UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

Corso di Laurea:

STATISTICA E GESTIONE DELLE IMPRESE

Tesi di Laurea

ANALISI DEI COMPORTAMENTI DI VISITA A UN SITO WEB

Relatore: Prof. Scarpa Bruno

Laureanda: Francesca Penzo

Matricola: 553656 – GEI

ANNO ACCADEMICO 2008/2009

Alla mia famiglia, che mi ha sempre sostenuto
e a tutti coloro che hanno contribuito
al raggiungimento di questo importante traguardo

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Internet ha cambiato sia il modo di concepire le informazioni sia il modo di renderle disponibili e di gestirle. Come accedere alle informazioni non è più il problema principale: l'obiettivo è cercare di scoprire, all'interno dei dati web, informazioni non note e rilevanti.

Ogni giorno i w*eb server*, che gestiscono il traffico di un sito, registrano nei *log file* il percorso di visita di ciascun utente che vi accede. Ogni visitatore lascia, quindi, una traccia del suo passaggio ed essa può fornire molte informazioni, se analizzata correttamente. Infatti, ogni file di log racchiude in sé il comportamento del navigatore nel sito: le sue preferenze, il tempo dedicato alla visita, le "esitazioni" e le sue scelte.

Analizzare i *log file* significa, quindi, comprendere quali siano i comportamenti di visita degli utenti e questo, unito alla capacità di offrire servizi personalizzati che soddisfino i loro bisogni, costituisce, per un'azienda, un utile strumento per competere efficacemente nel mercato, migliorando la comunicazione attraverso il Web e incrementando la soddisfazione dell'utente nella visita al sito.

In questo contesto, per le aziende assume sempre più importanza dotarsi di strumenti in grado di trasformare l'enorme mole di dati contenuti nei log file in informazioni utili per individuare modelli di navigazione degli utenti. Ciò ha permesso lo sviluppo di un'area specifica del *Data Mining*[1]*,* chiamata *Web Mining*.

In analogia con l'espressione *Data Mining*, che indica l'applicazione di particolari algoritmi per individuare "regolarità" nei dati di un *database*, il termine *Web Mining* fa riferimento all'applicazione di tecniche similari per estrarre informazioni dalle risorse presenti nel Web.

[1] Il Data mining può essere definito come il processo di estrazione (attraverso l'adozione di modelli statistici, algoritmi matematici di analisi e altri sistemi di rilevazione) delle informazioni da un database, con l'obiettivo di individuare le informazioni più significative e renderle disponibili e direttamente utilizzabili nell'ambito di processi decisionali.

La presente trattazione consiste nell'applicazione di alcune tecniche di *Web Mining* ai dati di navigazione di un sito con tre finalità differenti:

1. Segmentazione comportamentale: l'obiettivo è suddividere gli utenti in gruppi omogenei in base al comportamento di visita alle diverse aree del sito;
2. Analisi delle sequenze di visita: l'obiettivo è individuare le sequenze di pagine più ricorrenti e valutare l'importanza e la significatività delle associazioni tra insiemi di pagine;
3. Previsione dei comportamenti di visita: l'obiettivo è prevedere la conclusione del percorso di visita in una area d'interesse per l'azienda, in funzione delle aree visitate in precedenza .

Tutte le analisi sono state elaborate con il software R[2].

[2] R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Capitolo 1

Analisi esplorativa

1.1 Descrizione dei dati

L'insieme di dati a disposizione contiene una raccolta di *log file*[1] del sito web di un'azienda che fornisce soluzioni di *business intelligence* per le imprese; i dati si riferiscono ad un periodo di tempo pari a un anno, e si presentano in formato grezzo. Affinché il processo di data mining produca dei buoni risultati è necessario che questi dati vengano preparati attraverso varie operazioni di pulizia, andando ad identificare ed eliminare i record inutili per l'analisi[2].

