Università degli Studi di Napoli
Federico II

Web Usage e Web Structure Mining: contributi per l'integrazione e la visualizzazione

Marcello Pecoraro

Tesi di Dottorato di Ricerca in
Statistica Computazionale ed Applicazioni
XXI Ciclo

A mio padre

Indice

Indice delle figure

Indice delle tabelle

Introduzione

Già da una quindicina d'anni Internet è universalmente considerato un canale di comunicazione "completo" e non più uno strumento di nicchia, un mezzo per comunicazioni o applicazioni appannaggio di una ristretta cerchia di esperti.

A cogliere per primi questa evidenza sono stati, come sempre accade, i principali attori del sistema economico globale unitamente a piccole e dinamiche aziende hi-tech divenute, grazie a tale sensibilità, i pilastri portanti della cosiddetta net-economy. La realtà della Rete è così divenuta enorme e complessa sia per effetto dei colossali investimenti posti in essere da questi attori sia grazie a "liberi pensatori", menti rivoluzionarie e geniali che hanno visto in Internet un'espressione di libertà e di progresso socio-culturale (si pensi su tutti a Linus Torvald, padre di Linux, o più in generale ai promotori della filosofia Open Source), che hanno reso tale mezzo indispensabile anche nella vita quotidiana.

Le enormi dimensioni del fenomeno, la sua complessità e, non ultimo l'interesse economico che da ciò deriva hanno reso ben presto concreta l'esigenza di analizzare da un punto di vista statistico (o più correttamente di Data Mining) le caratteristiche del fenomeno medesimo, per gli scopi più svariati. Incremento del business, miglioramento del servizio, estensione

della platea di riferimento sono solo alcuni dei macro-obiettivi di analisi che ben presto si sono affermati.

Allo statistico "moderno", oltre agli obiettivi concreti appena citati, Internet si offriva e si offre tuttora come un campo di studio di dimensioni vastissime. La navigazione in Rete è infatti un'attività di natura assai varia, proprio per la molteplicità di bisogni a cui essa risponde. Si può accedere a Internet per studio, per lavoro, per svago, per socializzare, per comunicare, per acquistare un bene o un servizio e per numerose altre ragioni. Lo statistico, dunque, può essere interessato alla Rete non solo in quanto "nuovo fenomeno" che caratterizza l'attività umana, ma anche come veicolo di analisi e comprensione di una serie di dinamiche assai più antiche e consolidate. Si pensi al già citato bisogno di socialità, insito nella natura umana: esso ha ormai un luogo "virtuale" dove viene soddisfatto, e il boom dei social-network ne è la chiara testimonianza.

In più, oltre alle notevoli possibilità di spaziare nel campo o nel fenomeno da analizzare, al ricercatore statistico la Rete offre un altro importante vantaggio: un'enorme quantità di dati a disposizione.

Come verrà chiarito in questo lavoro, infatti, il World Wide Web ha una architettura assai complessa alle spalle che per funzionare necessita di immagazzinare una gran mole di dati. L'accesso (unito alla successiva attività di Mining) a questi dati rappresenta una fonte di conoscenza che probabilmente non ha eguali.

Si analizzeranno innanzitutto i tre filoni di ricerca che compongono il Web Mining: Web Usage Mining, Web Structure Mining e Web Content Mining, con particolare riferimento ai primi due, che rappresentano il "perimetro" concettuale analizzato. Particolare attenzione è dedicata, oltre all'illustrazione delle differenze tra queste branche, anche all'ambito

applicativo e soprattutto al reperimento e trattamento dei dati, problema che, anche esulando da un contesto squisitamente statistico, condiziona fortemente il successivo momento di analisi.

In seguito, verranno introdotte le tecniche statistiche adottate correntemente nel processo di Mining dei dati Web.

Infine si arriverà all'illustrazione vera e propria del lavoro, ispirato da un duplice scopo: il primo è quello di fornire un contributo all'integrazione (già caldeggiata da alcuni studiosi) tra Web Usage e Web Structure, anche perchè le categorie sono già concettualmente vicine. Il secondo obiettivo è quello di proporre un approccio alternativo alla visualizzazione del processo Web, combinando alcuni strumenti già noti in letteratura. A quest'ultimo principio si ispira anche la seconda strategia di analisi proposta, che però ha scopi differenti dalla prima, in quanto è proposto come strumento descrittivo di ausilio per una specifica valutazione di performance di un sito Web, la durata di visita e la descrizione della sua "sopravvivenza".

Capitolo I

Lo studio dei dati provenienti dalla Rete Internet: il Web Mining

1.1 INTERNET: LA NUOVA MINIERA DI DATI

L'introduzione e la diffusione di Internet hanno profondamente influenzato i comportamenti sociali. Questo nuovo canale di comunicazione, commercio e socializzazione è così vasto e complesso che, già a pochi anni dal suo avvento, si è manifestata la necessità di studiarne le caratteristiche. Dal punto di vista statistico, questo campo si è distinto sin da subito per la vastità di informazione a disposizione e per la sua eterogeneità. Tali peculiarità sono dovute al funzionamento stesso della Rete: l'architettura di Internet prevede infatti l'acquisizione di una grande quantità di dati, molto spesso "invisibili" agli occhi del navigatore. Questi dati sono, ad esempio, relativi

all'ammontare del traffico presso un determinato sito in una data finestra temporale, la provenienza delle richieste pervenute a un certo server e così via. Dal punto di vista statistico, l'approccio al problema è stato giocoforza quello del Data Mining, soluzione obbligata dovuta all'enorme quantità di dati potenzialmente processabili, in realtà nemmeno quantificabile con certezza. La vastità delle informazioni a disposizione provoca anche un ulteriore riflesso, che il ricercatore non può non tenere in considerazione: la necessità di focalizzare, sin da subito, la profondità del livello di analisi. Eventuali tentativi di "analizzare" Internet, o comunque di carpire il comportamento degli internauti, vengono infatti sempre ricondotti a una prospettiva più reale, a causa di una molteplicità di motivi. Innanzitutto, le capacità di immagazzinamento dei dati da parte dei comuni elaboratori non permette di ospitare nemmeno una piccola parte dei dati archiviati nei server, grandi e piccoli, che fanno funzionare la Rete nel pianeta. Inoltre, subentrerebbero problemi relativi alle capacità delle macchine odierne di processare una simile mole di dati. Laddove anche tutto ciò fosse possibile, andrebbero poi considerati i problemi di privacy e di autorizzazione all'uso di tali informazioni. Inoltre, le analisi di Web Mining di solito hanno obiettivi ben precisi, che molto raramente vanno al di là del singolo sito o portale Web. Nel presente lavoro, vengono presentate tecniche e contributi che, nella quasi totalità dei casi, fanno riferimento a quest'ultimo livello di analisi, ovvero il singolo sito Web.

1.2 IL WEB MINING

Il Web è costituito da un insieme vastissimo di files (per lo più pagine di testo) tra loro interrelati che risiedono su più server, macchine che hanno la funzione di fornire tali files quando un utente, attraverso il proprio browser, le richiede attraverso la digitazione di un indirizzo o la pressione di un click su un determinato link.

Il Web Mining può essere definito come il processo di Data Mining applicato ai dati provenienti dal Web. Si tratta di un campo di analisi particolarmente recente, dato che Internet si è affermato a fine anni '80 ed è divenuto uno strumento di uso comune solo a metà degli anni '90. Sono proprio di quegli anni i primi tentativi di comprensione e di analisi di questo fenomeno da un punto di vista statistico[1].

La dottrina, in virtù della vastità e dell'eterogeneità del campo di analisi, è solita distinguere, all'interno di tale ambito di ricerca, tre sotto-branche: Web Content Mining, Web Usage Mining, Web Structure Mining.

1.3 IL WEB CONTENT MINING

Il Web Content Mining è il processo di Web Mining volto all'analisi dei contenuti presenti sul Web. Gli oggetti di questo tipo di studio sono i vari elementi che compongono le pagine Web: il testo del documento, le immagini, i banner, i contenuti multimediali e quant'altro. Quando l'analisi si concentra in maniera spinta sugli elementi testuali della pagina, il Web

[1] O. Etzioni, "The World Wide Web: Quagmire or Gold Mine", in Communications of the ACM, 39(11):65-68, 1996

Content Mining presenta moltissime affinità con il Text Mining. Questo processo di estrazione di informazione utile parte, ulteriore particolarità di questo ambito, da un tipo di dati che di solito sono a disposizione di tutti gli utenti, poiché in linea di massima, tutte le pagine Web presenti in rete costituiscono fonte di analisi. Di solito i processi di Content Mining costituiscono i principali task dei motori di ricerca, che hanno l'obiettivo di catalogare e indicizzare i contenuti presenti su Internet. Le analisi vengono principalmente condotte mediante tecniche proprie da altre discipline quali l' Information Retrieval e il Natural Language Processing.

1.3.1 Il Web Content Mining: Il processo e le applicazioni

Un processo di Web Content Mining ha inizio con una fase di preprocessing dei dati, detta Content Preparation. In tale fase, si estrae innanzitutto il testo dai documenti HTML, dopodiché si "pulisce" il documento stesso, eliminando tutti i caratteri non necessari all'analisi. In seguito si calcolano alcune grandezze, quali la Collection Wide Word Frequencies (DF) e la Calculate per Document Term Frequencies (TF), ovvero le frequenze dei termini riferite al set di dati o al singolo documento. Attraverso il preprocessing si ottiene, per ogni documento preso in esame, un vettore sparso contenente i pesi di ogni termine. Il sistema più usato a questo proposito è il TFDIF, accompagnato solitamente da pesi aggiuntivi assegnati alle keywords o alle parole che compaiono nei titoli.

Una volta pronto il set di dati, si possono applicare le procedure tipiche del Data Mining, quali la classificazione, la clustering o l'analisi delle associazioni. La classificazione dei documenti avviene attraverso tecniche

supervisionate in cui prioritariamente si procede alla definizione delle categorie nelle quali verranno divise le pagine Web. L'identificazione di queste categorie avviene di solito al termine di una fase di training, condotta su un campione di apprendimento, nella quale i documenti sono solitamente catalogati a mano da un esperto.

La clusterizzazione dei documenti invece, si conduce in un'ottica non supervisionata. I documenti sono raggruppati/divisi in base a una misura di similarità/dissimilarità. A differenza della classificazione, non si conosce a priori quali saranno i gruppi in cui i documenti verranno divisi.

Esistono poi altri task tipici di un'analisi di Web Content, tra cui la Topic Identification e la Concept Hierarchy Creation. La prima è il risultato di una combinazione di clustering e classificazione volto all'assegnazione di un'etichetta (topic) ai nuovi documenti sulla base delle categorie precedentemente individuate. Inoltre, il processo abbisogna di periodici aggiornamenti allo scopo di identificare prontamente l'eventuale necessità di introdurre nuove categorie.

La seconda invece è volta alla comprensione delle gerarchie e delle relazioni tra le categorie individuate. La struttura relazionale tra i contenuti può essere piatta (flat), gerarchica (tree) o a rete (network). Inoltre, altri fattori che influenzano la catalogazione dei documenti sono il numero massimo di categorie a cui può appartenere un documento o la dimensione delle categorie stesse.

Altre applicazioni tipiche del Content Mining sono quelle volte a quantificare la rilevanza di un contenuto o di un documento in un dato contesto. I criteri di Relevance sono i seguenti:

1. Query Based
2. Document
3. User Based
4. Role/Task Based

Il primo metodo è quello più comunemente adottato nel campo dell'Information Retrieval. Sostanzialmente, esso associa le parole immesse nella query alle keywords dei documenti archiviati. Inoltre, questa tecnica possiede, tra gli altri vantaggi, la possibilità di arricchire il procedimento di ricerca attraverso l'adozione di parametri quali la popolarità (come il PageRank di Google) della pagina o la posizione dei termini cercati.

Il secondo criterio misura quanto il documento è rilevante in un contesto specifico. Tipicamente questo risultato viene visualizzato al termine di una query, nella quale la lista dei documenti è ordinata in base alla rilevanza.

La User Based Relevance invece, è una sorta di rilevanza personalizzata, tanto è vero che è il criterio più adottato dai sistemi di personalizzazione dei contenuti. Il procedimento parte dalla creazione di un profilo utente, le cui caratteristiche costituiscono la base per il confronto con gli altri documenti. Il processo non richiede l'adozione di query.

L'ultimo criterio rappresenta un'estensione di quello precedente. La differenza principale consiste nel fatto che le peculiarità dei profili possono anche essere delineate a seguito dell'attività di più utenti, non necessariamente di un singolo.

Riepilogando, le applicazioni di un processo di Web Content Mining sono riconducibili alle seguenti aree:

1. Identificazione dell'argomento di un documento
2. Catalogazione dei documenti
3. Identificazione delle similarità tra documenti che risiedono presso più server
4. Rilevanza dei contenuti

1.4 IL WEB STRUCTURE MINING

La seconda branca del Web Mining prende il nome di Web Structure Mining: con questo termine si è soliti indicare i processi di analisi volti a studiare la struttura di un sito Web o alcuni "frammenti" della Rete, dal punto di vista dell'organizzazione dei contenuti e, soprattutto, della struttura dei link (Hyperlink structure). L'informazione estratta serve in questo caso ad individuare difetti nella progettazione del sito, che ad esempio provocano zone difficilmente raggiungibili, o al contrario per identificare zone importanti del sito o della rete, che per interesse o per popolarità ricevono molti link e quindi, molte visite.

Il Web può essere visto come un'enorme rete in cui i nodi sono rappresentati dalle pagine, e gli hyperlink sono gli archi orientati che le collegano.

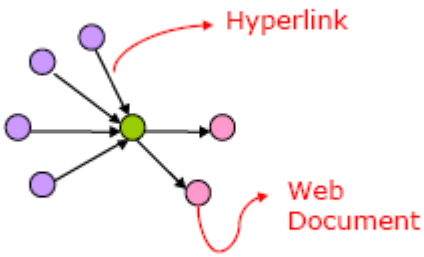

Figura 1.1 La rappresentazione classica di una struttura Web

Il concetto della Rete come grafo, è il fulcro del Web Structure Mining; per questo motivo, di seguito se ne riassume brevemente la terminologia:

- Web Graph: Un grafo orientato che rappresenta il Web o una sua porzione;
- Node: ogni pagina Web rappresenta un nodo della Rete;
- Link: è un arco orientato che collega due nodi di un Web Graph;
- In-degree: il numero dei nodi che puntano a un determinato nodo p;
- Out-degree: è il numero di link che partono dal nodo p verso gli altri nodi della Rete;
- Directed Path: una sequenza di link che dal nodo p consente di raggiungere il nodo q;
- Shortest Path: il percorso più breve che dal nodo p permette di raggiungere il nodo q;
- Diameter: considerando l'insieme di tutti i nodi del grafo considerato, il diameter è la massima distanza più breve (Shortest Path) tra questi nodi;

Un altro concetto fondamentale che riguarda lo studio delle strutture Web è senza dubbio la relazione tra le pagine/documenti. Efe e Raghavan (2000), [2] identificano alcune strutture Web tipiche:

- Endorsement: il collegamento diretto e unidirezionale tra un nodo (pagina o documento) e un altro;
- Transitive Endorsement: un endorsement indiretto, cioè nel quale vi è una o più pagine intermedie che collegano i due punti della Rete considerati;
- Mutual Reinforcement: il collegamento bidirezionale tra due nodi;
- Co-citation: il riferimento unidirezionale (collegamento) di un nodo a due o più nodi della Rete;
- Social Choice: i collegamenti (citazioni) in entrata che una pagina riceve da due o più altre pagine;

1.4.1 Il Web Structure Mining: il processo e le applicazioni

Un'analisi di Web Structure può essere condotta a due livelli:

1. Document Level (Intra page);
2. Hyperlink Level (Inter Page);

Il livello di analisi più interessante è il secondo, che viene svolto sia per incrementare l'efficacia di un processo di navigazione, sia per aumentare le

[2] Kemal Efe, Vijay Raghavan, C. Henry Chu, Adrienne L. Broadwater, Levent Bolelli, Seyda Ertekin (2000), The Shape of the Web and Its Implications for Searching the Web, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet- Proceedings at http://www.ssgrr.it/en/ssgrr2000/proceedings.htm, Rome. Italy, Jul.-Aug. 2000.

probabilità di accedere alle pagine cosiddette authority, pagine che incorporano informazioni interessanti relative al topic cercato.

Un'analisi di Hyperlink structure può essere condotta per vari motivi[3], ma ciascuno di essi rappresenta solitamente una fase preliminare alla progettazione dell'applicazione finale; di seguito sono riassunti tali scopi, secondo lo schema proposto da Desikan:

- Knowledge Models: sono modelli conoscitivi tesi innanzitutto a rappresentare la Rete oggetto di studio. Di solito questo task è preliminare a tutti i tipi di analisi;
- Analysis scope: è un'analisi volta a comprendere se il task che si sta per intraprendere è applicabile a un dato set di nodi;
- Properties: è la fase di descrizione delle caratteristiche di uno o più nodi;
- Measures and Algorithms: con le measures si definiscono delle grandezze che costituiscono gli standard per l'analisi di qualità, rilevanza e distanza tra nodi. Attraverso gli algoritmi si calcolano in maniera efficiente tali measures;

Come già accennato in precedenza, ciascuna di queste tecniche funge da analisi preliminare all'implementazione di un'applicazione vera e propria. Le applicazioni di Web Structure hanno diverse finalità, tra le quali rientrano:

1. Determinazione della qualità di una pagina: rientra in questa categoria il famoso PageRank di Google. Più in generale, appartengono a questa famiglia le applicazioni che restituiscono punteggi, ranking o

[3] P. Desikan, J. Srivastava, V. Kumar, P.-N. Tan, "Hyperlink Analysis – Techniques & Applications", Army High Performance Computing Center Technical Report, 2002.

indicazioni della pertinenza della pagina rispetto a un determinato topic;

2. Classificazione delle pagine: queste applicazioni determinano il contenuto di una pagina, identificano quelle tra loro correlate, stabiliscono quali pagine "marcare" nel processo di crawling;
3. Identificazione di strutture Web interessanti: rientrano in questa categoria, le applicazioni attraverso le quali si scoprono pagine duplicate, o strutture Web complesse come le sopracitate Social Choice;

1.4.2 Il PageRank di Google e il modello Hubs-Authorities

L'ormai celebre sistema di ranking del Web "PageRank" ideato da Google è sicuramente l'esempio più popolare di applicazione di un processo di Web Structure Mining per determinare la qualità di una pagina. Questa applicazione ha due scopi fondamentali: il primo è di natura "interna", nel senso che serve a Google stesso per determinare l'ordine di apparizione delle pagine Web nella lista dei risultati di una ricerca richiesta dall'utente al motore. Il secondo è di natura "esterna", nel senso che Google rende pubblico il ranking che il suo sistema calcola. Ciò ha una duplice funzione: serve agli utenti per conoscere un primo giudizio di importanza e popolarità della pagina che sta consultando, e serve come strumento di visibilità del sito o della pagina, che risultando importante per Google, acquisisce un ruolo di rilievo nel Web, con tutte le conseguenze che ne derivano, ad esempio in termini di ritorni pubblicitari.

Anche se Google non ha mai divulgato in maniera completa l'architettura del sistema PageRank, che tra l'altro è soggetto a continui e importanti aggiornamenti, la determinazione del ranking di una pagina avviene per effetto di due componenti principali. La prima è la pertinenza dei contenuti di una pagina ai topic indicati nei titoli e/o nelle dichiarazioni passate al motore di ricerca. Una seconda deriva invece dalla numerosità e dalla popolarità delle pagine che puntano alla pagina considerata.

Si fissa innanzitutto una misura di distanza, che serve a determinare il range entro il quale vengono considerate le pagine che puntano alla pagina da valutare. La popolarità di queste pagine limitrofe viene pesata e l'insieme di queste misure concorre alla determinazione del ranking della pagina considerata. Il principio è dunque che la popolarità di una pagina dipende da quanto lo sono le pagine che puntano verso di essa.

Il modello Hubs-Authorities proposto da Kleinberg fa riferimento a un particolare tipo di schema del Web, detto Bipartite Core. In questo schema figurano due tipi di nodi, gli Hubs e le Authorities. I primi sono, come il termine hub suggerisce, degli ottimi 'indirizzatori', nel senso che da quella pagina si riesce tramite links a raggiungere molte autorità, cioè molte pagine importanti relativamente ai topic considerati. Le authorities, come già parzialmente chiarificato, sono pagine molto interessanti rispetto a un argomento; in più si possono definire anche in senso inverso. In tal modo, un'authority è una pagina puntata da molti hubs e i due tipi di nodi si rinforzano a vicenda.

1.5 IL WEB USAGE MINING

Il Web Usage Mining è invece orientato alla comprensione delle dinamiche d'uso di uno o più siti Internet da parte dei navigatori. Appartengono a questo filone gli studi volti alla conoscenza delle abitudini di navigazione, facendo riferimento ad aspetti quali le pagine visualizzate con maggior frequenza, i percorsi di visita più ricorrenti (clickstream analysis), i tempi di connessione, le pagine che producono più entrate o più uscite dal sito. Un processo di Mining di questo tipo può essere espletato in osservanza di due obiettivi principali, spesso perfezionati sequenzialmente: il primo è denominato pattern discovery, ovvero il processo esplorativo che conduce alla scoperta dei pattern di navigazione più interessanti. Il secondo prende il nome di pattern analysis, ed è l'estensione naturale del primo in quanto è volto alla comprensione delle caratteristiche dei pattern scoperti mediante il precedente processo. Il tipico output di un'attività di Usage Mining è costituito da una segmentazione o una clusterizzazione degli utenti del sito, sulla base delle abitudini di navigazione e delle componenti socio-demografiche che sovente si riescono a reperire tramite i dati di registrazione o i form compilati per richieste specifiche.

