_2^2}$$

dove $m_4 = \dfrac{\sum_{i=1}^{N}(x_i - m)^4}{N}$ è il momento quarto della distribuzione

e m_2^2 è il quadrato del momento secondo $m_2 = \dfrac{\sum_{i=1}^{N}(x_i - m)^2}{N}$

L'indice β può assumere ogni valore reale positivo. In particolare:

- $b = 3$ se la variabile è perfettamente normale
- $b < 3$ se la distribuzione è iponormale (rispetto alla normale ha frequenza minore per valori molto distanti dalla media)

- $b > 3$ se la distribuzione è ipernormale (rispetto alla normale ha frequenza maggiore per valori molto distanti dalla media)

2.3 Analisi esplorativa bivariata

Anche in questo caso è possibile utilizzare dei grafici per visualizzare ed indagare sulle relazioni tra le due variabili. Ciò può essere fatto attraverso un diagramma di dispersione.

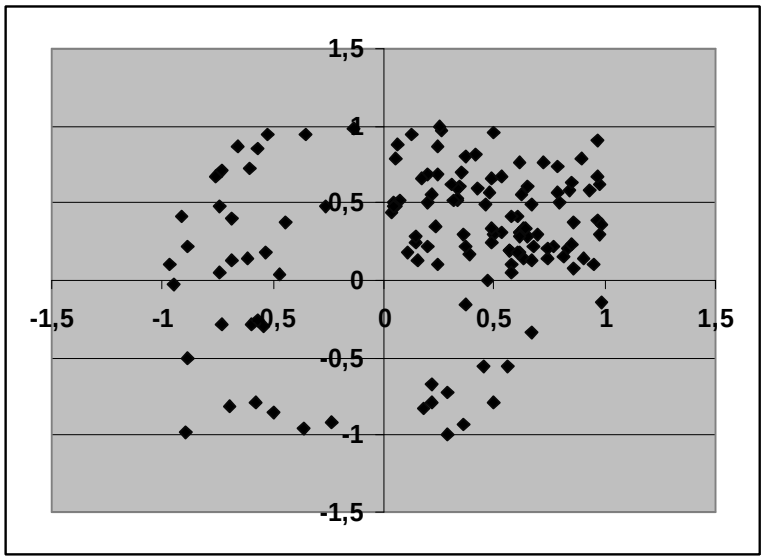

Figura 2.1 – Esempio di diagramma di dispersione

Anche nell'analisi bivariata si possono utilizzare indici statistici che riassumono le distribuzioni di frequenza, migliorando l'interpretazione dei dati, sia pure perdendo delle informazioni.

In questo paragrafo si esamineranno solamente gli indici sintetici sulle variabili quantitative. Gli indici relativi alle variabili qualitative verranno mostrati nel prossimo paragrafo relativo all'analisi multivariata.

Chiariamo innanzitutto il significato dei termini *concordanza* e *discordanza*. Con il termine concordanza si indica la tendenza delle modalità (poco) elevate di una variabile

ad associarsi a modalità (poco) elevate dell'altra. Con il termine discordanza si indica invece la tendenza di modalità meno elevate di una delle due variabili ad associarsi a modalità elevate dell'altra.

Per misurare il grado di concordanza o discordanza tra due caratteri si impiega la **covarianza**, definita da:

$$COV(X,Y) = \sum_{i=1}^{h} \sum_{j=1}^{k} [x_i - m_X] \cdot [y_i - m_Y] \cdot p_{XY}(x_i, y_j)$$

dove m_x e m_Y indicano, rispettivamente, le medie aritmetiche delle variabili *X* e *Y*. I dati assumono rispettivamente le *H* modalità $x_1,...,x_H$, e le *K* modalità $y_1,...,y_K$, mentre $p_{XY}(x_i, y_j)$ è la frequenza relativa.

- $COV(X,Y) > 0$ se c'è una relazione diretta tra le variabili;
- $COV(X,Y) < 0$ se c'è una relazione indiretta tra le variabili;
- $COV(X,Y) = 0$ se non c'è nessuna relazione tra le variabili (le variabili sono incorrelate);

Quindi valori positivi della covarianza denotano concordanza, mentre valori negativi segnalano discordanza.

La covarianza è un indice assoluto, per ottenere un indice relativo dobbiamo rapportare la covarianza al prodotto degli scarti quadratici medi, ottenendo il **coefficiente di correlazione lineare** tra le variabili *X* e *Y*, che si definisce come:

$$r_{XY} = \frac{COV(X,Y)}{s(X) \cdot s(Y)} \qquad -1 \le r_{XY} \le 1$$

Quando $r_{XY} = 0$ i due caratteri non sono legati da alcun tipo di relazione e le variabili *X* e *Y* si dicono *indipendenti in correlazione*.

2.4 Analisi esplorativa multivariata

Partiamo con l'analisi di dati **quantitativi**. Faremo uso della notazione matriciale per la matrice di dati, in quanto consente di formulare misure sintetiche in modo più compatto. Ricordiamo che ragionando in termini multivariati possiamo estendere i risultati al caso bivariato.

Indicando con X la matrice di dati, con n righe e p colonne possiamo calcolare le misure di sintesi nel seguente modo.

Le **medie aritmetiche** delle variabili saranno descritte dal vettore p-dimensionale:

$$\overline{X} = \frac{1}{n}\mathbf{1}X$$

dove **1** indica un vettore riga con tutti gli elementi pari a 1.

Consideriamo ora la matrice degli scarti dalla media:

$$\widetilde{X} = X - \frac{1}{n}JX$$

dove J è una matrice con tutti gli elementi pari a 1.

Possiamo trovare la **matrice di varianza-covarianza**, quadrata di dimensione p attraverso:

$$S = \frac{1}{n}\widetilde{X}'\widetilde{X}$$

Tale matrice S contiene, sulla diagonale principale le varianze di ciascuna variabile, mentre gli elementi fuori dalla diagonale rappresentano le covarianze fra le p variabili considerate. La matrice S inoltre è simmetrica e definita positiva, infatti:

$x'Sx > 0 \qquad \forall \underset{\_}{x} \neq 0$

A volte può essere conveniente sintetizzare la matrice varianza-covarianza con uno scalare che rappresenti la variabilità complessiva del sistema.

Una prima misura è la **traccia di *S***, ossia la somma degli elementi sulla diagonale principale:

$$tr(S) = \sum_{s=1}^{p} s_s^2$$

si può dimostrare che $tr(S) = \sum_{s=1}^{p} l_s$, cioè alla somma degli autovalori della matrice stessa.

La seconda misura è fornita dal determinante di *S*, e viene chiamata **varianza generalizzata di Wilks**:

$$W = |S|$$

Passiamo ora all'analisi esplorativa di dati multivariati di carattere **qualitativo**. Gli indici descritti in seguito si possono applicare anche ai dati qualitativi di livello nominale. Anche in questo caso possiamo estendere i risultati al caso bivariato.

X/Y	Y_1	…	Y_j	…	Y_J	Totale
X_1	n_{11}	…	n_{1j}	…	n_{1J}	$n_{1\bullet}$
…	…	…	…	…	…	…
X_i	n_{i1}	…	n_{ij}	…	n_{iJ}	$n_{i\bullet}$
…	…	…	…	…	…	…
X_I	n_{I1}	…	n_{Ij}	…	n_{IJ}	$n_{I\bullet}$
Totale	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet J}$	n

Tabella 2.3 – Contingenze teoriche

Considerando la tabella di contingenza teorica precedente, diciamo che:

n_{ij} indica la frequenza associata alla coppia di modalità (x_i, y_j), *i = 1, 2, ..., I; j = 1, 2, ..., J*, delle variabili *X* e *Y*.

$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ indica la frequenza marginale della riga *i*-esima. Denota il numero totale di unità statistiche che assumono la modalità *i*-esima di X.

$n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$ indica la frequenza marginale della colonna *j*-esima. Denota il numero totale di unità statistiche che assumono la modalità *j*-esima di Y.

Tra le frequenze della tabella vale la relazione di marginalizzazione seguente:

$$\sum_{i=1}^{I} n_{i\bullet} = \sum_{j=1}^{J} n_{\bullet j} = \sum_{i=1}^{I} \cdot \sum_{j=1}^{J} n_{ij} = n$$

Per sviluppare indici descrittivi della relazione tra variabili qualitative risulta conveniente introdurre il concetto di **indipendenza statistica**.

Due variabili *X* e *Y* si dicono indipendenti, in riferimento alle *n* unità statistiche analizzate, se si verifica la seguente condizione:

$$n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \qquad \forall \mathrm{i} = 1, 2, ..., \mathrm{I}; \;\; \forall \mathrm{j} = 1, 2, ..., \mathrm{J}$$

Se ragioniamo in termini di frequenze relative ciò equivale a stabilire che $p_{XY}(x_i, y_j) = p_X(x_i) p_Y(y_j)$ per ogni *i* e per ogni *j*.

Indici di connessione

Per misurare la discrepanza tra frequenze osservate e frequenze teoriche si può utilizzare la statistica $\boldsymbol{X^2}$ **di Pearson**, che fornisce una misura "globale" per la verifica dell'ipotesi di "indipendenza stocastica" tra *X* e *Y*. Questa misura è definita come:

$$X^2 = \sum_{i=1}^{I} \cdot \sum_{j=1}^{J} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

dove $n_{ij}^* = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ indica le frequenze teoriche, mentre gli n_{ij} sono le frequenze osservate.

La statistica X^2 risente della numerosità delle osservazioni, ovvero al divergere di *n*, tende a crescere indefinitamente. Per superare questo problema si possono ricavare misure alternative, funzioni della statistica precedente.

La prima misura alternativa proposta è il **coefficiente phi**:

$$f = \sqrt{\frac{X^2}{n}} = \sqrt{\sum_{i=1}^{I} \cdot \sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - 1}$$

Il quadrato di f è denominato **contingenza quadratica media** (mean contingency). Bisogna far rilevare che f non è un indice normalizzato (compreso nell'intervallo [0,1]).

La seconda misura proposta è l'**indice di Cramér**. Questo indice si calcola rapportando X^2 al valore massimo che può assumere nella specifica tabella. Tale estremo è dato dalla radice quadrata del più piccolo tra i valori $(I-1)$ e $(J-1)$, l'indice risulta quindi:

$$V = \sqrt{\frac{X^2}{n \min[(I-1),(J-1)]}}$$

L'indice risulta normalizzato, e quindi $0 \le V \le 1$ per qualunque tabella $I \times J$, in particolare $V = 0 \Leftrightarrow X$ e Y sono indipendenti. Inoltre $V = 1$ solo nel caso di massima dipendenza tra i caratteri X e Y.

Indici di dipendenza

Gli indici di connessione illustrati sopra sono tutti funzioni della statistica X^2 e differiscono solamente per il criterio di normalizzazione adottato. Tuttavia il loro impiego presenta degli inconvenienti, poiché tali misure di associazione risultano scarsamente interpretabili nella maggioranza delle applicazioni concrete. Sono stati quindi proposti altri indici dotati di un chiaro significato operativo, nello specifico contesto d'indagine in cui essi vengono applicati.

Una prima tipologia di tali indici è definita in relazione alla riduzione proporzionale nella probabilità di commettere un errore di previsione. Si considerino due variabili nominali X e Y, e si supponga di voler ottenere una previsione ottima della modalità assunta da Y, sulla base della categoria di X. In altri termini, si immagini che un'unità statistica sia scelta a caso dalla popolazione e si voglia prevedere quale sia la modalità di Y in corrispondenza di essa. La previsione può essere fatta: (i) non avendo alcuna informazione aggiuntiva; (ii) conoscendo la corrispondente modalità di X. Ovviamente ci si attende che la probabilità di commettere un errore di previsione nella prima

circostanza sia maggiore o uguale a tale probabilità nella seconda circostanza (cioè che valga P(i) > P(ii)).

Una misura del grado di associazione tra le due variabili X e Y è pertanto definita da:

$$l_{Y|X} = \frac{\text{Prob. di errore nel caso (i) - Prob. di errore nel caso (ii)}}{\text{Prob. di errore nel caso (i)}} = \frac{P(i) - P(ii)}{P(i)}$$

L'indice $l_{Y|X}$ rappresenta la riduzione proporzionale nella probabilità di commettere un errore di previsione, passando dalla situazione (i) alla situazione (ii), più informativa.

Illustriamo ora le quantità che compaiono nel calcolo dell'indice facendo riferimento al caso in cui le *n* unità statistiche che costituiscono la popolazione di riferimento siano classificate in una tabella di contingenza a due vie.

Nel caso (i), non avendo informazioni su *X*, la previsione della modalità di *Y* in corrispondenza di un'unità scelta a caso dalla popolazione può essere basata solo sulla corrispondente distribuzione marginale di frequenze: $\{n_{\bullet 1}, n_{\bullet 2}, ..., n_{\bullet J}\}$. Infatti le quantità $n_{\bullet 1}/n, n_{\bullet 2}/n, ..., n_{\bullet J}/n$ rappresentano le probabilità che dalla popolazione sia estratto un elemento con modalità di *Y* data, rispettivamente, da $Y_1, Y_2, ..., Y_J$.

In questo caso è opportuno scegliere come previsione di Y la classe a cui è associata la massima frequenza marginale di colonna, ossia $n_{\bullet(\max)}$. Quindi, la probabilità di errore nel caso (i) è data da:

$$P(i) = 1 - \frac{n_{\bullet(\max)}}{n}$$

dove $n_{\bullet(\max)} = \max\{n_{\bullet 1}, n_{\bullet 2}, ..., n_{\bullet J}\} = \max_j (n_{\bullet j})$

Nel caso (ii) si dispone dell'informazione aggiuntiva che l'unità estratta presenta la modalità X_i (*i* = 1, 2, …, *I*); pertanto, la previsione ottima della corrispondente modalità di *Y* sarà basata sulla distribuzione di frequenze nella *i*-esima riga della tabella

di contingenza: $\{n_{i1}, n_{i2}, ..., n_{iJ}\}$. Pertanto il previsore di Y è dato dalla classe cui è associata la massima frequenza nella i-esima riga della tabella, cioè $n_{i(\max)}$, e la probabilità complessiva di errore nel caso (ii) risulta:

$$P(ii) = 1 - \sum_{i=1}^{I} \frac{n_{i(\max)}}{n}$$

dove $n_{i(\max)} = \max\{n_{i1}, n_{i2}, ..., n_{iJ}\} = \max_{j}(n_{ij})$

Sostituendo nell'espressione di $l_{Y|X}$ si ottiene che:

$$l_{Y|X} = \frac{\sum_{i=1}^{I} n_{i(\max)} - n_{\bullet(\max)}}{n - n_{\bullet(\max)}}$$

Si può far notare che l'indice $l_{Y|X}$ assume valori nell'intervallo [0,1]. In particolare:

- $l_{Y|X} = 0 \Leftrightarrow$ la conoscenza di X non riduce la probabilità di errore nella previsione di Y.
- $l_{Y|X} = 1 \Leftrightarrow$ la conoscenza di X consente di prevedere esattamente la modalità assunta da Y.

L'indice assume il valore estremo *1* se e solo se c'è massima dipendenza di Y da X. All'opposto l'indice assumerà il valore estremo *0* se e solo se tutte le frequenze massime di riga giacciono nella stessa colonna della tabella di contingenza (come quando le variabili sono indipendenti).

Eppure, $l_{Y|X}$ può essere pari a *0* anche se i caratteri sono tra loro associati, questo perché l'indice misura un aspetto particolare dell'associazione: la dipendenza di Y da X. Pertanto anche se assume il valore *0* non bisogna escludere che vi siano altre forme di associazione tra i fenomeni esaminati.

Allo stesso modo si definisce l'indice di dipendenza di *X* da *Y*:

$$l_{X|Y} = \frac{\sum_{j=1}^{J} n_{(\max)j} - n_{(\max)\bullet}}{n - n_{(\max)\bullet}}$$

dove $n_{(\max)j} = \max\{n_{1j}, n_{2j}, ..., n_{Ij}\} = \max_i(n_{ij})$ e $n_{(\max)\bullet} = \max\{n_{1\bullet}, n_{2\bullet}, ..., n_{I\bullet}\} = \max_i(n_{i\bullet})$

Sottolineiamo, infine, che in generale $l_{Y|X} \neq l_{X|Y}$, poiché alle variabili è attribuito un significato logico differente. L'indice opera in modo asimmetrico rispetto alle variabili in esame, distinguendo il caso in cui una variabile sia dipendente dal caso in cui la stessa sia indipendente.

Indici modellistici

Gli indici modellistici, a differenza di quelli precedentemente considerati, risultano indipendenti dalle distribuzioni marginali dei caratteri. Ciò conduce all'importante vantaggio di essere interpretabili anche in un'ottica modellistica di tipo inferenziale. Nel seguito si farà riferimento alle probabilità di casella, indicate con p.

Considerando una tabella 2×2, relativa alle variabili *X* e *Y*, rispettivamente sulle righe e colonne della tabella, il **rischio relativo** per la variabile *Y* è definito da:

$$RR = \frac{p_{1|1}}{p_{1|0}} = \frac{P(Y=1 \mid X=1)}{P(Y=1 \mid X=0)}$$

ossia il rapporto delle probabilità di successo di *Y* (modalità *1*) nei due livelli della variabile *X* (modalità *0* e *1*). Questa quantità può assumere qualsiasi numero reale non negativo, in particolare:

- $RR = 1$ quando *Y* e *X* sono indipendenti
- $RR \in [0,1)$ se $p_{1|1} < p_{1|0}$, ossia se la probabilità di successo risulta maggiore nella riga 0 rispetto alla riga 1.

- $RR \in (1,+\infty)$ se la probabilità di successo è maggiore nella riga 1.

Continuando a fare riferimento a tabelle 2×2, introduciamo un'altra misura di associazione che costituisce un parametro fondamentale per i modelli statistici per l'analisi dei dati qualitativi: l'**odds ratio**. Ora si mostreranno i passaggi per la costruzione dell'indice.

Indichiamo con $p_{1|1}$ la probabilità di successo nella riga 1 e con $p_{1|0}$ la probabilità di successo nella riga 0. All'interno della riga 1 l'odds di successo è definito da:

$$odds_1 = \frac{p_{1|1}}{p_{0|1}} = \frac{P(Y=1 \mid X=1)}{P(Y=0 \mid X=1)}$$

Gli odds risultano non negativi, con valore maggiore di 1 quando un successo (modalità 1) è più probabile di un insuccesso (modalità 0). Esemplificando, se risulta $odds = 5$ significa che un successo è cinque volte più probabile di un insuccesso; quindi ci si aspetta di osservare cinque successi per ogni insuccesso. Invece $odds = 1/5 = 0{,}20$ significa che un insuccesso è cinque volte più probabile di un successo.

Per la riga 0 l'*odds* di successo risulta:

$$odds_0 = \frac{p_{1|0}}{p_{0|0}} = \frac{P(Y=1 \mid X=0)}{P(Y=0 \mid X=0)}$$

Andiamo quindi a definire l'odds ratio come:

$$J = \frac{odds_1}{odds_0} = \frac{p_{1|1} / p_{0|1}}{p_{1|0} / p_{0|0}}$$

Si può dimostrare, usando la definizione di probabilità congiunta, che:

$$J = \frac{p_{11} \cdot p_{00}}{p_{10} \cdot p_{01}}$$

Pertanto la quantità J è anche detta *rapporto dei prodotti incrociati* perché uguaglia il rapporto dei prodotti delle probabilità delle caselle diametralmente opposte.

Operativamente, è possibile sostituire le probabilità con le frequenze osservate. Ciò conduce alla seguente:

$$J = \frac{n_{11} \cdot n_{00}}{n_{10} \cdot n_{01}}.$$

Vediamo ora alcune proprietà dell'odds ratio.

1. $J \in [0,+\infty)$
2. X e Y indipendenti $\Rightarrow J = 1$
   - $1 < J < \infty \Rightarrow$ associazione positiva
   - $0 < J < 1 \Rightarrow$ associazione negativa
3. l'odds ratio tratta le variabili in modo simmetrico, pertanto non è necessario identificare una variabile come risposta e l'altra come esplicativa. Al contrario, il rischio relativo richiede questa distinzione, è quindi è più appropriato per analisi asimmetriche;
4. si dimostra che *odds ratio* = $RR\left(\frac{1-p_{1|0}}{1-p_{1|1}}\right)$
5. dal punto di vista modellistico, l'odds ratio può essere considerato l'analogo qualitativo del coefficiente di correlazione lineare.

In generale, per tabelle $I \times J$ gli odds ratio possono essere definiti con riferimento a ciascuna delle $\binom{I}{2} = I(I-1)/2$ coppie di righe in combinazione con ciascuna delle

$\binom{J}{2} = J(J-1)/2$ coppie di colonne. In una tabella $I \times J$ ci sono $\binom{I}{2}\binom{J}{2}$ odds ratio di questo tipo[9].

Indubbiamente, per grosse tabelle, il numero di odds ratio da calcolare diventa enorme, e risultano convenienti delle semplificazioni.

2.5 Metodi computazionali

In questo paragrafo andremo ad esaminare i principali metodi computazionali, dove non si richiede necessariamente una formulazione in termini di modello probabilistico. Queste metodologie, spesso ideate e sviluppate in ambito informatico, si contrappongono ai metodi statistici presentati nel paragrafo successivo, dove si assume invece un modello probabilistico che descrive il meccanismo generatore dei dati osservati.

Si andranno a descrivere inizialmente i concetti di prossimità e distanza fra unità statistiche, che stanno alla base di molte delle metodologie sviluppate nel paragrafo. La parte successiva è dedicata ai metodi di classificazione delle unità statistiche, che si possono dividere in :

- **supervisionati**: la classificazione si confronta con la presenza di una variabile di riferimento (target o risposta), le cui modalità sono note.

- **non supervisionati**: non vi sono variabili di confronto. Sarà l'analisi di classificazione che determinerà la natura ed il numero dei gruppi e collocherà le unità statistiche in essi.

[9] Il coefficiente binomiale tra *n* ed *x* si indica con $\binom{n}{x}$ e si calcola mediante il rapporto $\frac{n!}{x!(n-x)!}$

Misure di prossimità e distanza fra le unità statistiche

Per classificare e raggruppare le unità statistiche in gruppi omogenei è necessario introdurre la nozione di prossimità. Gli indici di prossimità tra coppie di unità statistiche forniscono le informazioni preliminari indispensabili per poter individuare gruppi di unità omogenee.

Un indice di prossimità tra due generiche unità statistiche u_i e u_j è definito come una funzione dei rispettivi vettori riga nella matrice dei dati:

$$IP_{ij} = f(x_i', x_j') \qquad i, j = 1, 2, ..., \text{n}$$

Nel caso in cui le variabili considerate siano quantitative gli indici di prossimità utilizzati sono le distanze, gli indici di distanza e gli indici di dissimilarità. Se invece i caratteri sono di tipo qualitativo verranno utilizzati gli indici di similarità. Esistono infine indici di prossimità che vengono utilizzati nel caso in cui le variabili siano miste, ovvero alcune qualitative e altre di tipo quantitativo.

L'indice di prossimità più utilizzato per le variabili quantitative è la distanza euclidea. La **distanza tra due punti** corrispondenti ai vettori riga $x, y \in \Re^p$ è una funzione $d(x, y)$ che gode delle seguenti proprietà:

non negatività:	$d(x, y) \geq 0$		$\forall x, y \in \Re^p$
identità:	$d(x, y) = 0$	$\Leftrightarrow$	$x = y$
simmetria:	$d(x, y) = d(y, x)$		$\forall x, y \in \Re^p$
disuguaglianza triangolare:	$d(x, y) \leq d(x, z) + d(y, z)$		$\forall x, y, z \in \Re^p$

Per il raggruppamento delle unità statistiche, generalmente si considera la distanza tra tutte le unità statistiche presenti nella matrice dei dati. L'insieme di tali distanze viene rappresentato in una matrice delle distanze.

Una generica **matrice delle distanze** è strutturata nel modo seguente:

$$\Delta = \begin{pmatrix} 0 & \dots & \dots & d_{1i} & \dots & \dots & d_{1n} \\ \dots & 0 & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0 & \dots & \dots & \dots & \dots \\ d_{i1} & \dots & \dots & 0 & \dots & \dots & d_{in} \\ \dots & \dots & \dots & \dots & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 0 & \dots \\ d_{j1} & \dots & \dots & d_{ni} & \dots & \dots & 0 \end{pmatrix}$$

dove il generico elemento d_{ij} è una misura della distanza tra le entità *i* e *j*.

La distanza euclidea è definita come la radice quadrata della differenza tra i rispettivi vettori, nello spazio euclideo:

$${}_2d_{ij}^2 = \left[\sum_{s=1}^{p} \left(x_{is} - y_{js} \right)^2 \right]^{\frac{1}{2}}$$

La distanza euclidea è fortemente influenzata da differenze elevate tra i valori, essendo funzione del quadrato delle stesse.

Per superare tale limitazione, tale distanza viene calcolata, non sulle variabili originarie, bensì su opportune trasformazioni di queste. La scelta più comune consiste nella standardizzazione delle variabili. A seguito della standardizzazione, ogni variabile statistica contribuisce al computo della distanza con uguale peso.

Si noti che, se le variabili statistiche sono standardizzate, con media nulla e varianza unitaria, risulta che:

$${}_2d_{ij}^2 = 2\left(1 - r_{ij}\right) \qquad (i, j = 1,\dots, p)$$

$$r_{ij} = 1 - {}_2d_{ij}^2 / 2 \qquad (i, j = 1,\dots, p)$$

dove r_{ij} indica il coefficiente di correlazione fra le unità statistiche *i* e *j*, calcolato impiegando come "osservazioni" le modalità assunte dalle differenti variabili in corrispondenza delle unità *i* e *j*.

Nel caso di caratteri qualitativi si utilizzano invece gli indici di similarità.

Dato un insieme finito di unità statistiche $u_i \in U$, si dice **indice di similarità** un'applicazione $S(u_i, u_j) = S_{ij}$ da $U \times U$ in $\Re$ che soddisfa le seguenti proprietà:

non negatività:	$S_{ij} \geq 0$	$\forall u_i, u_j \in U$
normalizzazione:	$S_{ii} = 0$	$\forall u_i \in U$
simmetria:	$S_{ij} = S_{ji}$	$\forall u_i, u_j \in U$

Gli indici di similarità sono definiti con riferimento agli elementi di un insieme (unità statistiche), anziché ai corrispondenti vettori riga, e assumono valori nell'intervallo chiuso [0, 1], anziché un qualunque valore non negativo (come accade invece a una distanza).

Osserviamo che il complemento a uno di un indice di similarità è detto indice di dissimilarità e rappresenta una classe di indici di prossimità più ampia delle distanze, che devono soddisfare anche la disuguaglianza triangolare.

Dal punto di vista operativo, i principali indici di similarità fanno riferimento a matrici dei dati contenenti variabili dicotomiche. Casi più generali, con variabili a più modalità qualitative, possono essere ricondotti al precedente mediante la tecnica della **binarizzazione**.

Attraverso la binarizzazione si rendono “metriche” le variabili qualitative, trasformando ogni variabile qualitativa in tante variabili binarie quante sono le modalità della stessa. Per esempio, se una variabile qualitativa *X* ha *r* modalità, saranno costruite *r* variabili binarie. Poniamo *i* come generica modalità. La corrispondente variabile binaria varrà 1

tutte le volte che, in corrispondenza della variabile *X*, l'unità statistica assume il valore i e 0 altrimenti.

Per capire il funzionamento di tali indici di similarità, consideriamo come esempio il comportamento di due visitatori di un sito web, nei confronti delle p = 20 pagine che essi possono visitare all'interno del sito.

Rappresentiamo tutte le possibilità della loro condotta nella tabella 2.4, tenendo conto che il comportamento di tali persone si esplica nella visita o meno delle pagine in questione (le pagine sono infatti variabili dicotomiche che assumono valore 1 se vengono visitate oppure 0 nel caso contrario).

Visitatore B / **Visitatore A**	**1**	**0**	**Totale**
1	$CP = 3$	$PA = 4$	7
0	$AP = 6$	$CO = 7$	13
Totale	9	11	$p = 20$

Tabella 2.4 – Comportamento di visita di due utenti

Sulle 20 pagine considerate, 3 sono state visitate da entrambi i visitatori; 3 rappresenta quindi la frequenza assoluta di fenomeni contemporaneamente presenti nelle due unità statistiche (*CP* significa co-presenze o positive matches); 7 è la frequenza dei fenomeni assenti in entrambe le unità (*CO* significa co-assenze o negative matches); infine 4 e 6 indicano la frequenza delle pagine che uno solo dei due visitatori, alternativamente, visita (*PA* significa presenza-assenza, allo stesso modo *AP* significa assenza-presenza, la prima lettera si riferisce al visitatore A e la seconda al visitatore B). Queste ultime due frequenze denotano gli aspetti di diversità tra i due visitatori e vanno quindi trattate

nello stesso modo, essendo simmetriche. Mentre l'importanza delle altre due frequenze, co-presenza e co-assenza, nel calcolo degli indici di similarità non è identica. Risultano infatti più importanti le co-presenze, perché concorrono a determinare la similarità tra i due visitatori, al contrario le co-assenze risultano meno importanti da questo punto di vista.

Esaminiamo brevemente i principali indici di similarità sviluppati in letteratura.

Indice di similarità di Russel e Rao: $$S_{ij} = \frac{CP}{p}$$

Indice di similarità di Jaccard: $$S_{ij} = \frac{CP}{CP + PA + AP}$$

Indice di similarità di Sokal e Michener: $$S_{ij} = \frac{CP + CA}{p}$$

Per questo ultimo indice si può dimostrare che il suo complemento a uno corrisponde alla media del quadrato della distanza euclidea fra i due vettori binari associati alle unità statistiche:

$$1 - S_{ij} = \frac{1}{p}\left({}_2 d_{ij}^2\right)$$

Questa relazione mostra che il **complemento a uno dell'indice di Sokal e Michener** è una distanza. In effetti, tale indice è uno degli indici di similarità più usati. È noto anche come coefficiente di "simple matching" o anche "binary distance".

Cluster analysis

Andiamo ora ad illustrare uno dei principali metodi computazionali, che si propone di effettuare la classificazione delle unità statistiche in gruppi (cluster).

L'obiettivo della cluster analysis, data una matrice dei dati *X* composta da *n* osservazioni (righe) e *p* variabili (colonne), è quello di mettere insieme le unità statistiche in gruppi il

più possibile omogenei al loro interno (coesione interna) ed eterogenei tra di loro (separazione esterna).

Vi sono numerosi modi per effettuare un'analisi di raggruppamento. Risulta quindi doveroso definire chiaramente i modi in cui la stessa viene svolta. In particolare, le scelte da effettuare dovranno riguardare:

1. **la scelta delle variabili da utilizzare** che deve tener conto di tutti gli aspetti rilevanti per il conseguimento degli obiettivi prefissati e, quindi, di tutte le variabili necessarie a tal fine, tenendo presente che l'utilizzo di variabili poco significative porta inevitabilmente a un peggioramento dei risultati. Questa scelta è un problema cruciale poiché condizionerà fortemente il risultato finale. In generale si può affermare che una classificazione può considerarsi soddisfacente quando non mostra un'eccessiva sensibilità a piccoli cambiamenti dell'insieme di variabili utilizzate;

2. **il metodo di formazione dei gruppi**: a questo proposito si distinguono metodi gerarchici e metodi non gerarchici. I metodi **gerarchici** consentono di ottenere una successione di raggruppamenti (detti partizioni) con un numero di gruppi da *n* a 1, partendo dalla più semplice in cui tutte le unità sono distinte, fino a quella in cui tutti gli elementi appartengono ad un unico gruppo. I metodi **non gerarchici** permettono invece di raggruppare le *n* unità statistiche in un numero di gruppi fissato (soggettivamente) a priori;

3. **l'indice di prossimità da utilizzare**: a seconda della natura delle variabili a disposizione, deve solitamente essere definita una misura di **prossimità fra le unità statistiche**, da utilizzare per calcolare la matrice delle distanze fra di esse. Sottolineiamo nuovamente l'importanza di una eventuale standardizzazione delle variabili, per evitare che alcune pesino più di altre nella determinazione dei risultati finali. Oltre a stabilire una misura di prossimità fra le unità statistiche, è necessario stabilire, nel caso dei metodi gerarchici, come verrà calcolata la **prossimità fra i gruppi** ottenuti nelle diverse fasi della procedura;

4. **la determinazione dei criteri di valutazione dei gruppi ottenuti**: valutare il risultato di raggruppamento ottenuto significa verificare che i gruppi siano coerenti con l'obiettivo primario della cluster analysis e che soddisfino quindi le condizioni di coesione interna e separazione esterna. Di fondamentale importanza è, a tal fine, la scelta del numero dei gruppi. Esiste un trade-off tra l'ottenimento di gruppi omogenei, caratteristica che è tipicamente funzione crescente del numero dei gruppi scelto, e la necessità di ottenere una rappresentazione parsimoniosa, che richiede, al contrario, un numero ridotto di gruppi.