La matrice di dati ripulita e riordinata contiene i dati relativi a 26226 sessioni utente[3]. A ciascun utente, identificato da un codice numerico, corrispondono nella matrice tante righe quante sono le pagine visitate durante la sua sessione di visita. Ogni riga contiene l'identificativo utente, l'indirizzo della pagina visitata, la data di accesso e altre informazioni descriventi la visita. Il numero totale di pagine presenti nel sito è pari a 308.

Tra le variabili a disposizione, per le analisi ne sono state considerate solo due: l'identificativo utente e l'indirizzo della pagina visitata.

In questo modo si è costruita una nuova matrice di dati, chiamata “utenti”, con 26226 righe, una per ciascun utente, e 308 colonne, una per ciascuna variabile discreta indicante il numero di visite effettuato da ogni utente alla pagina

[1] *Log file* (file di registro) è un termine del linguaggio informatico usato per indicare la registrazione cronologica che i *web server* operano sul traffico generato dai siti Internet. I dati a disposizione contengono informazioni derivanti da diversi tipi di log file: *transfer*, *referrer* e *agent*, in quanto riportano l’url della pagina visitata (transfer log), la pagina di provenienza (*referrer log*) e informazioni riguardanti il tipo e la versione di *browser* utilizzato dall'utente, il sistema operativo usato, la risoluzione video (*agent log*).

[2] Si veda l’allegato 1

[3] Il termine *sessione utente* indica la successione delle pagine richieste e visualizzate da un utente durante la navigazione all’interno di un sito web.

corrispondente. Per le pagine che non sono state visitate dall'utente nel suo percorso di navigazione, la variabile discreta assume valore zero.

Di seguito è riportato uno stralcio della matrice dei dati:

	azienda	contatti	eventi	home	pubblicazioni	servizi	settori	svago	altro
1	0	0	0	0	1	0	0	0	0
2	0	0	0	0	1	0	0	0	0
3	0	0	0	0	1	0	0	0	0
4	0	1	0	2	0	5	5	0	0
5	0	0	0	1	0	0	0	0	0
6	3	0	0	0	0	0	0	4	0
7	0	0	0	1	0	0	0	0	0
8	0	1	0	1	0	0	0	0	0
9	0	0	0	1	0	0	0	0	0
10	0	0	0	0	0	0	2	0	0
11	1	0	0	0	0	0	0	0	0
12	0	1	0	0	0	1	0	0	1
13	0	0	0	1	0	0	0	0	0
14	0	0	0	0	0	0	2	0	0
15	0	0	0	0	0	0	1	0	0
16	0	0	0	0	0	0	1	0	0

Tabella 1.1: Stralcio della matrice dei dati

1.2 Analisi esplorativa

L'analisi esplorativa dei dati ha evidenziato un'elevata dispersione delle visite, dovuta all'eccessivo numero di pagine considerate. Perciò si è deciso di raggruppare le pagine in aree, rispecchiando la reale suddivisione in sezioni del sito. Di seguito sono riportate le aree individuate corredate da una breve descrizione per agevolare l'interpretazione delle successive analisi:

HOME : è l'*home page* del sito

AZIENDA: è l'area che raggruppa le pagine di presentazione dell'azienda

CONTATTI: è l'area che contiene le informazioni utili per contattare l'azienda

EVENTI: consiste nel calendario degli incontri e convegni sponsorizzati dall'azienda

PUBBLICAZIONI: contiene articoli e altre pubblicazioni che riguardano le attività dell'azienda

SERVIZI: raggruppa le pagine che descrivono i servizi offerti dall'azienda

SETTORI: presenta le varie soluzioni di business intelligence per ogni settore di cui si occupa l'azienda

SVAGO: è un'area di "relax", dove gli utenti possono anche esprimere le loro opinioni sul sito

ALTRO: è l'area che contiene le pagine che non rientrano nelle categorie precedenti.