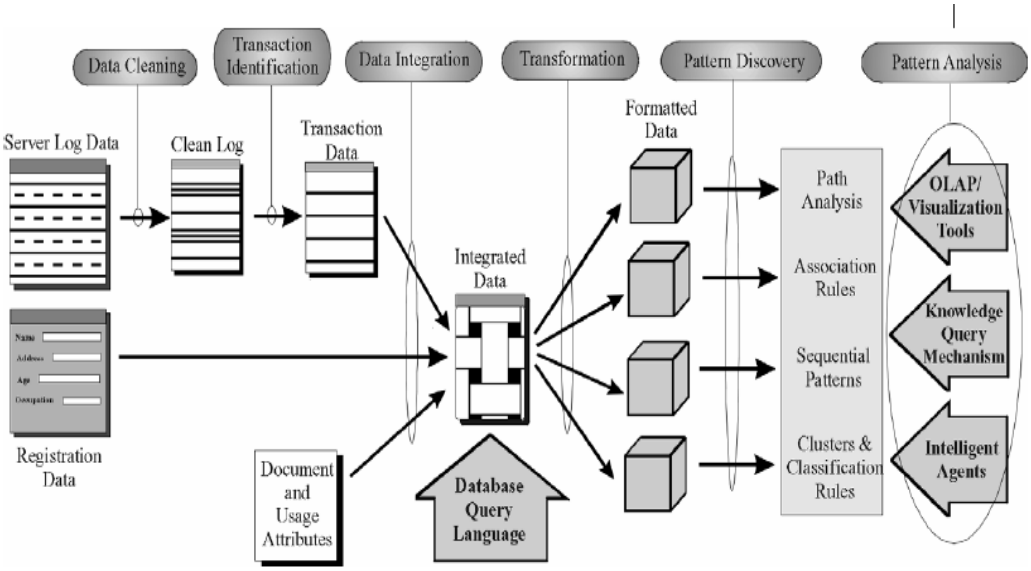

Figura 1.2 Architettura completa di un sistema di Web Usage Mining

1.6 LE FONTI DEI DATI DEL WEB MINING

Come accennato in precedenza, lo studio dei dati Web ai fini di estrapolare conoscenza utile rappresenta un campo di ricerca interessante anche per la vastità e l'eterogeneità dei dati a disposizione. Per condurre un'analisi in questo campo, è tuttavia necessario familiarizzare con la terminologia e con le "grandezze" che compongono il Web, poiché molte di queste costituiscono i dati di partenza per molti obiettivi di ricerca.

Allo scopo, si farà riferimento alla nomenclatura proposta dal World Wide Web Consortium (W3C):

- User (Utente): il singolo navigatore che durante la sua visita richiede una o più pagine Web ad un server attraverso il proprio browser;

- Page view (Pagina): la singola pagina Web visualizzata, comprensiva di tutti gli elementi che la compongono (testi, immagini, suoni, contenuti multimediali, ecc.).

- Click-stream (Sequenza di click): l'insieme di click sequenziali effettuati da un utente all'interno di una sessione di navigazione allo scopo di sfogliare una serie di pagine appartenenti allo stesso sito Web;

- User session (Sessione Utente): la sequenza di richieste di pagine effettuate dal singolo utente nella singola sessione, riferita però all'intero Web. E' facile osservare come il concetto di sessione utente è, almeno potenzialmente, assai più ampio di quello di clickstream;

- Server session (Sessione Server): la sequenza delle pagine richieste ad un singolo server;

- Episode: un sottoinsieme di user session o di server session idoneo ad identificare uno o più episodi significativi in una determinata finestra temporale;

Una volta esaminate, sia pur rapidamente, le principali grandezze del Web, occorre approfondire anche la natura di questi dati, le provenienze, le principali differenze e i problemi connessi all'immagazzinamento ai vari livelli.

Una prima, fondamentale distinzione, si può operare con riferimento ai livelli di collezionamento dei dati. Questi possono essere raccolti a tre livelli: il livello Server, il livello Client e il livello Proxy. A livello di Server, i dati sono memorizzati allo scopo di registrare tutta l'attività che riguarda il medesimo server. Di solito, l'output prodotto da tale attività di registrazione

è una serie di *log files*, file che contengono una serie di informazioni quali le pagine restituite, il tempo di connessione, gli IP dai quali provengono le richieste e così via. La completezza e la facilità di reperimento di questi dati costituiscono i due grandi vantaggi dei *log files*, che in più sono invisibili e non invasivi rispetto ai navigatori. I problemi sono da ricondurre principalmente al funzionamento dell'accoppiata Server-Browser quando si naviga in Rete. Infatti, per velocizzare il processo di consultazione di Internet, le pagine visitate di recente vengono memorizzate sul computer locale dell'utente, senza che il browser debba nuovamente richiederle al server. Tale processo prende il nome di *caching* e dal punto di vista della raccolta dei dati risulta penalizzante in quanto non rimane traccia della richiesta (dunque della consultazione) di tali pagine. Un problema analogo può manifestarsi quando l'utente sfrutta i tasti "Indetro" e "Avanti" del proprio browser. In alcuni casi, dipendenti dal browser stesso e dai suoi settaggi, può accadere che delle pagine raggiunte in questa maniera si perda traccia. Per ovviare ai suddetti fenomeni, sovente si fa uso dei cosiddetti *cookies*, piccole stringhe di testo che vengono memorizzate sul computer dell'utente, ne registrano l'attività e dialogano col browser fornendo tali dati. Se da un lato l'uso dei cookies può costituire un valido stratagemma per porre rimedio ai problemi di *tracking* della navigazione, dall'altro tale metodo non è esente da problemi e malfunzionamenti. In primis, infatti, tali cookies possono essere cancellati dai computer, sia dall'utente in maniera "diretta", sia attraverso scansioni da parte di software che hanno la funzione di tutelare la privacy eliminando qualsiasi tipo di dati personali.

Un'altro metodo con cui raccogliere i dati Web è il collezionamento a livello Client. Sostanzialmente, in questo modo si fa un uso intensivo

dell'elaboratore dell'utente, in particolare del browser con cui accede alla Rete. Il browser infatti, viene controllato dal cosiddetto *agent*, applicazione che ne monitora l'attività. Tale monitoraggio può avvenire in maniera più o meno consapevole da parte dell'utente. Esistono infatti agent che vengono volontariamente installati dai fruitori del computer, mentre, all'estremo opposto, ci sono applicazioni "spia", praticamente invisibili agli occhi del navigatore. Proprio per questo motivo, tali applicazioni sollevano forti dubbi per quanto riguarda la privacy degli internauti, e vengono classificati come *malware* (software "cattivi", più correttamente *spyware*) dai programmi che preservano l'integrità dei computer (antivirus, antispyware e antimalware). Inoltre, quando l'utente è consapevole di essere "osservato", subentrano dubbi sulla validità statistica dei dati raccolti, proprio perché il navigatore potrebbe essere condizionato dalla consapevolezza di essere in qualche modo controllato, e in ogni caso richiede un notevole sforzo collaborativo da parte dell'utente.

L'ultimo livello, che in realtà si configura come "intermedio" tra Server e Client è il livello Proxy. La presenza di questa piattaforma intermedia è spiegata dal bisogno, manifestatosi all'indomani dell'aumento del traffico in Internet, di navigare in maniera più veloce. I server Proxy sfruttano in maniera evoluta il principio del caching esposto pocanzi: per questo i dati che si raccolgono a livello proxy risultano molto simili a quelli che si riscontrerebbero a livello server. La principale differenza consiste nella capacità dei Proxy di veicolare un'insieme di richieste di utenti in maniera più efficiente. Ciò, dal punto di vista della raccolta dei dati risulta dannoso in quanto molto spesso, proprio a causa di questa particolarità di funzionamento, all'interno dei log files non rimane traccia dell'IP del singolo

utente, o, più precisamente, l'insieme degli utenti che si avvalgono del medesimo servizio Proxy saranno contrassegnati da un indirizzo IP collettivo.

1.7 IL PREPROCESSING DEI DATI NEL WEB USAGE MINING

In un processo di Web Usage Mining la fase di preprocessing riveste un ruolo fondamentale. Ciò in quanto, a differenza di altri tipi di analisi, in questo contesto l'estrazione di conoscenza si realizza prevalentemente grazie all'uso di dati "impliciti", non derivanti cioè da indagini, questionari o forme di esplicita rilevazione presso la popolazione di riferimento. Di seguito si analizzerà tale fase con particolare riferimento ad un problema di Web Usage Mining che, come precedentemente evidenziato, ha nei file di log il proprio input principale.

La fase di preprocessing può essere brevemente sintetizzata in una sequenza-tipo, quale quella proposta da Cooley, Mobasher e Srivastava (1999):

- Pulizia dei dati (Cleaning data): è il primo step della sequenza, nel quale I file di log vengono "puliti". La richiesta di una pagina Web comporta, nella quasi totalità dei casi, il download di più elementi che possono risultare come una serie di richieste differenti. Con questa attività si provvede dunque a rendere omogeneo l'output eliminando le richieste di parti strumentali alla visualizzazione della pagina, quali ad esempio le immagini.

  Generalmente, si conserva esclusivamente Il riferimento al nome della pagina effettivamente richiesta;

- Identificazione delle Sessioni e dei pageview (Session and page view identification): è probabilmente la fase più delicata del preprocessing poichè bisogna isolare le sessioni e le single pagine e non sempre ciò risulta agevole. Sovente si integrano le informazioni provenienti dal server con quelle provenienti dai cookies. Alcuni dei problemi di identificazione derivano da esigenze di tutela della privacy degli utenti, altri derivano strettamente dal funzionamento della Rete. Un ottimo esempio è costituito dal problema dell'individuazione della fine di una sessione poichè nei log file non vi è indicazione di tale avvenimento. A tale scopo, per prassi si stabilisce che un tempo di inattività (timeout) pari ad esempio a 30 minuti identifica il termine di una sessione;
- Completamento dei path (Traversal path completeness): I dati vengono processati ulteriormente al fine di eliminare punti di discontinuità nel percorso dovuti alle mancate registrazioni di alcune pagine;

- Identificazione degli episodi (Episode identification): è l'ultima fase, che consiste nell'isolare degli episodi significativi, rappresentati da sottoinsiemi delle sessioni di navigazione;

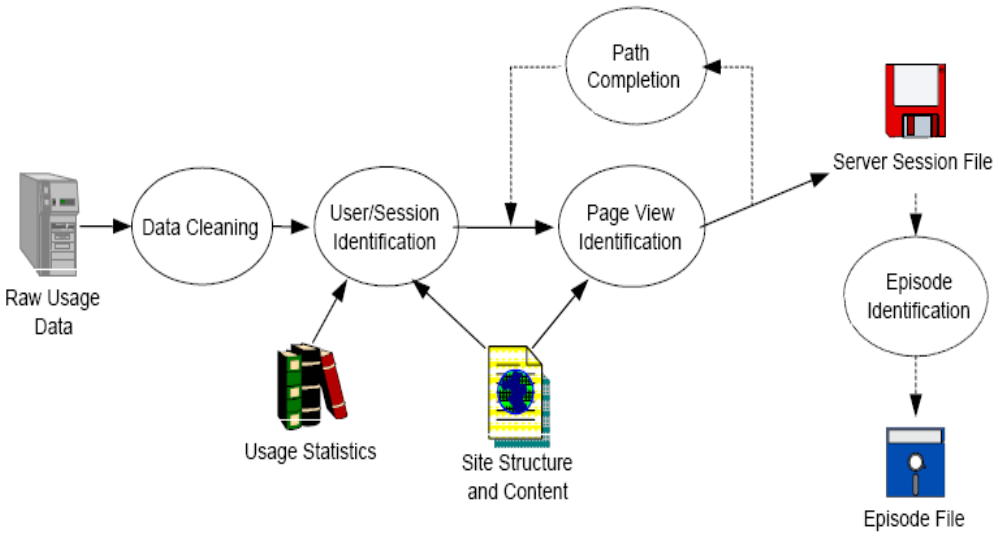

Figura 1.3 Il processo di estrazione degli Episode

1.8 PROBLEMI E SOLUZIONI NELL'ACQUISIZIONE DATI

Tra queste fasi, la più importante e delicata è senza dubbio quella di ricostruire con precisione le richieste e gli IP di provenienza. Ciò non è sempre facile, per una serie di ragioni che saranno esposte qui di seguito. Innanzitutto, è opportuno chiarire la forma "classica" dell'output di un log file. Nel formato ECLF, per ciascuna riga del file sono riportati i seguenti dati:

- IP Address: l'indirizzo IP dell'host remoto;
- Rfc931: il nome di login remote attibuito all'utente;
- Authuser: l'username scelto dall'utente in sede di registrazione;
- Date: giorno e ora della richiesta dell'utente;
- Request: la richiesta dell'utente, così come viene inoltrata dal client;
- Status: il codice http restituito al client a seguito della richiesta;

- Bytes: il numero di bytes trasferiti;
- Referer: l'URL su cui si trovava il client prima di inoltrare la richiesta corrente;
- User_agent: il software (browser e/o sistema operativo) in uso presso il client durante la richiesta;

La lettura e l'interpretazione di queste informazioni è spesso ostacolata da alcune caratteristiche proprie del funzionamento della Rete, che Srivastava, Cooley, Deshpande e Pang-Ning Tan (2000) riconducono a quattro categorie fondamentali:

- Singolo Indirizzo IP – Più utenti: non necessariamente a un singolo indirizzo IP corrisponde un solo utente. Se la navigazione avviene tramite server Proxy, oppure se l'accesso avviene tramite macchine ad accesso multiplo, ecco che il singolo indirizzo IP non è più indicatore univoco del singolo utente;

- Più sessioni – Singolo indirizzo IP: alcune applicazioni, soprattutto quelle progettate per garantire la privacy durante la navigazione, provvedono a non assegnare lo stesso indirizzo IP allo stesso utente; più precisamente si assegna lo stesso IP a più richieste, non necessariamente provenienti dallo stesso utente;

- Più indirizzi IP – Singolo Utente: anche quando gli indirizzi IP sono molteplici, non si è certi di trovarsi di fronte ad altrettanti utenti. Un navigatore può connettersi da più postazioni, risultando così di volta in volta un utente diverso

- Più Browser – Singolo Utente: un utente può inoltre navigare in Internet attraverso più browser, col risultato che potrebbe sfuggire o risultare un utente diverso dagli Agent, le applicazioni che monitorano il traffico a livello Client;

Inoltre, ci sono alcune situazioni specifiche che possono causare ulteriori problemi. E' il caso, ad esempio, delle Web TV o dei servizi in streaming. Per ottimizzare le prestazioni, alcuni sistemi predispongono la rotazione degli indirizzi IP, col risultato che essi si "confondono" dal punto di vista del monitoraggio dei dati di navigazione.

Per ovviare ad alcuni di questi problemi, gli esperti informatici predispongono una serie di accorgimenti. Alcuni, come i cookies o gli Agent, sono stati menzionati in precedenza; ci sono tuttavia altri sistemi, come la richiesta obbligatoria di identificazione attraverso login prima di ogni sessione di navigazione sul sito. Ancora, si può utilizzare congiuntamente un'applicazione Agent e monitorare gli IP oppure incorporare un identificativo dell'IP nei dati di sessione.

Un altro metodo per acquisire i dati di navigazione è necessario qualora questi vengano passati attraverso la cosiddetta Common Gateway Interface (CGI). Questo metodo prevede che le variabili e le richieste dell'utente passino come informazioni "legate" alla fine dell'URI. Se da un lato il monitoraggio di questi dati risulta vantaggioso per l'immediatezza con cui questi dati risultano fruibili, dall'altro anche questa tecnica presenta alcuni problemi: ad esempio, le variabili passate attraverso il metodo POST vengono ignorate. Le informazioni salvate

nella sessione non vengono replicate nell'indirizzo e non vengono considerate da questo tipo di monitoraggio.

Esistono, anche in questo caso, dei correttivi volti ad ovviare alle disfunzioni appena menzionate. Si può infatti cercare di passare quanti più dati possibile attraverso il traffico HTTP. Ciò è vantaggioso perché è un sistema che funziona per ogni configurazione Web, ma pone seri limiti di sicurezza per quanto riguarda il passaggio dei cosiddetti "secure data", proprio perché tutte le informazioni che il browser scambia col server sono visibili.

Un ulteriore ostacolo al corretto immagazzinamento dei dati di navigazione è costituito dal fenomeno del caching, ossia il salvataggio in locale delle pagine al fine di velocizzare la navigazione stessa. Queste pagine vengono richiamate principalmente quando l'utente fa ricorso ai pulsanti "Indietro" e "Avanti" del proprio browser. Anche in questo caso, è possibile predisporre azioni che riducono l'incidenza di questo fenomeno. La soluzione più efficace è senza dubbio l'introduzione di pagine dinamiche, che però rallenta la navigazione producendo un maggiore traffico dati. Un altro accorgimento è quello di ridurre il più possibile il tempo di scadenza della pagina (fino a impostarlo con un timing negativo).

Capitolo 2

Metodi statistici a supporto del Web Mining

2.1 UN APPROCCIO STATISTICO AL MONDO DI INTERNET

Nel capitolo precedente si è chiarito come il Web Mining sia un processo eterogeneo, poiché per perfezionarlo di solito si combinano tecniche proprie di più discipline. Per di più i ricercatori devono possedere un bagaglio di conoscenze di base che esulano dal contesto statistico e da quello del Data Mining. Innanzitutto, è necessario familiarizzare col funzionamento della Rete Internet, per comprendere le grandezze che la compongono, il loro significato e, soprattutto le procedure di acquisizione dei dati Web e i limiti

di queste. Inoltre, alcuni dei task tipici di alcune branche del Web Mining sono presi a prestito da discipline quali l'Information Retrieval.

Il fulcro di ogni analisi di Web Mining rimane però l'adozione di tecniche statistiche volte a ricavare conoscenza da questi dati. In questo capitolo si focalizzerà l'attenzione su queste tecniche, con particolare riferimento a quelle usate in un contesto di Web Usage Mining.

In tale contesto, si cerca soprattutto di pervenire a una profilazione dei navigatori, individuando categorie simili di utenti e di implementare modelli in grado di collocare nuovi individui in una di queste categorie. Queste attività costituiscono di solito il background necessario alla costruzione di applicazioni di raccomandazione dei contenuti, di visualizzazione di suggerimenti personalizzati, o ancora costruire sistemi di e-commerce intelligenti, in grado di proporre prodotti correlati o complementari rispetto a quelli osservati.

Una fase preliminare all'analisi vera e propria può essere condotta attraverso la consultazione delle statistiche descrittive, spesso ricavate in maniera automatizzata da utility ad uso dei webmaster. Celebre è il motore Urchin di Google Analytics. Altri metodi di indagine preliminari sono l'interrogazione della base di dati con query SQL, o ancora la consultazione di tools grafici.

2.2 LE TECNICHE DI DATA MINING PER LO STUDIO DEI DATI WEB

Le analisi più approfondite richiedono invece l'uso di tecniche proprie del Data Mining adattate al particolare contesto dei dati Web.

Le tecniche principali che vengono usate in questo contesto sono moltplici: di seguito se ne illustreranno quattro, sicuramente tra le principali: le regole associative, gli algoritmi di clustering, quelli di segmentazione e le reti neurali.

2.3 LE REGOLE ASSOCIATIVE

Le regole associative, introdotte da Agrawal, Imielinski e Swami nel 1993, sono utilizzate per identificare la presenza contemporanea di due o più item i in un set di transazioni t. Il procedimento di ricerca degli item che più frequentemente si presentano insieme nella base dati considerata si basa sull'algoritmo Apriori, introdotto dagli stessi autori, e da sue modificazioni. L'algoritmo si basa su due misure fondamentali, il supporto e la confidenza, anche se in dottrina esistono misure alternative o complementari, quali ad esempio il lift.

Dati due item, A e B il supporto della regola A => B è dato dal numero di transazioni in cui A e B si presentano assieme rispetto al numero totale delle transazioni stesse.

$$Sup(A \Rightarrow B) = P(A \cap B) / N°trans.$$

(2.1)

Come si può notare dalla formula, il supporto non è altro che la probabilità di osservare congiuntamente i due item in esame all'interno del dataset.

La confidenza invece è espressa dal numero di volte che gli item A e B si presentano assieme nel dataset rispetto al numero di volte in cui si presenta l'item A.

$$Sup(A \Rightarrow B) = P(A \cap B) / P(A)$$

(2.2)

Dal punto di vista probabilistico, dunque, la confidenza è assimilabile al concetto di probabilità condizionata di presenza dell'item B dato che è presente l'item A.

Sulla base di queste due misure, l'algoritmo si basa sull'identificazione dei cosiddetti Large itemset, insiemi di item che rispettano i vincoli di supporto e confidenza prefissati. L'elevatissimo numero di transazioni da cui sovente è composto il dataset renderebbe arduo, se non impossibile, ricavare i Large itemset. Questo ostacolo computazionale viene superato grazie all'uso dell'algoritmo Apriori. In sostanza, tale algoritmo evita di "provare" tutte le possibili combinazioni di item che si potrebbero generare, in base al seguente principio: se uno (o più) Candidate itemset non rispetta(no) le soglie di supporto e confidenza prefissati allora sicuramente non soddisfano questo tipo di vincolo tutte le regole generate a partire da questi item. Questo assunto permette di velocizzare notevolmente il processo di ricerca.

Va precisato inoltre che la fissazione di vincoli di supporto e confidenza non è l'unico modo di procedere per ricavare delle regole. Possono infatti essere fissati degli altri vincoli, come quelli sintattici: si cercano, in questo modo, le regole che presentano determinati item. Si stabilisce inoltre se questi item debbono comparire all'antecedente della regola (parte sinistra), al

conseguente (parte destra) o se non è importante la posizione. Ancora, si possono seguire approcci alternativi, come quello degli odds ratio, introdotto da Giudici e Passerone (2000).

Una volta generate le regole, esse non sono ancora "utilizzabili" ai fini dell'analisi. Infatti, tali regole possono essere numerosissime: in tal caso, o si decide di innalzare le soglie precedentemente fissate, o ci si avvale di strumenti di visualizzazione delle regole che possano aiutare ad individuare quelle più interessanti. Ancora, di solito è necessario un ulteriore step volto a eliminare le regole banali ovvero quelle già note. Infine, quando si dispone di un set di regole "finale", di solito l'analisi non si conclude ancora: è buona regola, infatti, analizzare la significatività di tali regole. Le regole possono essere sottoposte a test quali il DOC (Difference of Confidence) che confronta la confidenza della regola in questione con una il cui antecedente è generato casualmente.