Cominciamo col descrivere i **metodi gerarchici di classificazione**, che permettono di ottenere una famiglia di partizioni, ciascuna associata ai successivi livelli di raggruppamento fra le unità statistiche, calcolati sulla base dei dati a disposizione. Le diverse famiglie di partizioni possono essere rappresentate graficamente, mediante una struttura ad albero, detto albero di classificazione gerarchica o dendrogramma. Tale struttura associa a ogni passo della procedura gerarchica, che corrisponde a un numero g fissato di gruppi, una e una sola classificazione delle unità statistiche nei g gruppi.

Nella figura 2.2 possiamo vedere come si rappresenta graficamente un dendrogramma, per semplicità si supponga che esistano solamente 5 unità statistiche a disposizione.

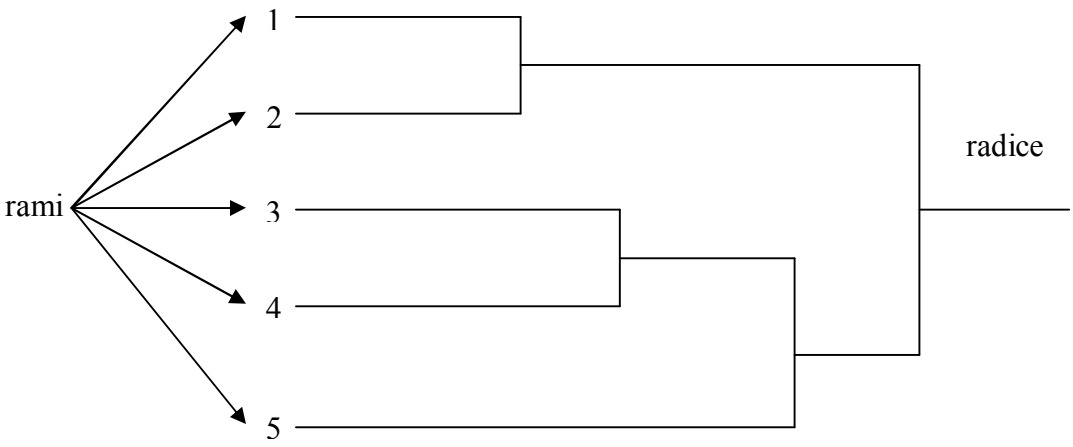

Figura 2.2 – Struttura del dendrogramma

I rami dell'albero descrivono classificazioni successive delle unità statistiche. Alla radice dell'albero, tutte le unità statistiche sono contenute in una sola classe. Le successive divisioni in rami individuano divisioni successive delle unità in cluster. Infine, i rami terminali indicano la partizione finale delle unità statistiche.

Se la formazione dei gruppi avviene dai rami alla radice (nella Figura 2.2, da sinistra verso destra), vale a dire, se si parte dalla situazione in cui ogni unità statistica appartiene a un gruppo a sé stante e si procede a un raggruppamento di tali unità, i metodi di classificazione gerarchica vengono detti **agglomerativi**. Invece, se la costruzione dei cluster avviene dalla radice ai rami dell'albero i corrispondenti metodi gerarchici vengono detti **scissori**.

Le successive partizioni individuate da un dendrogramma sono "nidificate". Ciò significa che, nei metodi gerarchici, gli elementi che vengono uniti (o divisi) a un certo passo resteranno uniti (divisi) fino alla fine del processo di classificazione. Questo modo di procedere ha il vantaggio di ridurre il numero di partizioni da confrontare, rendendo la procedura computazionalmente più efficiente, ma anche lo svantaggio di non poter "correggere" errori di classificazione commessi nei passi precedenti.

A questo punto possiamo descrivere, con riferimento ai metodi agglomerativi, la procedura statistica per ottenere un dendrogramma. Risulta opportuno schematizzare la procedura nelle fasi riassunte di seguito.

1. **Inizializzazione**: date n unità statistiche da classificare, ogni elemento rappresenta un gruppo (si hanno, in altri termini, n cluster). I cluster verranno indicati con un numero che va da 1 a n.

2. **Selezione**: vengono selezionati i due cluster più "vicini" rispetto alla misura di prossimità fissata inizialmente. Per esempio, rispetto alla distanza euclidea.

3. **Aggiornamento**: si aggiorna il numero dei cluster (che sarà pari a n - 1) attraverso l'unione, in un unico cluster, dei due gruppi selezionati nel punto precedente. Conseguentemente, si aggiorna la matrice delle distanze, sostituendo, alle due righe (colonne) di distanze relative ai due cluster, nei confronti di tutti gli altri, una sola riga di distanze, "rappresentativa" del nuovo gruppo. I metodi agglomerativi differiscono per il modo in cui viene definita tale rappresentatività.

4. **Ripetizione**: si eseguono i passi (2) e (3) n - 1 volte.

5. **Arresto**: la procedura si arresta quando tutti gli elementi vengono incorporati in un unico cluster.

In base ai diversi modi in cui vengono calcolate le distanze fra il gruppo neo-formato e le altre unità statistiche, si distinguono diversi metodi gerarchici di classificazione. Introduciamo i diversi metodi considerando la distanza fra due gruppi C_1 e C_2, con numerosità n_1 e n_2, uno dei quali potrebbe essere un nuovo gruppo.

Per prima cosa bisogna distinguere fra i metodi che richiedono esclusivamente, come input, la matrice di distanza, e i metodi che richiedono anche la matrice dei dati. Cominciamo con il primo tipo.

Metodo del legame singolo (single linkage): la distanza tra due gruppi è definita come il minimo delle $n_1 \cdot n_2$ distanze tra ciascuna delle unità del gruppo C_1 e ciascuna delle unità del gruppo C_2. In formula:

$$d(C_1, C_2) = \min(d_{rs}) \quad \text{dove} \quad r \in C_1, s \in C_2$$

Metodo del legame completo (complete linkage): la distanza tra due gruppi è definita come il massimo delle $n_1 \cdot n_2$ distanze tra ciascuna delle unità di un gruppo e ciascuna delle unità dell'altro gruppo. In formula:

$$d(C_1, C_2) = \max(d_{rs}) \quad \text{dove} \quad r \in C_1, s \in C_2$$

Metodo del legame medio (average linkage): la distanza tra due gruppi è definita come la media aritmetica delle $n_1 \cdot n_2$ distanze tra ciascuna unità di un gruppo e ciascuna unità dell'altro gruppo. Formalmente:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{r=1}^{n_1} \cdot \sum_{s=1}^{n_2} d_{rs} \quad \text{dove} \quad r \in C_1, s \in C_2$$

Il secondo tipo di metodi richiede anche la matrice dei dati. Illustriamo in seguito i principali.

Metodo del Centroide: la distanza tra due gruppi C_1 e C_2 di numerosità n_1 e n_2 è definita come la distanza tra i rispettivi centroidi (medie aritmetiche), $\bar{x}_1$ e $\bar{x}_2$. In formula:

$$d(C_1, C_2) = d(\bar{x}_1, \bar{x}_2)$$

II metodo del centroide e il metodo del legame medio presentano delle analogie: il metodo del legame medio considera la media delle distanze tra le unità di ciascuno dei due gruppi, mentre il metodo del centroide calcola le medie di ciascun gruppo, e in seguito misura le distanze tra di esse.

Metodo di Ward: questo metodo minimizza, nella scelta dei gruppi da aggregare, una funzione obiettivo che parte dal presupposto che una classificazione ha l'obiettivo di creare gruppi che abbiano la massima coesione interna e la massima separazione esterna.

La *devianza totale* delle *p* variabili viene scomposta in *devianza nei gruppi* e *devianza fra i gruppi*:

$$D_T = D_N + D_F$$

Formalmente, data una partizione di *g* gruppi:

- la devianza totale delle *p* variabili corrisponde alla somma delle devianze delle singole variabili rispetto alla corrispondente media generale $\bar{x}_s$:

$$D_T = \sum_{s=1}^{p} \cdot \sum_{i=1}^{n} \left(x_{is} - \bar{x}_s\right)^2$$

- la devianza nei gruppi è data dalla somma delle devianze di ciascun gruppo:

$$D_N = \sum_{k=1}^{g} W_k$$

dove W_k rappresenta la devianza delle *p* variabili nel gruppo k-esimo (di numerosità n_k e centroide $\bar{x}_k = [\bar{x}_{1k},...,\bar{x}_{pk}]$), descritta dalla seguente:

$$W_k = \sum_{s=1}^{p} \cdot \sum_{i=1}^{n_k} \left(x_{is} - \bar{x}_{sk}\right)^2$$

- la devianza fra i gruppi è data dalla somma delle devianze (ponderate) delle medie di gruppo rispetto alla corrispondente media generale:

$$D_F = \sum_{s=1}^{p} \cdot \sum_{k=1}^{g} n_k \left(\bar{x}_{sk} - \bar{x}_s\right)^2$$

Nel metodo di Ward, ad ogni passo della procedura gerarchica si aggregano tra loro i gruppi che comportano il minor incremento della devianza nei gruppi, D_N (e, quindi,

maggior incremento di D_F), ovvero consentono di ottenere la maggiore coesione interna possibile (e, quindi, la maggior separazione esterna possibile).

Per **valutare i metodi gerarchici** è necessario verificare che le partizioni conseguano l'obiettivo primario della cluster analysis, secondo il quale i gruppi ottenuti siano caratterizzati da coesione interna e separazione esterna. Ad ogni passo della procedura gerarchica viene valutata la bontà della corrispondente partizione ottenuta, in modo tale da poter scegliere quale sia più consona al raggiungimento degli obiettivi dell'analisi.

I principali indici per la valutazione si basano sulla scomposizione della devianza totale delle *p* variabili. Si definisce valida una classificazione caratterizzata da una bassa devianza nei gruppi e da un elevato valore della devianza fra i gruppi. Nel caso di *g* gruppi, un indice sintetico che misura la rispondenza a tale criterio è il seguente:

$$R^2 = 1 - \frac{D_N}{D_T} = \frac{D_F}{D_T} \qquad R^2 \in [0,1]$$

Una prima analisi dell'indice ci porterebbe a concludere che la partizione dei gruppi è ottimale se l'indice è prossimo a 1. Ciò tuttavia andrebbe a scapito della parsimonia della classificazione che, in generale, dovrebbe essere una delle finalità principali di una valida analisi statistica. Pertanto, la massimizzazione di R^2 non può costituire l'unico criterio su cui basarsi per la definizione del numero ottimale di gruppi. Tale criterio infatti condurrebbe a una classificazione costituita da *n* gruppi formati da una sola unità (tale per cui $R^2 = 1$).

Una misura alternativa all'indice R^2 è la Root-Mean-Square Standard Deviation o, semplicemente, **RMSSTD**. Tale indice considera solamente la parte della devianza nei gruppi "aggiuntiva" che si forma al corrispondente passo della procedura di classificazione gerarchica.

Considerando il passo *h*-esimo ($h = 2,...,n-1$) della procedura, l'indice RMSSTD è definito dalla seguente:

$$RMSSTD = \sqrt{\frac{W_h}{p(n_h - 1)}}$$

dove W_h è la devianza nel gruppo che si è costituito al passo *h* della procedura; n_h è la sua numerosità e *p* è il numero di variabili considerate.

Dal punto di vista interpretativo, un forte incremento di RMSSTD rispetto al passo precedente mostra che i due gruppi che si sono uniti sono fortemente eterogenei e, pertanto, sarebbe opportuno arrestare la procedura al passo precedente.

Un altro indice che, similmente a RMSSTD, misura il contributo "aggiuntivo" del passo *h*-esimo della procedura è il cosiddetto R^2 semiparziale (**SPRSQ**). Tale indice è definito da:

$$SPRSQ = \frac{(W_h - W_r - W_s)}{D_T}$$

dove *h* è il nuovo gruppo, ottenuto al passo *h* come fusione dei gruppi *r* ed *s*, D_T è la devianza totale delle osservazioni, mentre W_h, W_r e W_s indicano, rispettivamente, le devianze interne ai gruppi *h*, *r* ed *s*. In altri termini, SPRSQ misura l'incremento della devianza all'interno del gruppo ottenuto unendo i gruppi *r* e *s*. Un brusco innalzamento indica che si stanno unendo gruppi eterogenei e, pertanto, è opportuno arrestarsi al passo precedente.

L'indice R^2 ed uno a scelta fra i coefficienti RMSSTD e SPRSQ consentono quindi di valutare adeguatamente il grado di omogeneità (o coesione) dei gruppi ottenuti in ciascun passo di una classificazione gerarchica e di scegliere la partizione più soddisfacente.

Abbiamo finora descritto i metodi gerarchici di classificazione, andiamo a illustrare nel seguito i metodi non gerarchici.

I **metodi non gerarchici di classificazione** permettono di ottenere una sola partizione delle *n* unità statistiche in *g* gruppi (con *g* generalmente minore di *n*) il cui numero (*g* appunto) viene definito a priori da colui che svolge la classificazione.

A differenza di quanto accade nei metodi gerarchici, si perviene ad un unico raggruppamento che soddisfa determinati criteri di ottimalità, quali il raggiungimento della ripartizione che consente di ottenere la massima coesione interna, per un numero di gruppi prefissato.

Per ogni valore di *g*, ovvero per ogni numero dei gruppi in base al quale si intendono classificare gli *n* elementi iniziali, l'algoritmo non gerarchico classifica ciascuno di questi elementi fondandosi esclusivamente sul criterio prescelto e giunge, di conseguenza, a risultati diversi per diversi valori attribuiti a *g*.

In generale, negli algoritmi di classificazione non gerarchici viene seguita una procedura di analisi che si può schematizzare nelle seguenti fasi.

1. **Scelta del numero dei gruppi**, *g*, e conseguente scelta di una classificazione iniziale delle *n* unità statistiche in tali gruppi.

2. **Valutazione del "trasferimento"** di ciascuna unità statistica dal gruppo di appartenenza a un altro gruppo. Ciò al fine di massimizzare la coesione interna dei gruppi. Viene calcolata la variazione nella funzione obiettivo causata dallo spostamento e, se questa è rilevante, il trasferimento diviene permanente.

3. **Ripetizione** del punto precedente finché non viene soddisfatta una regola di arresto.

Gli algoritmi non gerarchici sono generalmente molto più veloci di quelli gerarchici, poiché ricorrono ad una struttura di calcolo, di tipo iterativo, che non richiede la determinazione preliminare della matrice delle distanze. Inoltre, per il modo in cui vengono costruiti, risultano tipicamente più stabili, rispetto alla variabilità campionaria. Gli algoritmi non gerarchici si rivelano perciò adatti per data set di grandi dimensioni.

Tuttavia, il numero di modi in cui è possibile suddividere *n* elementi in *g* gruppi non sovrapposti è molto grande, specie per dati reali, ed è impossibile ottenere e confrontare tutte queste combinazioni. Pertanto, per questo motivo, risulta difficile massimizzare globalmente la funzione obiettivo e, quindi, gli algoritmi di classificazione non gerarchica dovranno accontentarsi di soluzioni vincolate, spesso corrispondenti a massimi locali.

Dobbiamo evidenziare che gli aspetti critici connessi ai metodi non gerarchici di classificazione consistono soprattutto nella necessità di definire preliminarmente il numero dei gruppi e la configurazione di partenza dei gruppi, per inizializzare l'algoritmo iterativo di classificazione.

Il metodo di segmentazione non gerarchica più utilizzato è il metodo delle ***k*-medie** (*k*-means), con *k* che indica il numero dei gruppi stabilito (coincide con *g*).

La scelta del numero dei gruppi costituisce il punto critico della classificazione e influenza i passaggi successivi. Il criterio più utilizzato per decidere il numero dei gruppi consiste nell'effettuare ripetutamente l'analisi con diversi valori di *g* e nel valutare, successivamente, la soluzione migliore (sulla base di criteri opportuni di misurazione della bontà della partizione quale, per esempio, l'indice R^2).

Una simile metodologia presenta però l'inconveniente di non garantire che l'algoritmo sia effettivamente in grado di giungere alla soluzione ottimale. Una possibile soluzione alternativa consiste nel far precedere l'analisi non gerarchica da una di tipo gerarchico, in modo che quest'ultima possa dare suggerimenti circa il possibile valore di *g* che ottimizza la classificazione.

Presentiamo sinteticamente l'algoritmo di classificazione delle *k*-medie, che effettua una classificazione degli *n* elementi di partenza, in *g* gruppi distinti, con g fissato a priori, secondo il seguente flusso operativo:

1. **Scelta dei semi iniziali** (seeds): dopo aver determinato il numero dei gruppi, vengono definiti *g* punti nello spazio *p*-dimensionale che costituiscono i

centroidi (misure di posizione, di solito medie) dei cluster nella partizione iniziale. I centroidi dovrebbero essere sufficientemente distanti tra loro, affinché migliorino le proprietà di convergenza dell'algoritmo. Una volta definiti i centroidi, si costruisce una partizione iniziale delle unità statistiche, allocando ciascuna unità al gruppo il cui centroide risulta più vicino.

2. **Calcolo della distanza di ogni unità statistica dai centroidi** (medie) dei *g* gruppi: la distanza tra una generica unità statistica ed il centroide del gruppo a cui è stata assegnata deve essere minima e, nel caso in cui non lo fosse, l'elemento corrispondente verrà riassegnato al cluster il cui centroide è più vicino. Quando avviene tale spostamento vengono ricalcolati i centroidi del vecchio e del nuovo gruppo di appartenenza.

3. **Ripetizione del passo precedente** fino al raggiungimento della convergenza dell'algoritmo; in altri termini, il precedente punto viene ripetuto fino a raggiungere un'adeguata stabilizzazione dei gruppi. Per calcolare la distanza tra le unità statistiche ed i centroidi dei gruppi viene utilizzata la distanza euclidea: all'iterazione *t*, la distanza tra l'unità *i*-esima ed il centroide del gruppo *l* (con $i = 1,2,...,n$ e $l = 1,2,...,g$) sarà pari a:

$$d(x_i, \bar{x}_l^{(t)}) = \sqrt{\sum_{s=1}^{p} \left(x_{is} - \bar{x}_{s,l}^{(t)}\right)^2}$$

dove $\bar{x}_l^{(t)} = [\bar{x}_{1,l}^{(t)},...,\bar{x}_{p,l}^{(t)}]'$ è il centroide del gruppo *l* calcolato all'iterazione *t*.

Il metodo delle *k*-medie persegue l'obiettivo della ricerca della partizione degli *n* elementi iniziali in *g* gruppi (con *g* prefissato) che soddisfi un criterio di coesione interna fondato sulla minimizzazione della devianza nei gruppi; pertanto la bontà della soluzione ottenuta con questo algoritmo può essere controllata attraverso il calcolo dell'indice R^2.

L'algoritmo è particolarmente indicato quando si vogliono conoscere le caratteristiche di ciascun gruppo, come espresse da opportune misure di sintesi, quali i centroidi.

Un possibile svantaggio del metodo delle *k*-medie consiste nella presenza di notevoli distorsioni dei risultati nel caso in cui nei dati vi fossero dei valori anomali o *outliers*. In questo caso si deve partire da un numero di gruppi molto elevato per verificare l'esistenza di questi valori poiché, con molta probabilità, le unità non anomale tenderanno a concentrarsi in pochi gruppi, mentre gli outliers rimarranno isolati nella classificazione formando dei gruppi anche contenenti un solo elemento.

Analisi di segmentazione

L'analisi di segmentazione attua un raggruppamento delle unità statistiche in un'ottica asimmetrica. In pratica, si assume che fra le variabili a disposizione ve ne sia una (variabile risposta) che si possa considerare come dipendente dalle altre (variabili esplicative). L'obiettivo dell'analisi di segmentazione è la classificazione delle unità statistiche in gruppi fra loro omogenei, con riferimento alle modalità della variabile risposta.

Mentre la cluster analysis effettua una classificazione "non supervisionata" delle unità statistiche, in funzione di tutte le variabili a disposizione, l'analisi di segmentazione attua una classificazione in funzione di tutte le variabili esplicative a disposizione, "**supervisionata**" dalla presenza di una variabile *target*, o risposta, per la quale sono note a priori le modalità.

Dal punto di vista dei risultati, anche l'analisi di segmentazione produce una partizione delle unità statistiche. In effetti, l'output dell'analisi è solitamente rappresentato mediante una struttura ad albero, detta **albero decisionale**, che è molto simile, nella struttura, ad un albero di classificazione gerarchica (dendrogramma). Per un esempio si osservi la figura 2.3, dove si classificano le diverse specie di iris in base alla lunghezza dei petali. Ad ogni divisione dell'albero viene riportato il criterio di valutazione *improvement*, in questo caso è stato utilizzato l'indice di eterogeneità di Gini[10]. Nella prima divisione dell'albero il miglioramento è pari a 0,3106. Significa che l'impurità

[10] Una breve spiegazione dell'indice è riportata nel paragrafo 2.2.

(eterogeneità) che risulta dalla divisione nei due rami è 0.3106 in meno rispetto all'impurità del primo nodo.

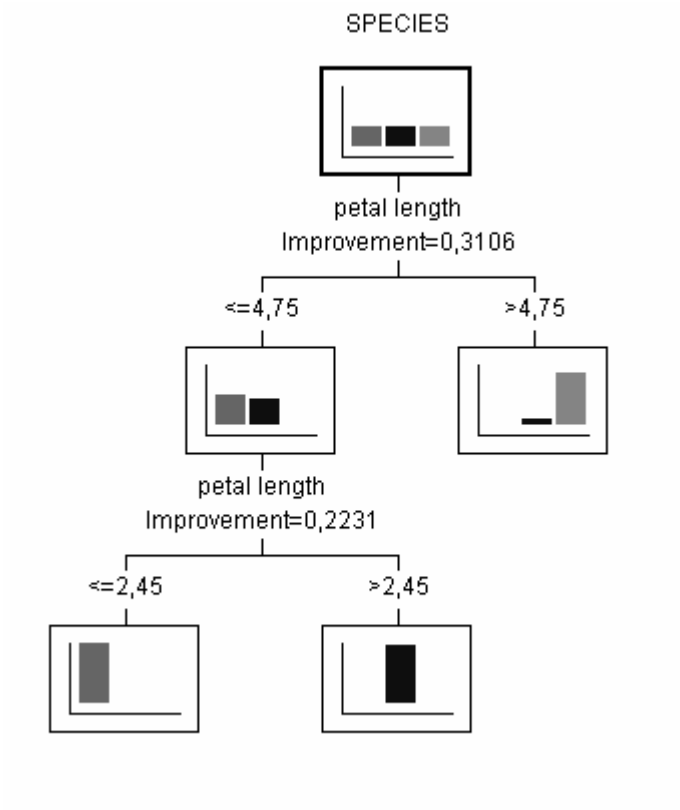

Figura 2.3 – Albero decisionale di un'analisi di segmentazione

Nonostante le similitudini "grafiche", vi sono importanti differenze fra un'analisi di raggruppamento gerarchica e un'analisi di segmentazione. La prima, che abbiamo già introdotto, riguarda la presenza di una variabile "supervisore". Una seconda importante differenza riguarda il funzionamento della regola di partizione, ai vari livelli della procedura. Mentre negli alberi decisionali la segmentazione viene tipicamente attuata utilizzando una sola variabile esplicativa alla volta (quella massimamente predittiva), negli alberi di classificazione gerarchici la regola divisiva (o agglomerativa) fra i gruppi viene stabilita in base a considerazioni sulla distanza fra di essi, calcolata utilizzando tutte le variabili a disposizione.

Infine, mentre l'analisi di raggruppamento ha finalità prevalentemente descrittive, di riduzione della complessità del data set, l'analisi di segmentazione ha come scopo

principale la produzione di regole di classificazione delle unità statistiche, utilizzabili ai fini previsivi.

Esponiamo brevemente gli aspetti salienti della metodologia di segmentazione, facendo principalmente riferimento al caso di alberi di classificazione per variabili risposta qualitative e tipo di segmentazioni binarie.

L'analisi di segmentazione può essere definita come una procedura ricorsiva[11], attraverso la quale un insieme di *n* unità statistiche viene progressivamente suddiviso in gruppi, secondo una regola divisiva che mira a massimizzare l'omogeneità interna ai gruppi ottenuti. A ogni passo della procedura, la regola divisiva è specificata da una partizione dei valori di una delle variabili esplicative. Pertanto, ad ogni passo, la scelta di una regola divisiva implica la scelta di quale variabile esplicativa utilizzare e di come partizionarla. Facciamo notare che, come nei metodi di classificazione gerarchica, il tipo di partizione effettuato a un certo stadio è influenzato dalle scelte precedenti.

Il termine della procedura di segmentazione è stabilito da un criterio di arresto, da fissare preliminarmente. Al raggiungimento di tale termine, la procedura conduce a una regola di classificazione delle unità statistiche, che è funzione delle modalità delle variabili esplicative considerate nel processo. Tale regola potrà essere impiegata successivamente per classificare nuove osservazioni.

L'analisi di segmentazione costituisce un approccio di tipo non parametrico, che non richiede assunzioni sulla distribuzione di probabilità della variabile risposta. Effettivamente, a conferma di questa flessibilità, l'analisi di segmentazione è generalmente applicabile, qualunque sia la natura della variabile dipendente o delle variabili esplicative.

Gli alberi decisionali vengono spesso distinti in **alberi di regressione** o **alberi di classificazione** a seconda che la variabile risposta sia, rispettivamente, quantitativa o qualitativa.

[11] Procedimento che consiste nella applicazione ripetuta di una serie di operazioni, usando ogni volta come base di partenza il risultato dell'esecuzione precedente.

Dal punto di vista interpretativo, gli alberi decisionali creano diagrammi di facile lettura ed è generalmente semplice seguire un sentiero dell'albero per poter spiegare particolari classificazioni. Determinano facilmente i gruppi significativi e forniscono un'immagine sintetica e intuitiva delle relazioni tra le variabili esplicative. Tuttavia, il processo di crescita dell'albero è computazionalmente costoso, poiché a ogni nodo bisogna comparare tutti i possibili divisori per scegliere il migliore. La complessità della loro costruzione può inoltre comportarne una intrinseca instabilità. Infine, sottolineiamo che le regole decisionali prodotte sono molto dipendenti dai dati osservati, e sono difficilmente generalizzabili a contesti differenti.

A partire dall'idea di analizzare un insieme di unità dividendole con una procedura ricorsiva in sottoinsiemi, sono stati avanzati molti algoritmi ricorsivi di segmentazione. Tra questi, i più diffusi sono contraddistinti dagli acronimi **CART** e **CHAID** che rispettivamente significano "Classification and regression trees" e "Chi-squared automatic interaction detection"[12]. Prima di descrivere le caratteristiche di questi due algoritmi, verranno introdotti i due principali aspetti della metodologia di segmentazione: i criteri divisivi e di arresto.

II principale elemento distintivo di un albero decisionale è il modo in cui viene scelta la **regola divisiva** delle unità appartenenti a un gruppo, corrispondenti a un nodo dell'albero decisionale. Come già anticipato, ciò equivale alla scelta del predittore migliore fra quelli disponibili, nonché alla scelta della partizione migliore, fra quelle a esso corrispondenti. Generalmente entrambe le scelte vengono effettuate calcolando, in corrispondenza di ogni predittore e partizione, un indice di efficacia della partizione che viene successivamente massimizzato.

Come **indice di efficacia** si utilizza una funzione $\Phi(s,t)$, detta funzione criterio di segmentazione, che fornisce una misura della diversità tra i valori della variabile risposta nei gruppi (figli) generati dalla suddivisione s, e quelli nel gruppo (genitore) t. La massimizzazione della funzione porta a individuare la suddivisione che genera

[12] L'algoritmo CART venne proposto da Breiman *et al* (1984), mentre il CHAID fu presentato da Kass (1980).

sottoinsiemi massimamente eterogenei tra loro e massimamente omogenei al loro interno.

Una delle funzioni criterio più spesso adottate si basa sul concetto di **impurità** delle unità statistiche presenti nei gruppi. Il concetto di impurità corrisponde al concetto di eterogeneità delle unità statistiche, con riferimento alle modalità della variabile risposta.

Sia $I(t)$ un indice statistico di impurità, ovvero di eterogeneità di un gruppo, con riferimento alle osservazioni riguardanti la variabile risposta. Come indice si può utilizzare, ad esempio, l'indice di Gini o l'indice entropico descritti all'inizio del capitolo. La funzione **criterio di segmentazione** è misurata dalla riduzione della disomogeneità ottenuta segmentando il gruppo genitore *t* nel modo *s*:

$$\Phi(s,t) = I(t) - \sum_{r=1}^{m} I(t_r)$$

dove t_r sono i gruppi figli generati dalla segmentazione *s* (*m* = 2 per la segmentazione binaria).

Al livello successivo, i sottogruppi creati dalla precedente spaccatura sono a loro volta divisi in base alla regola che risulta per loro migliore.

L'albero decisionale potrebbe crescere, in assenza di **criteri di arresto**, fino a quando ogni nodo contenga osservazioni identiche, in termini di modalità della variabile dipendente. Ciò potrebbe non costituire una segmentazione ottimale.

Sono necessari pertanto dei criteri di arresto, da applicarsi a seguito di ogni suddivisione, che possano determinare l'interruzione della crescita di un albero decisionale.

I metodi di segmentazione più diffusi solitamente utilizzano regole di arresto basate su soglie minime di numerosità dei nodi terminali, oppure sul numero massimo di passi del processo.

Il metodo CART, invece, utilizza una strategia alternativa all'implementazione dei criteri di arresto, basata sul concetto di **potatura** (pruning). In tale approccio, si costruisce dapprima l'albero di maggiori dimensioni, nel quale ogni nodo contiene solo un elemento oppure elementi appartenenti alla stessa classe. Successivamente l'albero viene "potato" secondo una regola che massimizza la capacità selettiva, a parità di complessità.

Un criterio generale per la scelta della migliore regola classificatoria fa riferimento al **tasso di errata classificazione**. La percentuale di classificazioni errate, vale a dire, di unità classificate in una modalità differente dal valore osservato, è detto tasso di errata classificazione, e costituisce la principale misura di performance di un albero decisionale. A parità di semplicità di rappresentazione (ovvero di numero di nodi terminali) verrà scelta la partizione che minimizza il tasso di errata classificazione.

Nell'algoritmo CART la potatura procede con una procedura ricorsiva a ritroso: per ogni classificazione successiva (partendo da quella corrispondente alla maggiore dimensione possibile dell'albero) viene eliminato il nodo con minore performance, ovvero con maggiore valore della **funzione di perdita** $R_a(T)$ definita dalla seguente:

$$R_a(T) = R(T) + aN(T)$$

La funzione di perdita viene misurata congiuntamente in termini di errore di classificazione e di complessità. Per una partizione finale T, $R(T)$ ne indica il tasso di errata classificazione, $N(T)$ il numero di nodi terminali, e a è una costante che stabilisce la penalizzazione per la complessità desiderata.

Un criterio di pruning alternativo si basa sulla minimizzazione dell'errore di previsione. Si tratta di impiegare un secondo data set, detto di validazione, che contiene osservazioni differenti da quelle di addestramento. Gli alberi decisionali da confrontare vengono costruiti sulla base dei dati nell'insieme di apprendimento, ma confrontati in termini di errore di classificazione delle nuove unità contenute nell'insieme di validazione.

II più noto algoritmo di segmentazione è il metodo CART, diventato uno dei metodi più popolari per costruire gli alberi decisionali. L'algoritmo CART costruisce un albero binario dividendo le osservazioni a ogni nodo, dopo aver deciso quale tra le variabili esplicative è la migliore discriminante; ammette, come variabile risposta, sia variabili qualitative sia quantitative, e così pure come variabili esplicative.

Adotta, per la determinazione dell'albero ottimale, anziché criteri espliciti di arresto, la tecnica della potatura. La funzione criterio per la segmentazione adottata dall'algoritmo CART è la distanza tra i valori effettivi della variabile risposta y_i, ed i valori attesi nell'ipotesi di omogeneità (coincidenti con la media del gruppo).

L'algoritmo CHAID si basa, per la costruzione della regola divisiva, sul test del chi-quadrato. Ammette, come variabile risposta e variabili esplicative, solo variabili qualitative. La principale differenza rispetto all'algoritmo CART è che CHAID preferisce bloccare la crescita dell'albero al livello ottimale, mediante un criterio di arresto esplicito, basato sulla significatività del test di omogeneità del chi-quadrato.

Per quanto riguarda la funzione criterio della segmentazione, CHAID utilizza, come indice di impurità di un gruppo la distanza tra le frequenze osservate e attese, nell'ipotesi di omogeneità. Ciò conduce a ottenere, come funzione criterio della segmentazione, l'indice c^2 di Pearson. Inoltre, si può ricavare implicitamente un criterio di arresto, basato sul test di significatività dell'ipotesi di omogeneità, che viene rifiutata per valori elevati della statistica di Pearson.

Da ultimo, sottolineiamo che nell'algoritmo CHAID sono possibili regole divisive dei nodi multiple, anziché binarie. Ciò può rendere più veloce il raggiungimento della configurazione ottimale.