La matrice di dati ottenuta in seguito al raggruppamento delle pagine in aree mostra comunque un'elevata dispersione: dall'analisi della Tabella 1.2 si nota come la percentuale di valori nulli nella distribuzione di ciascuna variabile di area sia superiore di molto al 50% tranne che per la variabile "pubblicazioni".

Il numero medio di pagine visitate per utente è 1,8 (deviazione standard pari a 3,39). Il valore è piuttosto basso, e si spiega considerando che su un totale di 26226 utenti, 21644 hanno visitato una sola pagina nella loro sessione (Figura 1.1).

variabili	media	deviazione standard	% valori nulli
azienda	0,16	1,22	95,9
contatti	0,03	0,21	97,7
eventi	0,03	0,22	97,2
home	0,21	0,62	83,9
pubblicazioni	0,52	1,09	54,8
servizi	0,2	0,96	89,0
settori	0,2	1,65	70,8
svago	0,06	0,53	97,1
altro	0,02	0,19	98,7

Tabella 1.2: Media, deviazione standard e percentuale di valori nulli della distribuzione di ciascuna variabile di area

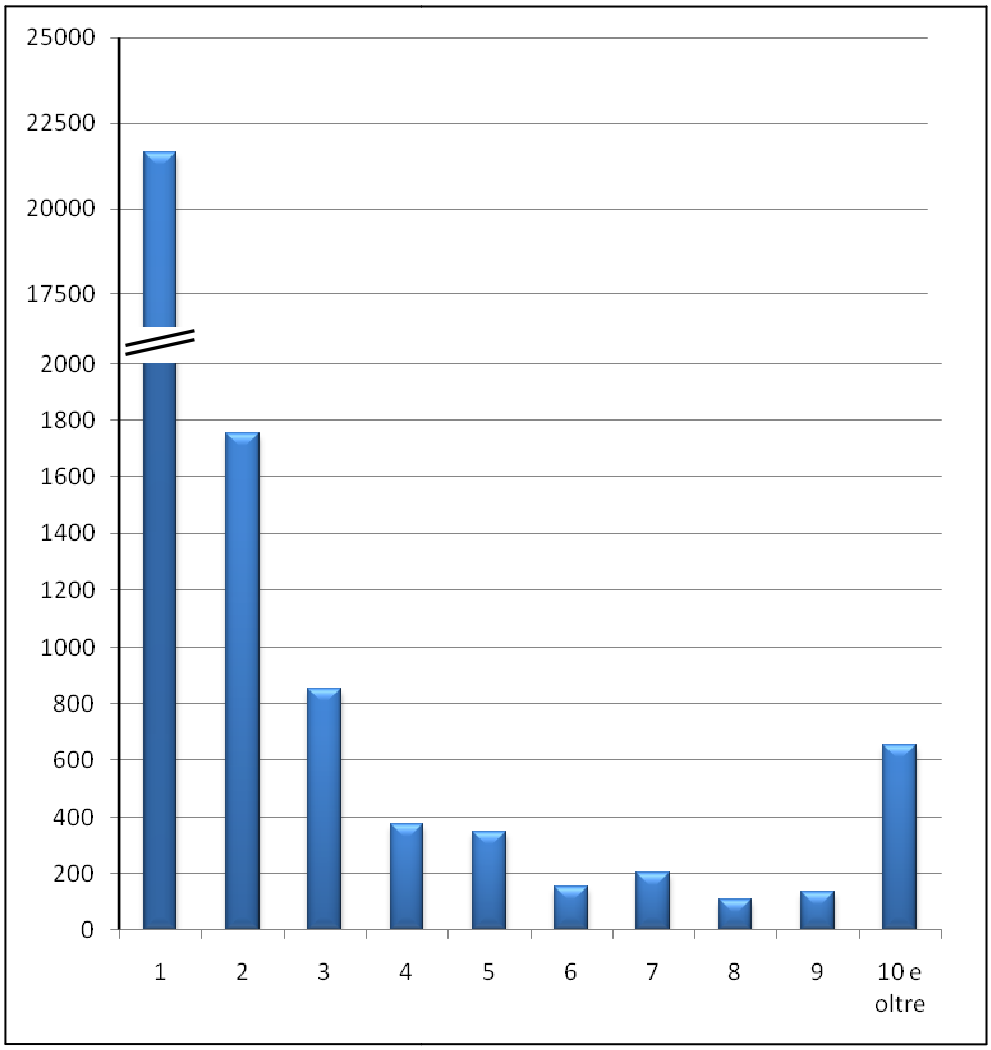