2.3.1 L'uso delle Regole Associative nel Web Usage Mining

Nel contesto del Web Usage Mining le regole associative sono usate soprattutto per identificare le pagine che più frequentemente sono visitate simultaneamente nell'insieme delle sessioni (transazioni). Questo tipo di analisi può essere condotta ponendo come indifferente l'ordine con cui tali pagine vengono consultate oppure, come più frequentemente accade, tenendo conto della sequenzialità della consultazione. La metodologia più

frequentemente utilizzata in questo caso prende il nome di clickstream analysis e sarà trattata in seguito.

Un'altra applicazione importante è la Market Basket Analysis riferita ai beni acquistati presso siti di e-commerce. Anche i risultati prodotti da questa analisi sono del tipo If-then, con regole che esprimono la probabilità di acquisto congiunto dei due prodotti.

Nel contesto del Web Usage Mining, tuttavia, si vuole pervenire alla profilazione dell'utenza attraverso la scoperta di ricorrenze nell'uso di uno o più siti Web. Le regole associative, come del resto le altre tecniche di Data Mining utilizzate in tale contesto, sono perciò utilizzate sempre con un forte orientamento alla scoperta di "pattern" interessanti all'interno del dataset di riferimento. Per questo motivo, di seguito si approfondirà esclusivamente l'uso di questa tecnica per scopi di profilazione, o, più genericamente, di identificazione di pattern di navigazione.

In questo tipo di applicazioni il dataset verrà preparato, mediante operazioni di preprocessing quali ad esempio quelle precedentemente esposte, in una forma matriciale "tipo" nella quale sulle righe compaiono le sessioni di visita (le transazioni) mentre sulle colonne si trovano, a seconda dell'obiettivo dell'analisi, tutte le pagine (gli items) possibili, oppure lo stato dei click: primo click, secondo click, terzo click, con l'indicazione della pagina consultata per ogni stato della sequenza. Si noti come le due matrici ottenute siano profondamente diverse: la prima è una matrice di presenza-assenza, avente tutte le righe della medesima lunghezza. Nella seconda, le righe non avranno in maniera tutte la stessa lunghezza ed anche il contenuto delle celle sarà profondamente differente.

Una volta preparata la base di dati si procede all'identificazione dei pattern di visita secondo due logiche principali: quella dei Sequential Patterns e quella delle Maximal Frequent Forward Sequencies. Col primo approccio si prendono in considerazione dei sottoinsiemi di sessioni, dette sequenze, maggiormente "popolari", nel senso che supportano i vincoli di supporto e confidenza imposti dal ricercatore. A seconda degli scopi dell'analisi e della volontà di colui che conduce l'indagine, possono essere considerate o meno le sequenze che contengono delle replicazioni di pagine.

L'altro approccio è volto all'identificazione di sequenze di navigazione "in avanti". Il principio ispiratore di questa tecnica è, infatti, quello che l'utente torna indietro tra le pagine, o comunque visualizza più volte la stessa pagina, per confusione o per errori di navigazione. Secondo questa logica dunque, i ritorni indietro o le duplicazioni sono da considerarsi "rumore" che come tale va eliminato dall'analisi.

A titolo di esempio, si faccia riferimento a un sito web la cui struttura presenta le pagine A,B,C,D,E come forward sequence. Partendo da tale condizione, allora le sequenze di visita {A,B,B,A,C,D} e {B,C,D,C,B,A,B} diverranno rispettivamente {A,B,C,D} e {B,C,D}. Anche questa tecnica utilizza il criterio della massima frequenza, nel senso che prende in considerazione solo gli episodi che godono di una certa ricorrenza all'interno del dataset.

Esistono comunque numerosi altri approcci che sfruttano i principi delle regole associative uniti a diverse varianti. In questo senso, merita menzione l'approccio di Aria, Mola e Siciliano che propongono l'utilizzo delle prediction rules in luogo delle association rules in maniera da tenere conto di

altri fattori che pesano le pagine e, di conseguenza, le associazioni che le legano.

2.4 GLI ALGORITMI DI CLUSTERING

Gli algoritmi di clustering sono utilizzati allo scopo di partizionare un insieme di unità statistiche in una serie di gruppi quanto più possibile omogenei (simili) al loro interno e quanto più possibile eterogenei (dissimili) tra loro. Questo partizionamento ha dunque l'obiettivo di identificare dei cluster, dei sottoinsiemi di individui aventi caratteristiche simili tra loro e al contempo diverse dal resto della popolazione o, più precisamente, dagli altri gruppi identificati.

Partendo da una matrice di *n* individui e *p* variabili, il primo passo da seguire è quello di selezionare le variabili che concorreranno all'analisi. E'uno step cruciale, in quanto valutazioni erronee compiute in questa fase possono avere pesanti ripercussioni sui risultati finali. In linea di massima, un modello di clustering può considerarsi soddisfacente quando si dimostra robusto, nel senso che non è eccessivamente sensibile ai cambiamenti di dataset o di variabili utilizzate.

Un'altra scelta preliminare all'analisi vera e propria è quella relativa al metodo di formazione dei gruppi (detto anche metodo di aggregazione). La dottrina distingue tra due grandi famiglie: i metodi gerarchici e quelli non gerarchici. I primi si basano su processi iterativi che considerano, a seconda dei casi, tutte le *n* unità come distinte in *n* gruppi iniziali per pervenire, dopo un certo numero di aggregazioni a un solo gruppo di numerosità *n*. I metodi

gerarchici però possono funzionare anche in maniera inversa: in tal caso al primo step del processo iterativo, le unità si considerano come appartenenti ad unico gruppo di numerosità n, per poi pervenire ad un certo numero di gruppi finali. I metodi non gerarchici invece pervengono a un'organizzazione delle unità statistiche in un numero di gruppi fissato a priori dal ricercatore.

Un'altra scelta da effettuare riguarda l'indice di prossimità da utilizzare che serve per calcolare la matrice delle distanze tra le unità in analisi. Allo scopo, di solito si procede alla standardizzazione delle variabili, utile nei casi in cui si vuole evitare che alcune variabili pesino più di altre. E'inoltre da precisare che nel caso in cui si opti per l'uso di un metodo gerarchico, va anche definito come calcolare la misura di prossimità tra i gruppi individuati.

Lo step finale, che si concretizza a procedura ultimata, consiste nella valutazione del risultato ottenuto, tenendo presente soprattutto lo scopo ultimo per il quale si pone in essere una procedura di clustering, vale a dire ottenere dei gruppi quanto più possibile omogenei internamente (gruppi costituiti da unità quanto più possibile simili tra loro dal punto di vista delle variabili considerate) e massimamente eterogenei tra loro. Di solito, la scelta ottimale si configura come trade-off tra la precisione e la "purezza" interna fornite da un numero di partizioni elevate e la semplicità e la chiarezza interpretativa garantite dalla formazione di pochi gruppi.

Una volta chiariti i principali step che caratterizzano un algoritmo di clustering, è utile soffermarsi sul funzionamento dei metodi gerarchici e di quelli non gerarchici.

2.4.1 I Metodi Gerarchici

Con riferimento ai primi, si è detto che questi si distinguono, a seconda di come procedono al partizionamento delle unità, in metodi agglomerativi (nei quali, passo dopo passo, si aggregano le *n* unità che inizialmente sono tutte distinte fino a pervenire a un unico gruppo di numerosità *n*) e metodi scissori (nei quali si procede all'inverso, partendo cioè da un unico gruppo di numerosità *n* per arrivare a formare n gruppi distinti di numerosità 1). Per chiarire il funzionamento delle procedure di partizionamento, di solito è più immediato avvalersi di una rappresentazione grafica, detta dendrogramma o albero di classificazione gerarchica.

Attraverso il dendrogramma si ha una descrizione visiva del processo di agglomerazione/scissione delle unità: da questo schema si può notare un'altra peculiarità dei metodi gerarchici, vale a dire la nidificazione delle partizioni create.

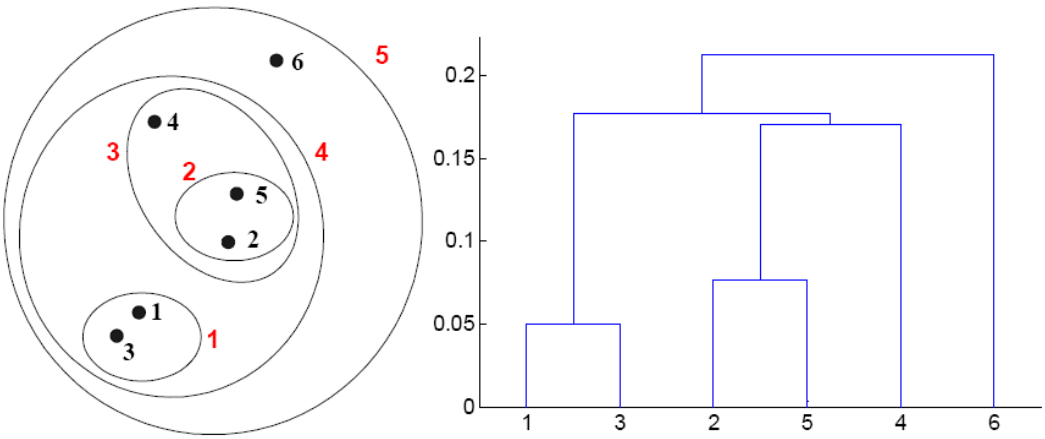

Figura 2.1 Una soluzione di clustering gerarchica e un dendrogramma

In virtù di tale caratteristica se due individui sono uniti (divisi nel caso delle procedure scissorie) resteranno uniti (divisi) anche in tutte le successive iterazioni che compongono il procedimento. Se da un lato ciò rende agevole la scelta del numero di cluster (ovvero del taglio) da selezionare, dall'altro denota come eventuali errori di assegnazione di un individuo a un gruppo risultino "irrimediabili" in quanto non è possibile riallocare l'individuo stesso nel corso della procedura.

Per quanto riguarda le distanze, le più conosciute e usate sono tre: il metodo del legame singolo, quello del legame completo e quello del legame medio.

Secondo il primo metodo, la distanza tra due cluster è data dalla distanza che intercorre tra i due individui più vicini, cioè la distanza minima tra i due gruppi.

$$d(C_1, C_2) = \min(d_{r;s})$$

con $r \in C_1$ e $s \in C_2$

(2.3)

Il metodo del legame completo, invece, stabilisce che la distanza tra due gruppi corrisponde alla massima tra le distanze tra tutti gli individui appartenenti ai medesimi gruppi.

$$d(C_1, C_2) = \max(d_{r;s})$$

(2.4)

Infine, il metodo del legame medio considera la media aritmetica tra le unità appartenenti ai due gruppi.

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{r=1}^{n_1} \sum_{s=1}^{n_2} d_{r,s}$$

(2.5)

Vi sono poi metodi che utilizzano come distanza alcuni indicatori ben noti in statistica come le misure di tendenza centrale. Con il criterio della media (o della mediana) si quantifica la distanza tra gli elementi del gruppo calcolando la media (mediana) di tutte le distanze riscontrabili tra gli elementi.

Si noti che questo primo gruppo di metodi, richiede, ai fini dell'applicazione, la sola matrice delle distanze.

Esistono, invece, metodi che per essere applicati richiedono anche la matrice dei dati. E'il caso del metodo dei centroidi, che restituisce la distanza tra i cluster in termini di distanza tra i rispettivi centroidi (medie aritmetiche). E'facile notare come questo metodo risulti molto simile a quello del legame medio precedentemente illustrato: l'unica differenza risiede nel fatto che il metodo del centroide considera le medie dei due gruppi per poi calcolare la distanza, mentre col legame medio si fa riferimento alla media delle distanze tra le unità dei due gruppi.

Un altro metodo che richiede la matrice dei dati è quello di Ward, che richiama fortemente l'idea alla base dei processi di clusterizzazione: l'obiettivo è, infatti, quello di pervenire alla formazione di gruppi quanto più possibile omogenei internamente e quanto più possibile eterogenei tra loro. I concetti di omogeneità interna e di eterogeneità esterna si traducono, concretamente, in devianza entro i gruppi (detta anche devianza within) e

devianza tra i gruppi (devianza between) che insieme compongono la devianza totale.

$$DEVtot = DEVw + DEVb$$

(2.6)

Formalmente, tali grandezze si esprimono come segue:

$$DEVw = \sum_{k=1}^{g} W_k \qquad \text{con} \qquad Wk = \sum_{s=1}^{p} \sum_{i=1}^{n_k} \left(x_{is} - \bar{x}_{sk} \right)^2$$

e

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

(2.7)

Il metodo di Ward consiste nell'aggregare, ad ogni passo della procedura, i gruppi il cui inserimento comporta il minor incremento della devianza interna (in altre parole il minor incremento dell' "impurità" interna al cluster) al gruppo stesso. Si noti come ciò comporta, automaticamente, anche un incremento della devianza between.

I calcoli inerenti la quantificazione dell'omogeneità interna e l'eterogeneità esterna ai cluster rivestono un ruolo fondamentale, oltre che nelle procedure di creazione dei gruppi, anche nella valutazione dei metodi gerarchici. Un indice indicativo della bontà del risultato ottenuto è ad esempio l'R^2, costruito nel modo seguente:

$$R^2 = 1 - \frac{DEVw}{DEVt} = \frac{DEVb}{DEVt}$$

(2.8)

E'importante precisare che il solo riferimento a questo indice risulta insufficiente e può addirittura trarre in errore. Per come è costruito tale indice, infatti, esso risulta pari ad 1 quando la configurazione trovata è di un cluster per ogni osservazione, soluzione evidentemente insoddisfacente e per di più in contrasto con i criteri di parsimonia propri di un risultato ottimale. Per questa ragione sovente si usa un altro indice (in sostituzione o meglio ancora in aggiunta al precedente) detto RMSSTD (acronimo di Root-Mean-Square Standard Deviation) che considera esclusivamente la devianza nei gruppi che si aggiunge nel passo in esame di una procedura. L'interpretazione di quest'indice è dunque abbastanza agevole, facendo riferimento anche alla formula sottostante, riferita a un generico passaggio *h*: se l'aggregazione proposta allo step in esame produce un aumento troppo marcato di tale misura, significa che il nuovo gruppo presenterebbe una devianza within assai più elevata rispetto al cluster precedente, pertanto la nuova aggregazione andrebbe evitata.

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

(2.9)

2.4.2 I Metodi non Gerarchici

I metodi non gerarchici conducono a una partizione delle unità statistiche in un insieme di gruppi, il cui numero viene fissato a priori da colui che esegue la clusterizzazione. Da ciò deriva che la procedura conduce ad uno e un solo risultato, che rispetta determinati criteri di ottimalità. Inoltre, il vincolo del numero dei gruppi influenza fortissimamente la soluzione che varia in maniera sensibile a seconda della numerosità prescelta e della misura adottata.

La distanza tra le unità statistiche nonché tra i centroidi si utilizza la metrica euclidea. In una generica iterazione, la distanza tra l'i-esima unità e il centroide l (dove i=1,2,…..,n e l=1,2,….,g) è pari a:

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

(2.10)

La procedura propria dei metodi non gerarchici può essere schematizzata come segue:

1. Scelta del numero dei gruppi: si scelgono i *g* gruppi in cui verranno raggruppate le *n* unità;
2. Misura dell'apporto derivante dagli spostamenti: in questa fase viene quantificata la variazione in termini di coesione interna ai gruppi derivante dalla riallocazione delle unità da un cluster all'altro. Se il beneficio è significativo, si procede alla riallocazione;

3. Confronto del risultato con la soglia d'arresto: la soluzione viene a questo punto confrontata con la soglia d'arresto. La procedura termina se il requisito è soddisfatto altrimenti si ripete il passo 2;

Tra i metodi di partizionamento non gerarchici più conosciuti e utilizzati spicca sicuramente quello delle k-medie, dove k è il numero di gruppi prescelto. Anche tale processo è riconducibile a diverse fasi:

1. Selezione dei seeds (semi) iniziali: in questo primo passo, successivo solo alla selezione del numero *g* dei gruppi, si definiscono k punti nello spazio p-dimensionale che rappresentano i centroidi dei cluster nella configurazione iniziale; è buona regola che tali centroidi siano sufficientemente distanti tra loro, per massimizzare il rendimento dell'algoritmo. Il passo si conclude con una prima allocazione delle unità statistiche che saranno "attirate" dai centroidi e formeranno i primi cluster;
2. Calcolo della distanza unità-centroide: in questo step si calcola, per ogni unità, la distanza col centroide di riferimento. Se questa non risultasse la minima possibile, ovvero ci fosse un altro centroide più vicino (meno distante) l'unità viene riallocata nel cluster relativo a questo secondo centroide. Il passo si conclude con il ricalcolo dei centroidi (che di solito corrispondono a medie) dei gruppi interessati dallo spostamento;
3. Verifica della convergenza dell'algoritmo: a questo punto si controlla se il risultato raggiunto è sufficientemente stabile.

2.4.3 Un confronto tra i due metodi di Clustering

I due differenti modi di procedere propri di questi metodi, di fatto concorrenti tra loro, rendono necessario un confronto orientato soprattutto a comprendere quando è conveniente usare gli algoritmi gerarchici piuttosto che gli altri.

Dal punto di vista della velocità di esecuzione, risultano più veloci le procedure di tipo non gerarchico, in quanto non prevedono il calcolo preliminare della matrice delle distanze. Anche dal punto di vista della stabilità della soluzione alla variabilità campionaria risulta da preferirsi la clusterizzazione non gerarchica. I vantaggi appena elencati rendono questi algoritmi utili nelle situazioni nelle quali si ha a che fare con dataset di grandi dimensioni; tuttavia in questi casi bisogna accontentarsi di massimi "locali", ovvero vincolati a condizioni restrittive dettate dall'impossibilità di calcolare e confrontare le innumerevoli combinazioni possibili in cui le n unità possono essere combinate per dare vita ai g gruppi.

Dall'altra parte, le procedure di classificazione gerarchiche possono rappresentare la soluzione ideale nei casi in cui si desidera sintetizzare le caratteristiche dell'insieme delle osservazioni in gruppi ottenendo inoltre una indicazione sintetica della loro composizione, espressa dal concetto di centroide. Un vincolo, però, può essere costituito dall'arbitrarietà con la quale si stabilisce il numero k dei gruppi in cui sarà partizionato il collettivo, che influenza in modo significativo l'analisi e la stabilità dei risultati. Un espediente usato nella pratica, che sovente ovvia allo svantaggio appena citato, è quello di svolgere preventivamente una classificazione non gerarchica per ottenere indicazioni circa il numero di gruppi da imporre,

successivamente, come input per la procedura gerarchica. Un ultimo svantaggio che vale la pena tenere presente nell'uso di questa tecnica deriva dalle distorsioni che l'algoritmo produce qualora il set di dati in esame fosse caratterizzato dalla presenza "diffusa" di valori anomali (outliers). Anche in questo caso, comunque, si possono ridurre questo tipo di svantaggi avviando la procedure con un numero g di gruppi assai elevato: in questo modo, l'algoritmo stesso tenderà a individuare gli outliers, che generalmente formeranno gruppi a se stanti, mentre le osservazioni "normali" tenderanno a concentrarsi in cluster di numerosità più elevata.

2.5 GLI ALGORITMI DI SEGMENTAZIONE SUPERVISIONATA

Nel paragrafo precedente si sono esaminate le caratteristiche di funzionamento dei metodi di classificazione "non supervisionata"; in quello presente si prenderanno in considerazione gli algoritmi di segmentazione, che sono metodi di classificazione di tipo "supervisionato". A differenza di quelli appena esposti, nei quali si presume che le variabili giochino tutte lo stesso ruolo nella descrizione del fenomeno esaminato, nell'approccio supervisionato si classificano le unità secondo una variabile target (detta anche "di risposta") di cui sono note a priori le modalità che essa può assumere. Per queste ragioni, le tecniche in questione sono dette anche asimmetriche, in contrapposizione alle precedenti che si dicono, all'opposto, simmetriche.

Le tecniche di segmentazione sono procedure iterative che forniscono come output una serie di partizioni che rappresentano sottoinsiemi del collettivo di origine contrassegnate secondo le modalità che la variabile di risposta

assume in corrispondenza di ciascuna partizione. In questa trattazione si farà riferimento alla cosiddetta segmentazione binaria, metodo nel quale la variabile di risposta può assumere solo due modalità. I gruppi che l'algoritmo definisce saranno in questo caso contrassegnati, a fine procedura, come appartenenti all'una o all'altra modalità. I vantaggi derivanti dall'uso di questa tecnica sono da ricercarsi in due motivazioni principali: l'intuitività con la quale i risultati sono esposti, derivante dall'output grafico che l'algoritmo produce, dalla caratteristica forma ad albero (grafo aciclico orientato) e l'agevole interpretazione delle regole che discriminano l'appartenenza all'una o all'altra categoria.

Tale albero descrive dunque la dipendenza di una variabile da un insieme di variabili esplicative in problemi di classificazione e regressione. Se la variabile di risposta considerata è di tipo qualitativa si parla di classificazione ad albero, se è di tipo quantitativa si parla di regressione ad albero. La segmentazione binaria parte dal considerare l'insieme delle osservazioni come un aggregato unico, sovente detto "nodo padre" o "nodo radice", che costituisce il punto di partenza dell'albero. Da questo si dipartono, a due a due, i vari nodi che compongono la struttura, collegati tra loro da archi orientati che indicano la direzione in cui l'albero si sviluppa. I nodi si dicono interni (generalmente rappresentati come cerchi) quando da essi si formano altri due nodi; viceversa si parla di nodi terminali quando la struttura, oltre i medesimi nodi, non si sviluppa ulteriormente (graficamente si rappresentano come quadrati). Si indicano infine come branche i sottoinsiemi della struttura che si ottengono a seguito dell'operazione di potatura (scissione di una parte dell'albero).

La procedura è di tipo iterativo e mira, attraverso una serie di partizionamenti binari di creare sottogruppi sempre più fini (di numerosità minore a quella del nodo radice o del nodo "padre" da cui nascono) e al contempo più omogenei al loro interno. L'omogeneità viene calcolata in termini di distribuzione delle osservazioni che compongono il nodo rispetto alla variabile di risposta.