Reti neurali

In questi anni si è sviluppato un notevole interesse verso le reti neurali, non solo per la loro rilevanza teorica e scientifica in campi diversi come la matematica, la neurologia e la psicologia, ma anche per la concreta prospettiva di applicazioni pratiche a seguito

dell'evoluzione tecnologica degli strumenti informatici. In generale tali applicazioni spaziano dal riconoscimento e la classificazione di "forme" e/o di comportamenti, all'estrazione di significati da associazioni di dati apparentemente casuali, sino alla costruzione di regole predittive.

Le reti neurali comprendono una vasta classe di modelli sviluppati nell'ambito delle scienze cognitive che, grazie a un processo di apprendimento mirante a simulare il comportamento dei sistemi nervosi viventi, riescono a risolvere complessi problemi di classificazione e previsione. La capacità di valutare un gran numero di fattori, la tolleranza verso dati imperfetti, come la presenza di dati mancanti o altri problemi di qualità dei dati, ne fanno uno strumento particolarmente adatto per analizzare dati del mondo reale.

Il sistema nervoso, da cui prendono spunto le reti neurali, è costituito da un gran numero (cento miliardi nell'uomo) di cellule nervose, dette neuroni, connesse tra di loro in modo tale da formare un'immensa rete neurale, che è responsabile della maggior parte dei processi funzionali dell'organismo. Le reti neurali artificiali sono costituite da decine, centinaia al più migliaia di neuroni artificiali, reciprocamente connessi. Alcuni neuroni sono di "input" e ricevono i dati del problema da risolvere, altri neuroni sono di "output" e forniscono la soluzione del problema, altri ancora sono neuroni intermedi o inter-neuroni, denominati anche neuroni nascosti (*hidden*), perché sono interni ad una "scatola nera" che mostra solamente input e output.

Una rete di n neuroni può avere un numero enorme di architetture, in funzione delle connessioni reciproche: generalmente, nelle applicazioni si ricorre ad architetture standard relativamente semplici.

La rete neurale è composta di elementi di elaborazione equivalenti ai neuroni biologici, connessi tra loro con collegamenti equivalenti alle sinapsi, in modo da formare uno schema assimilabile ad un semplicissimo tessuto nervoso.

Il neurone artificiale i, può assumere diversi stati a seconda del tipo di struttura dello stesso:

- Gli stati tra 0 e 1 se il neurone è binario;
- Gli stati tra -1 e 1 se si tratta di un neurone bipolare;
- Tutti gli stati tra un valore minimo e un valore massimo, tipicamente tra 0 e 1, se si tratta di un neurone a valori continui.

Lo stato $S_i(t)$ del neurone evolve nel tempo discreto 1,2,...,*t* in funzione degli input che riceve da altri neuroni.

Ogni neurone riceve *n* input $S_j(t)$ da altrettanti neuroni *j*, attraverso sinapsi di valore W_{ij} (pesi sinaptici), con un input netto:

$$NET_i = \sum_{j=1}^{n} \left[W_{ij} S_j(t) - J_i \right]$$

dove J_i è una soglia caratteristica. Tale soglia può essere eliminata aggiungendo un $(n+1)$-esimo input fittizio o *bias* con valore $S_{(n+1)} = 1$ e peso sinaptico $W_{(n+1)} = -J_i$. In tal caso si ha un input netto (denominato spesso potenziale P_i):

$$NET_i = P_i = \sum_{j=0}^{n} \left[W_{ij} S_j(t) \right]$$

Lo stato successivo $S_i(t+1)$ viene calcolato con un opportuna **legge di attivazione**:

$$S_i(t+1) = F(P_i)$$

dove *F* è anche denominata **funzione di trasferimento**.

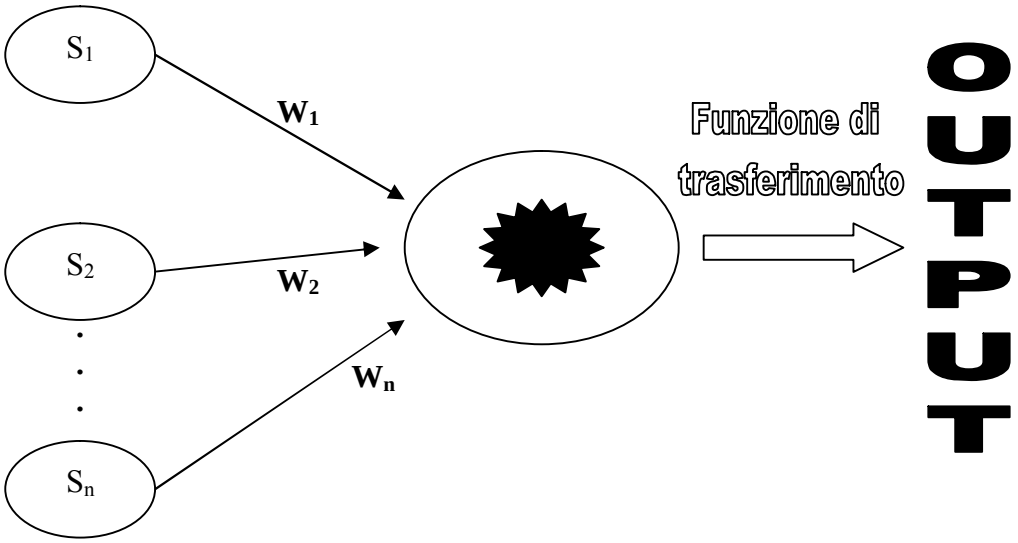

Figura 2.4 – Schema di un neurone artificiale

Pertanto il neurone è un operatore matematico che calcola una somma pesata dei suoi input, alla quale applica una legge di attivazione che genera l'output.

La funzione di trasferimento influenza la velocità e la capacità della rete di "adattarsi ai dati", e nel caso particolare di funzioni di trasferimento non lineari, come la sigmoide (o logistica), fa sì che la rete possa adattarsi a modelli non lineari.

Analogamente quindi ai neuroni biologici, la funzione di attivazione ha due principali caratteristiche: deve tenere conto della soglia e non deve mai superare un livello di saturazione.

Le funzioni di trasferimento più comuni sono:

1. **Funzione identità, lineare senza saturazione:**
   $F(P) = P$ o più in generale
   $F(P) = KP$.

2. **Funzione lineare con saturazione:**
   $F(P) = \min(\max(S), \max(0, KP))$.

3. **Funzione a gradino o di heavyside:**

   $F(P) = signum(P)$ con valori binari (0,1) o bipolari (-1,1)

   Nel primo caso si ha $F(P) = 0$ per $P \leq 0$, $F(P) = 1$ per $P > 0$; nel secondo caso si ha $F(P) = -1$ per $P \leq 0$, $F(P) = 1$ per $P > 0$.

4. **Funzione logistica o sigmoide:**

   $F(P) = \frac{1}{1 + e^{-kP}}$ con valori continui tra 0 e 1. Dove *k* è un parametro positivo che regola la pendenza della funzione.

   Tra le funzioni di attivazione, la funzione sigmoidale è quella che più si avvicina al funzionamento dei neuroni reali ed è la funzione di trasferimento più utilizzata.

Inizialmente, alle reti vengono assegnati dei pesi attraverso un processo aleatorio (in particolare valori normalizzati, ovvero compresi tra 0 e 1 o tra -0.5 e +0.5); esistono diversi metodi con cui le reti modificano automaticamente questi pesi fino ad assegnare loro quei valori che consentono di rispondere nel modo desiderato ed un certa stimolazione esterna. Tutti i **metodi di apprendimento** si dividono in due classi: metodo supervisionato e non-supervisionato.

L'**apprendimento supervisionato** si basa sulla disponibilità di una collezione di coppie [dati del problema / soluzione corrispondente] che viene solitamente ripartita in due insiemi:

- un *training set*, utilizzato per l'apprendimento;
- un *valid set* per verificare, ad addestramento concluso, che la rete non si limiti a memorizzare i casi del training set, ma possa anche trovare una soluzione appropriata per casi analoghi ma ad essa ignoti, manifestando così la capacità di generalizzazione.

Durante l'apprendimento, per ogni coppia [input / output desiderato] del *training set*, la rete riceve in input i dati **X** e restituisce un output effettivo **Y** generalmente diverso,

almeno nelle fasi iniziali, dall'output desiderato **D**. Ne consegue un errore dato dalla differenza [output effettivo - output desiderato]. L'errore quadratico medio su tutti gli esempi del *training set* è dato da:

$$E = 1/2 \sum_{i=1}^{n} (D_i - Y_i)^2$$

L'algoritmo di apprendimento deve allora modificare la rete in modo da tendere alla minimizzazione dell'errore quadratico medio, apportando progressivamente ai pesi sinaptici piccole e adeguate variazioni positive e negative.

Nell'ambito dell'apprendimento supervisionato, la principale architettura di rete è quella nota come **multilayer perceptron** (percettrone multistrato). Si tratta di una rete feed-forward (dove i segnali si propagano esclusivamente nel senso input-output), con più strati nascosti, uno di input ed uno di output, totalmente interconnessa.

Ogni neurone di uno strato riceve gli input da tutti i neuroni dello strato precedente, tramite connessioni di peso sinaptico W_{ij}, ed invia il suo output a tutti i neuroni dello strato successivo.

Le funzioni di trasferimento possono essere in teoria diverse da un neurone all'altro; in pratica, per semplicità, si adotta una funzione, solitamente di tipo sigmoide, comune a tutti i neuroni.

Occorre rilevare inoltre che, mentre per una data applicazione il numero dei neuroni di input e di quelli di output è perfettamente definito, non esiste alcun criterio rigoroso per definire il numero ottimale di strati intermedi o quello dei neuroni di questi strati. Tale scelta deve essere generalmente operata in base all'esperienza acquisita in applicazioni analoghe, anche se spesso si ritiene sufficiente aumentare solamente il numero dei neuroni mantenendo un unico strato intermedio.

Aumentando il numero dei neuroni e degli strati nascosti aumenta notevolmente il tempo di addestramento e la rete ha la tendenza ad "imparare troppo" dal *training set*, perdendo la capacità di generalizzazione. Per contro, uno scarso numero di strati e di

neuroni intermedi, non consentono alla rete un apprendimento adeguato dal *training set*. E' pertanto necessario, per definire il numero ottimale di neuroni e di strati intermedi, adottare l'espediente di modificare, durante l'addestramento, il numero di neuroni dello strato intermedio, in base al comportamento espresso dalla rete stessa nella capacità di apprendimento e di generalizzazione.

Le reti multilayer perceptron risultano di particolare interesse anche grazie alla scoperta di un semplice ma potente algoritmo di apprendimento, denominato **Error Back Propagation**, o semplicemente **BP**.

Tale algoritmo cerca di minimizzare l'errore quadratico medio, relativo al *training set* di esempi dati, scendendo lungo la superficie d'errore e seguendo la direzione di massima pendenza, in cerca di una "valle" sufficientemente profonda.

L'algoritmo Back Propagation è suddiviso in tre fasi:

1. Feed Forward: è la fase di propagazione diretta del segnale dall'ingresso all'uscita, in cui, noti gli ingressi ed i pesi, viene calcolato l'input della rete;

2. Error Back Propagation: è la fase di retro-propagazione dell'errore, dove l'errore calcolato dalla differenza tra l'uscita (calcolata al punto precedente) ed il target viene trasmesso a tutti i neuroni della rete;

3. Wheight Update: è la fase di modifica dei pesi, resa possibile dalle informazioni ottenute nelle prime due fasi.

Per ogni strato di neuroni si calcola la somma pesata degli input, il valore di attivazione di ogni singolo neurone e l'output. Dopo aver eseguito tale passaggio per tutti gli strati, si devono modificare i pesi in modo tale che l'output della rete, cioè l'output dell'ultimo strato di neuroni, si avvicini sempre di più a quello desiderato (processo di apprendimento).

L'algoritmo BP ha il merito della semplicità, ma anche alcuni difetti, tra i quali una convergenza generalmente molto lenta verso la soluzione desiderata, nonché il rischio di intrappolarsi in minimi relativi o in oscillazioni nell'intorno di un minimo.

Questi inconvenienti diventano particolarmente acuti nella soluzione di complessi problemi pratici, che generalmente coinvolgono reti di grandi dimensioni; il tempo di apprendimento cresce con le dimensioni della rete molto più che linearmente e anche la qualità dell'apprendimento ne soffre. Manca inoltre un criterio rigoroso per definire il valore più opportuno del coefficiente di *learning rate*, che determina la velocità di convergenza della rete: maggiore è il valore del *learning rate*, maggiore sarà la velocità di convergenza e maggiori i rischi di convergenza verso un minimo locale e non assoluto, a cui non corrisponde un output corretto.

Proprio per tali motivi sono state proposte numerose varianti dell'algoritmo BP, applicate spesso con successo; alcune utilizzano conoscenze maggiori della superficie d'errore (conoscenze di secondo ordine), mentre altre adottano un *learning rate* variabile, adattandolo continuamente alle caratteristiche locali della superficie d'errore[13].

L'**apprendimento non supervisionato** non fa riferimento ad alcuna casistica precostituita di esempi; la rete impara a rispondere in modo ordinato agli stimoli esterni che le vengono sottoposti, auto-organizzando la propria struttura in modo tale che stimoli simili attivino neuroni vicini e stimoli tra loro diversi attivino neuroni lontani. Il mondo degli input esterni viene così proiettato in una mappa topologica bidimensionale, nella quale input simili attivano zone vicine e input differenti attivano zone lontane: ne consegue una classificazione automatica dei vari tipi di input. Un input ad *n* dimensioni viene così ridotto alle due dimensioni di una mappa piana.

[13] Esistono numerosi algoritmi di apprendimento, rimandiamo il lettore interessato ad approfondire l'argomento attraverso la letteratura specifica sulle reti neurali. Si vedano ad esempio Bishop (1995), Ingrassia e Silani (2000) e Zani (2000).

L'apprendimento non supervisionato è inoltre di tipo competitivo, nel senso che per ogni input vengono attivati più neuroni, ma solo uno di essi (quello con attivazione maggiore) vince la competizione e viene premiato con una modifica dei suoi pesi sinaptici. Questo tipo di apprendimento è utilizzato, in particolare, dalle reti di Kohonen.

Le **reti di Kohonen** sono dei particolari tipi di reti neurali che permettono di classificare oggetti senza alcun tipo di supervisione e nascono dallo studio della topologia della corteccia del cervello umano. Tali tipi di rete, denominate anche **SOM** (**Self Organizing Maps**), hanno destato un notevole interesse sia perché, nonostante la semplicità architetturale ed operativa, sono applicabili in molti problemi pratici, sia perché sono dotate di due peculiari caratteristiche: l'analogia con certe strutture neurobiologiche e la capacità di auto-organizzazione.

Per quanto riguarda l'analogia neurobiologica, rileviamo che sulla corteccia cerebrale si vengono a formare, in base all'auto-apprendimento, mappe corticali tali che neuroni vicini sono attivati da stimoli simili (mappa acustica, visuale o retino-tipica, somato-sensoriale, …). Ad esempio, nella mappa acustica suoni con frequenze vicine stimolano aree neurali adiacenti, nella mappa somato-sensoriale si ha una vera e propria proiezione bidimensionale del corpo, deformandolo in funzione dell'importanza relativa dei vari organi per un dato animale. Analogamente, le reti SOM proiettano un input multidimensionale su una superficie di neuroni artificiali, rappresentandolo in due dimensioni ed effettuandone quindi una notevole compressione, pur conservando le similarità originarie. Inoltre la capacità di scoprire, in modo autonomo, proprietà interessanti di un input multidimensionale accomuna le reti SOM alla capacità degli animali di adattarsi all'ambiente senza la necessità di una guida esterna, e per questo, tali tipi di reti neurali sono utili in processi di data mining, ed in particolar modo in problemi di classificazione.

Queste reti tengono conto non solo delle connessioni sinottiche tra neuroni ma anche dell'influenza che può avere un neurone sul vicino. E' stato infatti osservato che, nel caso biologico, i neuroni che sono fisicamente vicini a neuroni attivi hanno i legami più

forti mentre quelli ad una particolare distanza hanno legami inibitori. A questa caratteristica Kohonen attribuisce uno sviluppo nella capacità di realizzare delle mappe topologiche localizzate nel cervello.

Una rete di Kohonen è costituita da una serie di neuroni di input che, come per le reti multistrato, servono a calcolare la somma pesata di tutti gli input e da un singolo strato bidimensionale di neuroni (organizzato quindi come una griglia posta su un piano) che calcolano l'output della rete. Ciascun neurone di input è connesso a tutti i neuroni dello strato successivo bidimensionale. L'apprendimento è legato alle interconnessioni laterali tra neuroni vicini. L'algoritmo di apprendimento di questo tipo di rete è il seguente:

1. Si definiscono con w_{ij} ($i = 1,...,n-1$ dove n è il numero di input) il peso tra il neurone i-esimo di input ed il neurone j-esimo della griglia al tempo t. I valori dei pesi vengono inizialmente posti tra zero e uno. Si pone come valore di $N_i(0)$ il maggiore possibile ($N_i(\cdot)$ rappresenta il numero di neuroni vicini al j-esimo neurone).

2. Si presenta un input: $x_o(t), x_1(t),...,x_n(t)$ dove $x_i(t)$ rappresenta l'i-esimo input.

3. Si calcolano le distanze euclidee tra l'input e ciascun neurone di output j:
$$_2 d_j^2 = \left[\sum_{i=0}^{n-1} \left(x_i(t) - w_{ij}(t) \right)^2 \right]^{\frac{1}{2}}$$

4. Si seleziona il neurone j^* a cui corrisponde la distanza minima. Si verifica quindi una competizione tra tutti i neuroni (*competitive learning*), vinta da uno solo di essi.

5. Si modificano i pesi dal neurone di input al neurone j^* e a tutti i suoi vicini definiti all'interno della superficie definita da $N_i^*(t)$. I nuovi pesi sono:
$$w_{ij}(t+1) = w_{ij}(t) + h(t)[x_i(t) - w_{ij}(t)]$$

dove $h(t)$ è il coefficiente di *learning rate*, compreso tra 0 e 1, che viene fatto diminuire nel tempo con legge lineare o di altro tipo, partendo da $h(0)=1$, in modo da rallentare di volta in volta l'adattamento dei pesi, per passare da una prima mappatura approssimativa a delle mappature più fini avvicinando i vettori dei pesi a quelli di input. E' importante sottolineare che il vicinato di ogni neurone, che all'inizio comprende tutta la matrice neurale, deve progressivamente diminuire anch'esso nel tempo sino ad annullarsi. Un vicinato relativamente grande favorisce la cooperazione tra neuroni e l'instaurarsi di aree di neuroni vicini che sono eccitati da input simili. Un vicinato relativamente piccolo favorisce invece l'indipendenza tra i neuroni. Pertanto nelle fasi iniziali viene conseguita la plasticità della rete, mentre nelle fasi finali la sua selettività;

6. Dopo l'aggiornamento del peso sinaptico, il ciclo ricomincia, a partire dal punto 2. Il processo di apprendimento dal punto 2) al punto 6), deve essere eseguito dalle 100 alle 1000 volte circa.

L'algoritmo di apprendimento di questa rete è molto più semplice di quello delle reti multilayer perceptron viste in precedenza: nel caso delle reti di Kohonen si confronta semplicemente un *pattern* di input[14] ed il vettore dei pesi. Il neurone con il vettore dei pesi più vicino al pattern di input viene selezionato ed il suo vettore dei pesi viene modificato in modo da allinearlo a quello degli input, ossia in modo da disunire la distanza ${}_2 d_j^2$.

Possiamo inoltre osservare che vengono modificati anche i vettori dei pesi dei neuroni vicini a quello selezionato. Questo perché la rete dovrà organizzarsi in regioni costituite da un ampio set di valori attorno all'input con cui apprende. Di conseguenza, i vettori che sono spazialmente vicini ai valori di training saranno comunque classificati correttamente anche se la rete non li ha mai visti. Questo dimostra le proprietà di generalizzazione della rete.

[14] Modello, tipo di input.

Regole associative e sequenze

Una delle classi di metodi computazionali "locali" più diffusa riguarda le cosiddette regole associative e sequenze (*association and sequence rules*).

Le associazioni e le sequenze sono tecniche esplorative spesso usate nella **market basket analysis** (analisi del carrello della spesa) e nella **clickstream analysis** (analisi delle sequenze di visita ai siti web) per misurare l'affinità di prodotti acquistati da un particolare consumatore oppure, nel nostro caso, delle pagine viste da un visitatore di un sito. L'obiettivo della market basket analysis è quello di evidenziare gruppi di prodotti legati da analoghe abitudini d'acquisto. Analizzando le combinazioni di acquisto ed il numero di volte che queste sono ripetute, si ottiene la regola associativa (del tipo: "if *condition* then *result*") che esprime la probabilità di acquisto simultaneo di prodotti differenti.

Se la regola è ordinata, si ha una sequenza. Evidentemente, ciascuna di queste regole descrive un particolare *pattern* locale, che seleziona un insieme ristretto di variabili.

I vantaggi delle regole associative e di sequenza sono la loro estrema semplicità e capacità comunicativa; tra gli svantaggi, invece, si annoverano gli elevati tempi e costi di elaborazione ma, soprattutto, la necessità di ridurne il numero, restringendole a quelle significative, con degli opportuni criteri di significatività. Questi possono essere sviluppati utilizzando quanto elaborato dalla teoria statistica in tema di modelli associativi.

Per descrivere le sequenze associative e gli indici utilizzati in questo tipo di analisi facciamo riferimento in particolare alla nostra analisi di web mining.

Nell'analisi dei dati sulle sequenze di visita ai siti web, si è solitamente interessati a determinare le sequenze di pagine più ricorrenti. Le pagine accessibili all'interno di un sito web hanno infatti la caratteristica di essere organizzate in pagine ipertestuali, cioè collegate tra di loro tramite dei *link* che permettono all'utente di passare con facilità da un documento all'altro. Diventa pertanto importante determinare i modelli di accesso

degli utenti e, in questo ambito, la comprensione delle sequenze di visita è fondamentale.

La considerazione dell'ordine di visita richiede di precisare la terminologia e, quindi, la notazione. Nell'ambito dell'analisi delle sequenze si usa comunemente l'espressione $A \rightarrow B$, che indica che la pagina A è stata visualizzata prima della pagina B. Sottolineiamo che, dopo aver richiesto la pagina A, potrebbero essere state visualizzate altre pagine prima di arrivare alla pagina B. In altri termini, la sequenza può condurre da A a B indirettamente, passando per altre pagine. La letteratura esistente in tema di web mining, si occupa prevalentemente di questo tipo di sequenze, che si definiscono indirette.

L'espressione descrivente una sequenza indiretta assume, in generale, la forma "Se A, allora B"; il termine di sinistra è detto "corpo della regola" o "condizione" mentre il termine di destra è chiamato "testa della regola" o "risultato".

L'impostazione appena presentata trae origine dal fatto che i metodi di analisi delle sequenze di visita sono stati recentemente mutuati dalle metodologie impiegate nell'ambito della market basket analysis.

In tal caso, infatti, la singola sessione utente corrisponde a una transazione di vendita ovvero al singolo carrello della spesa mentre le pagine corrispondono ai prodotti contenuti nel carrello. I due ambiti applicativi sono piuttosto differenti: in particolare, la rilevanza dell'ordine con il quale i prodotti vengono inseriti nel carrello della spesa è certamente minore rispetto a quanto accade per la clickstream analysis.

Introduciamo ora gli indici comunemente utilizzati nel web mining per lo studio delle sequenze di visita.

Si consideri la sequenza indiretta $A \rightarrow B$ e si indichi con $N_{A \rightarrow B}$ il numero di sessioni utente in cui tale sequenza compare, almeno una volta. Sia N il numero complessivo delle sessioni utente. Si osservi che la regola $A \rightarrow B$ viene conteggiata una sola volta anche se dovesse ripetersi più volte all'interno della stessa sessione.

Il **supporto** per la regola $A \rightarrow B$ si ottiene dividendo il numero di sessioni utente che soddisfano tale regola per il numero totale di sessioni utente:

$$support(A \rightarrow B) = \frac{N_{A \rightarrow B}}{N}$$

Quindi 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 si può affermare che il supporto per la regola $A \rightarrow B$ esprime la probabilità che una sessione utente contenga le due pagine, in sequenza:

$$support(A \rightarrow B) = P(A \cap B)$$

La **confidenza** per la regola $A \rightarrow B$ si ottiene invece dividendo il numero di sessioni utente che soddisfano la regola per il numero di sessioni utente che contengono la pagina A:

$$confidence(A \rightarrow B) = \frac{N_{A \rightarrow B}}{N_A} = \frac{\frac{N_{A \rightarrow B}}{N}}{\frac{N_A}{N}} = \frac{\text{support}(A \rightarrow B)}{\text{support}(A)}$$

Quindi, l'indice di confidence esprime la frequenza (e quindi, al limite, la probabilità) che in una sessione utente in cui è stata visualizzata la pagina A possa essere successivamente visualizzata la pagina B:

$$confidence(A \rightarrow B) = P(B \mid A)$$

Gli indici appena descritti si possono generalizzare ed utilizzare anche quando si vogliono individuare delle sequenze costituite da un numero di pagine maggiori di due. Si consideri la sequenza $\{A \rightarrow B \rightarrow C \rightarrow D\}$, in questo caso la regola va interpretata logicamente come "Se $A \rightarrow B \rightarrow C$, allora D"; si noti che si tratta di sequenze indirette, per cui dopo aver visualizzato la pagina A possono essere state richieste altre pagine prima di arrivare a B.

Come risultato della trattazione tradizionale delle analisi delle sequenze, si ottiene che le sequenze di pagine che attraggono di più i visitatori sono quelle a cui sono associati gli indici di *support* più alti. 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, per altri versi estremamente flessibili e informativi, è rappresentato dal fatto che, in quanto indici descrittivi, permettono di trarre conclusioni valide solo per il data set osservato. In altri termini, non permettono di ricavare delle previsioni di comportamento affidabili, per nuovi utenti.

2.6 Metodi statistici

In questo paragrafo si elencano i principali metodi statistici impiegati nel data mining. Soffermarsi sul particolare modello descrivendolo in modo esaustivo esula dagli scopi della presente trattazione, che vuol essere, in questa parte, solo descrittiva.

Tra i principali metodi statistici si annoverano:

- la regressione lineare semplice;
- la regressione lineare multipla;
- il modello lineare normale;
- il modello di regressione logistica;
- i modelli log-lineari;
- i modelli grafici.

Si rimanda il lettore interessato a testi specifici sull'argomento, si vedano ad esempio Zani (2000) o Giudici (2001).

3

Web Usage Mining

3.1 Introduzione

Il **Web Mining** costituisce l'area del data mining che si occupa dell'estrazione di conoscenza dal World Wide Web.

Possiamo suddividere il Web Mining in tre sottoaree:

1. **Web Content Mining**: si concentra sulle informazioni grezze disponibili nelle pagine web ed ha come scopo la classificazione e l'ordinamento delle pagine in base al contenuto. La fonte dei dati consiste principalmente nei dati testuali delle pagine web (es. parole, ma anche *tags*[15]).

2. **Web Structure Mining**: si focalizza sulla struttura del sito ed ha come scopi la classificazione delle pagine web in base ai collegamenti, l'ordinamento delle pagine web attraverso una combinazione di contenuto e struttura ed il *reverse engineering*[16] dei modelli del sito web. La fonte dei dati consiste principalmente nell'informazione sulla struttura delle pagine web (es. collegamenti alle altre pagine).

3. **Web Usage Mining**: si occupa dell'estrazione di conoscenza dai *log file* del web server. Le principali applicazioni sono basate sulle tecniche per modellare gli

[15] I *tags* sono delle istruzioni del linguaggio HTML per contrassegnare i contenuti della pagina web, servono, ad esempio, per formattare in grassetto una parola o per aggiungere un collegamento ipertestuale.

[16] Fare *reverse engineering* significa manipolare e analizzare un software a partire dal codice finale senza bisogno dei sorgenti. In questo caso si tratta di scoprire la struttura di un sito basandosi solamente sui collegamenti tra le pagine.

utenti, come la personalizzazione del web ed i siti web adattivi. La fonte dei dati consiste nei *log* (testuali) rappresentati in formati standard che vengono raccolti quando gli utenti accedono ai web server.

Il processo di Web Usage Mining può essere considerato come un processo a tre fasi. Nella prima fase, i dati del web log vengono sottoposti al preprocessing al fine di identificare utenti, sessioni, *pageview*[17] e così via. Nella seconda fase, vengono applicati metodi statistici e di data mining (come regole associative e sequenze, analisi di raggruppamento e analisi di classificazione) per scoprire dei modelli. Questi modelli vengono memorizzati cosicché possano venire analizzati nella terza fase del processo.

Gli anni recenti hanno visto prosperare la ricerca nell'area del Web Mining e particolarmente del Web Usage Mining. Dai primi articoli pubblicati nella metà degli anni 90, sono stati pubblicati finora più di 400 articoli sul Web Mining; più o meno 150 di questi sono stati pubblicati prima del 2001; circa il 50% di questi saggi riguarda il Web Usage Mining. Il primo seminario dedicato interamente a questo argomento, WebKDD, è stato tenuto nel 1999. Dal 2000 le pubblicazioni sul Web Usage Mining sono state più di 150 e mostrano un sensazionale aumento di interesse per quest'area.

3.2 Fonti dei dati

Le applicazioni di Web Usage Mining sono basate sulla raccolta dei dati da tre fonti principali: *web server*, *proxy server* e *web clients*.

Web server

Dal momento che possono raccogliere una grande quantità di informazioni nei loro log file, i web server rappresentano sicuramente la più ricca e la più comune fonte dei dati. Ogni accesso ad una pagina web viene registrato nel log degli accessi del web server che la ospita. La registrazione di un web log consiste in campi che seguono un formato predefinito. I campi del Common Log Format sono:

[17] Per la definizione di *pageview* si veda il paragrafo 3.3 relativo al livello di astrazione dei dati.

remotehost rfc931 authuser date "request" status bytes

dove:

- *remotehost* rappresenta il nome dell'host[18] remoto oppure l'indirizzo IP[19] se il DNS[20] del nome dell'host non è disponibile;
- *rfc931* è il nome remoto con cui accede l'utente;
- *authuser* è il nome di identificazione dell'utente (login), disponibile quando accede a pagine protette da password;
- *date* rappresenta la data e l'ora della richiesta;
- *"request"* è la riga di richiesta esattamente come proviene dal client (il file, il nome, ed il metodo usato per recuperarlo);
- *status* è il codice di risposta HTTP restituito al client, che indica se il file è stato recuperato con successo o meno ed eventualmente quale messaggio di errore è stato restituito;
- *bytes* rappresenta la dimensione del contenuto del documento trasferito.

Se qualcuno di questi campi non può essere risolto dal server, al posto del campo viene messo un segno meno (-).

[18] Per *host* si intende ogni computer in una rete che fornisce servizi agli altri computer. Nel nostro caso il servizio fornito dall'*host* (che può essere un Internet Service Provider) è la possibilità di navigare in Internet.

[19] Ogni macchina connessa alla rete Internet possiede un indirizzo IP univoco, rappresentato da un numero composto da quattro parti, separate da punti: es. 124.12.234.201.

[20] Il DNS (Domain Name System) è il sistema che traduce gli indirizzi IP in nomi di dominio: es. da 124.12.234.201 a *workshop.matisse.com*.

Oltre al Common Log Format esistono altri formati standard di rappresentazione delle informazioni registrate dai web log, per un esempio degli standard più comuni si osservi la tabella 3.1.

Common Log Format picasso.wiwi.hu-berlin.de - - [10/Dec/2003:23:06:31 +0200] "GET /index.html HTTP/1.0" 200 3540
Extended Log Format picasso.wiwi.hu-berlin.de - - [10/Dec/2003:23:06:31 +0200] "GET /index.html HTTP/1.0" 200 3540 "http://www.berlin.de/" "Mozilla/3.01 (Win95; I)"
Cookie Log Format picasso.wiwi.hu-berlin.de - - [10/Dec/2003:23:06:31 +0200] "GET /index.html HTTP/1.0" 200 3540 "http://www.berlin.de/" "Mozilla/3.01 (Win95; I)" "VisitorID=10001; SessionID=20001"

Tabella 3.1 – Esempi di formati standard

Un altro modo di rappresentare l'informazione è fornito dal linguaggio LogML, un estensione del linguaggio XML[21] che permette di salvare le varie sessioni utente in file strutturati. Il primo vantaggio di questo tipo di rappresentazione è dato dal risparmio dello spazio richiesto per la memorizzazione (circa il 50% in meno). Inoltre, il processo di Web Usage Mining risulta notevolmente semplificato e può essere implementato più efficientemente attraverso l'utilizzo del LogML, che può anche essere considerato come una prima fase di pre-trattamento dei dati. Per una spiegazione dettagliata del

[21] Acronimo di eXtensible Markup Language, è un linguaggio ampiamente utilizzato per definire la formattazione dei dati.

linguaggio LogML e dell'utilità che può avere nel processo di Web Usage Mining si veda Punin *et al.* (2002).