Figura 1.1: Distribuzione di frequenza del totale di pagine visitate in una sessione utente

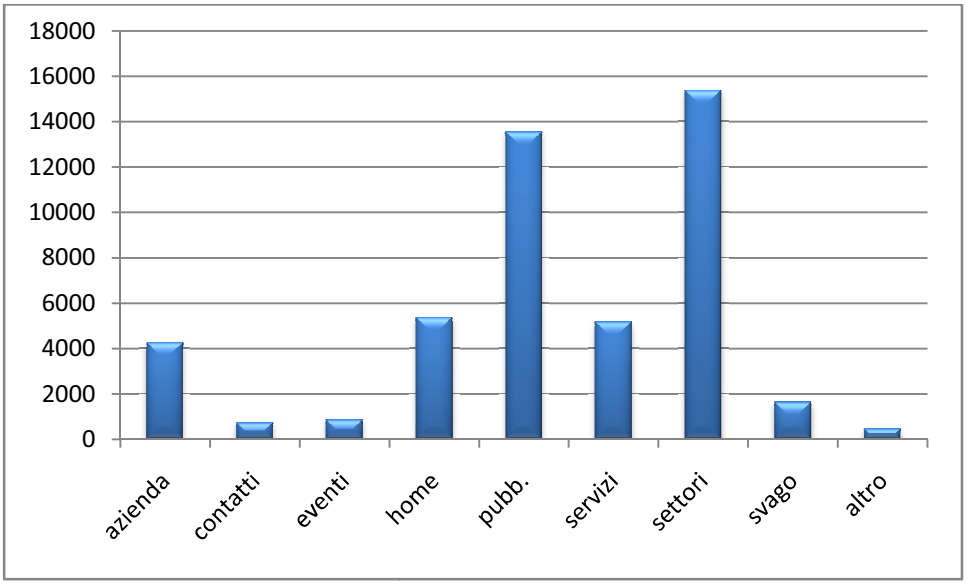

Figura 1.2 : Distribuzione di frequenza del numero di visite per area

Poiché la quota di sessioni utente di lunghezza unitaria (in termini di numero di pagine visitate) è particolarmente rilevante, e dato che l'interesse principale è raggruppare gli utenti in gruppi in base al loro percorso di navigazione, si può già operare una prima suddivisione delle unità in due gruppi: “visite1”, contenente le 21644 osservazioni relative alle visite ad un'unica pagina, e “visite2” contenente le rimanenti 4582 osservazioni.

1.3 Analisi esplorativa della matrice di dati “visite1”

L’indice di Gini (nella versione normalizzata) è pari a 0.75 e ciò indica una situazione di notevole mutabilità: dal diagramma a barre (Figura 1.3), si può notare che il 95% delle visite si concentra in quattro aree: “pubblicazioni”, “home”, “servizi” e “settori”.