Le partizioni hanno luogo grazie ai predittori: l'algoritmo esamina tutte le possibili dicotomizzazioni che si possono ottenere a partire dai predittori considerati e sceglie quella in grado di effettuare la divisione migliore, ossia quella che produce due sottogruppi quanto più possibile "puri": in seguito definiremo la sostanza del concetto di purezza.

2.5.1 La costruzione dell'albero

Si è detto in precedenza che l'albero si costruisce tramite una serie di partizioni ricorsive. In questo paragrafo si chiarisce tale procedimento, definendo anche misure e regole d'arresto della procedura.

Sia Y una multivariata detta "di risposta" di cui si vuole scoprire la dipendenza da un insieme X_m (X_1, X_2,….,X_m) di predittori. Si è chiarito già come ad ogni passo della segmentazione (nel caso considerato, quella binaria) si provvede alla ricerca della migliore divisione (split) che si può ottenere a partire dai predittori a disposizione. La migliore divisione è quella che massimizza una funzione obiettivo, detta misura di impurità, che nella maggior parte dei casi è una tra le seguenti:

- Tasso di errata classificazione:

$$i_Y(t) = 1 - \max_j p(j \mid t)$$

(2.11)

- Indice di eterogeneità (Gini):

$$i_Y(t) = 1 - \sum_j p(j \mid t)^2$$

(2.12)

- Indice di entropia:

$$i_Y(t) = 1 - \sum_j p(j \mid t) \log p(j \mid t)$$

(2.13)

dove p(j|t) è il numero di unità del nodo t che appartengono alla classe j. Le misure appena esposte si riferiscono ai contesti di classificazione, mentre nei problemi di regressione, l'impurità si traduce in una misura di varianza o di devianza di Y, quest'ultima sarà riferita alle sole unità del nodo:

$$i_Y(t) = 1 - \sum_{Xn \in t} (y_n - \bar{y}(t))^2$$

(2.14)

Il concetto di impurità, riferito al singolo nodo, riveste un'importanza ancora maggiore se si considera il cosiddetto decremento di impurità, ovvero la misura che a partire da essa si costruisce come differenza tra l'impurità del nodo padre e quella dei due nodi figli.

$$I_Y(t) = 1 - \sum_{t \in T} i_Y(t) p(t)$$

(2.15)

Approfondendo la costruzione delle partizioni, un predittore in scala numerica o ordinale G modalità distinte genera G-1 suddivisioni (split), mentre misurate in scala nominale ne genera $2^{G(G-1)}$-1. Tra l'insieme P delle partizioni ottenibili a partire dai predittori al nodo t, la prescelta sarà quella che, tra queste, minimizza il fattore locale di riduzione dell'impurità:

$$\max_{p \in P} \Delta i_Y(t, p) = \max_p \left\{ i_Y(t_k)\, p(t_k \mid t) \right\}$$

(2.16)

Questo procedimento, che è quello adottato dalla maggior parte degli algoritmi presenti nei software in commercio (CART, ID3, CN4.5), ma è abbastanza costoso in termini computazionali: per questo motivo sovente si fa ricorso al criterio a due stadi (two-stage). In base a tale criterio lo split ottimale non viene cercato tra tutti quelli che è possibile generare, ma solo tra un sottoinsieme di questi, quelli migliori in termini di capacità di riduzione del fattore globale di impurità. In sostanza dunque, l'algoritmo

cerca lo split in grado di ridurre al massimo l'impurità prima globale poi locale.

2.5.2 Le regole di arresto della procedura

L'albero delle decisioni risulta vantaggioso quando la sua struttura è facilmente interpretabile e quando la situazione che descrive non risulta troppo condizionata dal set di dati che è servito per produrla (overfitting). In ossequio a tali condizioni, è importante scegliere regole di arresto della procedure che impediscano un'espansione eccessiva dell'albero. Per quanto riguarda la misurazione della "taglia" della struttura, va chiarito che essa si può ottenere sia contando il numero dei nodi terminali che quello dei nodi interni.

Sono tre le principali regole d'arresto dell'algoritmo; si può, innanzitutto, imporre che la struttura non sia composta che da un numero massimo di livelli dell'albero. Si può altresì fissare una soglia di impurità tollerata, oltre la quale non si ritiene conveniente in quanto si otterrebbe una struttura troppo estesa e inutilmente particolareggiata. Si può, infine, arrestare la procedura nel momento in cui i nodi terminali raggiungono una numerosità minima prefissata, anche in questo caso per limitare la comparsa di gruppi la cui creazione è dovuta solo al particolare campione di apprendimento utilizzato.

2.5.3 Assegnazione della risposta ai nodi terminali

Una volta arrestata la procedura, si osserva un certo numero di gruppi finali, da cui non si dipartono ulteriori partizioni. Questi nodi si dicono terminali e sono quelli più in basso (o più in alto a seconda di come si rappresenta la struttura) e più lontani rispetto al nodo padre.

I nodi terminali vengono etichettati in base alla classe modale di appartenenza nel caso di problemi di classificazione, mentre il nodo sarà contrassegnato dal valore medio della variabile di risposta osservato all'interno del nodo stesso nei casi di regressione.

2.5.4 Dall'albero esplorativo a quello decisionale

La tecnica di segmentazione binaria descritta finora si pone in essere, inizialmente, a scopo esplorativo. Ad esempio, in un contesto applicativo come quello del Web Mining, gli algoritmi di segmentazione possono essere usati a scopo esplorativo per comprendere, a partire da un dataset a disposizione, cosa influenza maggiormente l'acquisto di un prodotto su un sito di e-commerce, piuttosto che la scoperta di ciò che condiziona la consultazione di una pagina o il download di un contenuto. Per scopi applicativi, invece, è fondamentale, oltre all'esplorazione e alla descrizione di un fenomeno, anche l'utilizzo di queste tecniche a scopi decisionali, ovvero per valutare se un nuovo utente, dato un set di caratteristiche osservate, comprerà il prodotto, visionerà quella pagina o scaricherà quel contenuto (ovviamente sulla base dell'esplorazione precedente).

E'necessario dunque che una volta terminata la fase esplorativa si arricchisca l'analisi ottenendo il cosiddetto albero decisionale.

Innanzitutto, affinchè questo passaggio avvenga correttamente, le misure considerate finora devono essere integrate con un'altra, atta a valutare l'accurattezza della previsione, denominata tasso di errata classificazione o previsione.

Il problema è che gli alberi che presentano un basso tasso di errata classificazione sono alberi molto complessi (espansi), poco affidabili nel prevedere l'appartenenza all'una o all'altra modalità della variabile di risposta. Per ovviare a questo deficit di interpretabilità e di predittività l'albero decisionale sarà meno complesso di quello esplorativo e sarà costruito a partire da un altro campione (normalmente complementare a quello di apprendimento) detto campione test.

Operativamente, vi sono tre modi per valutare, l'albero, stimando il tasso di errore. Essi sono:

1. Stima di risostituzione;
2. Stima test set;
3. Stima cross-validation;

Col primo metodo, in buona sostanza si pone a confronto il numero di unità mal classificate N(h) nell'albero T sul totale delle unità N.

Il secondo metodo si concretizza con la divisione del dataset in due parti: la prima, che generalmente è costituita dal 70% dei dati a disposizione, prende il nome di campione di apprendimento e viene utilizzata per la costruzione dell'albero. La seconda, detta campione test, viene in seguito fatta

"scivolare" all'interno dell'albero precedentemente ottenuto al fine di valutare la bontà della struttura stessa in termini di classificazione di nuove unità.

Il terzo metodo, infine, si applica solitamente quando le osservazioni a disposizione sono di numerosità troppo esigua perché si possa dividere in due il dataset e ricavarne due campioni di dimensioni soddisfacenti. In questi casi si procede dividendo i dati di partenza C in V campioni e costruendo tanti alberi a partire dai gruppi C escludendo di volta in volta un campione ($C\text{-}C_1$, $C\text{-}C_2$,….,$C\text{-}C_V$) e utilizzando poi il campione non utilizzato per validare la struttura. Quando si sono esaurite le combinazioni e sono stati creati tutti gli alberi, la media delle stime test set così ottenute fornirà la stima cross validation.

2.5.5 La semplificazione della struttura: la potatura degli alberi

In precedenza si è fatto cenno al trade-off precisione-complessità: se da un lato alberi molto espansi forniscono una descrizione esaustiva e dettagliata dei fenomeni che portano la variabile di risposta ad assumere una data modalità, dall'altro, in un'ottica decisionale, strutture così complesse risultano difficilmente interpretabili e soprattutto sono poco adattabili a nuovi set di dati. Quando si utilizza l'albero in un'ottica previsionale si rischia di spiegare, con un albero espanso, il campione di apprendimento piuttosto che le nuove osservazioni, caratterizzate da incertezza.

Per questi motivi si pone in essere, in questo tipo di procedure, un'ulteriore operazione, quella di pruning (potatura) della struttura, privandola delle parti inutili e/o dannose al funzionamento in termini decisionali. Il metodo più comune è quello proposto all'interno della metodologia CART, ancora oggi

comunemente utilizzato. Esso si basa su una misura detta di costo-complessità. Questo procedimento consiste nel provare a escludere dalla struttura tutte le sottobranche in cui l'albero è divisibile e calcola, per ognuna delle suddette branche, il costo-complessità (inteso come perdita informativa derivante dalla potatura) scegliendo quale punto di taglio il legame più debole (weakest link), ossia il punto che presenta il valore minore della misura considerata.

2.6 LE RETI NEURALI

Le reti neurali sono una famiglia di tecniche statistiche che "prendono a prestito" l'idea del funzionamento delle strutture cerebrali degli organismi viventi, intese come insieme di collegamenti neuronali che assolvono determinati compiti. Le reti neurali artificiali "imitano" questo schema di funzionamento, costituito da migliaia di neuroni, paragonabili a nodi di una rete. Questa rete presenta tantissime interconnessioni tra nodi che possono essere di natura differente: lo schema tipico, infatti, prevede la presenza di nodi di input, nodi di output e nodi intermedi. I primi hanno il ruolo di fornire alla struttura i dati da elaborare (una sorta di "sensori" di un sistema nervoso); i nodi intermedi (detti anche nascosti, hidden), quelli più numerosi e caratterizzati dal maggior numero di connessioni, costituiscono il cuore della rete, ossia la parte che elabora i dati ricevuti in input dal primo tipo di nodi. Infine vi sono i nodi di output, che hanno il ruolo di comunicare all'esterno i risultati del calcolo che la struttura ha effettuato. Concettualmente il sistema si configura come una black-box nella quale il procedimento non è esplicitato, ma forniti in input una serie di dati si riceve in uscita un output.

I neuroni di questa rete artificiale si caratterizzano per la capacità di assumere più stati. A seconda del numero e delle caratteristiche di questi stati distinguiamo:

1. I neuroni binari
2. I neuroni bipolari
3. I neuroni continui

Il primo tipo può assumere lo stato 0 o lo stato 1. Il secondo gli stati -1 e 1, mentre il terzo tipo una gamma continua di stati che di solito va da 0 a 1. Lo stato S(t) di ogni tipo di neurone evolve col passare del tempo (t) misurato in intervalli discreti.

Ogni neurone riceve una serie n di input, ciascuno caratterizzato da un proprio peso, detto peso sinaptico. Talora, l'input comprende la presenza di una soglia il cui effetto può essere anche eliminato inserendo un output fittizio che algebricamente annulla la soglia stessa. L'input complessivo di un nodo può essere formalizzato come segue:

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

(2.17)

Il passaggio allo stadio successivo avviene invece per effetto di una funzione di trasferimento. Sintetizzando, dunque, il neurone artificiale non è altro che un operatore che riceve una serie di input, ne calcola la somma "pesata" e

restituisce in output il risultato della sua legge di attivazione computato sulla base degli input ricevuti.

Lo schema di funzionamento appena esposta mostra come lo schema delle reti neurali risulti molto flessibile: dato un certo numero di input ed esplicitando un sistema di pesi, basta stabilire la funzione di trasferimento più opportuna per ottenere il risultato cercato. Si noti inoltre che la funzione di attivazione può essere indifferentemente di tipo lineare o non lineare. Lo stato successivo avviene tramite una legge di attivazione, cioè calcolando un'apposita funzione, detta funzione di trasferimento.

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

(2.18)

Le funzioni di trasferimento possono essere di vari tipi, tra i quali si ricorda:

1. Le funzioni identità (lineari senza saturazione)
2. Le funzioni lineari con saturazione
3. Le funzioni a gradino (heavyside)
4. Le funzioni logistiche

Tralasciando quelle del primo tipo, nelle quali dato un input p la funzione di trasferimento è semplicemente

$$F(p) = p \text{ oppure } F(p) = K \cdot p$$

(2.19)

Di solito, però, si usano funzioni più complesse, sulla natura delle quali occorre soffermarsi. Le funzioni lineari con saturazione restituiscono come valore il minimo tra i massimi della soglia S e della quantità K·p.

$$F(P) = \min(\max(S),\max(0,KP))$$

(2.20)

Le funzioni a gradino funzionano in maniera diversa a seconda che si abbiano in input valori binari (0,1) o bipolari (-1,+1). Nel primo caso, infatti, si ha:

$$F(P) = 0 \text{ per } P \leq 0 \text{ ; } F(P) = 1 \text{ per } P > 0;$$

(2.21)

Nel secondo invece:

$$F(P) = -1 \text{ per } P \leq 0 \text{ ; } F(P) = 1 \text{ per } P > 0.$$

(2.22)

Infine, le funzioni logistiche si esprimono come:

$$F(P) = \frac{1}{1+e^{kp}}$$

(2.23)

Tali funzioni possono assumere valori tra 0 e 1. Inoltre, rispecchiando molto il funzionamento dei veri neuroni, sono quelle più usate. Di solito, il procedimento parte assegnando a tutti i neuroni della rete un certo peso, compreso tra 0 e 1 o tra -0,5 e +0,5 a seconda del tipo di applicazione. Sono frequenti anche le situazioni in cui la rete apprende da sola quale sia lo schema di pesi ottimale per rispondere nel modo desiderato alle stimolazioni in input. Concretamente, si utilizzano le tecniche proprie dell'apprendimento supervisionato, che prevedono la distinzione del set di dati a disposizione in due sotto-campioni: il *training set* e il *validation set*. Il primo viene usato per l'appunto per scopi di apprendimento mentre con l'altra parte del dataset si verifica che lo schema di rete non sia adatto in maniera soddisfacente esclusivamente alle casistiche che sono servite alla costituzione della rete stessa, bensì che questa sia in grado di produrre risultati appropriati anche di fronte a nuovi dati, denotando così attitudini di stabilità e generalità.
Durante l'apprendimento la rete, dato un input X restituisce un output Y che si discosta più o meno marcatamente da un ottimo D chiamato output desiderato. Una misura atta a quantificare tale differenza è l'errore quadratico sui casi considerati:

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

(2.24)

Non è difficile comprendere come il processo di implementazione della Rete definitiva deve tendere alla minimizzazione di tale errore. Allo scopo, attraverso algoritmi di apprendimento si provvede alla modifica dei pesi

sinaptici mediante progressive variazioni in aumento o in decremento del valore degli stessi.

2.6.1 Lo schema di Rete

La configurazione di Rete più utilizzata nelle applicazioni pratiche è la cosiddetta multilayer perceptron. Essa si basa su di uno schema orientato, poichè esiste un verso di funzionamento, dall'input all'output (feed-forward). Inoltre, esistono più strati nascosti che separano l'input dall'output e la Rete è interamente interconnessa.

Come evidenziato precedentemente, ogni passaggio è caratterizzato dall'elaborazione di più input che, mediante un sistema di pesi, contribuiscono in maniera più o meno determinante alla formazione dell'output.

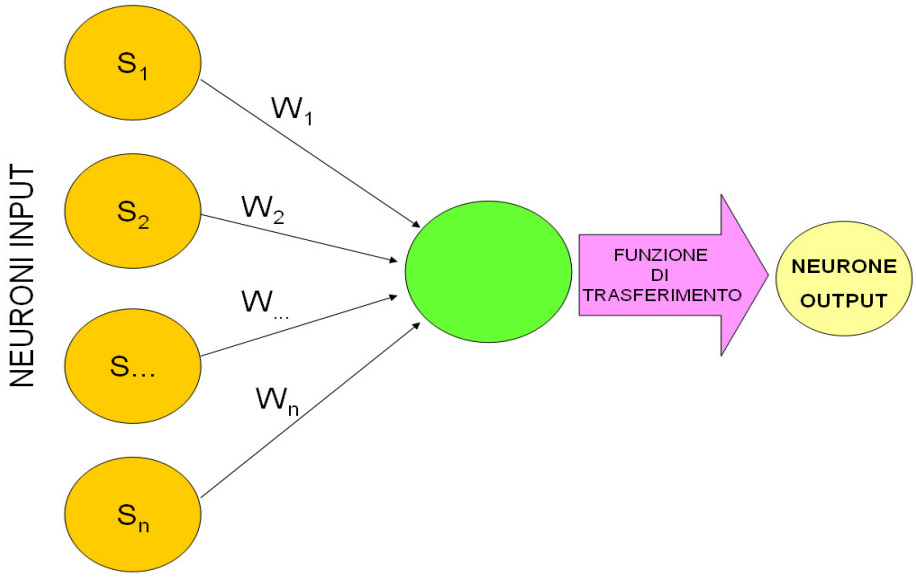

Figura 2.2 Schema esemplificativo di rete neurale

Si noti che nella figura precedente, a scopo esemplificativo, si è illustrata una rete che possiede un solo neurone nascosto che elabora gli *n* input S pesandoli attraverso i pesi W. In questo caso la funzione di trasferimento che porta alla definizione dell'output è una sola, ma potrebbe non essere così in quanto è possibile combinare o utilizzare diverse funzioni per i diversi nodi. Nella prassi però, anche in presenza di più neuroni intermedi, quindi di reti notevolmente più complesse, si preferisce utilizzare una sola funzione di trasferimento (in genere la sigmoide) che restituirà i vari output.

Inoltre, mentre in genere il numero di input e quello di output è più facilmente individuabile a partire dalle caratteristiche e dagli scopi dell'applicazione, di più difficile identificazione è il numero di neuroni intermedi. Non è raro, perciò, che uno schema come quello proposto nella figura precedente, ossia con un solo neurone *hidden* rappresenti una scelta applicativa semplice ed efficiente.

Ciò è dovuto soprattutto alla necessità di ricercare un compromesso ottimale tra le necessità di generalità e quella di apprendimento. Se si privilegia la fase di apprendimento infatti, si rischia di costruire una rete complessa che restituisce buone prestazioni in riferimento al training set considerato, ma è scarsamente generalizzabile. Per contro, se si costruisce una rete troppo semplice e "generica", essa verosimilmente necessiterà di una lunga fase di apprendimento e di ottimizzazione.

2.6.2 Un algoritmo di apprendimento: il back propagation

La diffusione del già citato schema multilayer perceptron è dovuta anche all'introduzione di un algoritmo di apprendimento semplice ma molto potente, chiamato *error back propagation*, che indicheremo anche con la sigla abbreviativa BP. Il principio alla base di tale metodo è di minimizzare l'errore quadratico medio relativo al training set in esame. Il back propagation rientra nella categoria dei metodi di apprendimento supervisionato, in quanto sfrutta informazioni provenienti da casistiche analoghe a disposizione del ricercatore.

Il funzionamento di tale algoritmo è sintetizzabile in tre fasi principali:

1. Feed Forward: è il primo passo, e consiste nella emanazione di un "segnale", cioè di un input che passa per tutta la rete arrivando alla fine, punto in cui la rete stessa restituisce l'output;

2. Error Back Propagation: a questo punto viene calcolato l'errore, attraverso la fase detta di retro-propagazione, nella quale viene misurata la differenza tra l'output restituita dalla rete e un target che viene trasmesso alla rete medesima;

3. Weight Update: è l'ultima fase, nella quale sulla base degli step precedenti si aggiornano i pesi che caratterizzano la rete in maniera da rendere quanto più possibile gli output simili ai target, ovvero i risultati ottimali a cui tendere.

Questo metodo è semplice ed efficace, e per questo motivo costituisce un riferimento nel campo della ricerca sulle reti neurali. Tuttavia, al crescere delle dimensioni della rete, anche un procedimento lineare e "snello" come

questo, diviene nella risoluzione di problemi pratici assai lento e complesso dal punto di vista computazionale. L'algoritmo infatti giunge a convergere molto lentamente, e le sue prestazioni diminuiscono rapidamente (e in maniera più che proporzionale) all'aumentare delle dimensioni della rete.

Per questi motivi la dottrina ha proposto molti correttivi a questa procedura, sostanzialmente riconducibili a due approcci: il primo consiste nell'aggiungere informazioni "esterne" relative all'errore (conoscenze di secondo ordine), mentre il secondo è caratterizzato dal calcolo di learning rate variabili in base a diverse situazioni di errore.

2.6.3 Le reti di Kohonen

Appartengono invece alla famiglia dei metodi di apprendimento non supervisionato le reti di Kohonen, dette anche Self Organizing Maps. Come l'ultimo termine suggerisce, queste reti hanno l'interessante capacità di auto-organizzarsi sfruttando dei principi simili a quelli che governano il funzionamento del cervello delle creature viventi. L'assonanza con la neuro-biologia è dovuta principalmente alla riproposizione dello schema cerebrale nel quale si possono distinguere aree neuronali deputate a determinate funzioni o alla ricezione di determinati stimoli. In tali aree vige il principio che neuroni tra loro vicini reagiscono in maniera analoga quando sollecitati da stimoli.

In più, un punto di contatto fondamentale con le applicazioni statistiche consiste nella capacità che tali reti hanno di proiettare input

multidimensionali sulla rete, che a sua volta è rappresentabile in due dimensioni.

Lo schema delle reti di Kohonen prevede, come quella precedentemente esposta, la presenza di una serie di neuroni di input che, attraverso un sistema di pesi, concorrono alla definizione di un output. La differenza principale sta nei neuroni che calcolano tale output: essi sono di tipo bidimensionale, e sono connessi a tutti i neuroni di input. L'apprendimento avviene attraverso i collegamenti laterali tra i neuroni vicini.