Proxy server

Molti fornitori di servizi Internet (ISPs, Internet Service Providers) offrono ai loro clienti servizi di Proxy Server per migliorare la navigazione attraverso il *caching*. Il *caching* è il processo di memorizzazione locale delle pagine richieste con più frequenza dagli utenti. Questo servizio si propone di rendere la navigazione più veloce, permettendo all'utente di non dover richiedere ogni volta la pagina al web server che ospita la pagina stessa: sarà direttamente il proxy a fornire all'utente la pagina. Il proxy verifica periodicamente che la pagina memorizzata risulti aggiornata.

In molti casi, i dati di navigazione raccolti a livello di proxy sono fondamentalmente gli stessi di quelli raccolti a livello di server. La differenza principale in questo caso risiede nel fatto che il proxy server raccoglie dati di *gruppi di utenti* che accedono ad *enormi gruppi* di web server.

Web clients

I dati di utilizzo possono essere rintracciati anche a livello di *client* usando *JavaScript*[22], *Java applet*[23], o anche modificando i browser. Queste tecniche evitano il problema dell'identificazione delle sessioni utente e i problemi causati dal caching (come l'utilizzo del pulsante *back* del browser). Inoltre, forniscono dettagliate informazioni riguardo al comportamento effettivo degli utenti. Tuttavia, questo approccio fa assegnamento eccessivamente sulla cooperazione degli utenti e fa emergere molti problemi riguardanti le severe leggi sulla privacy. Si confronti il paragrafo 3.7 relativo al problema della privacy.

[22] JavaScript è un linguaggio di programmazione che viene implementato dai browser degli utenti.

[23] Le Java applet sono piccole applicazioni scritte in linguaggio Java che vengono eseguite all'interno di pagine web.

3.3 Livelli di astrazione dei dati

Le informazione fornite dalle fonti descritte sopra possono essere usate per costruire o identificare alcuni livelli di astrazione dei dati, specificamente *users*, *server sessions*, *episodes*, *clickstream*, e *pageviews*. Allo scopo di fornire una certa concordanza sulla definizione di questi termini, la W3C[24] *Web Characterization Activity* (WCA) ha pubblicato una prima stesura delle definizioni dei termini rilevanti per analizzare il comportamento di navigazione.

User: singolo individuo che sta avendo accesso a file da uno o più web server attraverso un browser. Mentre questa definizione sembra banale, in pratica è molto difficile individuare singolarmente e ripetutamente gli utenti. Un utente può accedere al Web attraverso diversi computer, o usare più di un tipo di browser sullo stesso computer.

Pageview: consiste in ogni file che contribuisce alla visualizzazione della pagina sul browser dell'utente in un dato momento. La visualizzazione di pagine è di solito associata con una singola azione dell'utente (come un *click* del mouse) e può riguardare diversi file come *frames*[25], grafici e documenti. Quando si discute e si analizza il comportamento degli utenti ciò che risulta importante è la visualizzazione aggregata della pagina. L'utente non chiede esplicitamente che *n frames* ed *m* grafici vengano caricati nel suo browser, l'utente richiede una pagina web. Tutte le informazioni per determinare quali file costituiscono una *pageview* sono disponibili nel web server.

Clickstream: serie sequenziale di richieste di *pageview*. Tuttavia, i dati disponibili dal lato server non forniscono sempre abbastanza informazioni per ricostruire l'intera

[24] Il W3C (World Wide Web Consortium) è un consorzio si occupa della creazione degli standard Web. Il suo scopo è portare il Web al suo massimo potenziale, mediante lo sviluppo di tecnologie (specifiche, linee guida, software e tools) che possano creare un forum per informazioni, commercio, ispirazioni, pensiero indipendente e comprensione collettiva. Per saperne di più si visiti il sito **http://www.w3.org/**

[25] Per facilitare la navigazione da parte dell'utente, una pagina web può risultare divisa in diverse cornici. Un esempio concreto si ha quando l'indice del sito viene visualizzato sempre all'interno di una cornice, mentre in un'altra cornice vengono visualizzate le varie pagine richieste. Attraverso il termine *frame* si indica il file contenuto in una cornice della pagina web.

sequenza dei click per un sito. Qualche visualizzazione di pagina a cui si accede attraverso la cache a livello di proxy non sarà "visibile" a livello di server.

User session: *clickstream* per un singolo utente attraverso l'intero Web. Tipicamente, solo la parte di ogni sessione utente che ha accesso ad uno specifico sito può essere utilizzata per l'analisi, poiché le informazioni di accesso non sono disponibili pubblicamente dalla maggior parte dei web server.

Server session: insieme delle *pageview* in una sessione utente per un particolare sito (comunemente chiamata *visit*).

Un insieme di *server session* è il necessario input per ogni analisi del comportamento di navigazione. La fine di una server session è definita come il punto dove la sessione di navigazione dell'utente per quel sito è terminata. Tuttavia, questo è un semplice concetto che è molto difficoltoso rintracciare precisamente. Ogni sottosezione semanticamente significativa di una *user session* o di una *server session* è definita come un *episode* dalla W3C WCA.

3.4 Preprocessing

Il pre-trattamento dei dati gioca un ruolo fondamentale nelle applicazioni di Web Usage Mining.

Il preprocessing dei web log risulta di solito complesso e richiede molto tempo. Comprende tre diversi compiti: il data cleaning, l'identificazione e la ricostruzione delle sessioni utente, il recupero di informazioni riguardanti il contenuto e la struttura delle pagine.

Data cleaning

Questa fase consiste nel rimuovere tutti i dati rintracciati nel web log che sono inutili per gli scopi del data mining, ad esempio le richieste di contenuti grafici delle pagine (es. immagini *jpg* o *gif*); le richieste per qualsiasi altro file che potrebbe essere incluso

all'interno di una pagina web; o anche sessioni di navigazione eseguite da *robot* e *web spider*[26]. Mentre le richieste di contenuti grafici e di file risultano facili da eliminare, i percorsi di navigazione dei *robot* e *web spider* devono essere identificati esplicitamente. L'identificazione viene realizzata solitamente riferendosi al nome dell'host remoto, riferendosi all'*agent*[27] dell'utente, o controllando l'accesso al file '*robots.txt'*. Tuttavia, alcuni robots spediscono un falso *agent* dell'utente nella richiesta HTTP. In questi casi, si può utilizzare un'euristica basata sul comportamento di navigazione per separare le sessioni dei robots dalle sessioni degli utenti effettivi.

Identificazione e ricostruzione delle sessioni

Questa fase consiste nell'identificazione delle diverse sessioni utente dalle informazioni solitamente molto scarse disponibili dei log file e nella ricostruzione del percorso di navigazione degli utenti all'interno delle sessioni identificate. La maggior parte dei problemi incontrati in questa fase sono causati dal *caching* effettuato sia dai *proxy server* che dai *browser*. Il *caching* del *proxy* causa che un singolo indirizzo IP (quello appartenente al *proxy server*) sia associato a diverse sessioni utente, cosicché diventa impossibile usare l'indirizzo IP come identificativo dell'utente. Questo problema può essere parzialmente risolto con l'uso dei *cookie* o richiedendo all'utente di identificarsi quando entra nel sito. Il *caching* dei *web browser* è un problema più complesso. I log del *web server* non possono includere nessuna informazione riguardo all'uso del pulsante *back*. Questo può generare percorsi di navigazione inconsistenti nelle sessioni utente. Tuttavia, usando le informazioni aggiuntive riguardanti la struttura del sito è ancora possibile ricostruire un percorso consistente per mezzo di euristiche. Poiché il protocollo HTTP è senza *stato*, diventa virtualmente impossibile determinare quando un utente lascia effettivamente il sito per determinare quando una sessione dovrebbe venire considerata conclusa. A questo problema ci si riferisce come *sessionization*. In pratica

[26] I *robot* ed i *web spider* sono programmi utilizzati dai gestori dei motori di ricerca, servono per indicizzare le pagine web e verificarne periodicamente l'aggiornamento.

[27] L'*agent* è un identificativo dell'utente, solitamente contiene il nome del browser, la versione, il sistema operativo, etc. Se si tratta di un motore di ricerca contiene il nome del motore e la versione del *robot*.

viene stabilito un tempo limite di permanenza su una pagina, oltre il quale la sessione viene considerata conclusa.

Recupero del contenuto e della struttura

La maggior parte delle applicazioni di Web Usage Mining usa gli URL visitati come la principale fonte di informazioni per lo scopo del data mining. Tuttavia gli URL sono una povera fonte di informazioni poiché, per esempio, non trasmettono nessuna informazione riguardo al contenuto della pagina. Per ovviare a questo problema diventa fondamentale impiegare le informazioni riguardo al contenuto per arricchire i dati del web log. Se non si conosce anticipatamente un adeguata classificazione, si possono impiegare le tecniche di Web Structure Mining per svilupparne una. Come nei motori di ricerca, le pagine web vengono classificate secondo la loro area semantica per mezzo delle tecniche di Web Content Mining; questa classificazione dell'informazione può essere usata per arricchire le informazioni estratte dai log.

3.5 Tecniche

La maggior parte delle applicazioni commerciali di Web Usage Mining sfruttano tecniche di analisi statistica consolidate. In contrasto, la ricerca in quest'area è concentrata principalmente sullo sviluppo di tecniche di estrazione di conoscenza progettate specificamente per l'analisi di dati sull'utilizzo del web. La maggior parte dello sforzo di questa ricerca si focalizza su due paradigmi principali:

- regole associative e sequenze;
- clustering.

Regole associative e sequenze

Sono probabilmente la tecnica di data mining più elementare e, allo stesso tempo, la tecnica più utilizzata nel Web Usage Mining. Quando si applicano al Web Usage Mining, le regole associative vengono usate per trovare associazioni tra le pagine che

appaiono frequentemente insieme in una sessione utente. Le regole sequenziali vengono sfruttate per trovare modelli di navigazione *sequenziali* che appaiono frequentemente nelle sessioni utente. I modelli sequenziali tipici hanno la forma: il 70% degli utenti che *prima* visita *A.html* e *quindi* visita *B.html* successivamente, nella stessa sessione, ha anche avuto accesso alla pagina *C.html*. Per una descrizione più esaustiva di questa tecnica si rimanda il lettore al paragrafo 2.5.

Clustering

Tecniche che cercano gruppi di elementi simili tra grandi quantità di dati basate sull'idea generale di *funzione di distanza* che calcola la similarità tra i gruppi.

Alla descrizione di queste tecniche è stato dato ampio spazio nel paragrafo 2.5, al quale si rimanda il lettore interessato.

Si può aggiungere che il clustering è stato largamente utilizzato nel Web Usage Mining per raggruppare insieme sessioni simili. I primi studi che hanno suggerito di spostare il focus del Web Usage Mining dalle singole sessioni utente ai gruppi di sessioni utente sono stati realizzati da Xie *et al.* (2001).

3.6 Applicazioni

L'obiettivo generale del Web Usage Mining è quello di dedurre informazioni interessanti riguardo ai modelli di navigazione degli utenti. Questa informazione viene sfruttata per migliorare il sito dal punto di vista degli utenti. I risultati prodotti dal data mining sui web log vengono usati per vari scopi:

1. per personalizzare i contenuti web;
2. per migliorare la navigazione degli utenti attraverso il *prefetching* ed il *caching*;
3. per migliorare il web design;
4. per migliorare la soddisfazione del cliente in siti di e-commerce.

Trattiamo qui di seguito i primi tre scopi. Il commercio elettronico verrà trattato ampiamente nel paragrafo 3.9.

Personalizzazione del Contenuto Web

Le tecniche di Web Usage Mining si possono impiegare per fornire esperienze personalizzate all'utente. Per esempio, è possibile anticipare, in tempo reale, il comportamento dell'utente confrontando il modello di navigazione in corso con modelli tipici estratti dai passati web log. In quest'area le applicazioni più comuni sono i *sistemi di raccomandazione*; questi sistemi si propongono di suggerire collegamenti interessanti a prodotti che potrebbero risultare attraenti ai visitatori. Schafer *et al.* (2001) presentano uno studio sui sistemi di raccomandazione commerciali esistenti, attuati in siti web di commercio elettronico.

Prefetching e Caching

Si possono sfruttare i risultati prodotti dal Web Usage Mining per migliorare la performance dei web server e delle applicazioni basate sul web. Tipicamente, il Web Usage Mining può essere utilizzato per sviluppare opportune strategie di *prefetching* e *caching* in modo da ridurre il tempo di risposta del server.

Il *prefetching* è una caratteristica del browser che permette ad una pagina HTML di recuperare altri contenuti web quando la connessione del browser dell'utente è inattiva. Il contenuto del *prefetching* viene immagazzinato nella *cache* del browser ed appare quindi velocemente non appena l'utente accede alla pagina che contiene il contenuto immagazzinato.

Sostegno al Design

L'usabilità è uno dei maggiori punti di discussione nella progettazione ed implementazione dei siti web.

Il risultato prodotto dalle tecniche di Web Usage Mining può fornire linee guida per migliorare il design delle applicazioni web. Una recente tecnica che utilizza

stratogrammi[28] per valutare l'organizzazione e l'efficienza dei siti web dal punto di vista degli utenti è stata proposta da Berendt (2002). Un ulteriore passo avanti è rappresentato dai siti web adattivi, in questo caso, il contenuto e la struttura del sito viene riorganizzato dinamicamente in base ai dati estratti dal comportamento degli utenti.

3.7 Privacy

Il problema più importante che deve essere affrontato durante il processo di profiling dell'utente è la violazione della privacy. Molti utenti sono riluttanti a svelare informazioni personali sia implicitamente che esplicitamente, risultando esitanti a visitare siti web che usano i *cookie* (se loro sono consapevoli della loro esistenza) o evitando rivelazioni di dati personali nei moduli di registrazione.

In entrambi i casi, l'utente perde l'anonimato ed è consapevole che tutte le sue azioni verranno registrate e usate, in molti casi senza il suo consenso. In più, anche se un utente ha convenuto di fornire informazioni personali ad un sito, attraverso la tecnologia *cookie* tale informazione può venire scambiata tra siti diversi, avendo come risultato la sua divulgazione senza il permesso dell'utente.

P3P (Platform for Privacy Preferences) è una raccomandazione presentata del W3C che suggerisce un'infrastruttura per lo scambio di dati sulla privacy. Questo standard consente ai siti web di esprimere le loro prassi sulla privacy in un formato standard che può venire automaticamente recuperato ed interpretato dal browser dell'utente.

In questo modo, il processo di lettura delle politiche sulla privacy risulterebbe semplificato per gli utenti, perché le informazioni chiave riguardo a come i dati vengono raccolti da un sito possono essere trasmesse automaticamente ad un utente, e le discrepanze tra le pratiche del sito e le preferenze dell'utente riguardo la rivelazione di dati personali verrebbero segnalati automaticamente. P3P, tuttavia, non fornisce un meccanismo per assicurare che i siti agiscano effettivamente secondo le loro politiche.

[28] Gli stratogrammi sono una particolare modalità di visualizzazione dei modelli di navigazione degli utenti.

3.8 Software

Esistono molti strumenti commerciali che effettuano l'analisi sui dati dei log raccolti dai web server. La maggior parte di questi strumenti sono basati su tecniche di analisi statistica, mentre solo pochi prodotti sfruttano le tecniche di data mining. Una rassegna aggiornata degli strumenti commerciali disponibili per il Web Usage Mining è contenuta in Eirinaki *et al.* (2003). Nella maggior parte dei casi, gli strumenti di Web Usage Mining sono parte delle soluzioni integrate di CRM[29] per il commercio elettronico. Qualche volta, questi strumenti sono semplici analizzatori di web log. Un software sviluppato in un ambiente di ricerca, WUM, sembra aver raggiunto un interessante livello di maturità; WUM ha raggiunto attualmente la versione 7.0.

Web Utilization Miner (WUM)

Il software WUM è un sistema integrato per il Web Usage Mining sviluppato nell'ambito di un progetto di ricerca dell'Università Humboldt di Berlino (Faculty of Economics, Institute of Information Systems).

Lo scopo primario del software è analizzare il comportamento di navigazione degli utenti che visitano un sito web. WUM è uno strumento integrato che permette di effettuare operazioni di:

- pre-trattamento dei dati;
- interrogazione (querying);
- visualizzazione dei risultati.

Nella figura seguente viene schematizzato il processo di scoperta dei modelli di navigazione. Questo processo si realizza attraverso l'interazione tra l'esperto ed il software di data mining.

[29] Il CRM (Customer Relationship Management) è un processo che coinvolge tutta la struttura aziendale e che ha come *focus* la conoscenza del cliente e del mercato, finalizzata ad una più sicura crescita della redditività aziendale.

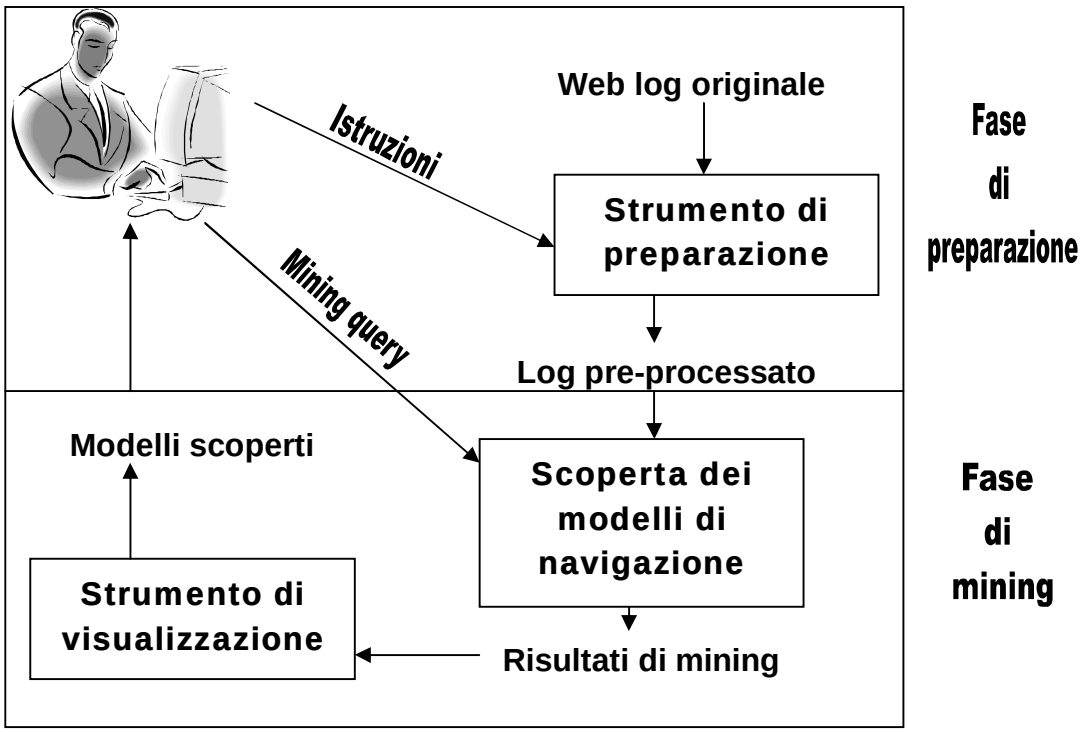

Figura 3.1 – Il processo di scoperta dei modelli di navigazione in un sito

Dopo la prima fase di pre-trattamento dei dati, un software di data mining dovrebbe testare se il sito è stato usato in conformità con gli obiettivi di progettazione. Ciò nonostante, questo semplice compito è più complesso di quello che la maggior parte dei software di data mining possono attualmente comprendere.

Andiamo a descrivere i requisiti desiderabili che un buon software di web mining dovrebbe possedere.

Requisito 1. Il programma dovrebbe capire la descrizione astratta del modello, così che il progettista possa istruire il software su quello che dovrebbe essere scoperto e quello che dovrebbe essere ignorato.

Immaginiamo un sito fittizio di commercio elettronico, dove la sequenza seguente risulta la più frequente:

`Welcome.html à orders.html`

Frequenza: 15%.

La nostra banale sequenza di esempio contiene utili suggerimenti. Ci dice che il percorso più frequente che conduce a "orders.html" parte dalla pagina "Welcome.html". Tuttavia, questa informazione non è sufficiente per valutare e migliorare il sito, lo sviluppatore del sito preferirebbe ottenere una risposta ad una domanda come: la maggior parte dei visitatori di "orders.html" partono dalla pagina "Welcome.html" oppure ci sono diversi percorsi arbitrari per raggiungerla? Può essere che essi interrompano la loro navigazione in corrispondenza di una determinata pagina, e in quale pagina? Per rispondere a queste domande, lo sviluppatore non ha bisogno solo di questa sequenza, ma anche dei percorsi più frequenti tra le estremità della sequenza.

Requisito 2. Un modello di navigazione dovrebbe essere qualcosa in più rispetto ad una sequenza di pagine a cui si accede frequentemente. Per riflettere il comportamento degli utenti, dovrebbe inoltre contenere statistiche sui percorsi che connettono le pagine frequentemente visitate insieme.

Conseguentemente, il data mining per la valutazione di un sito richiede una sofisticata interazione tra il progettista del sito ed il software di data mining.

Secondo il requisito 2, l'esperto necessita di informazioni su:

- percorsi frequenti;
- percorsi adiacenti le pagine che compongono i percorsi frequenti;
- statistiche sull'utilizzo di questi percorsi, così che possano essere distinti quelli preferiti da quelli usati raramente.

I *web log miner* dedicati costituiscono il nucleo del *MiDAS mining environment* e del *Web Utilization Miner (WUM)*. Entrambi i sistemi sono stati progettati in conformità con la crescente domanda per interazioni intensive con utenti umani, come descritto nella figura 3.1. Questa interazione è basata su di un potente linguaggio di data mining nel quale gli utenti esperti possono esprimere la loro conoscenza di fondo, guidare il

software e raffinare gradualmente o rifocalizzare il processo di scoperta, secondo i risultati ottenuti dopo ogni query.

Il linguaggio di mining di MiDAS e WUM usa "modelli" per descrivere le caratteristiche desiderabili dei percorsi di navigazione da scoprire. Tali caratteristiche possono includere una minima e/o una massima lunghezza o le pagine web che dovrebbero o non dovrebbero apparire nel percorso. Pertanto, in entrambi i software il requisito 1 è soddisfatto.

La più importante differenza tra MiDAS e WUM concerne la nozione di percorso di navigazione. In MiDAS, un percorso di navigazione è una sequenza di eventi che soddisfano i vincoli posti dall'esperto. Ciò implica che nei risultati del processo di mining appariranno solo le sequenze composte da eventi che si presentano insieme frequentemente. Tuttavia, secondo il requisito 2, il progettista ha anche bisogno di informazioni sui percorsi adiacenti agli eventi frequenti.

Rispecchiando questo bisogno, il concetto di percorso di navigazione è stato esteso nel WUM per includere sia le sequenze di eventi che soddisfano i vincoli dell'esperto, sia i percorsi che connettono questi eventi. Il risultato non è più una sequenza ma un albero composto da questi percorsi ed ogni pagina in ciascun percorso è annotata con il numero di visitatori che hanno raggiunto la pagina in questione attraverso quel determinato percorso. Il progettista può distinguere tra percorsi popolari e percorsi scelti raramente semplicemente esaminando il numero sulla rappresentazione grafica dei risultati della query e può inoltre identificare le pagine dove gli utenti terminano la loro navigazione (ogni volta che un percorso popolare cambia in un percorso seguito raramente). Perciò, WUM soddisfa entrambi i requisiti del Web usage mining per la valutazione dei siti.

3.9 Commercio Elettronico

Per molte società, la competitività nell'e-commerce richiede una presenza di successo sul web. I siti web vengono usati per determinare l'immagine dell'azienda, per

promuovere e vendere prodotti e per fornire assistenza ai clienti. Il successo di un sito influisce e riflette direttamente il successo dell'azienda nel mercato elettronico.

Basandoci sullo studio di Spiliopoulou *et al*. (2001) proponiamo una metodologia per migliorare il "successo" dei siti web, basata sullo sfruttamento della scoperta dei percorsi di navigazione.

Possiamo studiare la soddisfazione dei visitatori ad un sito selezionando un gruppo rappresentativo degli utenti e studiando sia il loro comportamento che intervistandoli direttamente, sulla base di specifici criteri che riflettono la nozione di soddisfazione. Questo approccio ha degli indubbi svantaggi. Primo, le spese generali per costituire un ambiente sperimentale sono troppo alte: la "regolare amministrazione" del successo non si può realizzare in questo modo. Secondo, il risultato di tale sforzo dovrebbe essere quello di massimizzare il successo piuttosto che calcolare soltanto un valore per una qualche misura dello stesso. Infine, la selezione di un gruppo di utenti rappresentativo non è scontata; il web è un mercato globale e non può essere valutato banalmente da un gruppo di utenti vicini e disponibili.

Da ciò, per migliorare il successo di un sito abbiamo bisogno di un diverso approccio con le seguenti proprietà: (i) deve tener conto di tutti i visitatori del sito; (ii) deve essere appropriato per essere eseguito frequentemente, meglio se abitualmente; (iii) deve condurre a concreti indicatori dei difetti del sito ed ai rimedi per attenuarli. Per questo fine, proponiamo di migliorare il successo analizzando le sequenze di visita.

Il data mining è per sua natura appropriato per questo tipo di analisi. Le attività di tutti gli utenti vengono registrate nel web server log ed il paradigma del data mining fornisce la metodologia per analizzarle. Tuttavia, il data mining non è di per se adeguato per migliorare il successo. Abbiamo bisogno di un modello del comportamento di navigazione appropriato, cosicché le sequenze scoperte possano fornire i necessari indicatori per migliorare il sito. Risulta necessaria anche una misura del successo del sito, che possa essere calcolata nel processo di data mining. Servono inoltre un software di data mining per realizzare la scoperta di sequenze di navigazione e una metodologia

per applicare la nostra misura, al fine di ottenere gli indicatori di miglioramento necessari.

L'interesse nel monitoraggio dell'utilizzazione di un sito è probabilmente tanto vecchio quanto lo stesso web. I primi strumenti assistevano gli amministratori dei siti web studiando e bilanciando il carico dei web server. I moderni strumenti per il monitoraggio degli accessi supportano la computazione di statistiche che possono servire come base per l'analisi del successo.

I progressi strettamente connessi con la nostra impostazione provengono da due campi: misure del successo per siti web commerciali e tecniche di data mining per analizzare l'uso del web.

Misurare il successo di un sito

La necessità di misurare il successo di un sito con rispetto agli scopi oggettivi dei suoi proprietari è riflesso in Berthon *et al.* (1996). Berthon *et al.* propongono due misure per il successo di un sito, l'efficienza di contatto (*contact efficiency*) e l'efficienza di conversione (*conversion efficiency*). La prima misura fornisce la percentuale di utenti che passano almeno una determinata quantità di tempo minimo esplorando il sito. La seconda misura fornisce la percentuale di utenti che, dopo aver esplorato il sito, hanno anche comprato qualcosa. Di conseguenza, il successo di un sito è definito come la sua efficienza nella "conversione" di visitatori in clienti e può essere misurato senza il coinvolgimento degli utenti.

La nozione di "successo" per i siti web

Le misure della qualità di un sito dovrebbero essere progettate rispettando gli obiettivi di business dei suoi proprietari. Per modellare il successo in questo contesto, è necessario intraprendere tre passaggi:

1. modellare i contenuti del sito accordandoli ai concetti che riflettono i suoi scopi oggettivi;

2. classificare gli utenti avendo riguardo alle loro attività mentre perseguono quegli obiettivi;

3. definire il "successo di un sito" come l'efficienza delle sue *parti* nel guidare gli utenti verso la realizzazione degli obiettivi del sito.

Un sito web può servire per molteplici scopi. Un sito commerciale offre tipicamente un meccanismo di ricerca fra il catalogo dei suoi prodotti ed un servizio di ordinazione per comprare i prodotti scelti. Inoltre, un distributore di software può anche proporre una *chat* o un *forum*, dove i clienti possono scambiarsi esperienze e assistersi vicendevolmente. Un motore di ricerca potrebbe adornare i risultati della ricerca con icone pubblicitarie, in questo modo adempiendo a due scopi: assistenza nella ricerca di documenti e marketing dei prodotti.

Quando si misura il successo di un sito, per prima cosa l'analista deve specificare il contesto nel quale questa analisi avrà luogo, es. l'obiettivo del sito attraverso il quale dovrebbe essere misurato il successo. Chiaramente, un sito per la commercializzazione di software risulta di successo verso lo scopo dell'acquisto online se le persone comprano software, pur non usando mai il forum interattivo.

La fase di specificazione del problema, che precede ogni ulteriore attività nel ciclo vitale della scoperta di conoscenza, riguarda la determinazione degli obiettivi verso dei quali viene realizzata l'analisi dei fattori di successo. Assumiamo che l'analisi riguardi un solo obiettivo e caratterizziamo questo obiettivo come "l'obiettivo del sito".

Scopi oggettivi del sito e pagine che li riflettono

Per rendere esplicito l'obiettivo del sito all'analisi del comportamento degli utenti, descriviamo le pagine del sito in termini della loro funzione per perseguire questo scopo.

Definizione 1. Una "pagina *action*" è una pagina la cui richiesta indica che l'utente sta perseguendo lo scopo del sito. Una "pagina *target*" è una pagina la cui richiesta indica che l'utente ha realizzato lo scopo del sito.

In un sito di commercio elettronico, la compilazione di un modulo di richiesta per il catalogo dei prodotti rappresenta una pagina *action*, mentre la compilazione del modulo di ordinazione dei prodotti risulta una pagina *target*. Per un software di archiviazione dei documenti, una pagina *action* potrebbe essere una richiesta del servizio di ricerca degli argomenti. La selezione e l'ispezione di un singolo documento dalla lista dei risultati potrebbe caratterizzarsi come una pagina *target*.

Assumiamo che una pagina *target* non possa essere raggiunta senza accedere prima ad una pagina *action*. Il presupposto è ragionevole poiché l'ottenimento di un documento o di un prodotto presuppone un meccanismo per acquisirlo.

Nella nostra definizione di pagine *action* e *target*, osserviamo un sito dal punto di vista del servizio che offre al fine di raggiungere i suoi scopi.

Successo come efficienza di contatto e di conversione

In un ambiente di marketing orientato al web, Berthon *et al.* (1996) anticipano che il successo di un sito è misurato dalla percentuale dei suoi visitatori che rimangono occupati nel visitarlo (*contact efficiency*) e nella percentuale di visitatori che alla fine diventano clienti (*conversion efficiency*). Conseguentemente classificano i visitatori sulla base delle attività che hanno realizzato, in particolare alle attività conformi allo scopo del sito – in questo caso, l'acquisto di prodotti. Sulla base di questa classificazione, definiamo un visitatore "short-time" come un utente che raggiunge il sito ma lo abbandona presto senza esplorarlo, mentre un "active investigator" è un utente che rimane più a lungo ed esplora il sito. Un sottoinsieme degli "active investigator" diventa "cliente", es. ordinano prodotti o realizzano qualche attività simile.

Passeremo ora dal concetto di *utente* al concetto di *sessione*, poiché un utente può varare sessioni multiple perseguendo diversi obiettivi. Quindi generalizzeremo i suddetti tipi di utenti in tipi di sessioni che riflettono il comportamento dei visitatori nella direzione degli scopi arbitrari del sito. Infine, introdurremo le relative misure di efficienza, non per il sito intero, ma per le sue parti.

Classifichiamo tutti gli utenti che accedono al sito come "visitatori". Un sequenza di attività realizzate dal visitatore ed osservate dall'analista come una singola unità di lavoro è chiamata "sessione".

Definizione 2. Una "sessione attiva" è una sessione che contiene almeno un'attività in direzione del raggiungimento dello scopo del sito. Tutte le altre sessioni vengono definite "inattive".

Riferendosi alla definizione di "pagine *action*", le sessioni attive sono quelle che contengono almeno un accesso ad una pagina *action*.

La nostra definizione ha due vantaggi rispetto alla distinzione tra "active investigator" e "short-time visitors" di Berthon *et al*. Per prima cosa, possiamo determinare singolarmente se una sessione utente è attiva o inattiva, senza riferirsi a criteri come tempo di permanenza o numero di pagine richieste. Tali criteri sono solo modestamente attendibili, perché potrebbero condurre ad equivocare la classificazione di un cliente esperto come un "short-time visitor" oppure considerare un utente disorientato come un "active investigator". In secondo luogo, la distinzione tra sessioni attive ed inattive è fatta sulla base degli obiettivi del sito. L'analisi del sito in direzione di altri scopi implica solamente di specificare le corrispondente pagine *action* e *target*: le sessioni attive e inattive vengono quindi ridefinite automaticamente.

***Definizione 3*.** Una "sessione cliente" è una sessione nella quale l'utente ha realizzato lo scopo del sito.

Secondo la definizione 2 ed il nostro presupposto che una pagina *target* è raggiungibile solamente da una pagina *action*, una sessione cliente è sempre una sessione attiva. Chiamiamo tutte le sessioni attive che non sono sessioni cliente come "sessioni non-cliente".