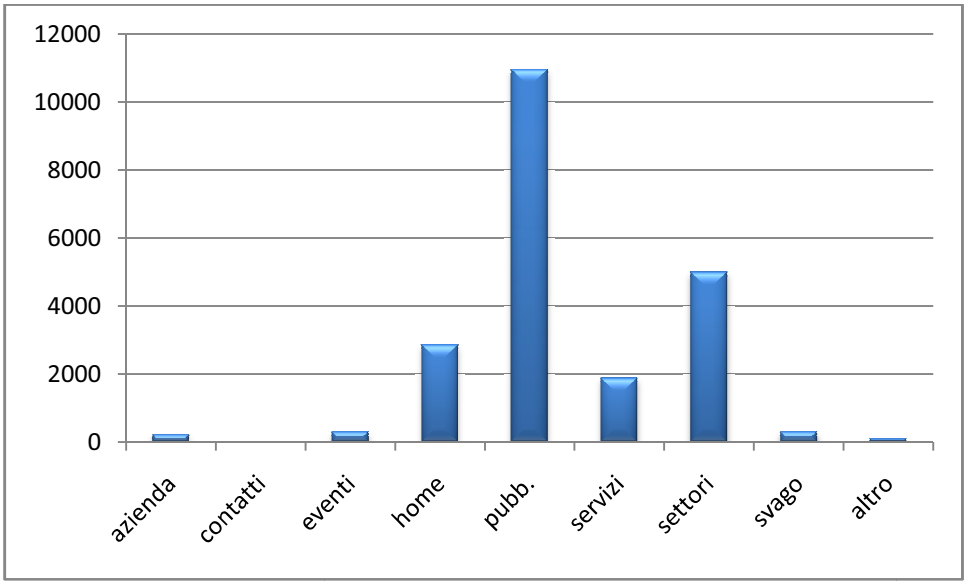

Figura 1.3: Distribuzione di frequenza del numero di visite per area

Per quanto riguarda la distribuzione di frequenza del numero di visite per area (Figura 1.3), si nota che circa la metà delle pagine visitate appartiene all’area “pubblicazioni”. Questo spiega la bassa percentuale di valori nulli nella Tabella 1.2 rispetto alla percentuale di valori nulli della variabile “settori”: nonostante l’area “settori” abbia registrato un numero di visite maggiore rispetto all’area “pubblicazioni”, quest’ultima ha una percentuale minore di valori nulli perché le

visite sono state effettuate in sessioni utenti differenti, mentre l'area "settori" è stata visitata più volte all'interno della stessa sessione.

Per approfondire l'analisi si è considerata la variabile, contenuta nella matrice di dati originaria, che riporta, ove disponibile, per ogni pagina visitata del sito dell'azienda l'indirizzo della pagina da cui proviene l'utente.

È emerso che gli utenti che hanno visitato un'unica pagina hanno avuto accesso al sito prevalentemente attraverso motori di ricerca, con lo scopo sia di cercare offerte di lavoro (seppur in minima parte) sia per recuperare informazioni riguardanti il data mining. Gli utenti vengono indirizzati al sito dell'azienda poiché nelle aree "servizi" e "settori" sono descritte le attività dell'azienda stessa e tra le tecniche che essa usa nei suoi servizi di business intelligence rientra pienamente il data mining. Inoltre, nell'area "pubblicazioni" sono riportati articoli e descrizioni di casi reali in cui l'azienda descrive l'uso delle tecniche sopra citate.

I dati contenuti nella matrice "visite1" non verranno considerati nelle successive analisi poiché l'interesse principale dello studio è l'analisi dei percorsi di navigazione; per gli utenti che hanno effettuato un'unica visita al sito non si può parlare propriamente di *percorso* di navigazione. Tuttavia, in ambito aziendale, non è da trascurare l'analisi anche di questo tipo di accessi al sito.