L'algoritmo tipo di queste reti è schematizzabile come segue:

1. Sia $w_{i,j}$ l'insieme di pesi assegnati in corrispondenza di ciascuna coppia tra neurone di input *i* e neurone di output *j* all'istante *t*;
2. Si definiscono $x_i(t)$ input (con i =1, 2,.....,n), nonchè il numero di neuroni $N_i(t)$ vicini al j-esimo neurone;
3. Si calcolano le distanze euclidee tra l'input e ciascun neurone di output j: $d_j^2 = \left[\sum_{i=1}^{n-1} \left(x_i(t) - w_{i,j}(t) \right)^2 \right]^{1/2}$;
4. Si seleziona il neurone *j** a cui corrisponde la distanza minima *d**. I neuroni sono dunque tra loro competitivi (*competitive learning*);
5. Si modificano i pesi al neurone *j** e a tutti i suoi vicini $N_i(t)$. In nuovi pesi si ricavano a partire dalla formula:

$$w_{i,j}(t+1) = w_{i,j} + \eta(t)[x_i(t) - w_{i,j}]$$

$\eta(t)$ rappresenta il learning rate, valore compreso tra 0 e 1. Inizialmente tale grandezza è fissata a 1 per poi regredire (attraverso una funzione lineare o non lineare) in modo da adattare i pesi dapprima in modo netto per poi affinarli in maniera più lieve e precisa. Anche il vicinato N viene all'inizio privilegiato includendo il più alto numero possibile di neuroni per favorire la cooperazione tra essi. Procedendo con le iterazioni dell'algoritmo si riduce la "popolazione" del vicinato, facendo in modo che i neuroni si comportino sempre più selettivamente, reagendo in maniera progressivamente diversa rispetto alla situazione iniziale che vede tutti i neuroni vicini reagire agli stimoli in maniera simile.

Una volta aggiornati i pesi il ciclo si ripete, ovviamente ad eccezione della prima fase di assegnazione di un valore iniziale ai pesi.

Questo tipo di algoritmo è anch'esso piuttosto semplice, ma le sue caratteristiche lo rendono capace di auto-organizzarsi, peculiarità che lo rende adatto alle generalizzazioni, ossia ai procedimenti che gli permettono di funzionare correttamente anche al termine della fase di costruzione, quando cioè nella rete entrano nuovi dati. In pratica, l'adattamento avviene in maniera automatica poiché prendendo il neurone i cui pesi più vicini all'insieme di input si uniformano tali pesi agli input ma soprattutto si uniformano anche i pesi dei neuroni etichettati come vicini. Ciò fa si che un'intera regione di neuroni sia "allenata" da un set di input di apprendimento, reagendo più efficacemente all'ingresso di nuovi dati.

CAPITOLO III

Strategie di analisi statistica a supporto dell'integrazione di Web Usage e Web Structure Mining

3.1 USAGE MINING E STRUCTURE MINING: DIFFERENZE E PUNTI DI CONTATTO

Nel primo capitolo si è introdotto il concetto di Web Mining, definendolo come il processo di Data Mining applicato al contesto del Web. All'interno di questo campo la dottrina distingue, tradizionalmente, tra tre differenti branche: Web Content Mining, Web Usage Mining e Web Structure Mining. Tralasciando il Content Mining, che non sarà trattato nel presente lavoro, ci si soffermerà sugli ambiti del Web Usage Mining (che indicheremo anche con l'acronimo WUM) e sul Web Structure Mining (WSM).

Si è detto inoltre che la "sigla" WUM indica l'insieme delle applicazioni volte all'identificazione delle abitudini di uso di uno o più siti web oggetto di

studio. Il termine Web Structure Mining indica invece i task di Data Mining che hanno come obiettivo la descrizione della struttura di un sito Web o di una porzione di Rete.

L'approccio tradizionale pone questi due ambiti come piuttosto differenziati e separati, dato che scopi e strumenti delle analisi sono abbastanza diversi. Diverse sono anche le fonti dei dati: nel Web Usage Mining si fa ricorso a tutto ciò che è in grado di ricostruire il profilo d'uso degli utenti di un sito. Concretamente si utilizzano i log-files e le informazioni socio-demografiche derivanti da form di registrazione e campi similari. Le analisi di Web Structure Mining partono invece dalla composizione delle pagine in termini di connessioni (link) in entrata e in uscita per compiere una valutazione circa la struttura del sito, la sua "posizione" nella Rete, i punti notevoli della struttura e quelli più isolati e/o poco raggiungibili. Diverse sono, come è facile intuire, anche le finalità per le quali si svolgono le due indagini: nel caso del WUM si è interessati a capire come si naviga sul sito analizzato: per quanto tempo, visitando quali pagine in quale sequenza, osservando secondo quali dinamiche si scarica un contenuto o si acquista un prodotto. In un contesto di WSM, invece, si studia il posizionamento del sito nella Rete, le connessioni in entrata e in uscita che esso possiede.

Anche all'interno di queste due categorie si identificano delle differenziazioni di scopo. Non è difficile comprendere, infatti, come lo studio del comportamento degli utenti (Usage) di un sito Web sia un concetto ampio, che ricomprende al suo interno una molteplicità di applicazioni pratiche.

In particolare, tali applicazioni sono riconducibili a cinque categorie:

1. Personalization: lo studio del comportamento dei navigatori viene utilizzato per proporre agli stessi dei contenuti compatibili coi loro gusti e le loro abitudini di consultazione;
2. System Improvement: le abitudini di uso servono anche per ottimizzare le prestazioni del sito, con particolare riferimento ai tempi di connessione/caricamento;
3. Business Intelligence: le evidenze relative ai comportamenti di visita fungono da input ai processi di Business Intelligence, quali ad esempio gli studi di Customer Attraction e Retention;
4. Usage Characterization: è probabilmente il task di matrice più statistica, poichè attiene alla segmentazione e alla profilazione degli utenti del sito;
5. Site Modification: è il task che presenta più punti di contatto col Web Structure Mining, poichè i dati relativi alle abitudini di navigazione vengono usati per razionalizzare la struttura del sito. E' uno degli ambiti di riferimento del presente lavoro, e per questo sarà ripreso nel corso della trattazione;

Anche col termine Web Structure Mining si indicano, in realtà, una serie di applicazioni differenti tra loro. Di seguito elenchiamo le principali, appartenenti al livello di analisi Hyperlink (si veda il cap.1):

1. Determinazione della qualità di una pagina: insieme di analisi che mira a restituire punteggi, ranking o indicazioni della pertinenza della pagina rispetto a un determinato topic;

2. Classificazione delle pagine: simili ad alcuni task di content mining, questo tipo di studi stabiliscono quali pagine "marcare" nel processo di crawling, e più in generale a identificare il contenuto delle pagine;
3. Identificazione di strutture Web interessanti: rientrano in questa categoria, le applicazioni attraverso le quali si scoprono pagine duplicate, o strutture Web complesse come le pagine popolari;

Queste distinzioni concettuali e applicative confermano e legittimano la distinzione che la dottrina individua tra i due contesti di analisi (tre considerando il Content Mining qui tralasciato).

In realtà, però, il Web Usage Mining e il Web Structure Mining hanno anche punti di contatto; non è difficile comprendere, ad esempio, come la struttura influenzi l'uso di un sito Web e come possa verificarsi anche l'ipotesi inversa (ovviamente, nel secondo caso, per effetto dell'intervento del Webmaster). A supporto di tale visione si pongono i lavori di Srivastava (2003) nei quali l'autore identifica quale evoluzione dello studio dei dati Web la combinazione dei task propri del Web Usage e del Web Structure Mining.

Questo nuovo tipo di applicazioni prende il nome di Web Process Mining, e rappresenta un modo più ampio ed eterogeneo di intendere un'applicazione di Web Mining. Come anticipato, infatti, con tale termine si identifica un filone di studi "crossover", a cavallo cioè tra WUM e WSM, che sfrutta peculiarità e vantaggi di entrambi gli ambiti per ottenere una visione più ampia e dettagliata del sito web oggetto di indagine. Più precisamente, col termine Web Process Mining si fa riferimento alle applicazioni di Data Mining nelle quali i dati Web (quelli tipicamente utilizzati in maniera

separata nelle analisi di Usage e di Structure) sono analizzati per estrapolare dei modelli di processo di un sito.

L'output tipico di questo genere di indagini è una descrizione del sito focalizzata sui cosiddetti Process Outcome Data, ovvero gli stati finali di visita (tipicamente le pagine di uscita). L'analisi del processo, inteso come percorso seguito dagli utenti in base a una struttura prefissata, permette agli amministratori del sito di comprenderne le dinamiche di uscita e, in base a queste, implementare strategie tese a scoraggiare o incoraggiare determinati fenomeni, tipicamente di abbandono o di acquisto. Gli outcome, cioè gli stati finali del processo di navigazione, sono generati, come è facile intuire, da sequenze di pagine le quali hanno ciascuna una propria probabilità di abbandono. La probabilità di abbandono può essere riferita ad un abbandono "desiderato", nel senso che l'outcome che genera è l'acquisto di un prodotto, il download di un contenuto ecc., oppure "indesiderato" quando l'outcome prodotto rappresenta un abbandono, un'uscita pura senza che l'utente abbia raggiunto alcuna posizione target all'interno del sito in questione.

Quello del Web Process Mining è un esempio tipico della concezione "unificata" delle due branche (WUM e WSM) che la dottrina teneva tradizionalmente separate all'interno della famiglia del Web Mining. Tale approccio sfrutta i dati dell'Usage proiettati sulla struttura (Sctructure) per quantificare le probabilità di abbandono e quelle di passaggio (traverse) da uno stato (pagina) all'altro tramite link. Graficamente il concetto può essere rappresentato come in figura:

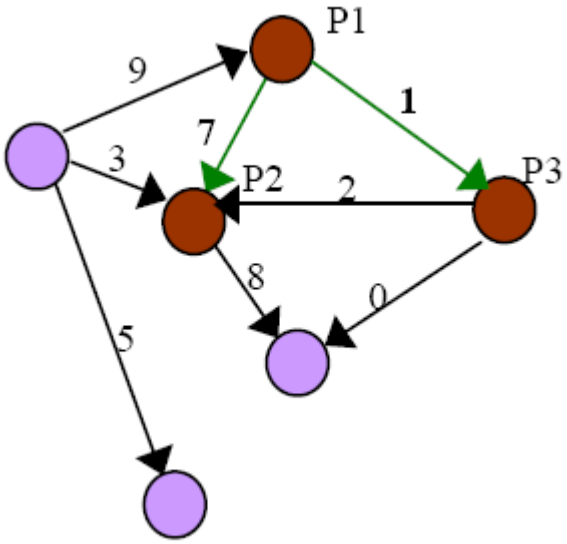

Figura 3.1 Process Mining di una struttura Web

Lo schema mostra un sito (o una sua parte) composto da 6 pagine, i cerchi, e da 8 link, le frecce orientate. Le cifre al di sopra di queste ultime rappresentano il numero di volte in cui tali link vengono sfruttati per passare alla pagina di destinazione. In questo esempio quindi la probabilità di passaggio dalla pagina 1 alla 2 sarà:

$$P_t(P_1 \Rightarrow P_2) = \frac{7}{7+1} = \frac{7}{8}$$

(3.1)

mentre la transition probability dalla pagina 1 alla 3 sarà:

$$P_t(P_1 \Rightarrow P_3) = \frac{1}{7+1} = \frac{1}{8}$$

(3.2)

Questo tipo di probabilità è comunemente utilizzata per calcolare una misura, detta Link prediction, nell'ambito della progettazione di siti web

adattivi[1], oppure per determinare la qualità di una pagina, obiettivo comune anche a indagini di Web Structure Mining.

Il presente lavoro segue la medesima filosofia dei questi studi, cioè quella di far confluire sia le fonti dei dati che gli obiettivi dell'analisi propri del Web Usage e del Web Structure Mining al fine di ottenere una visione più ampia e completa del fenomeno. Come già accennato, l'uso e la struttura di un sito Web possono essere considerate e analizzate in maniera separata essendo concettualmente differenti. Per obiettivi molto precisi e circoscritti, dunque, è consigliabile agire secondo i canoni del Web Usage Mining o del Web Structure Mining a seconda dei casi: per altri tipi di indagine, di più ampio spettro, l'opinione di chi scrive pende verso l'approccio integrato quale quello che sarà presentato a breve o quello pocanzi accennato, poichè presenta dei vantaggi rispetto all'esecuzione dei due task in maniera consecutiva e separata.

[1] Adaptive Web sites, siti dinamici "intelligenti" che si comportano in maniera diversa a seconda degli stili di navigazione. Appartengono a questa categoria i siti progettati in modo da proporre argomenti correlati a quelli scelti (la proposta del contenuto è computata di volta in volta dal sistema che osserva le pagine visitate in precedenza); molto simili sono anche i Recommendation Systems, cioè strumenti integrati all'interno dei siti che provvedono a raccomandare un prodotto abbinato all'acquisto appena effettuato o comunque correlato con un prodotto visualizzato. I sistemi più evoluti sono inoltre in grado di basare la proposizione del prodotto non solo in base all'utente presente, ma sono in grado di associare quest'ultimo a una categoria precedentemente profilata, raccomandandogli così la scelta effettuata da utenti dello stesso segmento.

3.2 USAGE E STRUCTURE MINING: UN APPROCCIO DESCRITTIVO INTEGRATO

Nell'approccio tradizionale (WUM) si parte dall'analisi dei log file e dalle informazioni provenienti dai form di registrazione per analizzare i profili utente e le caratteristiche di visita. Dunque informazioni sugli utenti per lo studio degli utenti. Parallelamente, nell'ambito del WSM si è interessati, partendo da una struttura Web nota, a analizzare le peculiarità di tale struttura: un'indagine sulla struttura a partire da dati (ad esempio lo schema dei link) di struttura.

Il nostro approccio, al contrario, cerca di avvicinare i due ambiti di analisi al fine di comprendere il funzionamento della struttura attraverso quanto espresso dagli utenti che navigano nella struttura/sito. La struttura del sito vista attraverso gli occhi dei suoi utenti.

Concretamente, si utilizza una fonte dei dati propria dei processi di Web Usage Mining, vale a dire le sequenze di visita di un sito Web. Esse vengono usate quale modalità esplicativa del funzionamento e della struttura del sito, col principio che le pagine frequentato solitamente assieme dovrebbero essere considerate in qualche modo come simili, e più in generale la similarità dovrebbe essere considerata nell'ottica di interventi sulla struttura (obiettivo tipico del Web Sctructure Mining) come ad esempio la riprogettazione dei link interni al sito.

3.3 UNO STRUMENTO PER LA RAPPRESENTAZIONE DELLA STRUTTURA WEB: IL MULTIDIMENSIONAL SCALING

Le informazioni derivanti dalle sequenze di visita degli utenti vengono utilizzate come input per un'analisi di Multidimesional Scaling.

Il Multidimensional Scaling è una tecnica usata per comprendere la sussistenza e la forza di relazioni di similarità/dissimilarità all'interno di un set di oggetti. Tale metodo è diffusamente adottato in quanto produce una rappresentazione grafica delle distanze fra gli oggetti in uno spazio geometrico, rappresentazione molto più comprensibile e immediata rispetto alla comprensione dei dati derivanti da un array di numeri.

In osservanza alla classificazione proposta da Borg e Groenen (1997), di seguito si riepilogano, in maniera schematica, i quattro ambiti di applicazione propri del Multidimensional Scaling (che sovente indicheremo con l'acronimo MDS):

1. Tecnica esplorativa: lo si usa quando non si dispone di evidenze a priori in grado di descrivere l'insieme di oggetti, pertanto le distanze e lo schema ottenuti forniscono utili indicazioni riguardo le differenze o le similarità tra le osservazioni in uno spazio di ridotta dimensionalità;
2. Test di ipotesi: si può far ricorso a questa tecnica anche per scopi confermativi rispetto a ipotesi avanzate riguardo il set di oggetti considerato;
3. Esplorazione e comprensione di strutture psicologiche: è l'ambito in cui questa tecnica si è sviluppata. Il concetto di somiglianza/similarità è assai importante in questi campi e l'uso dell'MDS è quasi imprescindibile in questo tipo di ricerche, soprattutto se si pensa che i giudizi degli individui riguardo un insieme di attributi rappresentano un esempio lampante di contesto multidimensionale. La soluzione proposta dal Multidimensional Scaling fornisce un contributo assai significativo alla quantificazione delle distanze fornite verbalmente attraverso i giudizi;

4. Modellazione dei giudizi di similarità: in questo tipo di indagini si mira, attraverso la modellizzazione, a comprendere la struttura latente rispetto ai giudizi di (dis)similarità;

I modelli di Multidimensional Scaling sono definiti specificando come, segnatamente attraverso quale funzione di trasformazione, i dati rappresentati dalle (dis)similarità tra gli oggetti diventano distanze in una configurazione MDS *X* di dimensionalità *m*.

$$f : p_{ij} \rightarrow d_{ij}(\mathrm{X})$$

(3.3)

La *f* della formula precedente è una funzione di rappresentazione attraverso la quale si specifica la relazione tra le prossimità *p* e le distanze *d*. La scelta del particolare tipo di funzione *f* equivale a scegliere il tipo di modello di MDS utilizzato. Tale funzione, dunque, date in input le prossimità *p* restituisce attraverso una trasformazione le distanze *d* che compongono una configurazione *X*.

Esiste una differenza fondamentale all'interno delle tecniche MDS. Si distingue, infatti, tra MDS metrico e non metrico. La prima famiglia di modelli presuppone una relazione lineare tra le distanze e le prossimità (dissimilarità) tra gli items. Questa relazione può essere di identità *(Absolute MDS)*: $d_{i,j} = p_{i,j}$. Può essere altresì frutto, come appena evidenziato, di una trasformazione di tipo lineare:

- Interval (transformation) MDS: $a+bp_{ij}$, con b>0;
- Ratio (transformation) MDS: bp_{ij} , con b>0;
- Spline (transformation) MDS: somma di polinomi p_{ij}

Nel caso in cui si voglia adottare un modello metrico, bisogna considerare che occorre una matrice dati completa (non devono figurare valori mancanti) e simmetrica.

A Shepard e Kruskall si deve invece il modello detto non-metrico. Al contrario dei precedenti, in questo caso non è richiesto che l'indice di dissimilarità rispetti proprietà metriche; si considera infatti la prossimità come funzione monotona della distanza, per cui l'ordinamento delle distanze è riflesso dell'ordinamento tra le dissimilarità.

E'importante sottolineare come il Multidimensional Scaling non produca una soluzione "perfetta", bensì, attraverso una serie di iterazioni, produca una configurazione che approssima nel modo migliore le distanze osservate. Posto che esiste più di una soluzione, è necessario stabilire un criterio di errore, chiamato funzione di Stress; la configurazione ottimale è, ovviamente, quella che riduce il più possibile il valore risultante da tale funzione. Una prima quantificazione dell'errore insito nel posizionamento di una coppia di oggetti è data dalla seguente:

$$e_{ij}^2 = \left[f\left(p_{ij}\right) - d_{ij}\left(\mathrm{X}\right) \right]^2$$

(3.4)

Tale formula, l'errore quadratico di rappresentazione, è per l'appunto il quadrato della differenza tra le prossimità ottenute e le distanze reali per una data coppia *i,j*. Sommando e_{ij}^2 per ogni coppia i,j si ottiene una misura assai nota in letteratura, il *raw stress.*

$$\sigma_r(X) = \sum_{i,j} \left[f\left(p_{ij}\right) - d_{ij}\left(\mathrm{X}\right) \right]^2$$

(3.5)

Chiaramente, essa fornisce una misura di scostamento della soluzione ottenuta rispetto alla reale distanza tra gli items considerati. Esiste anche una versione normalizzata di tale indice, ottenuta confrontando la precedente con la sommatoria dei quadrati delle distanze tra gli oggetti.

$$\sigma_1^2(X) = \frac{\sigma_r(X)}{\sum d_{ij}^2(X)}$$

(3.6)

Kruskall, invece, propose l'utilizzo della radice quadrata di tale misura, conosciuta come *Stress-1*.

$$Stress-1 = \sigma_1 = \sqrt{\frac{\sum \left[f\left(p_{ij}\right) - d_{ij}(\mathrm{X}) \right]^2}{\sum d_{ij}^2(X)}}$$

(3.7)

Il metodo proposto in questo lavoro si basa sull'uso di una soluzione di Multidimensional Scaling metrico atto a descrivere la struttura del portale Web oggetto di studio attraverso le dinamiche di uso del sito stesso. Il principio, che sarà trattato in dettaglio nel seguito, è quello di identificare come risulta la struttura, in termini di prossimità, "con gli occhi degli utenti". Inoltre, nella maggior parte delle applicazioni che sfruttano i principi propri del MDS, il modo in cui si reperiscono le opinioni degli utenti riguardo la differenza di due o più oggetti (nel nostro caso le pagine web che compongono il sito) è quello "diretto", vale a dire tramite modalità quali questionari, focus group ecc. nei quali l'utente dichiara in maniera esplicita la propria opinione, intesa nel nostro esempio come espressione della diversità tra le pagine.

Il nostro approccio invece propone l'utilizzo dei dati propri del Web Usage Mining, segnatamente l'insieme delle sequenze di visita degli utenti del sito: in questo senso, la matrice di (dis)similarità tra le pagine è formata da opinioni implicite (le sequenze di visita) in luogo di quelle esplicite solitamente utilizzate. Ecco, ancora una volta, il motivo per il quale si è ribadito precedentemente che il presente lavoro, dal punto di vista del Web Mining, si pone sulla scia degli studi di Process Mining e più generalmente è al limite tra l'Usage e lo Structure Mining.

La descrizione della struttura si completa poi attraverso un ulteriore elemento che arricchisce la portata informativa della soluzione adottata: le pagine sono infatti collegate tra loro non dal normale schema di link, bensì dall'output proveniente dalle evidenze delle regole associative. Tale descrizione non è esaustiva, nel senso che non contempla tutte le possibili direttrici di traffico rilevate dal server (ovvero tutte le sequenze verificabili in base allo schema di link che declina la struttura del sito). Ciò in ottemperanza a un vincolo metodologico e ad una scelta comunicativa: da un lato, infatti, il principio metodologico alla base delle association rules prevede che, tra tutte le possibili, l'algoritmo estragga solo quelle che rispettano dei vincoli prefissati.
Già questa circostanza, dunque, chiarisce il motivo dell'assenza della rappresentazione di tutti i patterns. Inoltre, si è stabilito di preservare la pulizia e l'interpretabilità dello schema proposto; pertanto, si è scelto di rappresentare solo le direttrici di traffico più "popolari", ovvero le sequenze maggiormente verificatesi nel lasso di tempo considerato. Per popolarità si intende il rispetto dei vincoli metodologici previsti nell'applicazione delle regole associative, vale a dire supporto e confidenza. Infine, a

completamento della descrizione di tali dinamiche, si è provveduto a inserire degli item "fantasma" a posteriori, vale a dire l'entrata e l'uscita dal sito, per verificare ulteriormente la prossimità tra alcune pagine e tali eventi.