Similmente alle sessioni attive ed inattive, dalle pagine che contiene, una sessione può essere caratterizzata singolarmente come una sessione cliente.

Nel seguito, classifichiamo gli utenti che danno vita a queste sessioni rispettivamente come "clienti" e "non-clienti". Utilizziamo questo sistema per semplicità di formulazione, nonostante il fatto che la stessa persona fisica potrebbe in linea teorica comportarsi una volta come un "cliente" ed una volta come un "non-cliente".

Usando come base i concetti di pagine action e sessioni attive, definiamo *l'**efficienza di contatto** di una pagina action* come il rapporto tra le sessioni che contengono questa pagina e tutte le sessioni del log. Bisogna notare che il log non è un semplice insieme ma è un insieme multiplo, poiché è possibile che vari utenti abbiano realizzato le stesse sequenze di attività, corrispondenti ad identiche sessioni.

Definizione 4. Supponiamo che *Sessions* denoti tutte le sessioni registrate nel log e che *A* sia una pagina *action* del sito web. Dunque, l'**efficienza di contatto** di *A* è data da:

$$contacteff(A) = \frac{card(\{\{s \in Sessions \mid A \in s\}\})}{card(Sessions)} \qquad (3.1)$$

dove $card(\cdot)$ sta per cardinalità e $\{\{...\}\}$ denota un insieme multiplo.

Poiché *A* è una pagina *action*, le sessioni incluse nel numeratore di $contacteff(A)$ sono ovviamente sessioni attive. Di conseguenza, l'efficienza di contatto di *A* è la percentuale di sessioni dove, usando la pagina *action A*, si è cercato di raggiungere lo scopo del sito. Calcolando questo valore per ogni pagina *action*, possiamo (i) identificare l'impatto di ogni pagina sul successo complessivo di un sito nell'attrarre visitatori e (ii) scoprire pagine con un'efficienza di contatto poco elevata.

Definizione 5. L'**efficienza di contatto relativa** ad una pagina *action A* è il rapporto tra le sessioni contenti questa pagina e l'insieme multiplo delle sessioni attive, chiamato $aSessions$:

$$Rcontacteff(A) = \frac{card(\{\{s \in aSessions \mid A \in s\}\})}{card(aSessions)} \qquad (3.2)$$

In questa definizione, il numeratore è lo stesso di quello dell'equazione (3.1), poiché una sessione che contiene *A* è proprio una sessione attiva. Questa misura esprime l'importanza relativa di ogni pagina *action* all'interno di un sito ed è appropriata per siti con molte pagine *action* o con un numero elevato di sessioni inattive.

Analogamente all'efficienza di contatto di una pagina action, definiamo l'efficienza di conversione di una pagina arbitraria per una pagina target. In questa definizione, abbiamo bisogno di considerare anche i percorsi utilizzati per raggiungere la pagina target. Per esempio, è importante sapere se la pagina target è stata raggiunta in 3 o in 13 passi. Se il sito è stato progettato come una netta gerarchia di pagine, dove ogni oggetto importante dovrebbe essere raggiunto entro un piccolo numero di passaggi, risultano indesiderabili i percorsi lunghi che conducono ad una pagina *target*. D'altra parte, se lo scopo oggettivo dell'analisi è l'esposizione dell'utente alla pubblicità, i percorsi più lunghi potrebbero essere più desiderabili di quelli brevi.

Definizione 6. Definiamo l'**efficienza di conversione** di una pagina *P* per una pagina target *T* su un gruppo di percorsi *G* da *P* a *T,* come il rapporto tra *G* e tutte le sessioni contenti *P*:

$$conveff(P,T,G) = \frac{card(G)}{card(\{\{s \in aSessions \mid P \in s\}\})} \qquad (3.3)$$

Dove un percorso è una parte di una sessione, composto da accessi consecutivi.

I percorsi *G* sono parti delle sessioni attive, poiché contengono *T*, la pagina target. Poiché essi contengono anche *P*, il numeratore sarà al massimo uguale al denominatore, di conseguenza il valore dell'indice sarà compreso tra 0 e 1.

Questa misura stima il successo di una pagina arbitraria nell'aiutare/guidare gli utenti verso la pagina *target*. La nostra generica definizione permette la stima di valori diversi di efficienza di conversione: es. su lunghi e su brevi percorsi o su tutti i percorsi, in questo ultimo caso troviamo un valore che denotiamo come $conveff(P,T,*)$. Con questa misura, possiamo studiare l'impatto di ogni pagina nel successo del sito e

identificare le pagine che hanno una bassa efficienza di conversione e richiedono miglioramenti. Tuttavia, per fare questo, dobbiamo identificare i gruppi di percorsi sui quali dovrebbe essere calcolata l'efficienza di conversione: questi gruppi sono i percorsi di navigazione che riflettono il comportamento degli utenti e vengono scoperti tramite un software di web mining.

Esempio 1. Vogliamo calcolare l'efficienza di conversione di una pagina *P* per una pagina target *T* su tutti i percorsi. La pagina *P* appare in 100 sessioni attive. In 20 sessioni tra queste, dopo la pagina *P* è stata visitata la pagina *A*, quindi è stata richiesta *T*. In altre 30 sessioni, la pagina *B* è stata visitata dopo *P*; in solo 10 di queste 30 sessioni è stata visitata successivamente la pagina *T*. Nelle rimanenti 40 sessioni, la pagina visitata dopo *P* è stata *C*; questi utenti non hanno mai raggiunto *T*.

Dai percorsi *PAT*, *PBT*, *PC*, solo *PAT* e *PBT* riguardano sia la pagina *P* che la pagina target *T*. Di conseguenza, il valore del numeratore è il numero di volte in cui *PAT* e *PBT* sono stati attraversati completamente. *PAT* è stato attraversato in 20 sessioni, *PBT* in 10 soltanto. Il valore del denominatore è 100, il numero di sessioni attive che contengono *P*. Quindi, $conveff(P,T,*) = \frac{20+10}{100} = 0.3$

Questo valore è abbastanza basso, almeno nel contesto di qualche applicazione. Per identificare se sono necessari dei miglioramenti alla pagina *P* o ad un'altra pagina visitata dopo *P*, dobbiamo ispezionare tutti i percorsi che provengono da *P* e conducono a *T* e identificare le pagine in cui gli utenti escono dal sito o seguono altri percorsi.

Il processo di Knowledge Discovery per l'analisi del successo

Il processo di KD viene tipicamente modellato come una serie di fasi, chiamate (i) specificazione del problema, (ii) raccolta e preparazione dei dati rilevanti, (iii) analisi dei dati con le tecniche di data mining, (iv) valutazione dei risultati in base alle misure stabilite precedentemente, (v) interpretazione dei risultati e (vi) comportamento secondo le decisioni strategiche. Completa il processo una fase ulteriore, la verifica dell'impatto delle azioni intraprese.

Per il problema particolare del miglioramento del successo di un sito web, modelliamo le fasi suddette nel modo seguente:

1. **Specificazione formale del problema.** L'obiettivo del miglioramento del successo del sito viene modellato attraverso i concetti di efficienza di contatto per le pagine *action* e di efficienza di conversione per le pagine *target*.

2. **Preparazione dei dati.** Per il concreto sito web, devono essere determinate le nozioni di "active investigator" e di "cliente". Inoltre, devono essere selezionate le pagine che effettuano il servizio di "pagine *action*" o "pagine *target*" rispettivamente. Le singole pagine possono essere riassunte in concetti più generici stabilendo gerarchie di concetti basati sul servizio che formano i servizi forniti dal sito. In questo caso, le pagine richieste nel web server log dovrebbero essere sostituite da richieste di concetti astratti. Infine, il web server log deve essere ripulito e gli accessi devono essere raggruppati per formare sessioni dei visitatori. A queste sessioni viene applicato il data mining.

3. **Data mining.** Il Web Usage Mining per l'analisi del successo si traduce nella scoperta di percorsi di navigazione che riflettono l'efficienza di contatto e l'efficienza di conversione delle pagine a cui si accede frequentemente. La scoperta di percorsi di navigazione viene realizzata sulla parte del web server log che contiene le sessioni "cliente". I percorsi scoperti riflettono il comportamento *desiderato* dei visitatori. Questi percorsi vengono quindi usati come base per analizzare le sessioni della rimanente parte del log, comprese le sessioni degli "active investigator" che non diventano clienti.

4. **Valutazione ed interpretazione dei risultati.** Le misure per la valutazione dei percorsi scoperti sono l'efficienza di contatto e di conversione introdotte in questa sezione. L'interpretazione dei risultati è basata sullo studio dei contenuti dei percorsi e delle loro statistiche e sulla conoscenza di fondo del proprietario del sito.

5. **Azioni da intraprendere in base ai risultati del data mining.** Nel contesto di miglioramento del successo, il data mining dovrebbe condurre a suggerimenti concreti per la ri-progettazione del sito o di una sua parte. Una volta che sono stati messi in pratica questi suggerimenti, si dovrebbe verificare l'impatto dei cambiamenti analizzando il web server log ottenuto dopo la ri-progettazione.

In questo paragrafo, abbiamo presentato un modello per migliorare il successo di un sito web attraverso le tecniche di data mining. Questo modello è stato progettato ponendo l'attenzione sui requisiti di funzionalità: (i) deve essere capace di tener conto di tutti gli utenti del sito, (ii) deve risultare appropriato per eseguire continui test al sito ed inoltre, ancora più importante, (iii) deve misurare il successo del sito e fornire indicazioni su come può essere massimizzato.

4
Applicazione

4.1 Introduzione

Nel presente capitolo andremo ad analizzare gli accessi al sito Internet della Biblioteca di Ateneo (**http://www.biblio.unimib.it**) attraverso le tecniche di Web Usage Mining descritte a livello teorico nel capitolo precedente.

Il sito considerato non è un sito di e-commerce, lo scopo principale del sito non riguarda la vendita di prodotti. Il sito fornisce un servizio agli utenti della biblioteca che, navigando al suo interno, possono accedere alle diverse aree: catalogo dei libri posseduti, risorse elettroniche remote, organizzazione della biblioteca, informazioni generali, statistiche, link a siti di interesse, etc.

Andiamo a descrivere sommariamente come si è sviluppato il lavoro. La prima fase del processo di data mining, che riguarda il pre-trattamento dei dati, è stata realizzata principalmente attraverso l'utilizzo del software WUMprep. Ci si è avvalsi inoltre di un Database Management System (DBMS) per l'esecuzione di query di eliminazione in linguaggio SQL.

Per analizzare il comportamento di visita degli utenti attraverso la scoperta dei modelli di navigazione è stato impiegato il software dedicato WUM (Web Utilization Miner), realizzato nell'ambito di un progetto di ricerca dell'Università di Berlino e descritto nel paragrafo 3.8.

Il lavoro è proseguito quindi con la classificazione degli utenti attraverso le tecniche di clustering, realizzata con l'ausilio del software statistico SPSS. La matrice dei dati, input essenziale dell'analisi cluster, è stata creata tramite l'esportazione dei dati dal software WUM e la rielaborazione degli stessi con una applicazione Java creata ad hoc.

Per tutelare la privacy dei visitatori del sito, gli indirizzi IP visualizzati in questo capitolo non sono quelli realmente presenti nei web log, al loro posto si è preferito inserire indirizzi IP casuali. Ciò nonostante, questo non andrà a condizionare la descrizione dell'analisi.

4.2 Preprocessing

Per comprendere la struttura del sito osserviamo la figura seguente. Nella cornice di sinistra (sommario) vengono sempre visualizzati i vari collegamenti alle pagine del sito, mentre nella cornice di destra si alternano le varie pagine richieste.

Figura 4.1 – Struttura del sito

I log file della biblioteca si presentano nell'*Extendend Log Format*. Nella figura seguente possiamo osservarne un estratto. Per questa analisi sono stati utilizzati i dati relativi alle pagine visitate nell'arco di tre settimane.

```
#Date: 2004-07-20 10:54:20
#Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status cs(User-Agent)
2003-01-20 10:54:20 151.116.55.114 - - 80 GET /elenco/U.HTM - 200 Mozilla/4.75+[en]+(WinNT;+U)
2003-01-20 10:54:20 184.20.124.251 - - 80 GET /intra/cen020215.pdf - 206 Mozilla/4.0+(+MSIE+5.5;+Windows+NT+5.0)
2003-01-20 10:54:23 80.25.51.43 - - 80 GET /erdisci.htm - 200 Mozilla/4.0+(+MSIE+5.5;+Windows+NT+4.0)
2003-01-20 10:54:32 87.35.24.201 - - 80 GET /elenco/P.HTM - 304 Mozilla/4.0+(+MSIE+5.01;+Windows+98)
2003-01-20 10:54:59 141.20.152.31 - - 80 GET /index.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 141.20.152.31 - - 80 GET /preload.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 141.20.152.31 - - 80 GET /blank.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 141.20.152.31 - - 80 GET /ultimora.htm - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 141.20.152.31 - - 80 GET /nuovopic.jpg - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 187.134.16.151 - - 80 GET /index.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 187.134.16.151 - - 80 GET /preload.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:54:59 187.134.16.151 - - 80 GET /blank.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:55:01 187.134.16.151 - - 80 GET /ultimora.htm - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.0)
2003-01-20 10:55:03 184.20.124.251 - - 80 GET /intra/cen020215.pdf - 304 Mozilla/4.0+(+MSIE+5.5;+Windows+NT+5.0)
2003-01-20 10:55:14 130.132.15.201 - - 80 GET /index.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.1)
2003-01-20 10:55:14 130.132.15.201 - - 80 GET /preload.html - 304 Mozilla/4.0+(+MSIE+6.0;+Windows+NT+5.1)
```

Figura 4.2 – Esempio del web log di origine

Data cleaning

Come si osserva dalla figura, i dati si presentano in formato grezzo. Affinché il processo di data mining produca dei buoni risultati è necessario che questi dati vengano *preparati* attraverso varie operazioni di pulizia, andando ad identificare ed eliminando i *record* che risultano inutili o perfino controproducenti per la nostra analisi.

Per prima cosa dobbiamo eliminare le sessioni dei *robot*, vale a dire dei programmi usati dai motori di ricerca che servono per indicizzare le pagine web e verificarne periodicamente l'aggiornamento. A tal fine, i web log vengono importati in un database la cui struttura è riprodotta dalle seguenti relazioni:

1. LOG(id, ip, agent, date, time, method, path, status)
2. ROBOTS(robotagent)

dove:

- **id** è un numero progressivo che rappresenta la chiave primaria[30] della relazione;
- **ip** rappresenta l'indirizzo IP dell'utente;
- **agent** identifica l'agente: il tipo di browser, il sistema operativo, etc. se si tratta di un utente, oppure il nome del *robot*, la versione, etc. se si tratta di un motore di ricerca;
- **date** è la data della richiesta;
- **time** è l'ora della richiesta;
- **method** rappresenta il metodo utilizzato per richiedere l'oggetto. Il tipo più comune è il metodo GET. Quando alla richiesta di una pagina si passano dei parametri (come nel caso in cui si compilano dei campi in un modulo) si può utilizzare il metodo POST. Esistono anche altri metodi che vengono utilizzati solitamente non da utenti ma da software come *robot* o *spider*, ad esempio il metodo HEAD;
- **path** rappresenta l'Uniform Resource Identifier (URI) del documento a cui si accede. In sostanza è la linea di richiesta così come proviene dal client, come può essere il nome della pagina (es. index.html);
- **status** è il codice di risposta dello stato HTTP restituito al client, che indica se il file è stato recuperato con successo o, in caso contrario, quale messaggio di errore è stato restituito;
- **robotagent** contiene il nome o una parte del nome della maggior parte dei robot conosciuti. In particolare il nome è preceduto e susseguito dal carattere '*', al fini di semplificare le successive query di eliminazione.

[30] La chiave primaria è un campo che identifica in modo univoco l'osservazione contenuta in un record della tabella.

Ora si tratta di eliminare i record del database che contengono richieste fatte da *robot*, per fare questo si eseguono le seguenti query di eliminazione in linguaggio SQL:

1. DELETE * FROM log WHERE agent LIKE '*crawler*' OR agent LIKE '*robot*' OR Agent LIKE '*spider*';

2. DELETE * FROM log WHERE ip IN (SELECT ip FROM log WHERE path LIKE '*robots.txt');

3. DELETE * FROM log WHERE id IN (SELECT log.id FROM log, robots WHERE agent LIKE robots.robotagent);

Andiamo a descrivere brevemente cosa vanno ad eliminare le query.

La prima query elimina dalla tabella le richieste fatte da robot "generici", è sufficiente che l'*agent* contenga la parola *crawler*, *robot* o *spider*.

Di norma, come prima richiesta, i robot dei motori di ricerca tentano di accedere al file 'robots.txt' presente solitamente sul web server, tale file specifica le politiche di indicizzazione del proprietario del sito. La seconda query elimina dalla tabella le richieste di coloro che accedono a questo file.

Nella tabella ROBOT sono stati inseriti i nomi degli *agent* della maggior parte dei motori di ricerca conosciuti. La terza query elimina tali richieste.

In alcuni casi il server non riesce a riconoscere o a risolvere il nome del documento che l'utente ha richiesto, in questo caso al posto del nome del file viene memorizzato il carattere meno (-).

Per eliminare queste richieste si esegue la seguente query:

4. DELETE * FROM log WHERE path="-";

Nel nostro caso le richieste eliminate finora sono pari al 2% della lunghezza iniziale del log.

A questo punto la tabella LOG viene importata in un foglio elettronico, al fine di eseguire le operazioni di ricodifica, in questo caso per uniformare i formati di data e ora.

```
143.172.143.58;30-Jun-2004;07:39:46;GET;/journal/subject.asp;200
143.172.143.58;30-Jun-2004;07:43:49;GET;/regolamento.htm;200
143.172.143.58;30-Jun-2004;07:45:30;GET;/repertorio/repertorio.htm;200
143.172.143.58;30-Jun-2004;07:45:48;GET;/repertorio/dirlavor.htm;200
143.172.143.58;30-Jun-2004;07:49:23;GET;/yourlink.htm;200
29.252.126.12;30-Jun-2004;07:56:05;GET;/stat01d.pdf;200
29.252.126.12;30-Jun-2004;07:58:09;GET;/normeej.htm;200
61.175.193.51;30-Jun-2004;07:58:14;GET;/journal/source.asp;200
29.252.126.12;30-Jun-2004;07:59:58;GET;/attiv00.htm;200
61.175.193.51;30-Jun-2004;08:05:19;GET;/cercagoogle.htm;200
61.175.193.51;30-Jun-2004;08:06:25;GET;/chisiamo.htm;200
61.175.193.51;30-Jun-2004;08:07:11;GET;/dir.htm;200
```

Si eliminano inoltre i campi non più necessari (*id* ed *agent*) e si esporta il foglio elettronico in un file di testo con i campi separati dal carattere punto e virgola, il risultato ottenuto finora si può osservare nella figura seguente:

Figura 4.3 – Esportazione dopo la prima fase di data cleaning

Il file così ottenuto è pronto per essere ulteriormente processato per mezzo del software di preprocessing WUMprep.

E' necessario configurare il software affinché riconosca i campi contenuti nel file. Per fare questo si modifica il file di configurazione 'logfileTemplate' inserendo la seguente stringa:

```
@host_ip@;@ts_day@-@ts_month@-
@ts_year@;@ts_hour@:@ts_minutes@:@ts_seconds@;@method@;@path@;@status@
```

Si esegue lo script 'logFilter'. Questo programma elimina tutte le richieste di immagini (*gif*, *jpg*, etc.) e le richieste successive allo stesso file da parte dello stesso utente fatte durante un brevissimo periodo di tempo (probabilmente causate da utenti impazienti).

Si profila ora un nuovo problema, che riguarda l'eliminazione delle richieste fatte dai *proxy server*. Come discusso nella parte teorica, i *proxy* non rappresentano un solo

utente, bensì un insieme di utenti che accedono alle pagine tramite l'indirizzo IP del *proxy*. A meno che non si utilizzino i *cookie* (e non è il nostro caso), è impresa ardua se non impossibile risalire ai singoli utenti. Si è perciò scelto di eliminare tali richieste per evitare che possano fuorviare i risultati delle analisi.

Si è scelto inoltre di eliminare anche le richieste fatte internamente alla rete 'biblio.unimib.it', poiché da una parte, non garantiscono di identificare i singoli utenti, dall'altra parte costituiscono un gruppo di utenti che conoscono "troppo bene" il sito e ciò non risulta utile per la nostra analisi, il nostro interesse deve essere focalizzato sull'utente generico (medio).

L'eliminazione di tali richieste è stata affrontata attraverso due fasi:

1. l'utilizzo dello script di WUMprep 'dnsLookup' per risolvere i nomi degli host dati gli indirizzi IP;

2. la successiva importazione in un database e l'esecuzione di una query di eliminazione.

La prima fase è molto semplice, si tratta di eseguire lo script 'dnsLookup' e attendere che il software, attraverso la funzione *domain name lookup* (ricerca nome di dominio), converta i vari indirizzi IP in nomi di dominio. Per eseguire questa operazione è necessario essere connessi alla rete Internet.

Una parte del file risultante è mostrato nella figura seguente:

```
veloxzone.com.br;30-Jun-2004;07:39:46;GET;/journal/subject.asp;200
veloxzone.com.br;30-Jun-2004;07:43:49;GET;/regolamento.htm;200
veloxzone.com.br;30-Jun-2004;07:45:30;GET;/repertorio/repertorio.htm;200
veloxzone.com.br;30-Jun-2004;07:45:48;GET;/repertorio/dirlavor.htm;200
veloxzone.com.br;30-Jun-2004;07:49:23;GET;/yourlink.htm;200
151.libero.it;30-Jun-2004;07:56:05;GET;/stat01d.pdf;200
151.libero.it;30-Jun-2004;07:58:09;GET;/normeej.htm;200
adsl-ull-21.net24.it;30-Jun-2004;07:58:14;GET;/journal/source.asp;200
151.libero.it;30-Jun-2004;07:59:58;GET;/attiv00.htm;200
adsl-ull-21.net24.it;30-Jun-2004;08:05:19;GET;/cercagoogle.htm;200
adsl-ull-21.net24.it;30-Jun-2004;08:06:25;GET;/chisiamo.htm;200
adsl-ull-21.net24.it;30-Jun-2004;08:07:11;GET;/dir.htm;200
```

Figura 4.4 – Risultato dopo la ricerca dei nomi di dominio

La seconda fase richiede l'importazione del file risultante in un database, la cui struttura può essere riassunta dalla seguente relazione:

DNS(hostname, date, time, method, path, status)

Si tratta ora di eseguire la query di eliminazione:

- DELETE * FROM dns WHERE dns.hostname LIKE '*biblio.unimib*' OR dns.hostname LIKE '*fw*' OR dns.hostname LIKE '*proxy*'

La query si propone di eliminare tutte le richieste provenienti dai *proxy server*. Probabilmente la query elimina anche qualche utente ma ciò non influisce eccessivamente sui risultati, se a questo punto si hanno a disposizione pochi dati si tratta solamente di aumentare il periodo di osservazione.

Dopo un'analisi esplorativa dei dati, si è scelto di eliminare anche le richieste fatte da host che superavano di molto la media di richieste giornaliere per utente, anche in questo caso si suppone che l'host sia un *proxy server*. In particolare sono state eliminati gli utenti che superavano il 95° percentile della distribuzione.

Dopo questa prima fase di data cleaning, la tabella risultante viene esportata in un file di testo per l'arricchimento semantico attraverso il software WUMprep.

Arricchimento semantico del web log

Nella maggior parte dei casi i nomi dei file non si rivelano rappresentativi del contenuto delle pagine. Per arricchire il web log con i concetti legati alle pagine è necessario che le pagine del sito vengano mappate all'interno di concetti che ne rappresentano il contenuto.

Attraverso lo script 'mapReTaxonomies' del software WUMprep i nomi dei file vengono sostituiti dai concetti che ne descrivono il contenuto[31].

Una volta eseguito il programma, il web log si presenta così:

```
veloxzone.com.br;30-Jun-2004;07:39:46;GET;risorse.elettroniche.ricerca.area;200
veloxzone.com.br;30-Jun-2004;07:43:49;GET;informazioni.regolamento;200
veloxzone.com.br;30-Jun-2004;07:45:30;GET;catalogo.repertorio.periodici.indice;200
veloxzone.com.br;30-Jun-2004;07:45:48;GET;catalogo.repertorio.periodici.diritto.lavoro;200
veloxzone.com.br;30-Jun-2004;07:49:23;GET;internet.link.utenti;200
151.libero.it;30-Jun-2004;07:56:05;GET;statistiche.legenda.2001;200
151.libero.it;30-Jun-2004;07:58:09;GET;risorse.elettroniche.norme.utilizzo;200
adsl-ull-21.net24.it;30-Jun-2004;07:58:14;GET;risorse.elettroniche.ricerca.editore;200
151.libero.it;30-Jun-2004;07:59:58;GET;documenti.ufficiali.attivita.2000;200
adsl-ull-21.net24.it;30-Jun-2004;08:05:19;GET;internet.cercagoogle;200
adsl-ull-21.net24.it;30-Jun-2004;08:06:25;GET;informazioni.chisiamo;200
adsl-ull-21.net24.it;30-Jun-2004;08:07:11;GET;organizzazione.direzione;200
```

Figura 4.5 – Risultato dopo la sostituzione dei concetti alle pagine

Per comprendere facilmente il passaggio effettuato si confronti la figura 4.5 con la figura 4.4.

Possiamo classificare le pagine del sito nelle seguenti 10 macroaree: *home*, *catalogo periodici*, *catalogo repertorio periodici*, *catalogo stone*, *risorse elettroniche*, *organizzazione*, *informazioni*, *internet*, *statistiche* e *documenti*.

[31] Nell'allegato 1 vengono elencate in dettaglio le 572 pagine del sito e i relativi concetti che le descrivono. Questo file rappresenta la base di partenza per l'impiego del programma 'mapReTanonomies'.

Ricordiamo che il catalogo OPAC[32] è escluso dall'analisi, in quanto i log file relativi vengono raccolti separatamente. Oltretutto, il catalogo può essere considerato come un sito a se stante, non suscettibile di modifiche.

Poiché alcune pagine andranno eliminate durante l'importazione successiva nel software WUM, si è scelto di far seguire al nome del concetto il suffisso "to.delete.before.analysis" in modo che possano essere riconosciute ed eliminate facilmente.

Il log viene formattato nel Common Log Format attraverso lo script 'transformLog' di WUMprep. Il risultato è presentato nella figura seguente.

```
veloxzone.com.br - - [30/Jun/2004:07:39:46 +0001] "GET risorse.elettroniche.ricerca.area HTTP/1.1" 200 1000
veloxzone.com.br - - [30/Jun/2004:07:43:49 +0001] "GET informazioni.regolamento HTTP/1.1" 200 1000
veloxzone.com.br - - [30/Jun/2004:07:45:30 +0001] "GET catalogo.repertorio.periodici.indice HTTP/1.1" 200 1000
veloxzone.com.br - - [30/Jun/2004:07:45:48 +0001] "GET catalogo.repertorio.periodici.diritto.lavoro HTTP/1.1" 200 1000
veloxzone.com.br - - [30/Jun/2004:07:49:23 +0001] "GET internet.link.utenti HTTP/1.1" 200 1000
151.libero.it - - [30/Jun/2004:07:56:05 +0001] "GET statistiche.legenda.2001 HTTP/1.1" 200 1000
151.libero.it - - [30/Jun/2004:07:58:09 +0001] "GET risorse.elettroniche.norme.utilizzo HTTP/1.1" 200 1000
adsl-ull-21.net24.it - - [30/Jun/2004:07:58:14 +0001] "GET risorse.elettroniche.ricerca.editore HTTP/1.1" 200 1000
151.libero.it - - [30/Jun/2004:07:59:58 +0001] "GET documenti.ufficiali.attivita.2000 HTTP/1.1" 200 1000
adsl-ull-21.net24.it - - [30/Jun/2004:08:05:19 +0001] "GET internet.cercagoogle HTTP/1.1" 200 1000
adsl-ull-21.net24.it - - [30/Jun/2004:08:06:25 +0001] "GET informazioni.chisiamo HTTP/1.1" 200 1000
adsl-ull-21.net24.it - - [30/Jun/2004:08:07:11 +0001] "GET organizzazione.direzione HTTP/1.1" 200 1000
```

Figura 4.6 – Risultato dopo la trasformazione del file nel Common Log Format

Identificazione e ricostruzione delle sessioni

A questo punto il log è pronto per essere importato nel software dedicato di Web Usage Mining **WUM**.

Dopo aver importato il log si procede con la creazione delle sessioni utente. Il software permette di impostare i parametri per identificare il termine delle sessioni. In base ad esperienze analoghe, riportate nella letteratura dedicata all'argomento, si è scelto di stabilire che la durata massima della visualizzazione di una pagina non risulti superiore ai 30 minuti.

[32] On-line Public Access Catalogue. Catalogo elettronico per la ricerca di monografie e periodici posseduti dalla biblioteca.

Delle circa 36000 richieste del log sono state create circa 6000 sessioni utente. Come si può facilmente desumere una sessione contiene in media 6 pagine visitate.

Dopo aver creato il database aggregato, il software permette la visualizzazione delle sessioni aggregate attraverso una struttura ad albero, un esempio è riportato nella figura seguente.

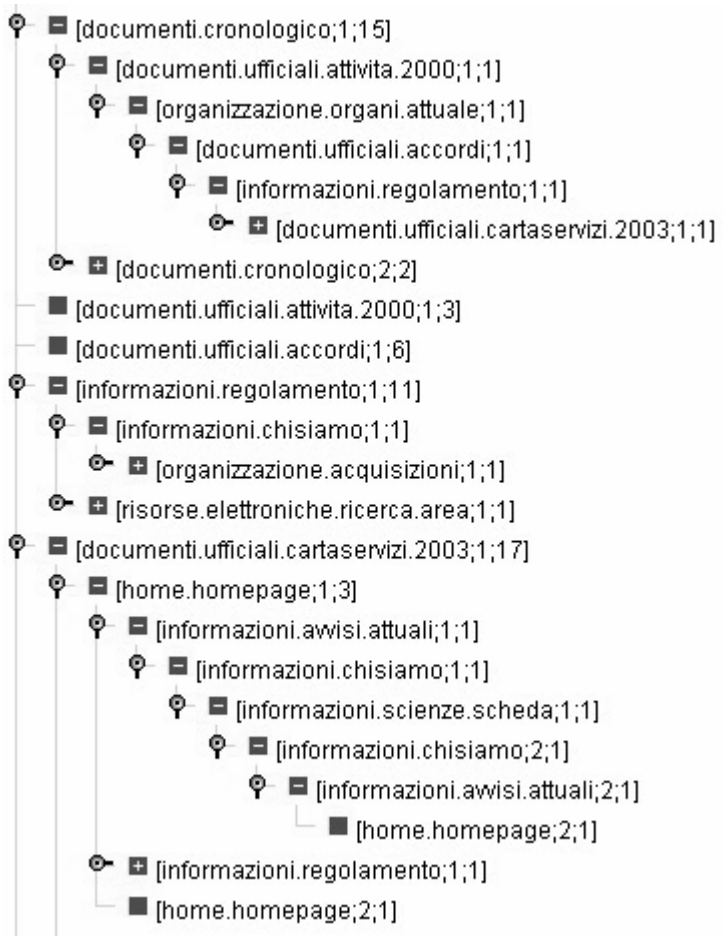

Figura 4.7 – Visualizzazione ad albero delle diverse sessioni aggregate

Il numeri accanto al nome delle pagine indicano rispettivamente l'*occurrence*, ossia il numero di volte che la pagina compare in quel determinato percorso dell'utente, e l'indice di *support*, vale a dire il numero di sessioni utente nelle quali compare la sequenza tra la pagina precedente (radice) e quella in questione.

4.3 Pattern discovery e pattern analysis

Il processo di scoperta dei modelli di navigazione è stato realizzato attraverso le due tecniche più comuni di Web Usage Mining, vale a dire le **regole associative e sequenze** ed il **clustering**. Per la descrizione teorica di queste tecniche si rimanda il lettore a confrontare il paragrafo 2.5.

Regole associative e sequenze

Il software WUM, con l'ausilio del linguaggio di web mining **MINT**, permette di calcolare gli indici di *support* e di *confidence* per le varie pagine del sito.

A tal fine sono state eseguite delle query in linguaggio MINT del tipo:

```
select t from node as a b, template a * b as t
where a.url startswith "documenti"
and b.url !startswith "documenti"
and a.occurrence = 1
and b.occurrence = 1
and ( b.support / a.support ) >= 0.1
and a.support > 5
```

Andiamo a descrivere cosa mette in pratica la query precedente. Vengono selezionate tutte le sequenze *t* tra le pagine *a* e *b*. In particolare le pagine *a* e *b* devono riguardare due aree diverse del sito ed in questo caso si vanno a scoprire le sequenze tra le pagine relative all'area "documenti" e quelle relative alle altre aree. Questo è espresso dai primi due vincoli.