1.4 Analisi esplorativa della matrice di dati "visite2"

Un'analisi esplorativa sulla matrice "visite2" mostra come la distribuzione delle nove variabili sia fortemente asimmetrica: per tutte tranne che "settori" la mediana è nulla. Per sei delle nove variabili, la percentuale di zeri nella distribuzione è superiore all'80%. Invece per "home" e "servizi" è circa il 70% e solo per "settori" è inferiore alla metà (42%). Nonostante la variabile "azienda" assuma per l'80% delle osservazioni valore nullo, il 99-mo percentile è elevato rispetto a quello delle altre variabili. Ciò fa supporre che gli utenti che accedono a quest'area siano caratterizzati da un numero medio di pagine visitate superiore a quello di altre aree con la stessa percentuale di valori nulli.

variabili	media	deviazione standard	99-mo percentile	massimo	% valori nulli
azienda	0,66	1,83	12	95	81,3
contatti	0,13	0,37	2	5	87,6
eventi	0,09	0,33	2	10	90,9
home	0,46	0,86	5	41	69,2
pubblicazioni	0,41	0,93	6	82	80,7
servizi	0,55	1,44	10	35	78,1
settori	2,04	2,60	15	52	41,8
svago	0,18	0,75	7	17	90,2
altro	0,05	0,25	2	8	94,8

Tabella 1.3: Alcune statistiche descrittive sulle distribuzioni univariate delle variabili di area

Confrontando, per la distribuzione di ciascuna variabile, il valore del massimo con quello del 99-esimo percentile (Tabella 1.3) si nota che alcuni valori sono molto distanti. Le osservazioni che presentano valori superiori al 99-mo percentile della distribuzione di almeno una variabile sono state rimosse dall'insieme di dati e studiate separatamente (§ 1.5) per evitare che i risultati delle analisi siano fuorviati dalla presenza di valori anomali. La matrice “visite2” così ottenuta e utilizzata nelle analisi contiene ora 4338 unità.

Dall'analisi della matrice di correlazione relativa ai dati senza osservazioni estreme (Tabella 1.4), si può osservare che, in generale, la correlazione tra le variabili non è molto elevata. Inoltre, è da notare che la variabile “pubblicazioni” risulta correlata negativamente con tutte le altre variabili.

	azienda	contatti	eventi	home	pubblicazioni	servizi	settori	svago	altro
azienda	1,00	0,23	0,28	0,22	-0,11	0,13	0,02	0,21	0,16
contatti	0,23	1,00	0,12	0,37	-0,12	0,07	-0,06	0,17	0,20
eventi	0,28	0,12	1,00	0,24	-0,07	0,13	0,06	0,24	0,07
home	0,22	0,37	0,24	1,00	-0,17	0,15	0,00	0,19	0,10
pubb.	-0,11	-0,12	-0,07	-0,17	1,00	-0,13	-0,28	-0,08	-0,05
servizi	0,13	0,07	0,13	0,15	-0,13	1,00	-0,03	0,12	0,05
settori	0,02	-0,06	0,06	0,00	-0,28	-0,03	1,00	0,02	0,00
svago	0,21	0,17	0,24	0,19	-0,08	0,12	0,02	1,00	0,10
altro	0,16	0,20	0,07	0,10	-0,05	0,05	0,00	0,10	1,00

Tabella 1.4: Matrice di correlazione di “visite2”

1.4.1 Analisi delle componenti principali

Per visualizzare graficamente le osservazioni della matrice "visite2" è necessario ridurre la dimensionalità dei dati, attraverso l'utilizzo delle componenti principali (Mardia, Kent e Bibby, 1979). Ciò avviene individuando delle combinazioni lineari delle variabili inizialmente osservate, che siano incorrelate tra loro ed abbiano varianza massima (in modo da non disperdere informazioni). La riduzione della dimensionalità avviene limitandosi ad analizzare le principali (per varianza) tra le nuove variabili.