L'impiego delle Regole Associative nel Web Mining è ormai consolidato, ragion per cui nel capitolo precedente si è dato ampio spazio all'illustrazione di tale metodologia. L'approccio seguito nel presente lavoro si ispira agli studi di Giudici e Blanc (2002) per la parte che riguarda l'uso di questa tecnica al contesto delle sequenze di visita a un sito Web. Gli autori infatti distinguono tra sequenze dirette e indirette. Con le prime indicano le sessioni in cui, date due pagine A e B, esse appaiono come consecutive; non vi sono, in altre parole, pagine "intermedie" consultate tra queste. In maniera residuale si ricava la definizione di sequenze indirette, i casi cioè in cui si visualizza assieme ma non consecutivamente A e B. Inoltre, gli autori inseriscono nell'analisi un item fittizio proprio al fine di valutare la sequenzialità diretta tra l'evento "Entrata" e la prima pagina effettivamente consultata. In aggiunta però il nostro approccio non si limita a un'analisi di questo tipo, bensì fornisce un'indicazione riguardante anche le similarità tra le pagine (dunque sulla struttura) evidenziata dalla soluzione di Multidimensional Scaling. L'output di questa tecnica funge anche da "mappa" del sito nella descrizione delle evidenze, contrariamente a Blanc e Giudici che la descrivono come un albero, come evidenziato nella seguente figura.

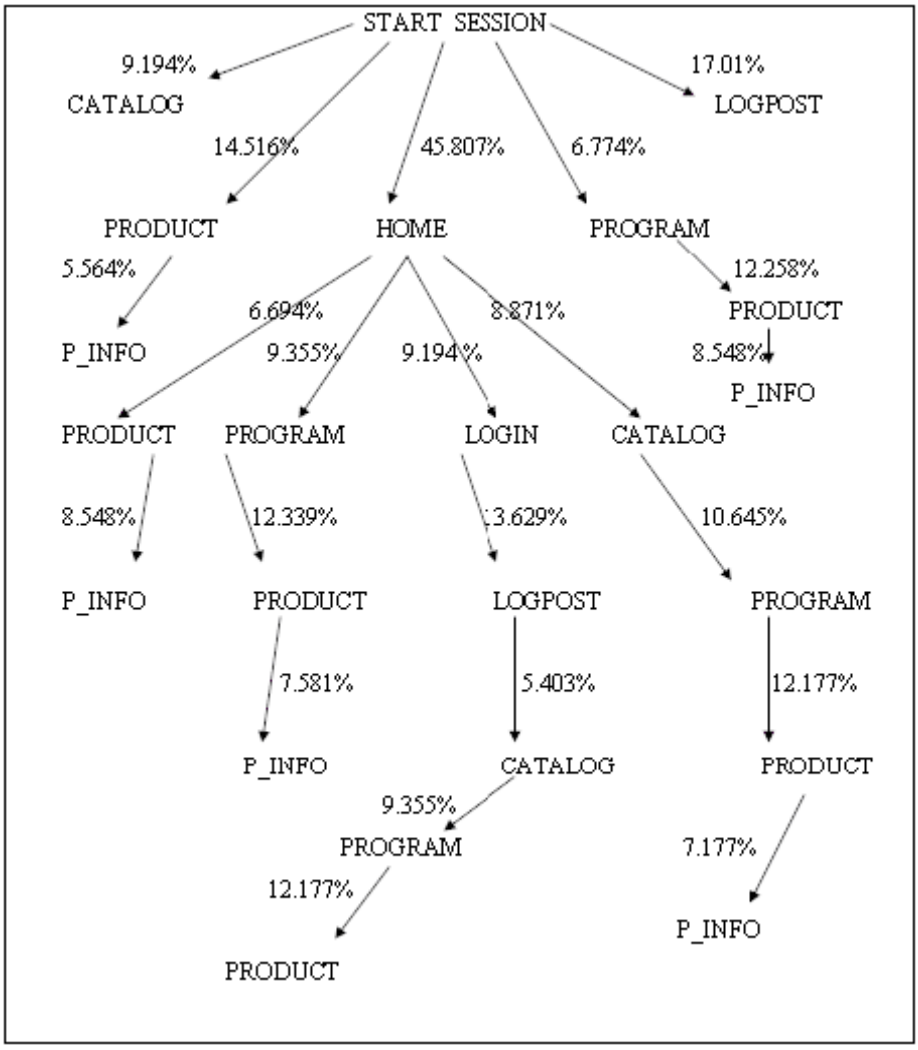

Figura 3.2 La descrizione grafica delle Sequence Rules (Blanc e Giudici, 2002)

3.4 LA DESCRIZIONE DELLA STRUTTURA DI UN PORTALE ATTRAVERSO LE SUE VISITE: IL CASO MSNBC.COM

In questo capitolo si esporranno i risultati concreti dei metodi proposti a livello teorico nel capitolo precedente. Il portale preso in esame è www.msnbc.com (più la sezione dedicata alle notizie del sito www.msn.com), il quale si compone di 17 sezioni che trattano argomenti di

informazione e di attualità di carattere piuttosto eterogeneo. Più precisamente tali sezioni sono: Front page, News, Tech, Local, Opinion, On-air, Misc, Weather, Msn-news, Health, Living, Business, Msnsports, Sports, Summary, BBS, Travel. I dati trattati sono stati collezionati nell'arco di una giornata, e si compongono di 989818 osservazioni che rappresentano le sequenze di visita degli utenti. Il livello di dettaglio è, per l'appunto, la singola sezione: pertanto, due pagine appartenenti alla stessa sezione sono etichettate all'interno del set di dati come due visite consecutive alla medesima area del portale. Le sessioni sono contrassegnate esclusivamente da un numero progressivo e sono, dunque, anonime.

Il dataset in oggetto è disponibile online sul Repository UCI Machine Learning (http://archive.ics.uci.edu/ml/).

La tabella seguente mostra lo schema originario del dataset in esame.

Come si può notare dallo schema, sulle righe sono riportate le sezioni visitate per ciascuna sessione, mentre ciascuna colonna indica un click della sequenza di visita presa in considerazione. A titolo di esempio si faccia riferimento alla prima sessione: l'utente che la ha generata è entrato sul sito attraverso la Frontpage. Successivamente, al secondo click, è passato alla sezione Sports. In seguito ha visitato due pagine appartenenti entrambe alla categoria delle News (terzo e quarto click). Infine, l'utente ha visitato la sezione Weather, abbandonando il portale proprio da tale punto del sito.

Tabella 1: Il dataset originario

Session	First section	Second section	Third section	Fourth section	Fifth section	Sixth section	…
1	Frontpage	Sports	News	News	Weather		
2	Frontpage	Opinion	Local	Tech	Opinion	Opinion	Living
3	Weather	Travel	Tech				
4	News	News	News	Local	On-air	Frontpage	
5	BBS	Travel	Business	Travel	Living	Living	Living
6	Frontpage	Sports	Local	Sports	News	Opinion	
…	…	…	…	…	…	…	…

Si noti che questo modo di rappresentare le sessioni di visita comporta, quale primo effetto, la presenza di numerose celle vuote all'interno della tabella-dati. Ciò, ovviamente, perchè le sessioni di visita non hanno tutte la stessa lunghezza. Nell'esempio preso in esame in questo lavoro, vi erano sessioni costituite da una visita ad un'unica sezione e sessioni formate da più di 50 click.

3.4.1 Il Pre-processing dei dati

L'organizzazione dei dati esplicitata pocanzi ha reso indispensabile uno step di pre-preprocessing dei dati stessi, in quanto così come si presentavano non erano utilizzabili quali input per un'analisi di Multidimensional Scaling (ricordiamo che per questi modelli occorrono matrici senza dati mancanti). Inoltre, l'estrema variabilità che caratterizzava la lunghezza delle sessioni che componevano il dataset ha determinato la scelta di considerare ai fini dell'analisi, esclusivamente i primi 50 click per ciascuna di tali sessioni.

La trasformazione scelta è la seguente: per ogni sessione di visita è stato contato il numero di volte in cui l'utente ha scelto ciascuna delle sezioni del sito. In questo modo si ottiene una matrice 989818 righe x 17 colonne. Le righe accolgono sempre le sessioni, mentre sulle colonne vi sono tutte le 17 sezioni in cui le pagine del sito sono state suddivise.

Tale trasformazione, che in effetti consiste nella computazione delle frequenze di visita per ciascuna sessione rende possibile il calcolo di una soluzione MDS utilizzando la distanza del chi-quadrato. Questa distanza è particolarmente adatta nei casi in cui i dati di partenza esprimono frequenze.

Le tabelle successive mostrano i risultati del passaggio di pre-processing. Il primo schema ripropone la situazione iniziale, mentre il secondo mostra l'organizzazione dei dati utilizzata come input per l'analisi di Multidimensional Scaling.

Tabella 2: Lo schema iniziale

Session	First Section visited	Second Section visited	Third Section visited	Fourth Section visited	Fifth Section visited	Sixth Section Visited	…
1	Frontpage	Sports	News	News	Weather		
2	Frontpage	Opinion	Local	Tech	Opinion	Opinion	Living

Tabella 3: L'organizzazione dei dati dopo il pre-processing

Session	Frontpage	News	Tech	Local	Opinion	On air	…
1	1	2	0	0	0	0	…
2	1	0	1	1	3	0	…

Tabella 4: Indicatori di adattamento della soluzione individuata

Stress and Fit Measures	
Normalized Raw Stress	0.029
Stress-I	0.171
Stress-II	0.313
D.A.F.	0.971
Tucker's Coefficient of Congruence	0.985

La soluzione trovata denota dei buoni risultati sia in termini di stress che di bontà di adattamento in genere. Si ricorda che lo mentre lo Stress dovrebbe essere quanto più possibile vicino allo zero, le altre misure di adattamento dovrebbero avvicinarsi a 1.

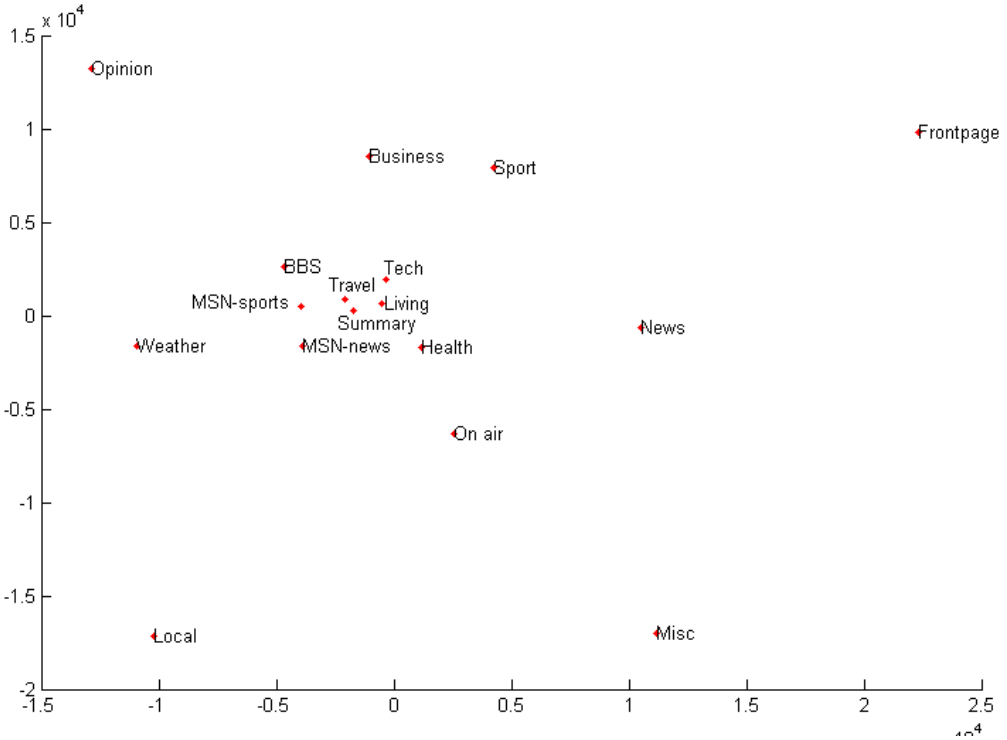

Figura 3.3 La soluzione MDS calcolata sull'intero Dataset

La figura 3.3 mostra la soluzione ottenuta a partire dall'intero set di dati.

Da tale configurazione si evince l'estrema differenziazione che caratterizza la prima pagina del sito (homepage), la sezione dell'informazione locale (Local) e di quella di opinione (Opinion). Distanti dal nucleo centrale risultano anche le pagine dedicate alle trasmissioni in diretta (On-air), all'informazione di carattere generale (News), al meteo (Weather) allo Sport e agli Affari (Business).

Questa configurazione rappresenta la mappatura "implicita" del portale, effettuata attraverso le opinioni osservate a partire dal comportamento dei navigatori e non direttamente riscontrate attraverso, ad esempio, questionari.

Tale soluzione, pur essendo metodologicamente corretta, rappresenta un risultato parziale, in quanto descrive in maniera alternativa la struttura del

sito. In un'ottica integrata, infatti, lo schema proposto non fornisce indicazioni riguardanti le relazioni e i collegamenti tra le pagine.

Per questo motivo la soluzione derivante dall'MDS viene successivamente arricchita da collegamenti tra pagine, che non sono i normali link tra le sezioni bensì derivano dai risultati dell'applicazione delle regole associative al dataset oggetto di studio.

In accordo con la visione di Giudici e Blanc si è provveduto ad estrarre le regole dirette e quelle indirette. Inoltre, per rendere più significativa l'analisi, si è scelto di restringere l'ambito di analisi alle sessioni che si componevano di non meno di 10 click, in maniera da riferire i risultati alle visite più approfondite.

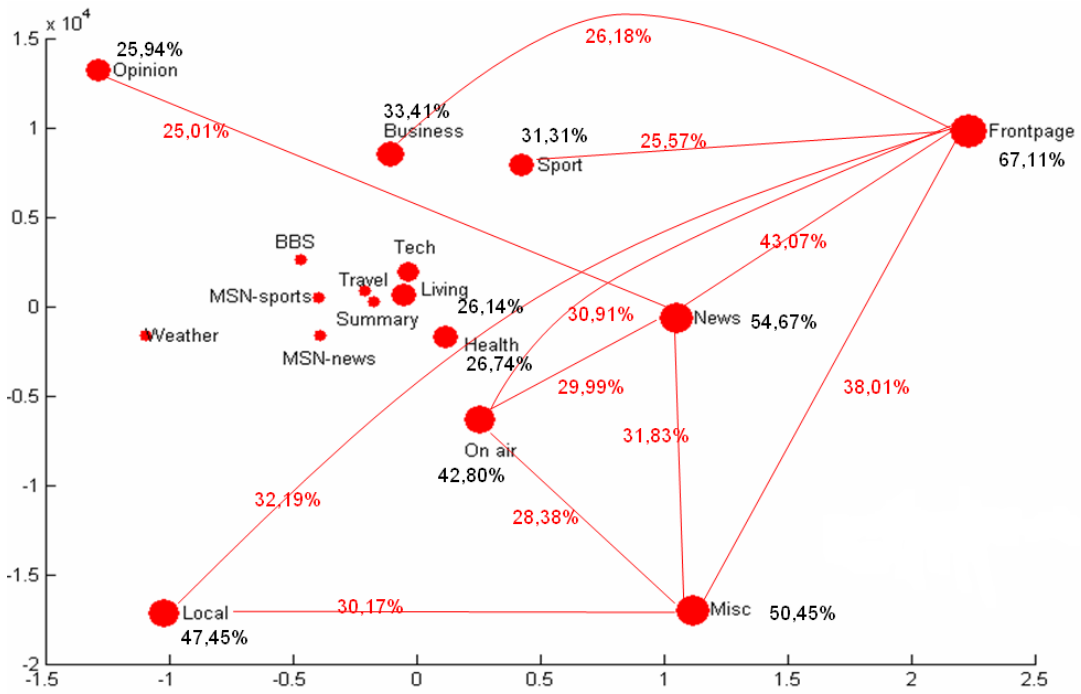

Figura 3.4 MDS e regole indirette (sessioni con almeno 10 click)

Nella rappresentazione mostrata nella figura 3.4 gli archi rappresentano i collegamenti più "forti" (dal punto di vista del supporto e della confidenza) mentre l'ampiezza dei punti rappresenta il supporto della singola sezione. Si noti come, eccezion fatta per la categoria Misc, le sezioni che risultavano maggiormente distanti rispetto alle altre appaiono ancora ben distanziate, formando categorie a se stanti.

E'facile comprendere, infine, come la pagina d'ingresso rappresenti lo snodo principale dal quale si diparte la maggior parte dei collegamenti del sito. Particolarmente forti risultano le associazioni tra la homepage e le categorie News, Misc e Local.

Seguendo questa impostazione si può aggiungere il livello di dettaglio desiderato, a seconda del tipo e della quantità di informazione che si desidera visualizzare.

L'altra strategia di analisi, quella che si basa sulle Direct Rules, è invece da preferirsi quando lo scopo applicativo necessita di una descrizione puntuale delle dinamiche di traffico. L'approccio delle Indirect Rules, per come è costruito, è ideale per sintetizzare la vasta mole di informazioni in "macro-direzioni" di traffico indicate dalle regole associative più forti, a prescindere dal fatto che le sezioni/pagine siano tra loro direttamente sequenziali.

Mediante lo studio delle Direct Rules invece, si traccia un profilo molto più dettagliato delle sequenze di visita. Contemporaneamente, però, la visualizzazione si complica notevolmente a causa della grande quantità di collegamenti da indicare.

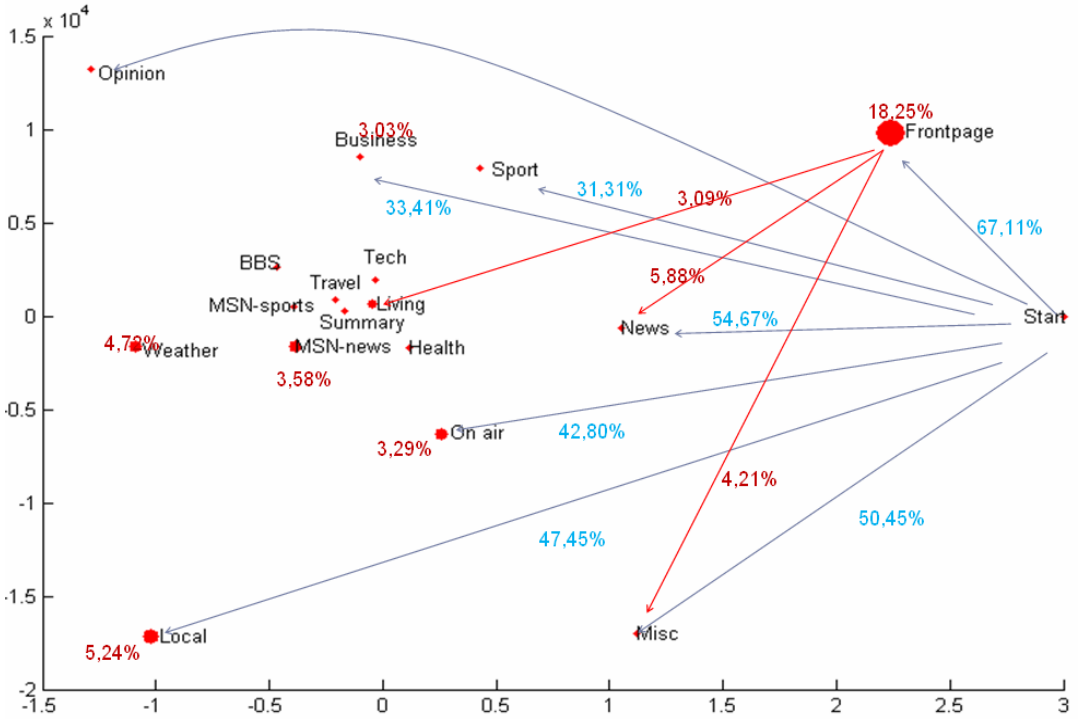

Figura 3.5 Rappresentazione delle relazioni più forti tra le sezioni (Direct Rules)

La figura precedente mostra le relazioni più evidenti tra le sezioni ottenute mediante le regole associative "dirette", vale a dire quelle che si possono ottenere a partire da item tra loro concettualmente adiacenti (nel caso delle applicazioni web la sequenzialità tra due item A e B, siano essi pagine, documenti o sezioni, è rappresentata dalla presenza di un collegamento che permette all'utente di passare direttamente, senza passaggi intermedi, dall'item A a quello B).

Si noti altresì che descrivendo le sequenze attraverso l'uso delle regole associative si torna a considerare il fattore temporale che caratterizza la visita, fatta di una sequenza di accadimenti che rispetta un certo ordine

temporale. Tale sequenzialità risultava infatti perduta dopo il passaggio di pre-processing servito a poter applicare il Multidimensional Scaling.

Inoltre, con questo approccio si perviene a una descrizione e a una rappresentazione della struttura e dell'uso di un sito Web anche laddove si ha a disposizone solo (come nell'esempio considerato) informazioni riguardanti le sequenze di visita.

Quest'ultima condizione rende ancora più profittevole l'adozione di tale strategia, poichè i dati relativi alle sessioni di visita costituiscono una base di cui di solito si dispone, e in più sono, per loro natura, informazioni aggiornabili con molta facilità e senza coinvolgere gli utenti o gli amministratori del sito.