Il terzo ed il quarto vincolo impongono che le pagine compaiano una sola volta nel percorso di navigazione. In questo modo consideriamo come significative sia le sessioni che le sottosessioni (*subsessions*) dei diversi percorsi di navigazione. In pratica vogliamo dare rilevanza non solo alle intere sessioni, ma anche alle diverse parti di una singola sessione.

Il quinto vincolo richiede che l'indice di **confidence** tra le due pagine sia pari almeno al 10% affinché la sequenza possa considerarsi significativa. Nell'ambito del software

WUM tale indice coincide con l'**efficienza di conversione** descritta nel paragrafo 3.9 del capitolo precedente.

L'ultimo vincolo riguarda l'indice di **support** e specifica che la pagina *a* deve essere stata visitata in almeno 6 sessioni, al fine di evitare che si prendano in considerazione percorsi con un alto livello di confidence poiché sono molto rari.

Nella tabella seguente vengono riportati i risultati dell'analisi, ordinati per livello di confidence. Tra parentesi si riportano i nomi dei file.

Ricordiamo che l'indice di confidence tra le pagine *a* e *b* si ottiene dividendo il numero di sessioni utente che soddisfano la regola $a \rightarrow b$ per il numero di sessioni utente che contengono la pagina *a*:

$$confidence(a \rightarrow b) = \frac{N_{a \rightarrow b}}{N_a} = \frac{\frac{N_{a \rightarrow b}}{N}}{\frac{N_a}{N}} = \frac{\text{support}(a \rightarrow b)}{\text{support}(a)}$$

Quindi, l'indice di confidence esprime la frequenza (e quindi, al limite, la probabilità) che in una sessione utente in cui è stata visualizzata la pagina *a* possa essere successivamente visualizzata la pagina *b*:

$$confidence(a \rightarrow b) = P(b \mid a)$$

Pagina *a*	**Pagina *b***	**Confidence (%)**
organizzazione.servizi.informatici (inf.htm)	documenti.lavoro.internet.sito (intra/intranet.htm)	30,2
organizzazione.servizi.informatici (inf.htm)	documenti.cronologico (crono.htm)	30,0
informazioni.medicina.servizi (sermed.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	27,2
organizzazione.servizi.pubblico (pre.htm)	documenti.ufficiali.cartaservizi.2003 (cartaservizi.htm)	25,9

Pagina *a*	**Pagina *b***	**Confidence (%)**
informazioni.medicina.servizi.old (med.htm)	catalogo.repertorio.periodici.indice (repertorio/repertorio.htm)	25,1
organizzazione.acquisizioni (acq.htm)	informazioni.prestito. interbibliotecario (faqill.htm)	25,1
organizzazione.acquisizioni (acq.htm)	informazioni.ricerchebibliografiche (faqref.htm)	24,8
organizzazione.servizi.pubblico (pre.htm)	informazioni.prestito.domicilio (faqpre.htm)	22,2
organizzazione.periodici (per.htm)	documenti.lavoro.progetto.serse (serse.htm)	21,0
catalogo.repertorio.periodici.indice (repertorio/repertorio.htm)	risorse.elettroniche.indice (journal/ej_intro.asp)	21,0
statistiche.web.2003.menu (/STATISTICHE/ Report2003/menu.htm)	internet.bibliolink (biblink.htm)	20,2
informazioni.medicina.scheda (bibliou8.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	20,0
organizzazione.servizi.pubblico (pre.htm)	informazioni.ricerchebibliografiche (faqref.htm)	18,5
organizzazione.servizi.pubblico (pre.htm)	informazioni.consultazione (faqcons.htm)	18,2
organizzazione.scienze (sci.htm)	informazioni.ricerchebibliografiche (faqref.htm)	18,1
documenti.lavoro.internet.sito (intra/intranet.htm)	statistiche.indice (stat.htm)	17,5
organizzazione.scienze (sci.htm)	risorse.elettroniche.indice (journal/ej_intro.asp)	17,5
catalogo.periodici.alfabetico (elenco/intro.htm)	organizzazione.periodici (per.htm)	17,4
organizzazione.scienze (sci.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	16,7
organizzazione.organigramma (orga.pdf)	statistiche.indice (stat.htm)	16,7
informazioni.medicina.servizi.old (med.htm)	catalogo.periodici.alfabetico.j (elenco/J.HTM)	16,6
catalogo.stone.indice (stone/stone.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	16,6
informazioni.centrale.piantina.livello.2 (piantina/index2.html)	risorse.elettroniche.basididati (ertipo.htm)	16,3
catalogo.stone.indice (stone/stone.htm)	informazioni.regolamento (regolamento.htm)	16,2

Pagina *a*	Pagina *b*	Confidence (%)
informazioni.ricerchebibliografiche (faqref.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	16,1
informazioni.centrale.piantina. classi.cdd.indice (piantina/tabella.html)	documenti.lavoro.cdd.progetto (intra/progcdd.htm)	16,0
statistiche.web.2003.menu (/STATISTICHE/ Report2003/menu.htm)	internet.percorsi (xcorsi.htm)	15,8
informazioni.prestito.domicilio (faqpre.htm)	documenti.ufficiali.cartaservizi.2003 (cartaservizi.htm)	15,1
organizzazione.servizi.pubblico (pre.htm)	internet.percorsi (xcorsi.htm)	14,8
documenti.ufficiali.cartaservizi.attuale (cartaservizi.pdf)	informazioni.regolamento (regolamento.htm)	13,3
statistiche.indice (stat.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	13,3
informazioni.ricerchebibliografiche (faqref.htm)	risorse.elettroniche.basididati (ertipo.htm)	12,9
documenti.lavoro.istruzioni. collocazione (intra/istrucol.htm)	informazioni.consultazione (faqcons.htm)	12,6
risorse.elettroniche.indice (journal/ej_intro.asp)	catalogo.periodici.alfabetico (elenco/intro.htm)	12,6
informazioni.foto (foto.htm)	organizzazione.servizi.pubblico (pre.htm)	12,5
documenti.lavoro.istruzioni. collocazione (intra/istrucol.htm)	catalogo.periodici.alfabetico (elenco/intro.htm)	12,5
informazioni.ricerchebibliografiche (faqref.htm)	internet.percorsi (xcorsi.htm)	12,2
risorse.elettroniche.indice (journal/ej_intro.asp)	internet.bibliolink (biblink.htm)	11,7
documenti.ufficiali.accordi (accordi.htm)	informazioni.consultazione (faqcons.htm)	11,6
internet.percorsi (xcorsi.htm)	risorse.elettroniche.basididati (ertipo.htm)	11,6
catalogo.periodici.alfabetico (elenco/intro.htm)	risorse.elettroniche.indice (journal/ej_intro.asp)	11,2
internet.percorsi (xcorsi.htm)	documenti.lavoro.internet.sito (intra/intranet.htm)	11,2
statistiche.indice (stat.htm)	internet.percorsi (xcorsi.htm)	10,8

Pagina *a*	Pagina *b*	Confidence (%)
internet.link.utenti (yourlink.htm)	risorse.elettroniche.basididati (ertipo.htm)	10,8
risorse.elettroniche.ricerca.area (journal/source.asp)	internet.bibliolink (biblink.htm)	10,6
documenti.ufficiali.accordi (accordi.htm)	informazioni.regolamento (regolamento.htm)	10,5
catalogo.repertorio.periodici.indice (repertorio/repertorio.htm)	risorse.elettroniche.basididati (ertipo.htm)	10,5
internet.link.utenti (yourlink.htm)	risorse.elettroniche.indice (journal/ej_intro.asp)	10,5
internet.bibliolink (biblink.htm)	risorse.elettroniche.areadisciplinare (erdisci.htm)	10,3
catalogo.repertorio.periodici.indice (repertorio/repertorio.htm)	internet.bibliolink (biblink.htm)	10,2

Tabella 4.1 – Sequenze di pagine con livelli significativi di confidence

Una prima indicazione che questo tipo di analisi suggerisce riguarda la creazione di collegamenti (*link*) dalle pagine *a* alle relative pagine *b*. In questo modo l'accesso alle pagine correlate risulta semplificato. Ricordiamo che lo scopo principale dell'analisi di Web Usage Mining è rappresentato dal miglioramento del sito dal punto di vista degli utenti.

Il Web Usage Mining può essere utilizzato anche per sviluppare opportune strategie di *prefetching* e *caching* in modo da ridurre il tempo di risposta del server. Il *prefetching* è una caratteristica del browser che permette ad una pagina HTML di recuperare altri contenuti web quando la connessione del browser dell'utente è inattiva. Il contenuto del *prefetching* viene immagazzinato nella *cache* del browser ed appare quindi velocemente non appena l'utente accede alla pagina che contiene il contenuto immagazzinato.

L'ulteriore suggerimento che si può indicare è quello di inserire in ogni pagina *a* l'istruzione in linguaggio HTML per recuperare il contenuto della relativa pagina *b*. L'istruzione dovrà essere inserita all'inizio del documento e risulterà del tipo:

```
<link rel="prefetch" href="document.html">
```

Sostituendo opportunamente al file di esempio "document.html" i nomi delle relative pagine *b*. Ad esempio nella pagina *a* "inf.htm" si aggiungerà l'istruzione (*cfr.* prima riga della tabella precedente):

```
<link rel="prefetch" href="intra/intranet.htm">
```

In questo modo, mentre l'utente legge il contenuto della pagina relativa ai servizi informatici (inf.htm), il browser preparerà nella sua *cache* la pagina relativa al documento di lavoro riguardante il sito internet (intra/intranet.htm). Il contenuto della pagine memorizzata apparirà quindi velocemente non appena l'utente la richiederà.

Clustering

L'analisi cluster permette di raggruppare utenti con caratteristiche simili in base ai diversi percorsi di navigazione. Si è dunque interessati ad individuare i diversi segmenti comportamentali.

L'input fondamentale per ogni analisi cluster è rappresentato dalla matrice dei dati. Per ottenere tale matrice è stato necessario rielaborare, attraverso una applicazione Java creata ad hoc[33], i dati esportati dal software WUM.

I dati esportati dal software si presentano in questo modo:

[33] Il codice sorgente dell'applicazione è riportato nell'allegato 2.

session	visitor	t_occ	p_id	p_occ	p_url
103967	1017131	1	3000182	1	home.homepage
103967	1017131	2	3000189	1	internet.bibliolink
104051	1017131	1	3000182	1	home.homepage
104057	1017131	1	3000189	1	internet.bibliolink
104057	1017131	2	3000181	1	catalogo.periodici
104057	1017131	3	3000195	1	informazioni.medicina
104057	1017131	4	3000210	1	informazioni.ricerchebiblio
104057	1017131	5	3000195	2	informazioni.medicina
104057	1017131	6	3000184	1	documenti.lavoro
104057	1017131	7	3000197	1	catalogo.repertorio.periodi
104057	1017131	8	3000192	1	internet.cercagoogle
104057	1017131	9	3000185	1	risorse.elettroniche
104060	1017131	1	3000182	1	home.homepage
104176	1017131	1	3000182	1	home.homepage
104210	1017131	1	3000185	1	risorse.elettroniche
104210	1017131	2	3000185	2	risorse.elettroniche
104210	1017131	3	3000185	3	risorse.elettroniche
104217	1017131	1	3000185	1	risorse.elettroniche
104217	1017131	2	3000185	2	risorse.elettroniche

Figura 4.8 – Esportazione dal software WUM

Descriviamo brevemente cosa rappresentano i campi:

- **session:** codice della sessione
- **visitor:** codice del visitatore
- **t_occ:** numero delle pagine visitate fino a quel punto della sessione
- **p_id:** codice della pagina
- **p_occ:** numero delle pagine visitate per una particolare area del sito
- **p_url:** nome dell'area del sito a cui appartiene la pagina visitata

In particolare, il nostro interesse riguarda solamente i campi *session*, *visitor*, *p_occ* e *p_url*. Attraverso la rielaborazione tramite l'applicazione Java "reOrganizeTable" si ottiene la matrice di dati, uno stralcio della quale è rappresentato nella figura seguente:

visitorID	catalogo_perioc	catalogo_repert	documenti_cror	documenti_lavo	documenti_uffic	informazioni_av
1017372	0	0	0	0	0	0
1017373	1	4	0	0	0	0
1017374	0	0	0	0	0	0
1017375	2	0	0	0	0	0
1017376	0	0	0	0	0	0
1017377	1	0	0	0	0	0
1017126	0	0	0	0	0	1
1017127	1	0	0	0	0	0
1017128	0	0	0	0	0	0
1017129	0	0	0	1	0	0
1017130	0	0	0	0	0	0
1017131	2	1	0	1	1	0
1017132	0	0	0	0	0	0
1017133	0	0	0	0	0	0
1017134	0	0	0	0	0	0
1017135	0	0	0	0	0	0
1017136	0	0	0	0	0	0
1017137	0	0	0	0	0	0
1017138	0	0	0	0	0	0
1017139	0	0	0	0	0	0
1017140	0	0	0	0	0	0

Figura 4.9 – Stralcio della matrice dei dati

Le righe della matrice rappresentano i diversi visitatori, le colonne rappresentano le diverse aree del sito. Si noti che in corrispondenza di ciascuna area di pagine si ha una variabile discreta, che mostra il numero di visite effettuato da ciascun visitatore all'area considerata. Per mancanza di spazio, nella figura di esempio vengono visualizzate solamente 6 delle 32 aree considerate. Si riportano nella tabella seguente le 32 aree analizzate:

catalogo.periodici
catalogo.repertorio.periodici
catalogo.stone
documenti.cronologico
documenti.lavoro
documenti.ufficiali
home.homepage
informazioni.avvisi.archivio
informazioni.avvisi.attuali
informazioni.avvisi.collaborazioni
informazioni.avvisi.notizie
informazioni.avvisi.software
informazioni.chisiamo
informazioni.consultazione

informazioni.foto
informazioni.generali
informazioni.medicina
informazioni.opac
informazioni.prestito.domicilio
informazioni.prestito.interbibliotecario
informazioni.regolamento
informazioni.ricerchebibliografiche
informazioni.scienze
informazioni.whoweare
internet.bibliolink
internet.cercagoogle
internet.link.utenti
internet.percorsi
internet.webstory
organizzazione
risorse.elettroniche
statistiche

Tabella 4.2 – Aree considerate

I dati così elaborati vengono importati nel software statistico **SPSS**. Per individuare i diversi segmenti comportamentali si utilizza la procedura **TwoStep Cluster Analysis**. L'algoritmo impiegato in questa procedura possiede diverse caratteristiche desiderabili che lo differenziano dalle tecniche tradizionali di clustering. In particolare, è un algoritmo scalabile che ha la capacità di creare cluster basandosi sia su variabili categoriche che continue. Un ulteriore vantaggio è rappresentato dal fatto che il numero di cluster viene selezionato automaticamente.

Andiamo ad osservare i risultati dell'analisi. La procedura trova 5 differenti cluster. Per prima cosa andiamo a vedere come si distribuiscono gli *N*=3411 visitatori all'interno dei cluster trovati.

Distribuzione cluster			
		N	**% del totale**
Cluster	1	1288	37,8%
	2	599	17,6%
	3	463	13,6%
	4	543	15,9%
	5	518	15,2%
Totale		3411	100,0%

Tabella 4.3 – Distribuzione delle osservazioni all'interno dei cluster

Per una migliore visualizzazione rappresentiamo la distribuzione all'interno di un diagramma a torta:

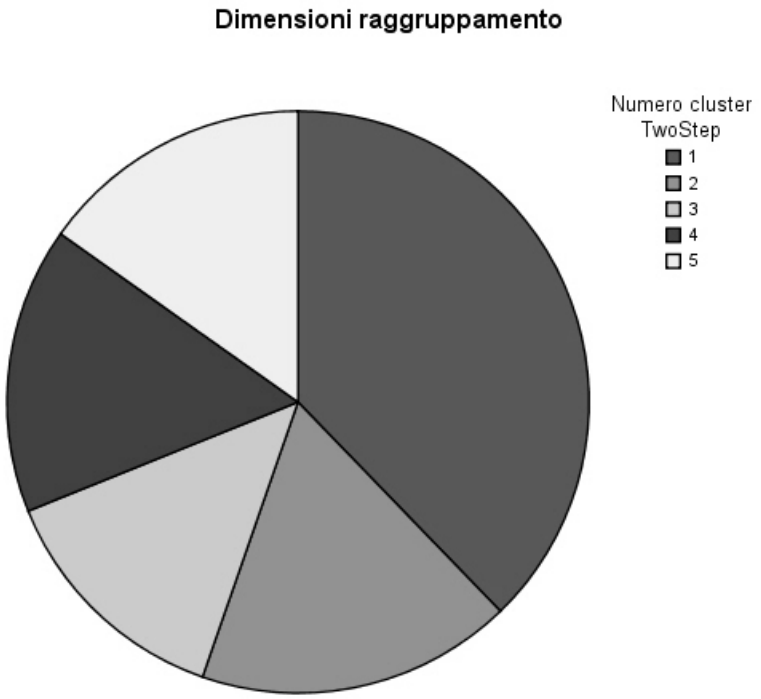

Figura 4.10 – Diagramma delle dimensioni del raggruppamento

Come si può facilmente osservare, la maggior parte dei visitatori si colloca all'interno del primo cluster (37,8%), mentre gli altri visitatori si distribuiscono quasi uniformemente all'interno dei restanti quattro cluster.

Andiamo ora a presentare i grafici dei diversi segmenti comportamentali, in ascissa viene calcolata la statistica c^2 di Pearson[34] che rappresenta l'importanza di ciascuna variabile all'interno dello specifico cluster.

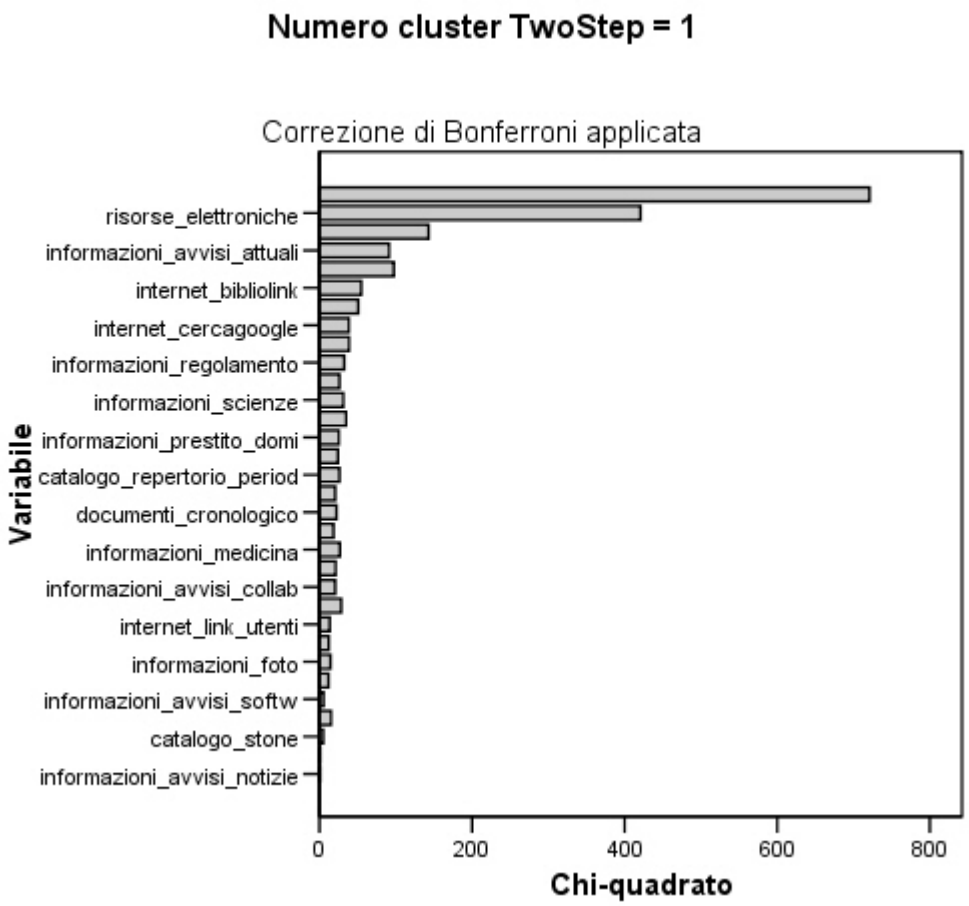

Figura 4.11 – Cluster 1

[34] Per una spiegazione della statistica c^2 si confronti il paragrafo 2.4 nella parte relativa agli indici di connessione.

Segmento 1: "Ricercatori informatizzati": i componenti di questo gruppo sono caratterizzati da un utilizzo del sito rivolto soprattutto alla navigazione di pagine relative alle risorse elettroniche. Sono inoltre interessati anche ai cataloghi delle riviste cartacee possedute dalla biblioteca e tendono a rimanere aggiornati tramite il frequente accesso agli avvisi e alle altre pagine di informazione. Probabilmente una buona parte di questo gruppo è rappresentata da docenti, ricercatori, dottorandi e tesisti.

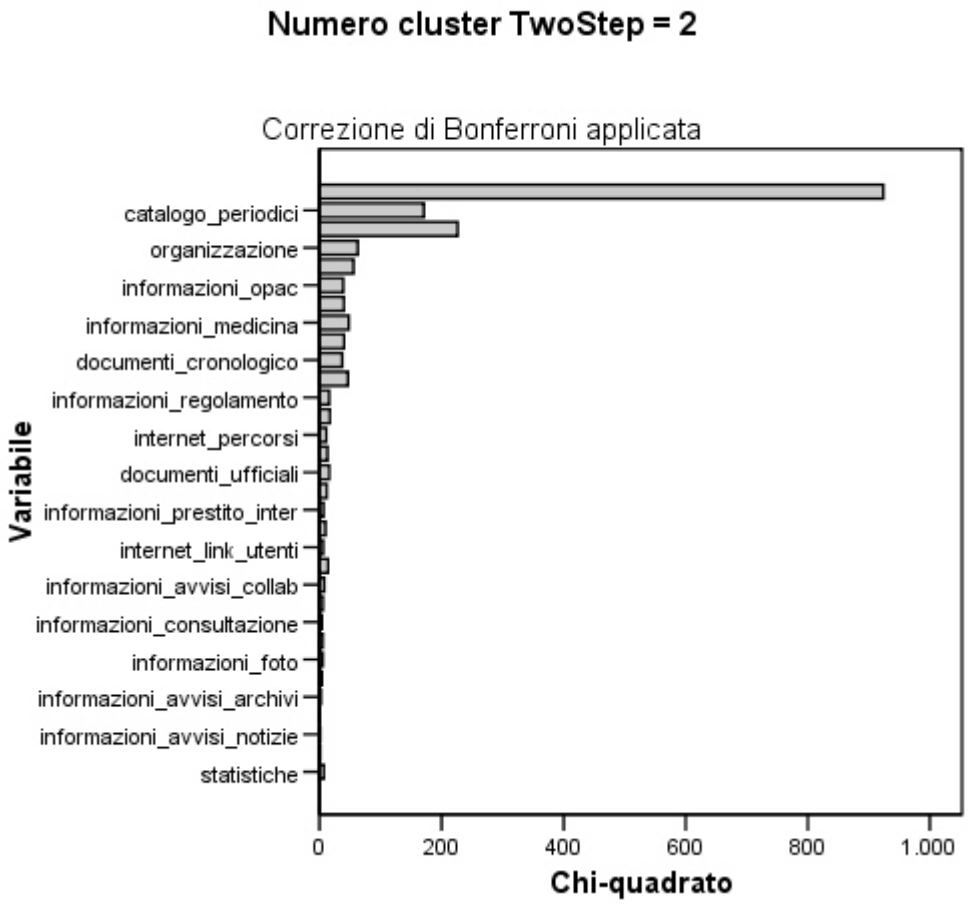

Figura 4.12 – Cluster 2

Segmento 2: "Ricercatori tradizionali": questi utenti accedono principalmente al catalogo dei periodici cartacei posseduti dalla biblioteca. Sono inoltre interessati all'organizzazione della biblioteca e alle informazioni relative al catalogo OPAC. Si noti che molti accedono anche alle informazioni sulla sede di medicina, ciò può far

supporre che una buona parte degli utenti di questo gruppo facciano parte della Facoltà di Medicina. Probabilmente anche questo gruppo è formato principalmente da docenti, ricercatori, dottorandi e tesisti.

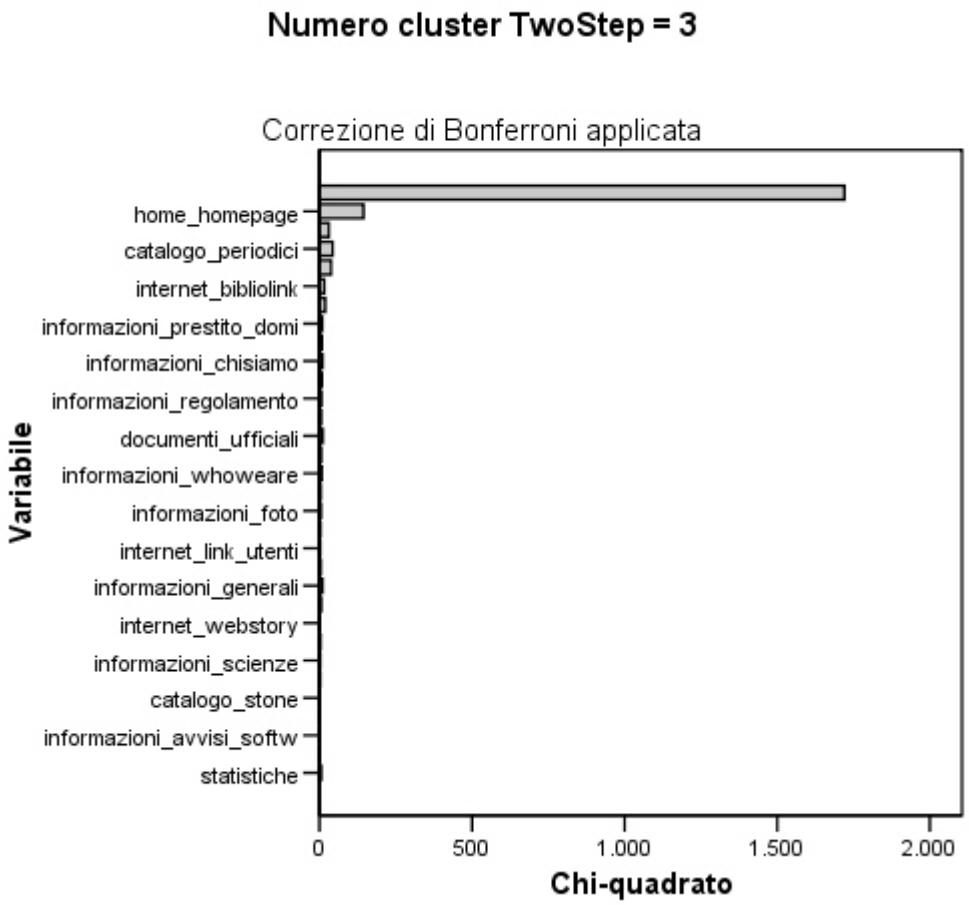

Figura 4.13 – Cluster 3

Segmento 3: "Diretti all'OPAC": i navigatori appartenenti a questo gruppo sono accomunati dal fatto di visitare quasi esclusivamente l'home page del sito. Questo significa che con ogni probabilità essi accederanno successivamente al catalogo OPAC, per ricercare monografie o periodici posseduti dalla biblioteca. Ricordiamo che il catalogo OPAC non è compreso nella nostra analisi. Si può notare il fatto che oltre alla home page le pagine con maggiore accesso siano il catalogo periodici e la pagina dei

link, forse per estendere la loro ricerca alle altre biblioteche. Si può supporre che questo gruppo sia formato principalmente dagli studenti universitari.

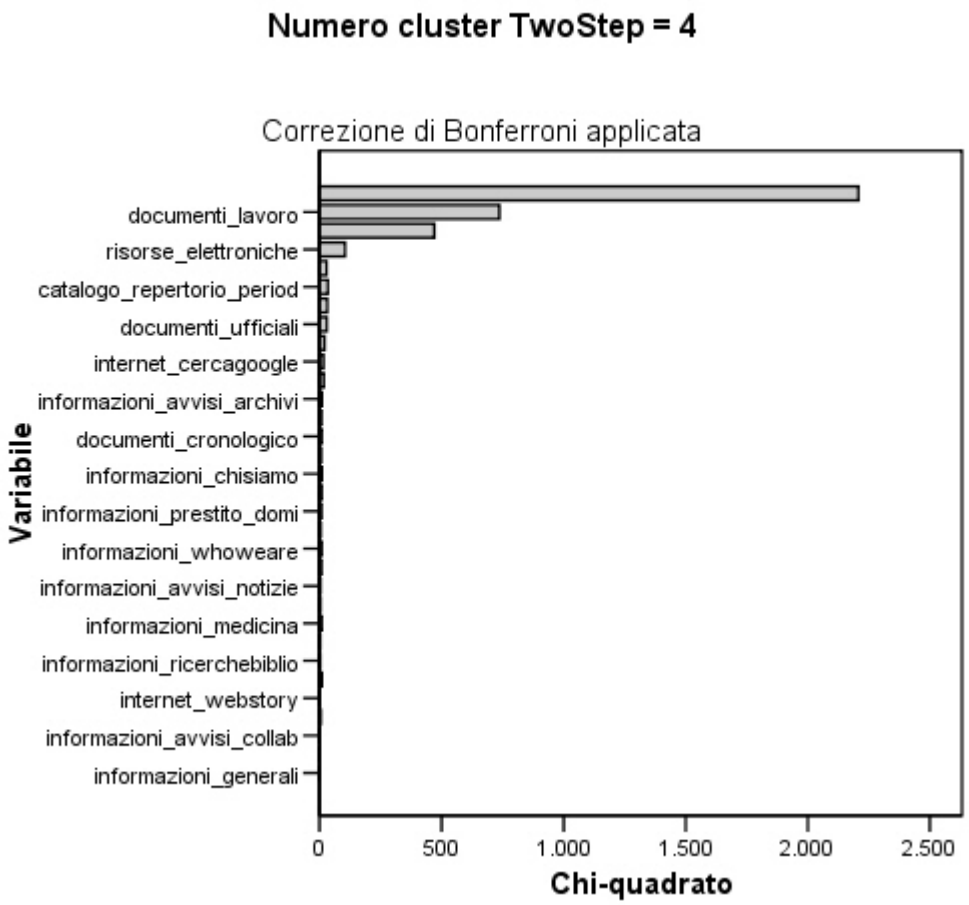

Figura 4.14 – Cluster 4

Segmento 4: "Bibliotecari": è evidente che gli appartenenti a tale gruppo siano in qualche modo dei professionisti che lavorano in ambito bibliotecario. Essi hanno una frequentazione media estremamente alta delle pagine riguardanti i documenti di lavoro. Costoro potrebbero essere visti come persone legate al sito per motivi di lavoro, probabilmente molti di essi operano in altre biblioteche ed accedono ai documenti per informarsi e per avere dei punti di riferimento.

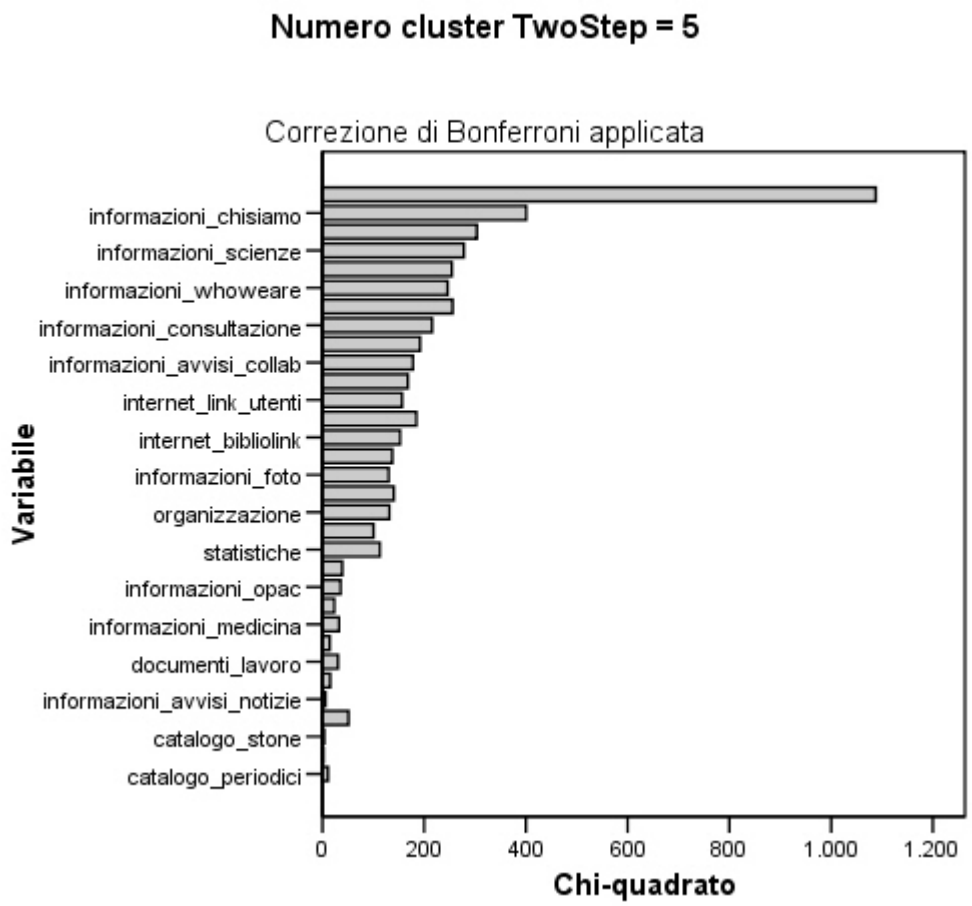

Figura 4.15 – Cluster 5

Segmento 5: "Curiosi": è il gruppo di utenti che sfruttano le aree informative del sito. La loro navigazione avviene principalmente sulle pagine riservate alle informazioni generali, sulle pagine relative all'organizzazione della biblioteca, sulle pagine relative alle statistiche e su quelle preposte a fornire *link* ad altri siti di interesse. Questo gruppo di utenti potrebbe essere composto per lo più da studenti universitari o da aspiranti tali, ma non è da escludere che anche professionisti legati al mondo delle biblioteche possano appartenere a questo gruppo.