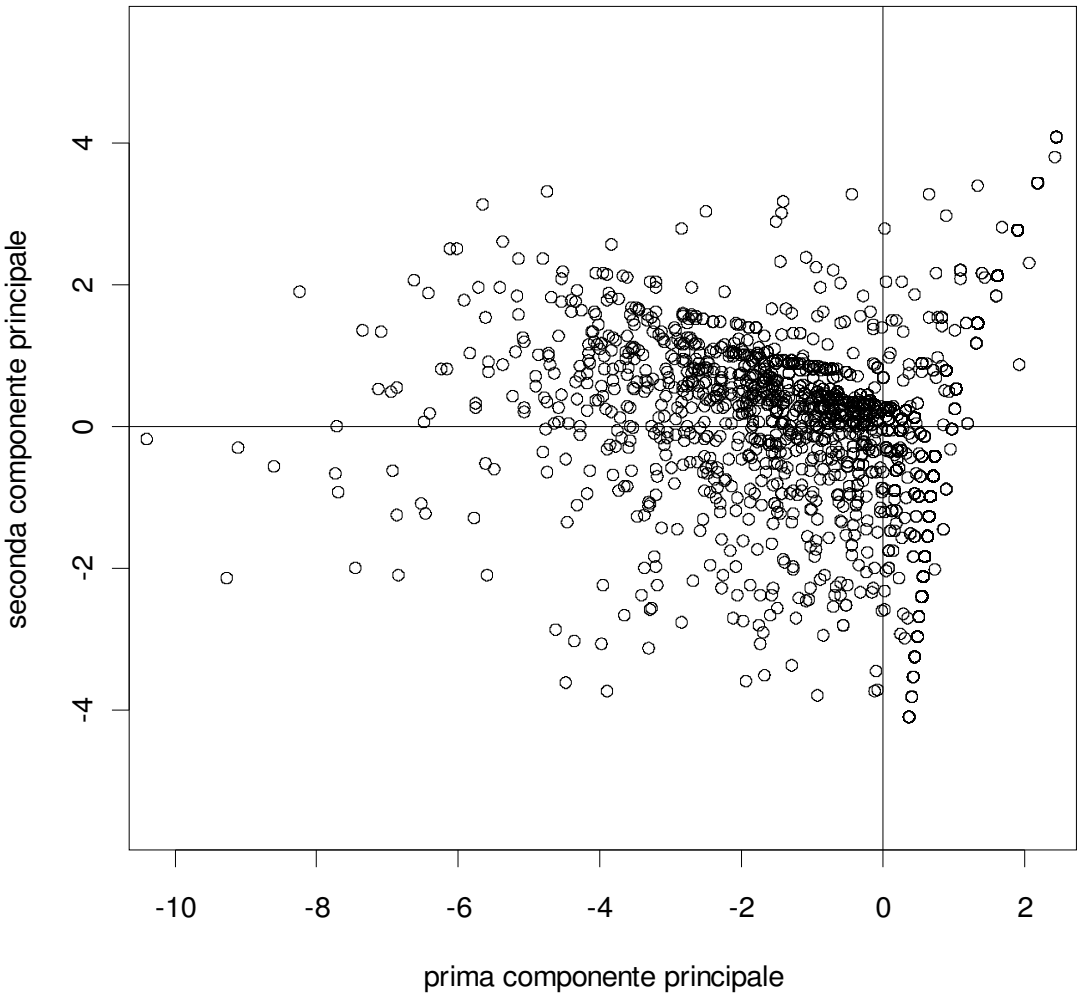

Figura 1.4: Rappresentazione bidimensionale dei dati mediante le componenti principali

Le prime due componenti principali spiegano il 35.6% della variabilità dei dati.

Per quanto riguarda la prima componente principale, la variabile "pubblicazioni" è l'unica ad avere peso positivo, tutte le altre hanno peso negativo, in particolar modo "home" e "azienda" (Tabella1.5). Per quanto riguarda la seconda componente principale, la variabile "pubblicazioni" ha un elevato peso positivo e "settori" un elevato peso negativo (Tabella1.5).

variabili	prima componente principale	seconda componente principale
azienda	-0.418	0.071
contatti	-0.402	0.219
eventi	-0.378	0.008
home	-0.445	0.067
pubblicazioni	0.259	0.610
servizi	-0.250	-0.019
settori	-0.071	-0.736
svago	-0.359	0.082
altro	-0.245	0.145