3.5 LA PERMANENZA E L'ABBANDONO DI UN SITO: IL SEQUENCE TREE

Appartiene invece al filone di studi proprio del Web Usage Mining il secondo contributo applicativo descritto nel presente paragrafo. Anche questo metodo, come il precedente, verrà analizzato con l'ausilio di un caso studio proposto di seguito. Il topic è di fondamentale importanza nell'ambito del Marketing e dell'analisi delle performance di un sito Web. Si tratta infatti del monitoraggio della permanenza, e, di riflesso, dell'abbandono in termini di conclusione delle sessioni di visita .

Il contesto di riferimento è, più precisamente, la predizione della durata "in vita" della sessione di visita in termini di successivi click che saranno realizzati a partire dallo stato considerato.

La base di dati a partire dalla quale si illustrerà il metodo è la stessa usata per la strategia appena descritta, ovvero il dataset proveniente dal portale msnbc.com, contenente quasi un milione di sequenze di visita.

Per questa applicazione però si considereranno le sessioni la cui durata si esaurisce entro il decimo click, poichè si è osservato che oltre tale soglia avviene il 90% delle uscite. Tali osservazioni verranno trascurate poichè assai poco rappresentative del fenomeno della prosecuzione della visita; inoltre, risultava oltremodo complessa la loro trattazione poiché si è osservata anche una elevatissima variabilità della durata delle sessioni oltre il decimo click (basti pensare che vi erano sessioni che arrivavano a 50 click e più). La percentuale di sessioni escluse per questo motivo si attesta attorno al 10% dell'intera base di dati.

Innanzitutto, si è scelto di considerare le sessioni che si limitavano ad un massimo di 10 click. Le sessioni escluse rappresentano solo il 10% del dataset originario e la loro inclusione rappresentava un ostacolo interpretativo e realizzativo; si pensi, ad esempio, che vi era anche un numero assai limitato di sessioni che superavano i 50 click.

Tabella 5: Frequenze del numero di click per sessione

Numero click	Frequenza assoluta	Frequenza relativa	Frequenza relativa cumulata
1	365610	36,94%	36,94%
2	153644	15,52%	52,46%
3	97590	9,86%	62,32%
4	70157	7,09%	69,41%
5	56350	5,69%	75,10%
6	44130	4,46%	79,56%

7	34009	3,44%	82,99%
8	27169	2,74%	85,74%
9	21681	2,19%	87,93%
10	17716	1,79%	89,72%
11	14210	1,44%	91,15%
12	11954	1,21%	92,36%
13	9980	1,01%	93,37%
14	8522	0,86%	94,23%
15	7094	0,72%	94,95%
16	6102	0,62%	95,56%
17	5064	0,51%	96,08%
18	4287	0,43%	96,51%
19	3773	0,38%	96,89%
20	3422	0,35%	97,24%
21	2905	0,29%	97,53%
22	2597	0,26%	97,79%
23	2232	0,23%	98,02%
24	1997	0,20%	98,22%
25	1678	0,17%	98,39%
26	1576	0,16%	98,55%
27	1314	0,13%	98,68%
28	1265	0,13%	98,81%
29	1096	0,11%	98,92%
30	1027	0,10%	99,02%
31	892	0,09%	99,11%
32	737	0,07%	99,19%
33	675	0,07%	99,26%
34	625	0,06%	99,32%
35	564	0,06%	99,38%
36	513	0,05%	99,43%
37	476	0,05%	99,48%
38	411	0,04%	99,52%
39	391	0,04%	99,56%
40	365	0,04%	99,59%

41	347	0,04%	99,63%
42	341	0,03%	99,66%
43	272	0,03%	99,69%
44	266	0,03%	99,72%
45	236	0,02%	99,74%
46	217	0,02%	99,76%
47	226	0,02%	99,79%
48	204	0,02%	99,81%
49	173	0,02%	99,82%
50	170	0,02%	99,84%
51	144	0,01%	99,86%
52	156	0,02%	99,87%
53	122	0,01%	99,88%
54	110	0,01%	99,90%
55	122	0,01%	99,91%
56	121	0,01%	99,92%
57	106	0,01%	99,93%
58 e oltre	685	0,07%	100,00%

Nel capitolo precedente si è fatto riferimento alla segmentazione binaria quale strumento di Mining dei dati provenienti dal Web. Tale metodologia può essere applicata nella sua accezione classica per comprendere e/o prevedere il comportamento di una variabile di risposta in funzione di una serie di predittori. Nell'ambito del Web Mining la segmentazione ad albero viene utilizzata per capire quali sono le variabili che influenzano maggiormente l'acquisto di un bene nei siti di e-commerce, piuttosto che il download di un contenuto o la visualizzazione di una pagina target.

Le strutture ad albero sono ampiamente considerate in questo campo anche perchè si può tenere conto di variabili sia quantitative che qualitative in modo da arricchire la profondità dell'analisi.

L'algoritmo consiste sostanzialmente in una procedura iterativa che divide ad ogni passaggio il collettivo esaminato in partizioni binarie attraverso una variabile, detta di split, che è quella che, nel passaggio considerato, riesce a bipartire meglio il collettivo generando due nodi "figli" più puri del nodo padre e più puri di quelli che avrebbero generato tutti gli altri criteri di ripartizione.

3.5.1 Il metodo

La strategia proposta di seguito è una proposta di adattamento dell'algoritmo di segmentazione binaria al particolare caso dell'uso degli alberi per descrivere i processi di visita a un sito, con riferimento alla durata della visita espressa in termini di numero di click che in media l'utente compie prima di abbandonare il sito a partire dallo stato considerato.

Per chiarire questo concetto si faccia riferimento al dataset utilizzato per esplicare la procedura d'analisi descritta nel paragrafo 3.4.

Dopo le opportune trasformazioni, si focalizza l'analisi sulle sessioni composte da un massimo di 10 click; ogni click descrive l'accesso a una delle 17 sezioni declinate nel paragrafo 3.4.

A partire da questo set di dati si vuole comprendere, oltre alle pagine dalle quali si lascia più spesso il sito, quanti click in media l'utente compie nella sua visita al portale. La variabile di risposta considerata nell'albero di segmentazione sarà dunque il numero di click restanti a partire dal nodo (n-esimo click) considerato. La prima particolarità del metodo proposto è proprio questa: i nodi dell'albero non sono posizionati gerarchicamente dal normale procedere dell'algoritmo di segmentazione, bensì la gerarchia rispetta la sequenzialità tipica di una visita ad un sito. Al primo step, ad esempio, si prendono in considerazione, ai fini dell'individuazione del primo

split, le scelte effettuate dagli utenti al primo click, cioè la prima sezione del sito che hanno visitato. I due nodi figli che si formano saranno etichettati mediante il numero medio di click compiuti dagli utenti che non abbandonano il sito dopo la consultazione della prima sezione. Contemporaneamente, a destra dell'albero, si riporta la distribuzione per sezione delle uscite dopo il primo click; è logico, infatti, che coloro i quali non hanno proseguito oltre la propria visita, vengano esclusi dalla procedura. Inoltre, il computo dettagliato degli abbandoni fornisce importanti indicazioni circa le pagine che fungono più spesso da uscita dal portale considerato.

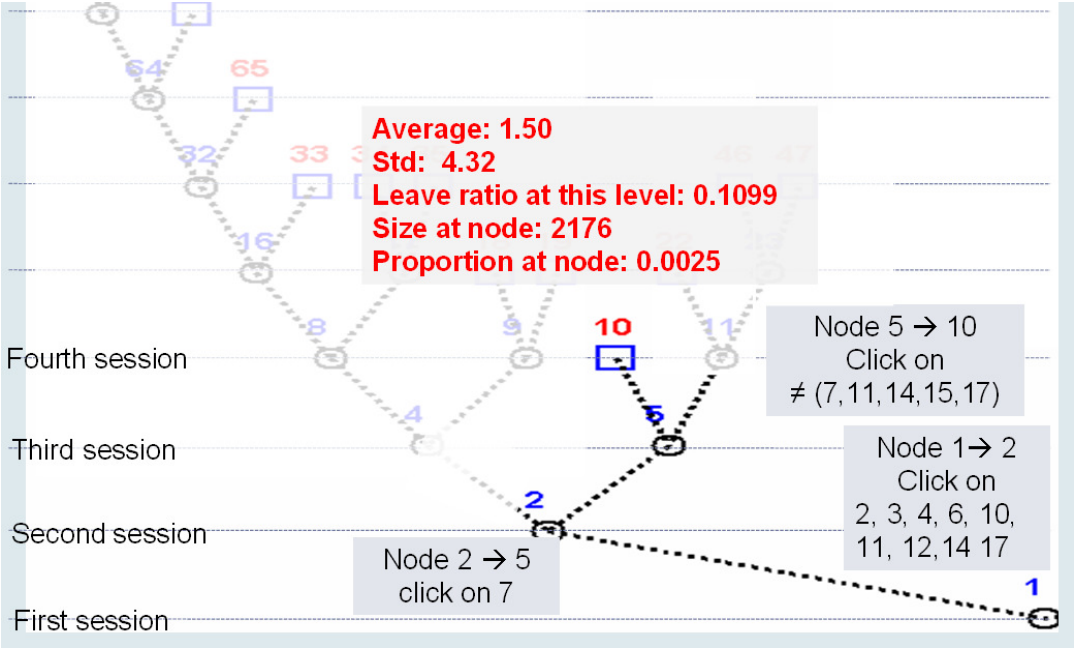

Figura 3.6 Dettaglio dell'output del sequence tree

La figura 3.6 chiarisce meglio i dettagli della procedura. Essa è un dettaglio della diramazione sinistra dell'albero risultante dall'applicazione

dell'algoritmo di segmentazione binaria calcolato coi vincoli di "sequenzialità" descritti pocanzi.

Si prenda ad esempio il nodo 10 evidenziato nella 3.6. Esso è un nodo di 4° livello, dunque è calcolato sulla base delle sessioni durate almeno 4 click. In questo nodo ci sono individui caratterizzati da sessioni che durano in media 1,5 click ancora (durano in pratica 5,5 click in media). Vengono computate e indicate anche la numerosità del nodo, la proporzione tra questa e il totale delle sessioni considerate, la deviazione standard e il tasso di abbandono al livello considerato (nel nostro caso il quarto), il cui dettaglio per sezione è ulteriormente declinato in un apposito box a destra dell'albero, visibile nella successiva figura che mostra l'intero albero (fig. 3.7).

La procedura continua fino all'iterazione 9, calcolando gli split a partire dalle sessioni non ancora concluse e tenendo conto della sequenzialità degli eventi "click".

Il secondo step della procedura, ad esempio, viene calcolato a partire dai primi due eventi di navigazione, i primi due click, raccolti nelle prime due colonne della base dati.

Notevole importanza riveste anche l'indicazione, visibile nei box della 3.6, delle scelte compiute dagli utenti caratterizzati dalle sessioni finite in quel nodo che continuano la visita.

La variabile di risposta è dunque rappresentata dal numero di click che in media l'utente che non abbandona la visita compie successivamente allo stato considerato. Pertanto, la variabile di risposta considerata è di tipo numerica. Misurando in termini di click, ciò equivale ad avere 10 diverse

variabili di risposta a seconda che l'utente abbia abbandonato il sito al primo, al secondo, all'n-simo click. Ad ogni livello, la variabile di risposta si "aggiorna", tenendo conto dello stato della sequenza in cui si va a misurare tale variabile.

La procedura iterativa proposta può essere sintetizzata come segue:

1. L'intero collettivo iniziale, tutte le sessioni che hanno meno di 10 click, viene diviso in due parti mediante l'intervento di una variabile di split che tiene conto delle scelte di navigazione effettuate al primo click.
2. Dal nodo iniziale abbiamo così ottenuto i due nodi figli. A questo punto vengono escluse dall'analisi tutte quelle sessioni che si concluse al primo click;
3. Per le sessioni concluse si calcola il leave ratio (tasso di uscita) comprensivo dell'indicazione, anche grafica, delle percentuali di uscita da ciascuna delle sezioni del sito. E'importante segnalare che tale misura è riferita al nuovo collettivo considerato, vale a dire l'insieme delle sessioni che durano al massimo 10 click;
4. Ad ogni ulteriore iterazione *n* entrano in gioco esclusivamente le scelte effettuate all'*n*-simo click del processo di navigazione. Al termine di ciascuna iterazione vengono escluse dal computo successivo le sessioni che si sono concluse in corrispondenza del click numero *n*.

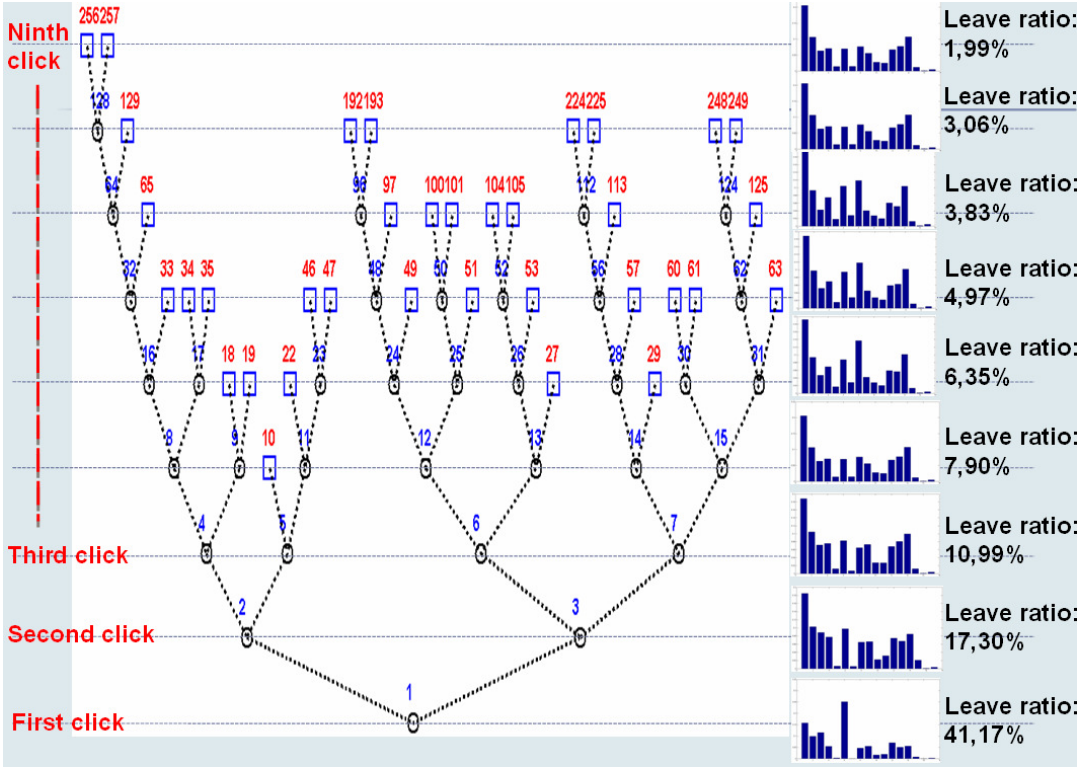

Figura 3.7 L'output principale del sequence tree

3.5.2 Gli scopi applicativi e le evidenze

Attraverso il sequence tree si possono ottenere informazioni di varia natura, anche se l'uso preminente del metodo è la conoscenza delle dinamiche di abbandono/permanenza su di un sito a partire dalla conoscenza delle sessioni di visita. La procedura, infatti, si pone come supporto alla conoscenza di alcune caratteristiche relative all'uscita da un sito, riassumibili come segue:

1. Indicazione del *leave ratio*: in corrispondenza di ogni click, che equivale ad uno "stato" della sequenza di visita, viene calcolato ed indicato il tasso di abbandono;

2. Dettaglio delle uscite: rappresenta un'ulteriore declinazione del leave ratio. Per ogni step della sequenza di visita, si evidenziano le sezioni che hanno prodotto più abbandoni. Esse rappresentano dunque le principali "vie d'uscita" dal sito in oggetto. Questa informazione è particolarmente importante nell'ottica di Marketing o di monitoraggio delle performance del sito stesso;
3. Indicazione del numero medio di pagine restanti: è la variabile di risposta considerata nell'algoritmo di segmentazione, ed è utile per stabilire, nodo per nodo, il comportamento degli utenti caratterizzato dalle loro sessioni di visita.
4. Indicazione del percorso compiuto da coloro che non abbandonano il sito allo stato considerato; questa informazione completa ulteriormente la descrizione della situazione oggetto di studio, in quanto illustra il comportamento di visita di chi continua la navigazione;

Si è chiarito che una prima indicazione è fornita dal *leave ratio* ad ogni click della sequenza, che descrive la pendenza della curva di abbandono.

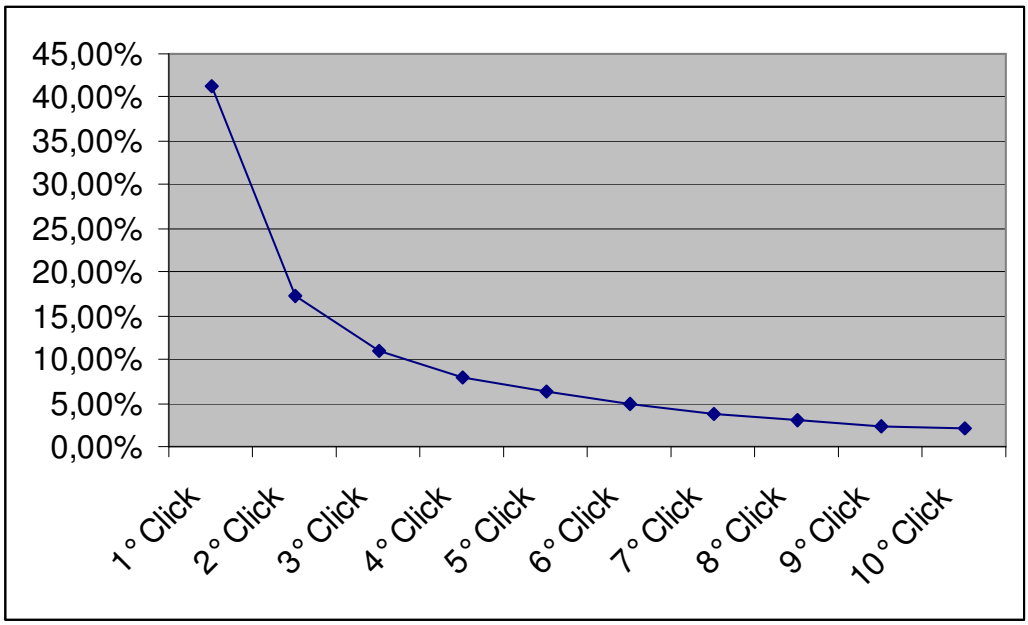

Figura 3.8 La curva del leave ratio

Più in dettaglio, però, significative informazioni riguardanti il fenomeno dell'uscita dal sito sono date dal dettaglio del *leave ratio* per singola sezione. Si può così costruire un ulteriore approfondimento che tiene conto sia del momento dell'uscita che delle sezioni che ad ogni step della frequenza di navigazione costituiscono le principali "porte di uscita" del sito.

A riguardo, nel caso del sito msnbc.com le frequenza di uscita per sezione si sono rivelate piuttosto costanti; una serie di queste, infatti, costituisce fonte di abbandono che non sembra risentire di variabilità rispetto ai vari momenti della sessione in cui tale fenomeno è misurato. Viceversa, lo stesso accade per un altro gruppo di sezioni, che fa registrare pochi abbandoni, anche in questo caso con un trend costante rispetto ai click considerati.

Tabella 6: Abbandono/Permanenza: dettaglio delle sezioni

ABBANDONO / PERMANENZA *Dettaglio sezioni*	
ABBANDONO	PERMANENZA
Frontpage	Opinion
News	Misc
On-air	Living
Weather	Business
Sports	Bbs
	Travel

Per quanto riguarda invece l'indicazione del numero di pagine restanti, come si può osservare dal riepilogo sottostante le classi di appartenenza dei vari nodi terminali rispecchiano la composizione sequenziale.

Il primo nodo terminale si incontra al livello 4, ed è il nodo 10, la cui classe di risposta è 1,50, mentre la deviazione standard è pari a 4,32.

Salendo di un livello si trovano 5 nodi terminali di livello 5, ossia gruppi di nodi popolati da sessioni durate almeno 5 click.

Già in questo step tali nodi cominciano a differenziarsi rispetto alla variabile di risposta. Se ad esempio gli individui caratterizzati dalle sessioni finite nel nodo 22 restano in media per altri 0,91 click, gli appartenenti al nodo 29 compiono in media altri 2,07 click. Anche allo step 6, al quale si trovano 12 nodi terminali, si riscontrano differenze anche marcate: si va infatti dai 0,64 click restanti misurati al nodo 46 ai 2,03 del nodo 33.

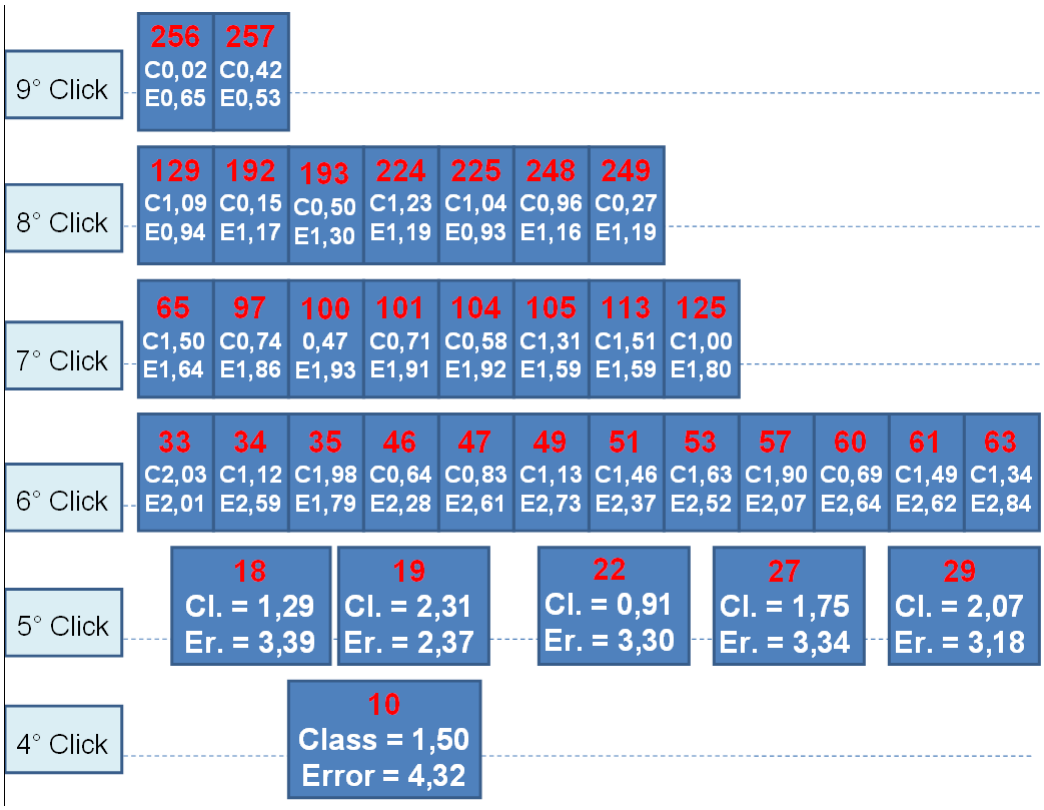

Figura 3.9 Il dettaglio dei nodi terminali

Tali informazioni rivestono un interesse ancora maggiore quando le si incrocia con il dettaglio dei percorsi compiuti in corrispondenza delle sessioni che caratterizzano ciascun nodo terminale. In questo modo si può collegare in maniera diretta la sezione e la permanenza (o l'abbandono).