L'analisi cluster condotta, e la conseguente divisione in gruppi di visitatori, ci suggerisce di creare dei percorsi personalizzati per i cinque diversi tipi di utente. Le

pagine di ciascun gruppo dovrebbero risultare collegate reciprocamente attraverso l'inserimento di *link*. Oltre a questo si potrebbero inserire nella home page i collegamenti alle pagine più importanti di ciascun gruppo, in modo che un utente trovi subito il servizio di interesse.

Analisi delle macroaree del sito

Prendendo in considerazione le macroaree del sito definite nel paragrafo 4.2 si può procedere con un'ulteriore analisi, calcolando gli indici di confidence e gli indici di correlazione tra le diverse aree.

Sono stati calcolati, attraverso l'impiego del software WUM, gli indici di confidence relativi alle diverse aree. Riassumiamo i risultati ottenuti nel grafo visualizzato in figura 4.16. Si tratta di un grafo orientato nel quale il verso indica la direzione della relazione, mentre i numeri sugli archi indicano il livello percentuale di confidence. Gli archi con un livello di confidence maggiore sono quelli più marcati.

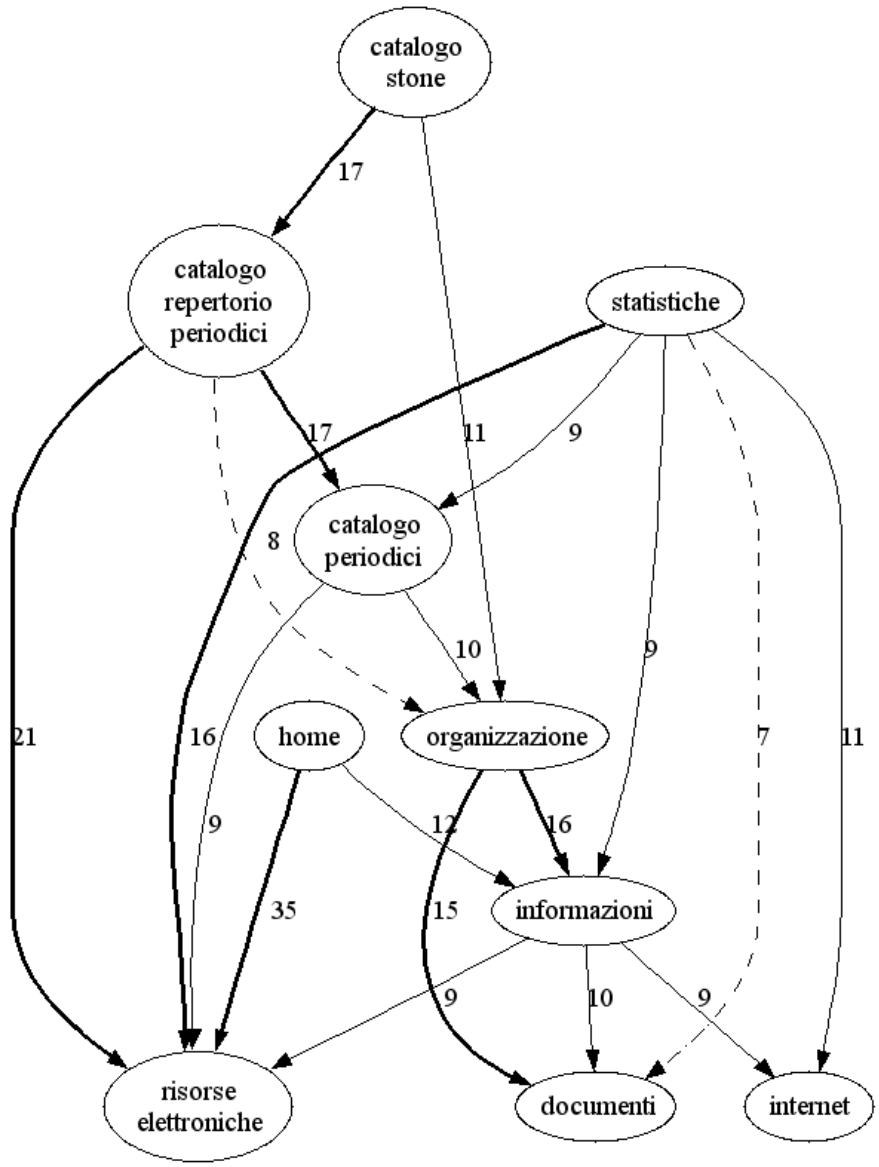

Figura 4.16 – Livelli significativi di confidence (%) per le diverse macroaree

Abbiamo calcolato anche il coefficiente di correlazione lineare[35] tra le diverse aree. Nel grafo visualizzato nella figura 4.17 possiamo osservare i risultati ottenuti. Si tratta di un grafo non orientato, i numeri sugli archi indicano i diversi coefficienti di correlazione lineare (in percentuale) calcolati per le diverse aree. Anche in questo caso valori più elevati sono contraddistinti da archi più marcati.

[35] Per una spiegazione del coefficiente di correlazione lineare si rimanda il lettore al paragrafo 2.3.

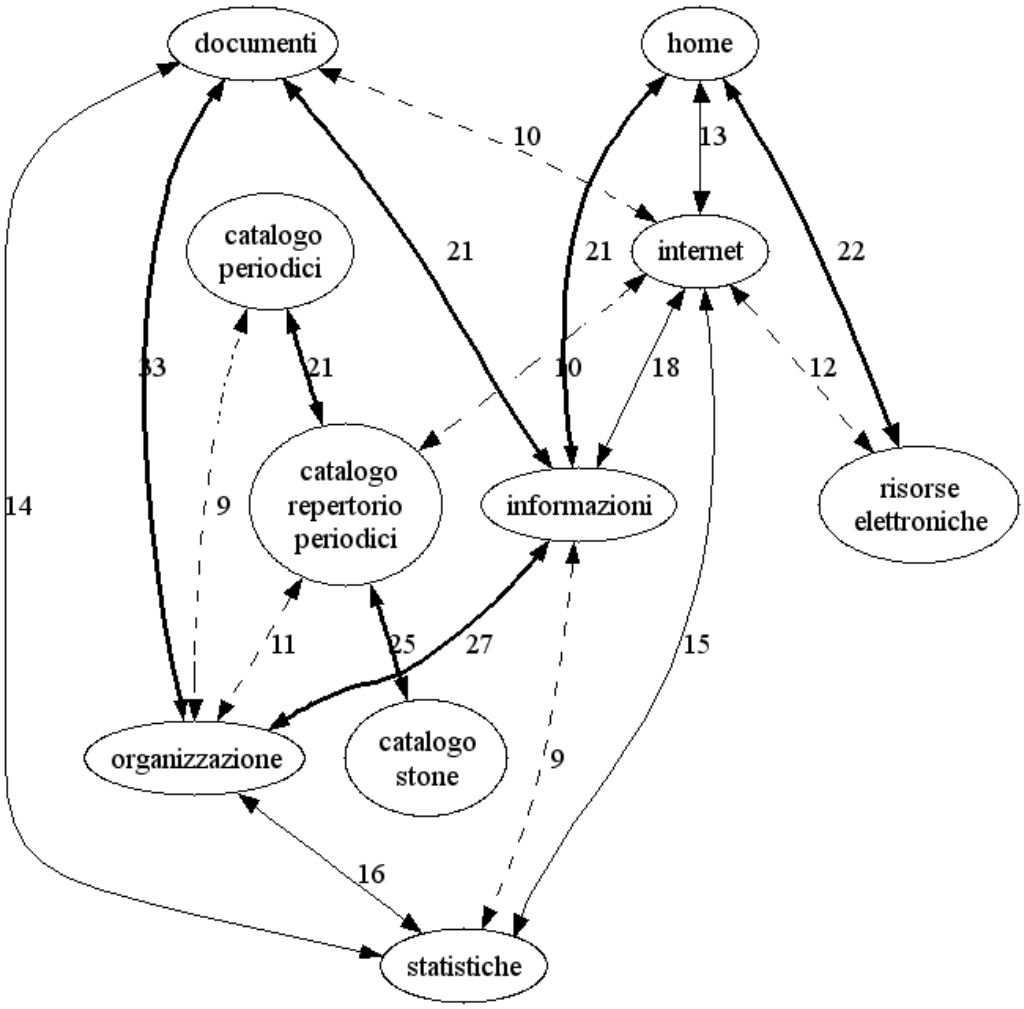

Figura 4.17 – Coefficienti di correlazione lineare (%) calcolati per le diverse aree

Da una prima osservazione dei due grafi, notiamo che le due misure impiegate non ci hanno condotto ad identici risultati, ma ci hanno condotto a risultati in qualche modo simili. Ciò è dovuto al fatto che l'indice di confidence prende in considerazione le *sequenze* di pagine, mentre il coefficiente di correlazione lineare considera solamente gli insiemi di pagine di ogni sessione utente, senza riferimenti temporali.

Dall'ispezione dei due grafi possiamo trarre alcune considerazioni. I gruppi di diverse aree maggiormente correlate tra loro sono:

1. *Documenti – Organizzazione – Informazioni*

2. *Internet – Informazioni – Statistiche*

3. *Catalogo Periodici – Catalogo Stone – Catalogo Repertorio Periodici*

Ordiniamo le visite alle aree correlate considerando anche le sequenze di visita:

- *Organizzazione* **à** *Informazioni* **à** *Documenti*
- *Organizzazione* **à** *Documenti*
- *Statistiche* **à** *Informazioni* **à** *Internet*
- *Statistiche* **à** *Internet*
- *Catalogo Stone* **à** *Catalogo Repertorio Periodici* **à** *Catalogo Periodici*

Possiamo quindi osservare la *direzione* di navigazione degli utenti mentre esplorano le diverse aree del sito alla ricerca di informazioni. Anche in questo caso, per migliorare la navigazione si potrebbero inserire dei collegamenti alle aree successive delle sequenze sopra riportate.

Conclusioni

In questa ricerca, è stato presentato il recente campo di studi interdisciplinare denominato Web Usage Mining.

Basandoci sulla letteratura riguardante l'argomento abbiamo analizzato il Web Usage Mining in relazione ai recenti sviluppi di quest'area, che sta ricevendo una crescente attenzione da parte degli esperti di Data Mining.

Misurare il successo di un sito, stabilire quali pagine devono essere modificate per ottenere dei vantaggi economici, in termini di fidelizzazione dei clienti o di acquisizione di nuovi clienti, sono tutti risultati di notevole interesse per i gestori dei siti di commercio elettronico. Inoltre, attraverso le tecniche di Web Usage Mining si possono stabilire i diversi segmenti comportamentali degli utenti e pertanto proporre campagne pubblicitarie personalizzate.

Per esemplificare le tecniche si è scelto di analizzare il sito Internet della Biblioteca di Ateneo impiegando le sequenze associative e l'analisi di raggruppamento, due tra le metodologie maggiormente utilizzate nel Web Usage Mining. Per quanto riguarda la scoperta dei percorsi di navigazione, ci si è avvalsi del software WUM, realizzato nell'ambito di un progetto di ricerca dell'Università di Berlino. Mentre l'analisi cluster è stata eseguita con l'ausilio del software statistico SPSS.

I risultati di questa analisi hanno condotto a delle raccomandazioni su come migliorare la struttura del sito, creando dei percorsi personalizzati in base al diverso gruppo cui appartiene un utente e inserendo dei collegamenti alle pagine che vengono visitate spesso in sequenza.

Analizzando questa tesi in retrospettiva e considerando il fatto che questo campo di studi è in continuo sviluppo ad opera di autorevoli ricercatori scientifici, spero che questo lavoro sia riuscito a dare una visione introduttiva dell'argomento.

Allegato 1 – Tassonomia

Pagina	Concetto
/000202.txt	documenti.ufficiali.verbali.2000.0202
/000216.txt	documenti.ufficiali.verbali.2000.0216
…	…
/040614.txt	documenti.ufficiali.verbali.2004.0614
/STATISTICHE/Report2003/123LogReport.htm	statistiche.web.2003.home
/STATISTICHE/Report2003/menu.htm	statistiche.web.2003.menu
/STATISTICHE/Report2003/p0.htm	statistiche.web.2003.pag1
/STATISTICHE/Report2003/p1.htm	statistiche.web.2003.pag2
…	…
/STATISTICHE/Report2003/p9.htm	statistiche.web.2003.pag10
/STATISTICHE/Report2003/print.htm	statistiche.web.2003.tutte
/STATISTICHE/Report2003/report.csv	statistiche.web.2003.to.delete.before.analysis
/accordi.htm	documenti.ufficiali.accordi
/acq.htm	organizzazione.acquisizioni
/aidawin.htm	informazioni.avvisi.software.aida
/archivio.htm	informazioni.avvisi.archivio
/attiv00.htm	documenti.ufficiali.attivita.2000
/attiv99.htm	documenti.ufficiali.attivita.1999
/avvisi.htm	informazioni.avvisi.attuali
/biblink.htm	internet.bibliolink
/bibliou2.htm	informazioni.scienze.scheda
/bibliou6.htm	informazioni.centrale.scheda
/bibliou8.htm	informazioni.medicina.scheda
/blank.html	home.homepage
/cartaservizi.htm	documenti.ufficiali.cartaservizi.2003
/cartaservizi.old.old.htm	documenti.ufficiali.bozza.cartaservizi.1999
/cartaservizi.pdf	documenti.ufficiali.cartaservizi.attuale
/cartaservizi.txt	documenti.ufficiali.bozza.cartaservizi.2001
/cdb.htm	organizzazione.organi.attuale
/cdb_00-02.htm	organizzazione.organi.2002
/cdd.htm	organizzazione.classificazione
/cercagoogle.htm	internet.cercagoogle
/cgeol.pdf	informazioni.scienze.catalogo.cartegelogiche
/chisiamo.htm	informazioni.chisiamo
/crono.htm	documenti.cronologico
/dir.htm	organizzazione.direzione
/ebsco.htm	catalogo.periodici.to.delete.before.analysis
/elenco/A.HTM	catalogo.periodici.alfabetico.a
/elenco/B.HTM	catalogo.periodici.alfabetico.b
…	…
/elenco/Z.HTM	catalogo.periodici.alfabetico.z

Pagina	Concetto
/elenco/az.htm	catalogo.periodici.menu.to.delete.before.analysis
/elenco/intro.htm	catalogo.periodici.alfabetico
/elenco/periodici.htm	catalogo.periodici.elenco.to.delete.before.analysis
/erdisci.htm	risorse.elettroniche.areadisciplinare
/ertipo.htm	risorse.elettroniche.basididati
/faqcons.htm	informazioni.consultazione
/faqill.htm	informazioni.prestito.interbibliotecario
/faqopac.htm	informazioni.opac
/faqpre.htm	informazioni.prestito.domicilio
/faqref.htm	informazioni.ricerchebibliografiche
/foto.htm	informazioni.foto
/grafcdd.htm	statistiche.grafici.cdd
/grafpre.htm	statistiche.grafici.prestiti
/index.html	home.to.delete.before.analysis
/indicatori.pdf	statistiche.indicatori.2001.2003
/inf.htm	organizzazione.servizi.informatici
/intra/MdG/mdghome.htm	organizzazione.direzione.direttore
/intra/aleph/000.pdf	documenti.lavoro.aleph.000
/intra/aleph/100.pdf	documenti.lavoro.aleph.100
…	…
/intra/aleph/900.pdf	documenti.lavoro.aleph.900
/intra/aleph/aleph.htm	documenti.lavoro.aleph.indice
/intra/aleph/deriva.pdf	documenti.lavoro.aleph.derivazione
/intra/aleph/derivaweb.pdf	documenti.lavoro.aleph.derivazioneweb
/intra/aleph/gesper.pdf	documenti.lavoro.aleph.catalogo.periodici.gestione
/intra/aleph/gestord.pdf	documenti.lavoro.aleph.ordini.gestione
/intra/aleph/item.pdf	documenti.lavoro.aleph.item
/intra/aleph/leader.pdf	documenti.lavoro.aleph.leader
/intra/aleph/multivolume.pdf	documenti.lavoro.aleph.multivolume
/intra/aleph/ordcanc.pdf	documenti.lavoro.aleph.ordini.cancellazione
/intra/aleph/ser040518.pdf	documenti.lavoro.aleph.formazione.programma
/intra/biblioguidau2.pdf	documenti.ufficiali.scienze.cartaservizi
/intra/cen011001.pdf	documenti.lavoro.centrale.backoffice.riorganizzazione
/intra/cen020214a.pdf	documenti.lavoro.prestiti.grafici.2001
/intra/cen020214b.pdf	documenti.lavoro.prestiti.grafici.1995.2001
/intra/cen020214c.pdf	documenti.lavoro.prestiti.interbibliotecario.grafici
/intra/cen020214d.pdf	documenti.lavoro.reference.utenti.distribuzione
/intra/cen020215.pdf	documenti.lavoro.sbn.collocazione
/intra/cen020408.pdf	documenti.lavoro.diritto.inpillole
/intra/cen020527.pdf	documenti.lavoro.bibliografica.ricerca.online
/intra/cen020529.pdf	informazioni.avvisi.collaborazioni.bando.2002.05
/intra/cen020622.pdf	informazioni.avvisi.collaborazioni.graduatoria.2002.06
/intra/cen020904.pdf	documenti.lavoro.aleph.guida.circolazione
/intra/cen020906.pdf	informazioni.avvisi.norme.accesso
/intra/cen020918a.pdf	documenti.lavoro.catalogo.consultazione.guida
/intra/cen020918b.pdf	documenti.lavoro.catalogo.consultazione.guida.breve
/intra/cen020923.pdf	documenti.lavoro.centrale.1998.2002
/intra/cen021005.pdf	documenti.lavoro.elenco.catalogo.periodici.doppi
/intra/cen021104.pdf	informazioni.avvisi.collaborazioni.bando.2002.11
/intra/cen021122.pdf	informazioni.avvisi.collaborazioni.graduatoria.2002.11
/intra/cen030409.pdf	informazioni.avvisi.collaborazioni.bando.2003.04

Pagina	Concetto
/intra/cen030430.pdf	informazioni.avvisi.collaborazioni.graduatoria.2003.04
/intra/cen030908.pdf	informazioni.avvisi.collaborazioni.bando.2003.09
/intra/cen030922.pdf	documenti.lavoro.sedecentrale.servizi.guidarapida
/intra/cen030929.pdf	informazioni.avvisi.collaborazioni.graduatoria.2003.09
/intra/cen031117.pdf	informazioni.avvisi.collaborazioni.graduatoria.2003.11
/intra/cen031204.pdf	informazioni.avvisi.collaborazioni.graduatoria.2003.12
/intra/cen040130.pdf	informazioni.avvisi.collaborazioni.graduatoria.2004.01
/intra/cen040223.pdf	informazioni.avvisi.collaborazioni.graduatoria.2004.02
/intra/cen040305a.pdf	documenti.lavoro.servizi.guidarapida
/intra/cen040305b.pdf	documenti.lavoro.opac.guidarapida
/intra/cen040305c.pdf	documenti.lavoro.consultazione.scaffale.guidarapida
/intra/cen040702.pdf	informazioni.avvisi.collaborazioni.bando.2004.07
/intra/convunimi.pdf	documenti.lavoro.convenzione.reciprocita.unimi
/intra/corsostudenti/index.htm	documenti.lavoro.corsostudenti.indice
/intra/corsostudenti/porta2.ppt	documenti.lavoro.corsostudenti.presentazione.intera
/intra/corsostudenti/porta2001.ppt	documenti.lavoro.corsostudenti.presentazione.intera
/intra/corsostudenti/sld001.htm	documenti.lavoro.corsostudenti.presentazione.slide.1
…	…
/intra/corsostudenti/sld009.htm	documenti.lavoro.corsostudenti.presentazione.slide.9
/intra/corsostudenti/tsld001.htm	documenti.lavoro.corsostudenti.presentazione.slide.to.delete.before.analysis
…	…
/intra/corsostudenti/tsld009.htm	documenti.lavoro.corsostudenti.presentazione.slide.to.delete.before.analysis
/intra/faq.html	internet.webstory.opac.faq
/intra/hidden.htm	internet.webstory.hidden
/intra/home9609.htm	internet.webstory.home.1996.09
/intra/home9610.htm	internet.webstory.home.1996.10
/intra/home9612.htm	internet.webstory.home.1996.12
/intra/home9701.htm	internet.webstory.home.1997.01
/intra/home9801.htm	internet.webstory.home.1998.01
/intra/intranet.htm	documenti.lavoro.internet.sito
/intra/istrucdd.htm	documenti.lavoro.istruzioni.cdd
/intra/istrucol.htm	documenti.lavoro.istruzioni.collocazione
/intra/istruric.htm	documenti.lavoro.istruzioni.ricollocazione
/intra/istrustu.htm	documenti.lavoro.istruzioni.collaboratori
/intra/joke.htm	documenti.lavoro.internet.scherzetto
/intra/katia/wound care.mht	various.to.delete.before.analysis
/intra/lsd.htm	documenti.lavoro.prestito.lsd
/intra/med020502.pdf	informazioni.avvisi.collaborazioni.medicina.bando.2002.05
/intra/med020510.pdf	documenti.lavoro.medicina.bibliografica.ricerca.online
/intra/med021010.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2002.10
/intra/med030124.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2003.01
/intra/med030213.pdf	informazioni.avvisi.collaborazioni.graduatoria.medicina.2003.02
/intra/med030303.pdf	documenti.lavoro.medicina.situazione.2003
/intra/med030603.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2003.06
/intra/med030618.pdf	informazioni.avvisi.collaborazioni.graduatoria.medicina.2003.06
/intra/med030901.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2003.09
/intra/med030924.pdf	informazioni.avvisi.collaborazioni.graduatoria.medicina.2003

Pagina	Concetto
	.09
/intra/med031124.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2003.11
/intra/med031217.pdf	informazioni.avvisi.collaborazioni.graduatoria.medicina.2003.11
/intra/med040120a.pdf	documenti.lavoro.medicina.servizi.guidarapida
/intra/med040120b.pdf	documenti.lavoro.medicina.document.delivery.guidarapida
/intra/med040210.pdf	documenti.lavoro.medicina.servizi.personale.ospedale.sangerardo
/intra/med040223.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2004.02
/intra/med040324.pdf	informazioni.avvisi.collaborazioni.graduatoria.medicina.2004.02
/intra/opac.htm	internet.webstory.opac
/intra/progcdd.htm	documenti.lavoro.cdd.progetto
/intra/progetti.htm	internet.webstory.progetti
/intra/progrivi.htm	documenti.lavoro.catalogo.periodici.progetto
/intra/rota.htm	documenti.lavoro.rotazione
/intra/sceltecdd.htm	documenti.lavoro.cdd.scelte
/intra/sci011119.pdf	documenti.lavoro.scienze.presenza.utenti
/intra/sci011120.pdf	documenti.lavoro.scienze.sondaggio.utenti
/intra/sci011122.pdf	documenti.lavoro.scienze.prestiti.classe.1
/intra/sci020204.pdf	documenti.lavoro.scienze.prestiti.classe.2
/intra/sci020415.pdf	documenti.lavoro.scienze.promemoria.uso.interno
/intra/sci020708.pdf	documenti.lavoro.scienze.situazione.prospettive.2002
/intra/sci021111.pdf	informazioni.avvisi.collaborazioni.bando.scienze.2002.11
/intra/sci021129.pdf	informazioni.avvisi.collaborazioni.graduatoria.scienze.2002.11
/intra/ser010601.pdf	documenti.lavoro.valutazione.backoffice
/intra/ser010701.pdf	documenti.lavoro.situazione.biblioteca.2000.2001
/intra/ser010801.pdf	documenti.lavoro.corso.formazione
/intra/ser011023.pdf	documenti.lavoro.software.aleph.scelta
/intra/ser020627.pdf	documenti.lavoro.servizi.informatici.sviluppo
/intra/ser020726.pdf	documenti.lavoro.software.sbn.ricordo
/intra/ser020904.pdf	documenti.lavoro.software.aleph.guida.circolazione
/intra/ser021127.pdf	documenti.lavoro.classificazione.scelta.strategica
/intra/ser021220.pdf	informazioni.avvisi.circolare.posti.disponibili.2002.12
/intra/ser021223.pdf	informazioni.avvisi.personale.bando.assunzioni.2002.12
/intra/ser030109.pdf	informazioni.avvisi.personale.bando.assunzioni.bibliografia.2002.12
/intra/ser030122.pdf	informazioni.avvisi.convenzione.reciprocita.insubria
/intra/ser030228.pdf	informazioni.avvisi.collaborazioni.bando.elaborazione.rer.2003.02
/intra/ser030304.pdf	documenti.lavoro.corso.ricerca.bibliografica.introduzione
/intra/ser030403.pdf	documenti.lavoro.gruppo.aleph
/intra/ser030515.pdf	documenti.ufficiali.riepilogo.economico.2002
/intra/ser030714.pdf	documenti.lavoro.formazione.colleghi
/intra/ser031028.pdf	documenti.lavoro.struttura.organizzativa
/intra/ser040213.pdf	documenti.ufficiali.riepilogo.economico.2003
/intra/ser040219.pdf	documenti.ufficiali.strategie.obiettivi.2004
/intra/trans.htm	internet.webstory.ragnatela
/intra/visioni.htm	documenti.lavoro.libri.visione
/intra/webstory.htm	internet.webstory.home

Pagina	Concetto
/journal/alpha.asp	risorse.elettroniche.alfabetico.tutti
/journal/alpha.asp?titolo=A	risorse.elettroniche.alfabetico.a
/journal/alpha.asp?titolo=B	risorse.elettroniche.alfabetico.b
…	…
/journal/alpha.asp?titolo=Z	risorse.elettroniche.alfabetico.z
/journal/ej_intro.asp	risorse.elettroniche.indice
/journal/scoinn.asp	risorse.elettroniche.ricerca.libera.contiene
/journal/score.asp	risorse.elettroniche.ricerca.libera.inizia
/journal/source.asp	risorse.elettroniche.ricerca.editore
/journal/subject.asp	risorse.elettroniche.ricerca.area
/med.htm	informazioni.medicina.servizi.old
/normeej.htm	risorse.elettroniche.norme.utilizzo
/notizie/0109cen.txt	informazioni.avvisi.notizie.centrale.2001.09
/notizie/0109med.txt	informazioni.avvisi.notizie.medicina.2001.09
/notizie/0109sci.txt	informazioni.avvisi.notizie.scienze.2001.09
/notizie/0109ser.txt	informazioni.avvisi.notizie.servizi.area.2001.09
/notizie/0110cen.txt	informazioni.avvisi.notizie.centrale.2001.10
/notizie/0110med.txt	informazioni.avvisi.notizie.medicina.2001.10
…	…
/notizie/0403cen.txt	informazioni.avvisi.notizie.centrale.2004.03
/notizie/0403med.txt	informazioni.avvisi.notizie.medicina.2004.03
/notizie/0403sci.txt	informazioni.avvisi.notizie.scienze.2004.03
/notizie/0403ser.txt	informazioni.avvisi.notizie.servizi.area.2004.03
/notizie/0404cen.txt	informazioni.avvisi.notizie.centrale.2004.04
/notizie/0404sci.txt	informazioni.avvisi.notizie.scienze.2004.04
/notizie/0404ser.txt	informazioni.avvisi.notizie.servizi.area.2004.04
/notizie/0405cen.txt	informazioni.avvisi.notizie.centrale.2004.05
/notizie/0406cen.txt	informazioni.avvisi.notizie.centrale.2004.06
/notizie/bandou2.pdf	informazioni.avvisi.collaborazioni.bando.scienze.2002.01
/notizie/bandou6.pdf	informazioni.avvisi.collaborazioni.bando.centrale.2002.01
/notizie/bandou8.pdf	informazioni.avvisi.collaborazioni.bando.medicina.2002.01
/orga.htm	organigramma.orphan.to.delete.before.analysis
/orga.pdf	organizzazione.organigramma
/pacc0328.txt	documenti.lavoro.bozza.convenzione.reciprocita.unimi
/per.htm	organizzazione.periodici
/percesi.pdf	documenti.lavoro.catalogo.periodici.cesi
/perhsg.pdf	documenti.lavoro.catalogo.periodici.ospedale.sangerardo
/piantina/cdd000.html	informazioni.centrale.piantina.classi.cdd.000
/piantina/cdd100.html	informazioni.centrale.piantina.classi.cdd.100
…	…
/piantina/cdd900.html	informazioni.centrale.piantina.classi.cdd.900
/piantina/cellaA.html	informazioni.centrale.piantina.celle.a
/piantina/cellaB.html	informazioni.centrale.piantina.celle.b
…	…
/piantina/cellaL.html	informazioni.centrale.piantina.celle.l
/piantina/helplibri.html	informazioni.centrale.piantina.help.libri
/piantina/helpprova.html	informazioni.centrale.piantina.help.indice
/piantina/helpriviste.html	informazioni.centrale.piantina.help.riviste
/piantina/index1.html	informazioni.centrale.piantina.livello.1
/piantina/index2.html	informazioni.centrale.piantina.livello.2
/piantina/tabella.html	informazioni.centrale.piantina.classi.cdd.indice

Pagina	Concetto
/pre.htm	organizzazione.servizi.pubblico
/preload.html	home.to.delete.before.analysis
/recipro00.htm	documenti.lavoro.reciprocita.unimi.dati.2000
/regolamento.htm	informazioni.regolamento
/repertorio/antico.htm	catalogo.repertorio.periodici.antico
/repertorio/consgen.htm	catalogo.repertorio.periodici.consultazione.generale
/repertorio/crimino.htm	catalogo.repertorio.periodici.criminologia
/repertorio/dirambie.htm	catalogo.repertorio.periodici.diritto.ambiente
/repertorio/dirammi.htm	catalogo.repertorio.periodici.diritto.amministrativo
/repertorio/dircanon.htm	catalogo.repertorio.periodici.diritto.canonico
/repertorio/dircivil.htm	catalogo.repertorio.periodici.diritto.civile
/repertorio/dircomeu.htm	catalogo.repertorio.periodici.diritto.comunita.europee
/repertorio/dircomme.htm	catalogo.repertorio.periodici.diritto.commerciale
/repertorio/dircompa.htm	catalogo.repertorio.periodici.diritto.comparato
/repertorio/dircosti.htm	catalogo.repertorio.periodici.diritto.costituzionale
/repertorio/direccle.htm	catalogo.repertorio.periodici.diritto.ecclesiastico
/repertorio/dirfallim.htm	catalogo.repertorio.periodici.diritto.fallimentare
/repertorio/dirindus.htm	catalogo.repertorio.periodici.diritto.industriale
/repertorio/dirinter.htm	catalogo.repertorio.periodici.diritto.internazionale
/repertorio/dirintpr.htm	catalogo.repertorio.periodici.diritto.internazionale.privato
/repertorio/dirlavor.htm	catalogo.repertorio.periodici.diritto.lavoro
/repertorio/dirpenal.htm	catalogo.repertorio.periodici.diritto.penale
/repertorio/dirpriva.htm	catalogo.repertorio.periodici.diritto.privato
/repertorio/dirproci.htm	catalogo.repertorio.periodici.diritto.procedura.civile
/repertorio/dirprope.htm	catalogo.repertorio.periodici.diritto.procedura.penale
/repertorio/dirpubbl.htm	catalogo.repertorio.periodici.diritto.pubblico
/repertorio/dirregio.htm	catalogo.repertorio.periodici.diritto.regionale
/repertorio/dirroman.htm	catalogo.repertorio.periodici.diritto.romano
/repertorio/dirsanit.htm	catalogo.repertorio.periodici.diritto.sanitario
/repertorio/dirsinda.htm	catalogo.repertorio.periodici.diritto.sindacale
/repertorio/dirsport.htm	catalogo.repertorio.periodici.diritto.sportivo
/repertorio/dirtribu.htm	catalogo.repertorio.periodici.diritto.tributario
/repertorio/dirurban.htm	catalogo.repertorio.periodici.diritto.urbanistico
/repertorio/filsocdi.htm	catalogo.repertorio.periodici.filosofia.diritto
/repertorio/medlegal.htm	catalogo.repertorio.periodici.medicina.legale
/repertorio/repertorio.htm	catalogo.repertorio.periodici.indice
/repertorio/sciecono.htm	catalogo.repertorio.periodici.scienze.economiche
/repertorio/scipolit.htm	catalogo.repertorio.periodici.scienze.politiche
/repertorio/scisocia.htm	catalogo.repertorio.periodici.scienze.sociali
/repertorio/stodirme.htm	catalogo.repertorio.periodici.storia.diritto
/rer.htm	organizzazione.risorse.elettroniche
/sci.htm	organizzazione.scienze
/scifinder/manualewin.htm	informazioni.avvisi.software.scifinder
/scmat.pdf	scienze.materiali.orphan.to.delete.before.analysis
/sermed.htm	informazioni.medicina.servizi
/sersci.htm	informazioni.scienze.servizi
/serse.htm	documenti.lavoro.progetto.serse
/stat.htm	statistiche.indice
/stat00.htm	statistiche.generali.2000
/stat01a.pdf	statistiche.generali.2001
/stat01b.pdf	statistiche.per.area.disciplinare.2001

Pagina	Concetto
/stat01c.pdf	statistiche.per.sede.2001
/stat01d.pdf	statistiche.legenda.2001
/stat02a.pdf	statistiche.generali.2002
/stat02b.pdf	statistiche.per.area.disciplinare.2002
/stat02c.pdf	statistiche.per.sede.2002
/stat02d.pdf	statistiche.legenda.2002
/stat03a.pdf	statistiche.generali.2003
/stat03b.pdf	statistiche.per.area.disciplinare.2003
/stat03c.pdf	statistiche.per.sede.2003
/stat03d.pdf	statistiche.legenda.2003
/stat98.htm	statistiche.generali.1998
/stat99.htm	statistiche.generali.1999
/statdiv00.htm	statistiche.per.area.disciplinare.2000
/statdiv98.htm	statistiche.per.area.disciplinare.1998
/statdiv99.htm	statistiche.per.area.disciplinare.1999
/statsolo98.htm	statistiche.dati.solo.1998
/statsolo99.htm	statistiche.dati.solo.1999
/stone/1936-1959.htm	catalogo.stone.anni.1936.1959
/stone/1960-1969.htm	catalogo.stone.anni.1960.1969
/stone/1970-1993.htm	catalogo.stone.anni.1970.1993
/stone/stone.htm	catalogo.stone.indice
/topo00.htm	documenti.lavoro.topografico.controllo.2000
/ultimora.htm	home.to.delete.before.analysis
/whoweare.htm	informazioni.whoweare
/xcorsi.htm	internet.percorsi
/yourlink.htm	internet.link.utenti

Allegato 2 – Sorgente Java

```
/* ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø
   ø  by Gianluca Tavella - gianluca.tavella@libero.it                     ø
   ø  Lug 2004                                                             ø
   ø  Tesi di laurea - Web Usage Mining                                    ø
   ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø ø

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~
 * Classe reOrganizeTable:    riorganizza la tabella 'seq' del database
 *                      'sessions' per la successiva analisi cluster
 *

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~
 */
/*
 * import dei package necessari
 */
import java.sql.*;
/*
 * CLASSE: reOrganizeTable
 *        utility per la creazione di una matrice organizzata ai fini
 *        della successiva analisi cluster tramite software statistico
 */
class reOrganizeTable{

   //////////////////////////////////////
   //// Variabili e oggetti di classe ///
   //////////////////////////////////////
    public String[] Concetti;
    public int MinVisitorID;
    public String normalbinary;
    public String visitorsession;

    /*
     *  Costruttore della classe
     *   In base ai parametri passati a riga di comando stabilisce
     *  quale metodo utilizzare per costruire la matrice.
     *  La prima stringa stabilisce il metodo:
     *  "normal" = crea la matrice con il numero di visite
     *          di ogni utente per ogni pagina;
     *  "binary" = crea la matrice con i valori 0 e 1 (1 se
     *          l'utente ha visitato almeno una pagina,
     *          0 altrimenti).
     *   La seconda stringa stabilisce se utilizzare come
     *  osservazioni:
     *   "visitor" = gli utenti;
     *  "session" = le sessioni.
     */
    public reOrganizeTable(String met, String vs){
       normalbinary = met;
       visitorsession = vs;
       if ((!normalbinary.equals("normal") && !normalbinary.equals("binary"))
          || (!visitorsession.equals("visitor")
```

```
          && !visitorsession.equals("session"))){

         System.out.println("I parametri inseriti non sono corretti.");
         System.out.println("La prima stringa stabilisce il metodo:");
         System.out.println("\"normal\" = crea la matrice con il numero di
visite");
         System.out.println("di ogni utente per ogni pagina;");
         System.out.println("\"binary\" = crea la matrice con i valori 0 e
1");
         System.out.println("(1 se l'utente ha visitato almeno una pagina, 0
altrimenti).");
         System.out.println("La seconda stringa stabilisce se utilizzare
come");
         System.out.println("osservazioni:");
         System.out.println("\"visitor\" = gli utenti;");
         System.out.println("\"session\" = le sessioni.");
         System.out.println("I parametri vanno inseriti con caratteri
minuscoli ");
         System.out.println("omettendo le virgolette (\").");
          System.exit(0);
       }



    }


    /*
     *   Metodo DistinctConcepts()
     *   Estrae i diversi concetti legati alle pagine dalla query
'UrlDistinti'
     *   Fa ritornare una matrice di oggetti di tipo String
     */
    public String[] DistinctConcepts(){
      try{
         /*
          * registrazione del driver jdbc (bridge jdbc-odbc)
          */
         Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");  //registra driver
         /*
          * creazione della connessione al database "sessions"
          */
         Connection connessione = DriverManager.getConnection
                              ("jdbc:odbc:sessions");
         /*
          * creazione del comando
          */
         Statement comando = connessione.createStatement
            (ResultSet.TYPE_SCROLL_INSENSITIVE,ResultSet.CONCUR_READ_ONLY);

         ResultSet risultatoQuery = comando.executeQuery
            ("SELECT * FROM UrlDistinti;");
         int i = 0;
         while (risultatoQuery.next()) {
            i++;
         }
         System.out.println("   -> Numero di diversi concetti:" + i);
         /*
          * se il ResultSet è vuoto restituisce una matrice con una sola
          * stringa vuota e si interrompe l'esecuzione del metodo,
          * altrimenti l'esecuzione del metodo prosegue
          */
         if (i == 0){
```

```
            String[] urldistinti = new String[1];
            urldistinti[0] = "";
            return urldistinti;
          }
          /*
           * posizionamento alla prima riga del ResultSet
           */
          risultatoQuery.first();
          /*
           * creazione dell'array di stringhe in base alla
           * dimensione del ResultSet
           */
          String[] URLdistinti = new String[i];
          /*
           * popolamento dell'array
           */
          for (int j = 0; j < URLdistinti.length; j++){
             URLdistinti[j] = risultatoQuery.getString(1);
             risultatoQuery.next();
          }
          return URLdistinti;
       /*
        * gestione esplicita delle eccezioni
        * nessuna azione correttiva intrapresa di fronte ad
        * eventuali eccezioni
        * visualizzazione di un messaggio
        */
       }catch (Exception e){
          System.out.println("Problemi di connessione al database,
riprovare.");
          System.out.println(e.getMessage());
          String[] urldistinti = new String[1];
          urldistinti[0] = "";
          return urldistinti;
       }
     }
     /*
      *   Metodo IstruzioneDiCreazioneTabella()
      *   crea la stringa con l'istruzione SQL per la creazione
      *   della tabella in base ai concetti relativi alle pagine
      */
      public String IstruzioneDiCreazioneTabella(){
        Concetti = DistinctConcepts();
        String istruzione = "CREATE TABLE MatrixCluster (" + visitorsession
+"ID INTEGER, ";
        for (int i=0; i < Concetti.length-1; i++){
           istruzione = istruzione + Concetti[i].replace('.','_') + " INTEGER,
";
        }
        istruzione = istruzione +
              Concetti[Concetti.length-1].replace('.','_') +" INTEGER);";
         return istruzione;
     }

    /*
     *   Metodo getFieldNumber
     *   fa ritornare il numero del campo relativo al concetto
     *  passato al metodo
     */
     public int getFieldNumber(String Concetto){
        for (int i=0; i < Concetti.length; i++){
           if (Concetto.equalsIgnoreCase(Concetti[i]))
              return i;
```

```
      }
      return 100000000;
   }
   /*
    *   Metodo CreateMatrix
    *   crea la matrice dei dati per l'analisi
    *   esegue le istruzioni SQL per la creazione della tabella
    */
   public void CreateMatrix(){
      try{
        /*
         * registrazione del driver jdbc (bridge jdbc-odbc)
         */
        Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");  //registra driver
        /*
         * creazione della connessione al database "sessions"
         */
        Connection connessione =
DriverManager.getConnection("jdbc:odbc:sessions");
        /*
         * creazione del comando
         */
        Statement comando = connessione.createStatement
           (ResultSet.TYPE_SCROLL_INSENSITIVE, ResultSet.CONCUR_READ_ONLY);

        ResultSet risultatoQuery;
        String vs;
        if (visitorsession.equals("visitor")){
               risultatoQuery   = comando.executeQuery
                 ("SELECT Max(Visitor), Min(Visitor) FROM
SequencesVisitor;");
              vs = "visitatori";
        }else{
              risultatoQuery = comando.executeQuery
                 ("SELECT Max(session), Min(session) FROM
SequencesSessions;");
              vs = "sessioni";
        }
        risultatoQuery.first();

        int MaxVisitorID = risultatoQuery.getInt(1);
        MinVisitorID = risultatoQuery.getInt(2);
        int NumberOfVisitors = MaxVisitorID - MinVisitorID;

        System.out.println("   -> Numero di " + vs + ": " +
NumberOfVisitors);

        int ColonneMatrice = Concetti.length;

        int[][] DataMatrix = new int[NumberOfVisitors+1][ColonneMatrice];

        System.out.println("Inizializzazione della matrice...");

        for (int i=0;i < NumberOfVisitors+1; i++){
           for (int j=1; j < ColonneMatrice; j++){
              DataMatrix[i][j] = 0;
           }
        }
        if (visitorsession.equals("visitor")){
           risultatoQuery = comando.executeQuery("SELECT * FROM
SequencesVisitor;");
        }else{
```

```
            risultatoQuery = comando.executeQuery("SELECT * FROM
SequencesSessions;");
         }
         System.out.println("Conteggio del numero di records...");

         int numeroRecords = 0;
         while (risultatoQuery.next()){
            numeroRecords++;
         }
         System.out.println("   -> Numero di records: " + numeroRecords);
         System.out.println("Inserimento dei dati nella matrice...");
         if (normalbinary.equals("normal")){
            risultatoQuery.beforeFirst();
            int VisitorID;
            String ConceptURL;
            int Occurrence;
            while (risultatoQuery.next()){

               VisitorID = risultatoQuery.getInt(1);
               ConceptURL = risultatoQuery.getString(3);
               Occurrence = risultatoQuery.getInt(2);
               DataMatrix   [VisitorID - MinVisitorID]
                        [getFieldNumber(ConceptURL)] = Occurrence;
            }

         }else{
            risultatoQuery.beforeFirst();
            int VisitorID;
            String ConceptURL;
            int Occurence;
            while (risultatoQuery.next()){

               VisitorID = risultatoQuery.getInt(1);
               ConceptURL = risultatoQuery.getString(3);
               Occurence = risultatoQuery.getInt(2);
               if (Occurence != 0){
                  Occurence = 1;
               }
               DataMatrix   [VisitorID - MinVisitorID]
                        [getFieldNumber(ConceptURL)] = Occurence;
            }
         }

         String addInTable;

         System.out.println("Inserimento dei dati nella tabella...");

         for (int i=0; i < NumberOfVisitors+1; i++){
            addInTable = "INSERT INTO MatrixCluster VALUES(" +
(i+MinVisitorID);
               for (int j=0; j < ColonneMatrice; j++){
                  addInTable = addInTable + ", " + DataMatrix[i][j];
               }
            addInTable = addInTable + ");";
            comando.executeUpdate(addInTable);
         }

         System.out.println("Processo riuscito!");
         System.exit(0);

      /*
       * gestione esplicita delle eccezioni
       * nessuna azione correttiva intrapresa di fronte ad eventuali eccezioni
```

```
       * visualizzazione di un messaggio
       */
      }catch (Exception e){
         System.out.println("Problemi di connessione al database,
riprovare.");
         System.out.println(e.getMessage());
      }
   }

   /*
    *   Metodo ReOrganizer()
    *   stabilisce la prima connessione al database e richiama gli altri
metodi
    */
    public void ReOrganizer(){
      try{

         /*
          * registrazione del driver jdbc (bridge jdbc-odbc)
          */
         Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");  //registra driver
         /*
          * creazione della connessione al database "sessions"
          */
         Connection connessione =
DriverManager.getConnection("jdbc:odbc:sessions");
         /*
          * creazione del comando
          */
         Statement comando = connessione.createStatement();

         deleteMatrixCluster();

         System.out.println("Creazione della struttura...");
         comando.executeUpdate(IstruzioneDiCreazioneTabella());

         System.out.println("Creazione della matrice...");

         CreateMatrix();
      /*
       * gestione esplicita delle eccezioni
       * nessuna azione correttiva intrapresa di fronte ad eventuali eccezioni
       * visualizzazione di un messaggio
       */
      }catch (Exception e){
         System.out.println("Problemi di connessione al database,
riprovare.");
         System.out.println(e.getMessage());
      }

   }

   /*
    *   Metodo deleteMatrixCluster
    *   eventuale eliminazione della tabella 'MatrixCluster' già esistente
    *   nel database
    */
   public void deleteMatrixCluster(){
      try{


         /*
          * registrazione del driver jdbc (bridge jdbc-odbc)
```

```
          */
         Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");  //registra driver
         /*
          * creazione della connessione al database "sessions"
          */
         Connection connessione =
DriverManager.getConnection("jdbc:odbc:sessions");
         /*
          * creazione del comando
          */
         Statement comando = connessione.createStatement();

         comando.executeUpdate("DROP TABLE MatrixCluster;"); //da eliminare

      /*
       * gestione esplicita delle eccezioni
       * nessuna azione correttiva intrapresa di fronte ad eventuali eccezioni
       */
      }catch (Exception e){
      }
   }

   /*
    *   Inizio dell'applicazione
    *   Crea un nuovo oggetto reOrganizeTable ed esegue il metodo ReOrganizer
    */
    public static void main(String args[]){
      try{
         reOrganizeTable rot = new reOrganizeTable(args[0],args[1]);
         System.out.println("Riorganizzazione della tabella per
clustering...");
         rot.ReOrganizer();
      }catch (Exception e){
         System.out.println("Inserire i 2 parametri.");
         System.out.println("La prima stringa stabilisce il metodo:");
         System.out.println("\"normal\" = crea la matrice con il numero di
visite");
         System.out.println("di ogni utente per ogni pagina;");
         System.out.println("\"binary\" = crea la matrice con i valori 0 e
1");
         System.out.println("(1 se l'utente ha visitato almeno una pagina, 0
altrimenti).");
         System.out.println("La seconda stringa stabilisce se utilizzare
come");
         System.out.println("osservazioni:");
         System.out.println("\"visitor\" = gli utenti;");
         System.out.println("\"session\" = le sessioni.");
         System.out.println("I parametri vanno inseriti con caratteri
minuscoli ");
         System.out.println("omettendo le virgolette (\").");
         System.exit(0);
      }
   }
}
```

Glossario

Action (pagina): pagina la cui richiesta indica che l'utente sta perseguendo lo scopo del sito.

Active investigator: utente che rimane a lungo sul sito e lo esplora. (*Cfr.* short-time visitor *e* cliente).

Alberi di classificazione: alberi decisionali (*cfr.*) con variabile risposta qualitativa.

Alberi di regressione: alberi decisionali (*cfr.*) con variabile risposta quantitativa.

Albero decisionale: rappresentazione grafica costruita suddividendo ripetutamente i dati secondo sottogruppi definiti dai valori delle variabili di risposta, per trovare sottoinsiemi omogenei. Tale suddivisione produce una gerarchia ad albero, dove i sottoinsiemi intermedi vengono chiamati *nodi* e quelli finali vengono chiamati *foglie* (*cfr.* analisi di segmentazione).

Analisi di segmentazione: analisi che attua un raggruppamento delle unità statistiche assumendo che fra le variabili a disposizione ve ne sia una (variabile risposta) che si possa considerare come dipendente dalle altre (variabili esplicative). L'obiettivo dell'analisi di segmentazione è la classificazione delle unità statistiche in gruppi fra loro omogenei, con riferimento alle modalità della variabile risposta. L'output dell'analisi è solitamente rappresentato mediante una struttura ad albero, detta albero decisionale (*cfr.*).

Analisi esplorativa dei dati: *vedi* analisi preliminare.

Analisi preliminare o esplorativa dei dati: elaborazione delle informazioni a disposizione al fine di descrivere in modo sintetico l'insieme dei dati a disposizione tramite rappresentazioni grafiche o indicatori statistici.

Association rules: *vedi* regole associative e sequenze.

BP: acronimo di *Error Back Propagation* (*cfr.*).

Caching: processo di memorizzazione locale delle pagine richieste con più frequenza dagli utenti.

CART: algoritmo ricorsivo impiegato nella analisi di segmentazione (*cfr.*). Acronimo di *Classification And Regression Trees*.

CHAID: algoritmo ricorsivo impiegato nella analisi di segmentazione (*cfr.*). Acronimo di *Chi-squared Automatic Interaction Detection.*

Clickstream analysis: analisi delle sequenze di visita ai siti web.

Cliente *(customer)*: utente che ha realizzato lo scopo del sito. (*Cfr.* active investigator *e* short-time visitor).

Cluster analysis: analisi di raggruppamento che si propone di mettere insieme le unità statistiche in gruppi il più possibile omogenei al loro interno (coesione interna) ed eterogenei tra di loro (separazione esterna). Fa parte dei metodi di classificazione non supervisionati (*cfr.*), può essere gerarchica (*cfr.* gerarchico) o non gerarchica (*cfr.* non gerarchico).

Concordanza: tendenza delle modalità (poco) elevate di una variabile ad associarsi a modalità (poco) elevate dell'altra (*cfr.* discordanza).

Confidenza (*confidence*): indice utilizzato nelle regole e sequenze associative (*cfr.*). Nella clickstream analysis (*cfr.*) l'indice di confidence per la regola $A \rightarrow B$ esprime la frequenza (e quindi, al limite, la probabilità) che in una sessione utente in cui è stata visualizzata la pagina A possa essere successivamente visualizzata la pagina B: $confidence(A \rightarrow B) = P(B \mid A)$. (*Cfr.* supporto).

Contact efficiency: *vedi* efficienza di contatto.

Conversion efficiency: *vedi* efficienza di conversione.

CRM: acronimo di *Customer Relationship Management (cfr.)*.

Customer Relationship Management: processo che coinvolge tutta la struttura aziendale e che ha come *focus* la conoscenza del cliente e del mercato, finalizzata ad una più sicura crescita della redditività aziendale.

Customer: *vedi* cliente.

Data cleansing (data cleaning): controllo di qualità dei dati disponibili ed (eventuale) pulizia preliminare dei dati.

Data mart: (database di marketing) database tematico, orientato all'attività di marketing che contiene dati di tipo descrittivo e di tipo comportamentale, utili per valutare attentamente i propri clienti, identificare esigenze e stili di comportamento, stabilire strategie commerciali differenziate.

Data mining: processo di selezione, esplorazione e modellazione di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori, con lo scopo di ottenere un risultato chiaro e utile al proprietario del database.

Data retrieval: attività consistente nell'estrazione da un archivio o da un database di una serie di dati, basandosi su criteri definiti a priori, in maniera esogena all'attività di estrazione stessa.

Data warehouse: raccolta di dati, orientata al soggetto, integrata, non volatile e variabile nel tempo, volta a supportare le decisioni del management.

Data webhouse: (1) data warehouse convenzionale fruibile attraverso il web, con interfacce utilizzabili da semplici browser; (2) data warehouse contenente i dati sul comportamento di coloro che interagiscono, attraverso i propri browser, con i siti Internet.

Database di marketing: *vedi* data mart.

Discordanza: tendenza delle modalità meno elevate di una delle due variabili ad associarsi a modalità elevate dell'altra (*cfr*. concordanza).

Efficienza di contatto (*contact efficiency*): percentuale di utenti che passano almeno una determinata quantità di tempo minimo esplorando il sito.

Efficienza di conversione (*conversion efficiency*): percentuale di utenti che, dopo aver esplorato il sito, hanno anche realizzato il suo scopo (es. comprando i prodotti).

Error Back Propagation: algoritmo di apprendimento utilizzato nelle reti multilayer perceptron (*cfr.*), dove la retro-propagazione dell'errore, calcolato dalla differenza tra l'uscita ed il target, viene trasmesso a tutti i neuroni della rete, permettendo di modificare i pesi sinaptici.

Gerarchico (metodo): metodo di classificazione che permette di ottenere una famiglia di partizioni, ciascuna associata ai successivi livelli di raggruppamento fra le unità statistiche. (*Cfr.* non gerarchico).

Impurità: concetto impiegato nell'analisi di segmentazione (*cfr.*) corrispondente al concetto di eterogeneità delle unità statistiche, con riferimento alle modalità della variabile risposta.

Knowledge Discovery in Databases: processo di estrazione della conoscenza da un database, dall'individuazione degli obiettivi di business iniziali fino all'applicazione delle regole decisionali trovate.

Kohonen: *vedi* reti di Kohonen.

Market basket analysis: analisi del carrello della spesa.

Matrice dei dati: rappresentazione dei dati in una forma tabellare, disegnata sulla base delle esigenze di analisi e degli obiettivi preposti.

Multilayer perceptron (percettrone multistrato): rete neurale ad apprendimento supervisionato di tipo feed-forward (dove i segnali si propagano esclusivamente nel

senso input-output), con più strati nascosti, uno di input ed uno di output, totalmente interconnessa. (*Cfr.* reti neurali).

Non gerarchico (metodo): metodo di classificazione che permette di ottenere una sola partizione delle *n* unità statistiche in *g* gruppi (con *g* generalmente minore di *n*) il cui numero viene definito a priori da colui che svolge la classificazione. (*Cfr.* gerarchico).

Non supervisionato: metodo che non si confronta con variabili di riferimento (target o risposta), opera quindi in funzione di tutte le variabili a disposizione. (*Cfr.* supervisionato).

OLAP: strumento, spesso di tipo grafico, che permette di visualizzare le relazioni tra le variabili a disposizione, seguendo la logica di analisi di un report a due dimensioni. Acronimo di *On Line Analitical Processing*.

Pagina action: *vedi* action (pagina).

Pagina target: *vedi* target (pagina).

Percettrone: primo modello di macchina per l'apprendimento automatico, da cui si svilupparono le reti neurali. (*Cfr.* reti neurali, multilayer perceptron).

Potatura (pruning): tecnica utilizzata nell'algoritmo CART (*cfr.*), relativa alla costruzione degli alberi decisionali (*cfr.*), attraverso la quale si costruisce dapprima l'albero di maggiori dimensioni, dove ogni nodo contiene solo un elemento oppure elementi appartenenti alla stessa classe. L'albero viene quindi "potato" secondo una regola che massimizza la capacità selettiva, a parità di complessità.

Prefetching: caratteristica del browser che permette ad una pagina HTML di recuperare altri contenuti web quando la connessione del browser dell'utente è inattiva. Il contenuto del *prefetching* viene immagazzinato nella *cache* del browser ed appare quindi velocemente non appena l'utente accede alla pagina che contiene il contenuto immagazzinato.

Regole associative e sequenze (*association and sequence rules*): tecniche esplorative spesso usate nella market basket analysis (*cfr.*) per misurare l'affinità di prodotti acquistati da un particolare consumatore; e nella clickstream analysis (*cfr.*) per misurare l'affinità delle pagine visitate da un utente di un sito.

Reti di Kohonen: tipi di reti neurali che permettono di classificare oggetti senza alcun tipo di supervisione e nascono dallo studio della topologia della corteccia del cervello umano. Sono denominate anche *SOM* (*Self Organizing Maps*).

Reti neurali: classe di modelli sviluppati nell'ambito delle scienze cognitive. Riescono a risolvere complessi problemi di classificazione e previsione grazie ad un processo di apprendimento in cui "imparano" la forma dei dati modificando i propri parametri interni.

Ricorsivo: procedimento che consiste nella applicazione ripetuta di una serie di operazioni, usando ogni volta come base di partenza il risultato dell'esecuzione precedente.

Sequence rules: *vedi* regole associative e sequenze.

Short-time visitor: utente che raggiunge il sito ma lo abbandona presto senza esplorarlo. (*Cfr.* active investigator *e* cliente).

SOM: acronimo di *Self Organizing Maps.* (*Cfr.* reti di Kohonen).

Supervisionato: metodo che si confronta con la presenza di una variabile di riferimento (target o risposta), le cui modalità sono note. (*Cfr.* non supervisionato).

Supporto (*support*): indice utilizzato nelle regole e sequenze associative (*cfr.*). Nella clickstream analysis (*cfr.*) il supporto per la regola $A \rightarrow B$ esprime la frequenza (e quindi, al limite, la probabilità) che una sessione utente contenga le due pagine, in sequenza: $support(A \rightarrow B) = P(A \cap B)$. (*Cfr.* confidenza).

Target (pagina): pagina la cui richiesta indica che l'utente ha realizzato lo scopo del sito.

Unità statistiche: elementi del collettivo di interesse ai fini dell'analisi (*cfr.* variabili statistiche)

Variabili continue: *vedi* variabili quantitative.

Variabili discrete: *vedi* variabili quantitative.

Variabili nominali: *vedi* variabili qualitative.

Variabili ordinali: *vedi* variabili qualitative.

Variabili qualitative: variabili relative a dati espressi in forma di aggettivo verbale. Danno origine a classificazioni in categorie e si possono distinguere in qualitative ordinali e qualitative nominali a seconda che sia possibile stabilire un ordinamento o meno tra le varie modalità.

Variabili quantitative: variabili legate a quantità intrinsecamente numeriche. Si distinguono in quantitative discrete, quando assumono un numero finito di valori, e quantitative continue, quando assumono un'infinità numerabile di valori.

Variabili statistiche: insieme delle caratteristiche di interesse per l'analisi, misurate su ciascuna unità statistica (*cfr.* unità statistiche).

Web Content Mining: area del Web Mining (*cfr.*) che si concentra sulle informazioni grezze disponibili nelle pagine web; la fonte dei dati consiste principalmente nei dati testuali delle pagine web e le tipiche applicazioni sono la classificazione e l'ordinamento delle pagine in base al contenuto.

Web Mining: area del data mining che si occupa dell'estrazione di conoscenza dal World Wide Web.

Web Structure Mining: area del Web Mining (*cfr.*) che si focalizza sulla struttura del sito; la fonte dei dati consiste principalmente nell'informazione sulla struttura delle pagine web (es. collegamenti alle altre pagine). Le tipiche applicazioni sono la

classificazione delle pagine web in base ai collegamenti e l'ordinamento delle pagine web attraverso una combinazione di contenuto e struttura.

Web Usage Mining: area del Web Mining che si occupa dell'estrazione di conoscenza dai *log file* del web server; La fonte dei dati consiste nei *log* (testuali) rappresentati in formati standard che vengono raccolti quando gli utenti accedono ai web server. Le tipiche applicazioni sono basate sulle tecniche per modellare gli utenti, come la personalizzazione del web ed i siti web adattivi.

Riferimenti ipertestuali

Web Usage Mining, sito realizzato per questa tesi di laurea. Contiene i risultati delle analisi, una guida al linguaggio MINT, la classe Java realizzata per la creazione della matrice dei dati, la struttura dei database impiegati, le varie query SQL, il glossario, la bibliografia, etc. **http://spazioinwind.libero.it/polar/wum**

Biblioteca di Ateneo, sito internet della Biblioteca di Ateneo dell'Università degli Studi di Milano – Bicocca. I log file del sito sono stati utilizzati come fonte principale della ricerca, descritta nella parte pratica (Capitolo 4). **http://www.biblio.unimib.it**

ActiveState, linguaggio di programmazione Perl. Per utilizzare il software WUMprep è necessario installare sul proprio PC la distribuzione ActivePerl. **http://www.activestate.com/**

Graphviz, software open source per disegnare diversi tipi di grafi, realizzato dai laboratori di ricerca dell' AT & T. **http://www.research.att.com/sw/tools/graphviz/**

Java Technology, sito di riferimento per il linguaggio di programmazione Java. Per utilizzare il software WUM è necessario scaricare il Java Runtime Environment (JRE) da questo sito ed installarlo sul proprio PC. **http://java.sun.com/**

KDnuggets, sito di riferimento per il Data Mining, il Web Mining ed il Knowledge Discovery. Contiene guide, software disponibili, pubblicazioni, news, etc. **http://www.kdnuggets.com**

WUM e **WUMprep**, software per il Web Usage Mining sviluppato nell'ambito del progetto Open Source Knowledge Discovery and Knowledge Management. **http://www.hypknowsys.org/**

World Wide Web Consortium (W3C), consorzio che si occupa della creazione degli standard Web. Il suo scopo è portare il Web al suo massimo potenziale, mediante lo

sviluppo di tecnologie (specifiche, linee guida, software e *tools*) che possano creare un forum per informazioni, commercio, ispirazioni, pensiero indipendente e comprensione collettiva. **http://www.w3.org/**

Università degli Studi di Milano – Bicocca: sito dell'Università degli Studi di Milano – Bicocca: **http://www.unimib.it**

Bibliografia

Berendt B. 2002. Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*. 6: 37–59.

Berry M., Linoff G. 1997. *Data Mining techniques for marketing, sales, and customer support*. New York. Wiley.

Berry M., Linoff G. 2001. *Data mining*. Milano. Apogeo.

Berthon P., Pitt L.F., Watson R.T. 1996. The world wide web as an advertising medium. *Journal of Advertising Research*, 36: 43–54.

Bishop C. 1995. *Neural networks for pattern recognition*. Oxford. Clarendon Press.

Breiman L., Friedman J.H., Olshen R., Stone C.J. 1984. *Classification and regression trees*. Belmont. Wadsworth.

Camillo F., Tassinari G. 2002. *Data mining, web mining e CRM: metodologie, soluzioni e prospettive*. Milano. FrancoAngeli.

Chang G. 2001. *Mining the world wide web : an information search approach.* Boston Kluwer Academic.

Del Ciello N., Dulli S., Saccardi A. 2000. *Metodi di Data Mining per il Customer Relationship Management.* Milano. FrancoAngeli.

Eirinaki M., Vazirgiannis M. 2003. Web mining for web personalization. *ACM Transactions on Internet Technology*, 1: 1-27.

Facca F.M., Lanzi P.L. 2003. Recent Developments in Web Usage Mining Research. *Data Warehousing and Knowledge Discovery: 5th International Conference, DaWaK*

2003 Prague, Czech Republic, September 3-5 in *Lecture Notes in Computer Science*, 2737: 140-150.

Giudici P. 2001. *Data Mining. Metodi statistici per le applicazioni aziendali*. Milano. McGraw-Hill.

Ingrassia S., Silani S. 2000. *Reti neuronali per l'analisi di dati complessi*. Milano. SAS.

Kass G.V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*. 29: 119-127.

Mena J. 1999. *Data mining your website*. Boston. Digital press.

Punin J. R., Krishnamoorthy M. S., Zaki M. J. 2002. LOGML: Log Markup Language for Web Usage Mining. *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points* in *Lecture Notes in Computer Science*, 2356: 88-112.

Schafer J. Ben, Konstan Joseph A., Riedl John. 2001. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*. 5: 115–153.

Spiliopoulou M. 2000. Web usage mining for Web site evaluation, *Communications of the ACM,* 8: 127-134.

Spiliopoulou M., Pohle C. 2001. Data Mining for Measuring and Improving the Success of Web Sites. *Data Mining and Knowledge Discovery*, 5: 85–114.

Srivastava J., Cooley R., Deshpande M., Tan P. 2000. Web usage mining: discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations*, 2: 12-23.

Thuraisingham B. 1999. *Data mining : technologies, techniques, tools and trends.* Boca Raton (Florida). CRC Press.

Xie Yunjuan, Phoha Vir V.. 2001. Web user clustering from access log using belief function. *K-CAP 2001 - Proceedings of the First International Conference on Knowledge Capture.* October 22-23: 202–208.

Zani S. 2000. *Analisi dei dati statistici*, volume 2: *Osservazioni multidimensionali*. Milano. Giuffrè.