3.5.3 Considerazioni, prospettive e lavori futuri

Questo lavoro vuole rappresentare una proposta di integrazione tra Web Usage e Web Structure Mining. Allo scopo, sono state proposte due strategie che rispondono a questo scopo, pur avendo contesti applicativi differenti.

Con il primo metodo, infatti, si cerca di fornire una descrizione alternativa della mappa del sito, composta non attraverso i canoni "già noti", cioè quelli derivanti dall'architettura che il Webmaster ha pensato, bensì con l'idea di similarità espressa dagli utenti mediante le loro visite. In più, i collegamenti tra le pagine sono anch'essi frutto delle relazioni più forti che si riscontrano nell'insieme delle sessioni, e non noti a priori mediante lo schema di link.

La seconda strategia è volta invece alla descrizione e alla valutazione delle dinamiche di permanenza/abbandono di un sito web, sempre a partire dall'insieme delle sessioni di visita. Il metodo presentato è una variante dell'albero di segmentazione che tiene conto della sequenzialità tipica delle sessioni di navigazione e di un altro vincolo assai significativo, ovvero la diversa durata di ciascuna sessione. Ciò comporta che la variabile di risposta assuma un significato diverso dalle applicazioni tradizionali, essendo dei click residui prima dell'abbandono, ma non in termini "assoluti", bensì a partire dallo stadio di appartenenza del nodo nella quale la si sta calcolando in quel momento.

Tale soluzione è senza dubbio originale, e presenta dei vantaggi applicativi interpretativi: come per il primo metodo proposto, il sequence tree abbisogna di pochi dati per funzionare, le sole sessioni di visita. In più, questo tipo di informazione si reperisce con facilità su ogni tipo di sito Web. Inoltre, laddove ve ne sia possibilità, si può sfruttare la capacità degli alberi di processare informazioni di diversa natura che vanno ad arricchire ulteriormente l'analisi. Il sequence tree fornisce alcune indicazioni

interessanti circa il tasso di abbandono (leave ratio), il numero di click restanti per stadio e il dettaglio delle uscite per sezione.

Il dettaglio dei percorsi presenta invece prospettive di miglioramento e ottimizzazione in quanto, anche se al momento sono già restituiti dall'albero, risultano di difficile interpretazione.

Un'interessante prospettiva di ottimizzazione sulla quale concentrarsi è fornita dal lavoro "Posterior Prediction Modelling of Optimal Trees" presentato di recente al COMPSTAT 2008 da Roberta Siciliano, Massimo Aria e Antonio D'Ambrosio. Il contributo proposto riguarda la prospettiva di interazione tra più predittori nella formazione dello split. Tale prospettiva potrebbe ovviare in maniera significativa alle difficoltà interpretative dei percorsi dovute alla necessità di affidarsi a split binari che riassumono tutte le 17 possibilità di scelta della sezione che nel caso studio si presentavano.

Per quanto concerne invece il primo contributo, l'integrazione tra MDS e regole associative applicate agli *implicit behaviors,* anche producendo tuttora risultati interpretativi soddisfacenti, si può lavorare verso l'estensione e l'integrazione del metodo in presenza di variabili e informazioni di altra natura, come il tempo di navigazione o i dati di carattere socio-demografico. Se è vero che uno dei pregi di questi metodi è la capacità di fornire conoscenza con informazioni minimali e facilmente reperibili, è altrettanto vero che il contenuto informativo derivante dall'adozione di tali strategie aumenterebbe in maniera significativa alla presenza di dati di natura "complementare".

Conclusioni

Nei campi di indagine propri del Data Mining, come l'ambito Web di cui si è discusso in questo lavoro, vi sono numerose problematiche che vanno al di là di quelle che lo statistico era abituato a fronteggiare nelle analisi classiche, nelle quali la scelta del metodo o del modello da adottare e la sua rigorosa applicazione costituivano il fulcro del suo lavoro. Con l'avvento del Data Mining, pur non cambiando la "sostanza" della ricerca statistica, è cambiato il peso che rivestono alcune fasi della ricerca, segnatamente la fase di raccolta e pre-processing dei dati e quella di visualizzazione dei risultati o delle evidenze. L'ambito del Web Mining costituisce una chiara testimonianza in merito: la mole di dati a disposizione è vastissima, disposta su più macchine e a più livelli. In più essa non è direttamente utilizzabile (salvo le dovute eccezioni) ma deve essere oggetto di un'intensa e accorta attività di Mining e di pre-processing. A valle dell'attività di analisi si pone poi un'altra criticità, ovvero quella della visualizzazione dei risultati della ricerca.

Non è un caso dunque che nell' illustrazione del lavoro appena presentato si ponga l'accento su questi due aspetti, che nel Web Mining sono assai critici. Dapprima infatti si è discusso delle strategie da adottare nell'ottica di analizzare il sito web come un insieme di processi integrati, in un paradigma

di Web Process Mining. A tale scopo si è proposta una strategia di realizzazione e di visualizzazione da applicare per descrivere simultaneamente uso e struttura di un sito quando si hanno a disposizione, o si preferisce analizzare, solo i dati relativi alle sequenze di visita. Non è un caso che tale metodo ha richiesto, come pocanzi si ricordava, un'attenta fase di pre-processing e di trasformazione del dato. Particolare enfasi è dedicata alla ricerca di risultati di visualizzazione chiari e facilmente interpretabili.

La scelta iniziale è ricaduta sull'uso del Multidimensional Scaling: tale tecnica è stata utilizzata per descrivere, attraverso lo studio della similarità tra le visite, la disposizione "concettuale" delle pagine, definendo una mappatura alternativa del sito Web, che non segue la disposizione logica che un webmaster compie nel definire la struttura, e non si configura nemmeno come mappatura schematica dei link. L'indicazione dei collegamenti tra le pagine si è omessa come scelta realizzativa per due motivi: innanzitutto perchè di solito, attraverso una semplice mappa del sito si dispone del quadro completo dei collegamenti alle pagine. La visualizzazione di tale quadro dunque non aggiunge notevoli contributi informativi, al contrario inficia in maniera anche significativa la chiarezza dello schema visivo proposto a supporto dell'interpretazione delle evidenze ottenute.

Ovviamente, non si è per questo trascurata l'importanza di indicare in qualche maniera le direttrici di traffico che caratterizzano il sito.

Allo scopo, si è optato per l'indicazione delle relazioni più significative tra le pagine, ottenute attraverso l'applicazione delle regole associative al dataset oggetto di studio. Nella finestra di visualizzazione tale informazione è andata ad arricchire la soluzione proposta, anche mediante un intuitivo sistema di colori che indica in quale passaggio di visita (primo click, secondo click, ecc.) tale relazione insiste.

A partire dallo stesso set di dati si è infine proposto un metodo di diagnostica e visualizzazione delle caratteristiche di visita di un sito Web, che pone l'attenzione sulla descrizione delle uscite, la previsione degli abbandoni e la visualizzazione della situazione nel suo complesso.
L'approccio metodologico proposto per questa seconda applicazione è quello della segmentazione ad albero. Anche in questo caso si è posto l'accento su uno dei problemi chiave del contesto del Web Mining, ovvero la visualizzazione della gran mole di dati trattata.

In definitiva, col presente lavoro ci si è posti lo scopo di approfondire il framework del Web Mining, evidenziarne le peculiarità ed i punti di contatto con altri ambiti applicativi, cercando di fornire un punto di vista innovativo ai problemi di analisi "integrata" della struttura e dell'uso di un sito. Mentre nel primo capitolo si è dato ampio spazio alla trattazione dei problemi di carattere informatico e sistemistico propri del contesto, nel seguito della trattazione si è esposta la varietà di tecniche statistiche che possono essere utilizzate per trattare i dati provenienti dalla Rete. Nel terzo capitolo, infine, è stato illustrato il cuore del lavoro di ricerca, ovvero la proposizione di due strategie che rispondono a due differenti esigenze: con la prima si è cercato di dare un ulteriore contributo sulla strada dell'integrazione e dell'analisi congiunta tra due contesti che finora sono stati separati dalla dottrina. Con la seconda, invece, si è cercato di proporre uno strumento alternativo per la comprensione e la valutazione delle performance di un sito Web in termini di durata delle visite e di comprensione delle dinamiche che portano all'abbandono del perimetro di un sito Web.
Entrambe queste proposte sono, ovviamente, migliorabili ed ampliabili, ma rappresentano comunque un punto di vista innovativo per quanto riguarda la

visualizzazione della mappa di un sito web, dato che sia il posizionamento delle sezioni che i collegamenti tra le stesse non sono definiti a priori e non corrispondono alla site-map declinata in fase di progettazione dal webmaster. Il sequence tree, pur presentando qualche difficoltà interpretativa, può rappresentare uno strumento utile ad una valutazione "alternativa" e integrata di alcune grandezze topiche dell'andamento di un sito web. I margini di miglioramento in tale direzione sono costituiti senza dubbio dal miglioramento nell'interpretazione dei percorsi discriminanti che caratterizzano l'abbandono/permanenza (in termini di permanenza media) sul sito. Un' altra strada percorribile è quella dell'implementazione di un sistema di visualizzazione di questi "KPI" integrati ricavabili da questo tipo di analisi.

Bibliografia

1. AGRAWAL, R., SRIKANT, R., *Fast Algorithms for Mining Association Rules,* Proc. of the 20th Int'l Conference on VeryLarge Databases, Santiago, Chile, Sept. 1994.
2. ARIA, M., MOLA, F., SICILIANO, R. (2002). Growing and Visualizing Prediction Paths Trees in Market Basket Analysis, in Härdle, W. et al. (eds.): Proceedings of COMPSTAT (Berlin, August 24-28, 2002), Physica Verlag.
3. BERENDT B., HOTHO A., MLADENIC D., VAN SOMEREN M., SPILIOPOULOU M., STUMME G. : *A Roadmap for Web Mining: from Web to Semantic Web*. Web Mining: from Web to Semantic Web. Proceedings of First European Web Mining Forum, EWMF 2003.
4. BORG I., GROENEN P., (2005). Modern Multidimensional Scaling, 2nd edition, Springer.
5. BLANC E., GIUDICI P. (2002), Statistical Models for web clickstream analysis, Technical Report.

6. BLANC E., GIUDICI P. (2002), Sequence Rules for web clickstream analysis, Advances in Data Mining. Springer-Verlag Berlin Heidelberg.
7. BRODERET A. et al, *Graph Structure in the Web*. In the Proc. 9th WWW Conference 2000.
8. CHAKRABARTI S. (2002). Mining the Web, Morgan Kaufmann.
9. CONVERSANO, C., MOLA, F. SICILIANO, R., (2003), How to harvest fruits from trees, Conferenza della Società Italiana di Statistica: "Analisi Statistica Multivariata per le scienze economiche-sociali, le scienze naturali e la tecnologia" (Napoli, June 2003), RCE Edizioni, Napoli, 119-131.
10. CONVERSANO, C., MOLA, F., SICILIANO, R. (2001). Partitioning and Combined Model Integration for Data Mining, presented at the Symposium on Data Mining and Statistics (Augsburg, November 2000), Journal of Computational Statistics, 16, 323-339, Physica Verlag, Heidelberg (D).
11. CONVERSANO C., SICILIANO R., (2005). "Statistical Data Editing", in Wang J. (eds.), Encyclopedia of Data Warehousing and Data Mining, IDEA Group. Inc., Hershey, USA, volume 2, pag. 359-361
12. COOLEY R., (2000) *"Web Usage Mining: Discovery and Usage of Interesting Patterns from Web Data*", Ph.D. Thesis, University ofMinnesota, Computer Science & Engineering,
13. COOLEY R., MOBASHER B., SRIVASTAVA J.(1999): *Data Preparation for Mining World Wide Web Browsing Patterns,* in Knowledge and Information Systems.

14. COOLEY R., MOBASHER B., AND SRIVASTAVA J (1997). Web mining: *Information and pattern discovery on the world wide web*. In International Conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach,IEEE.
15. COX A., COX T.F. (1994) Multidimensional Scaling, Chapman and Hall, London.
16. DESIKAN P., SRIVASTAVA J., KUMAR V., PANG-NING TAN (2002), "Hyperlink Analysis –Techniques & Applications", Army High Performance Computing Center Technical Report.
17. D'AMBROSIO A., PECORARO M., (2008), "Multidimensional Scaling as Visualization tool of Web Sequence Rules", submitted for "*Studies in Classification, Data Analysis, and Knowledge Organization*", Springer (Proceedings of First Joint Meeting of Sociètè Francophone de Classification and the Classification and Data Analysis Group of SIS – Caserta, Italy, June, 11-13th 2008).
18. GHOSH J., SRIVASTAVA J., ed. (2002), Proceedings of "*Workshop on Web Analytics*", Arlington, VA.
19. GHOSH J., SRIVASTAVA J., ed. (2001), Proceedings of "*Workshop on Web Mining*", Chicago, IL.
20. GIUDICI P. (2001), Metodi statistici per le applicazioni di Data Mining, McGraw-Hill Libri Italia, Milano.
21. MASAND B., SPILIOPOULOU M., SRIVASTAVA J., ZAIANE O (2002), ed. Proceedings of "*WebKDD2002 –Web Mining for Usage Patterns and User Profiles*", Edmonton, CA,
22. MING-SYAN CHEN, JONG SOO PARK, PHILIP S. YU, *Data Mining for Path Traversal Patterns in a Web Environment*, Proc. of Intern. on Distributed Computing Systems, 2000.

23. MOBASHER B., COOLEY R., SRIVASTAVA J.,(2000) *Automatic personalization based on Web usage mining*, Communications of the ACM, Vol. 43, Issue 8,.
24. MOLA, F., SICILIANO, R. (2002). Discriminant Analysis and Factorial Multiple Splits in Recursive Partitioning for Data Mining, in Roli, F., Kittler, J. (eds.): Proceedings of International Conference on Multiple Classifier Systems (Chia, June 24-26, 2002), 118-126, Lecture Notes in Computer Science, Springer, Heidelberg.
25. MOLA, F., SICILIANO, R. (1998). A general splitting criterion for classification trees, Metron, 56, 3-4.
26. MOLA, F., SICILIANO, R. (1997). A Fast Splitting Procedure for Classification and Regression Trees, Statistics and Computing, 7, Chapman Hall, 208-216.
27. MOLA, F., KLASCHKA, J., SICILIANO, R. (1996). Logistic Classification Trees, in A. Prat (Ed.): Proceedings in Computational Statistics: COMPSTAT '96 (Barcellona, August 24-28, 1996), , Physica-Verlag, Heidelberg (D), 373-378.
28. MOLA, F., SICILIANO, R. (1992). A two-stage predictive splitting algorithm in binary segmentation, in Y. Dodge, J. Whittaker. (Eds.): Computational Statistics: COMPSTAT '92, 1, Physica Verlag, Heidelberg (D), 179-184.
29. NASRAOUI O., (2005), *World Wide Web Personalization*, Invited chapter in Encyclopedia of Data Mining and Data Warehousing, J. Wang, Ed, Idea Group,.

30. PANDEY A. , SRIVASTAVA J., SHEKHAR S., (2001), "A Web Intelligent Prefetcherfor Dynamic Pages Using Association Rules – A Summary of Results, SIAM Workshop on Web Mining.
31. PAZZANI M., MURAMATSU J., BILLSUS D., (1996) "Syskill and Webert: Identifying Interesting Web Sites", in Proceedings of AAAI/IAAI Symposium,.
32. PECORARO M., SICILIANO R., (2008). *Statistical Methods for Profiling Users in Web Usage Mining*, in Handbook of research in Text and Web Mining Technologies, IGI Global.
33. PETRAKOS G., CONVERSANO C., FARMAKIS G., MOLA F., SICILIANO R., STRAVOPULOS P. (2004), *New Ways of Specifying Edits*, Journal of The Royal Statistical Society, serie A Statistics In Society, n.2, pag 249-274.
34. SICILIANO, R., ARIA, M., D'AMBROSIO, A. (2008). Posterior Prediction Modelling of Optimal Trees, in Proceedings in Computational Statistics 18th Symposium Held in Porto, Portugal, Brito, Paula (Ed.), Springer-Verlag, pp. 323-334
35. SICILIANO, R., ARIA, M., D'AMBROSIO, A. (2006.) Boosted Incremental Tree-based Imputation of Missing Data, In *Studies in Classification , Data Analysis, and Knowledge Organization*, a cura di S.Zani, A.Cerioli, M.Riani e M.Vichi, ed. Springer-Verlag, pp.271-278.
36. SICILIANO R., CONVERSANO C., (2005). "Decision Tree Induction", in Wang J. (eds.), Encyclopedia of Data Warehousing and Data Mining, IDEA Group. Inc., Hershey, USA, volume 2, pag. 242-248.

37. SICILIANO, R., ARIA, M., CONVERSANO, C. (2004). Harvesting trees: methods, software and applications. In Proceedings in Computational Statistics: 16th Symposium of IASC Held in Prague, August 23-27. 2004 (COMPSTAT2004), Eletronical Edition (CD) Physica-Verlag, Heidelberg.
38. SICILIANO, R., MOLA, F. (2000). Multivariate Data Analysis through Classification and Regression Trees, Computational Statistics and Data Analysis, 32, 285-301, Elsevier Science.
39. SICILIANO, R. (1999). Latent budget trees for multiple classification, in M. Vichi, P. Optitz (Eds.): Classification and Data Analysis: Theory and Application, Springer Verlag, Heidelberg (D).
40. SICILIANO, R. (1998). Exploratory versus Decision Trees, invited lecture at COMPSTAT '98 (Bristol, August 24-28, 1998), in R. Payne, P. Green (Eds.): Proceedings in Computational Statistics: 13th Symposium of COMPSTAT, Physica Verlag, Heidelberg (D), 113-124.
41. SICILIANO, R., MOLA, F. (1998). Ternary Classification Trees: a Factorial Approach, in M. Greenacre, J. Blasius (Eds.): Visualization of Categorical Data, cap. 22, , Academic Press, San Diego (CA), 311-323.
42. SRIVASTAVA J., COOLEY R.: , DESHPANDE M., PANG-NING TAN (2000). *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. SIGKDD Explorations.

RINGRAZIAMENTI

In un'ipotetica classifica de "Le cose che non so fare" i ringraziamenti occuperebbero sicuramente un posto di rilievo. Tuttavia, credo che questa sia l'ultima tesi di dottorato che scriverò, dunque è il caso di fare uno sforzo vincendo la paura di raccontarsi e soprattutto di farlo dimenticando qualcuno di importante.

Pensando a questo lavoro la prima persona che mi viene in mente è senza dubbio Roberta Siciliano. I suoi insegnamenti, accademici e non, li porterò sempre con me, dato che oltre alla Statistica mi ha insegnato a lavorare con passione e slancio. Accanto a lei colloco immediatamente quelli che qualcuno ha chiamato "Siciliano Boys", ovvero i ragazzi del suo gruppo, amici prima che compagni di studio. Le lunghe chiacchierate con Valerio, le sigarette che ho visto fumare ad Antonio, la mimica inconfondibile di Gianfranco, i trucchi del mestiere imparati da Massimo e il tocco femminile di Sonia, erano piccole consuetudini che non posso e non voglio dimenticare. A colorare le mie giornate in dipartimento c'erano poi tutti i professori, i ricercatori, i dottorandi ed il personale che formano la grande famiglia del Prof. Lauro, il quale mi ha accolto "in casa" sorvegliando con attenzione e affetto sul mio percorso di studio.

In verità, però, la persona alla quale devo maggiormente la realizzazione di questo lavoro è mia madre, che ha sempre appoggiato le mie scelte, soprattutto quelle che comportavano sacrificio. Sarà retorico, ma non credo di poter fare mai abbastanza per sdebitarmi di tutto quello che ha fatto per me. A completare il quadro delle mie "sostenitrici" silenziose c'è Francesca, che ha sempre scommesso su di me con una fiducia e una sicurezza che a volte mi chiedo su che basi si fondino; la risposta, comunque, credo di conoscerla.

Gli ultimi ringraziamenti vanno a coloro che mi hanno supportato e sopportato durante l'ultima parte della stesura della tesi: mi riferisco a Stefania Gentile di Intesa San Paolo, che con la sua disponibilità ha reso meno arduo il mio ruolo di studente-lavoratore. Insieme a lei ringrazio i nuovi colleghi, che giorno per giorno hanno vissuto insieme a me gli ultimi piccoli traguardi. Infine, ringrazio gli amici che ho conosciuto o ritrovato a Torino, i personaggi che allietano questo nuovo ed entusiasmante capitolo della mia vita. Come volevasi dimostrare ho "trascurato" un sacco di persone, in primis mio fratello Walter, ma anche Mimmo. Lucio, Robbertina, Flaminia, Marco, Angela, Enzo, Gianluca, Armando e Giapu avranno molto da ridire...ma l'ho detto, non sono bravo a scrivere i ringraziamenti, e in più è tardi e la mia fonte di ispirazione, il limoncello, va usato con misura perchè a Torino è un bene prezioso!!!