Tabella 1.5: Pesi per le prime due componenti principali

Sembrano non emergere gruppi distinti di osservazioni. Nonostante ciò, si può comunque tentare un'interpretazione del grafico di Figura 1.4: per quanto riguarda le osservazioni situate nel primo quadrante, esse si trovano in corrispondenza di valori positivi per entrambe le componenti, e quindi per queste osservazioni sembrano aver maggior peso le visite all'area "pubblicazioni". Per quanto riguarda le osservazioni del quarto quadrante, esse sono prossime a valori nulli della seconda componente, quindi l'accesso all'area "pubblicazioni" sembra essere inferiore rispetto a quello dell'area "settori". Per quanto riguarda il secondo e terzo quadrante, essi risultano essere più difficili da interpretare a causa della molteplicità di variabili con peso rilevante che compongono la seconda componente principale. Tuttavia si evidenzia un addensamento delle unità, soprattutto in corrispondenza del secondo quadrante.

1.5 Analisi esplorative dei dati anomali

Le osservazioni rimosse dalla matrice "visite2" si caratterizzano per una maggior variabilità del numero di visite per sessione utente. (Tabella 1.6 e Figura 1.5)

	mediana	media	deviazione standard
visite2	3	4,6	4,1
outlier	21	24,5	15,2

Tabella 1.6: Confronto sulla distribuzione del numero di visite per sessione utente

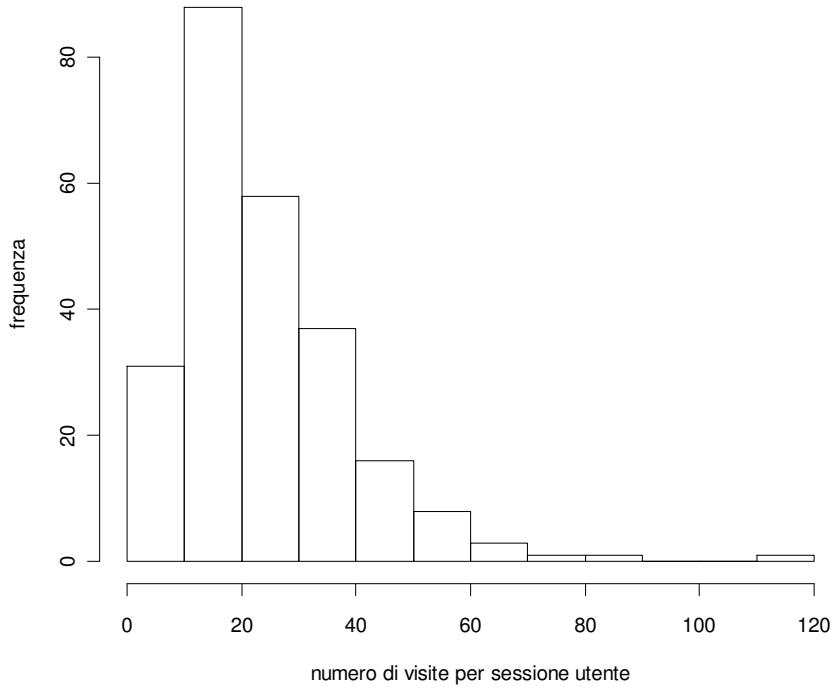

Figura 1.5: Istogramma del numero di visite per sessione utente

variabili	matrice completa	matrice senza osservazioni anomale	osservazioni anomale
azienda	81,3	83,1	49,2
contatti	87,6	88,7	67,6
eventi	90,9	92,4	63,5
home	69,2	70,7	41,8
pubblicazioni	80,7	80,9	77,5
servizi	78,1	79,8	48,4
settori	41,8	42,0	39,4
svago	90,2	92,1	56,6
altro	94,8	95,9	75,4

Tabella 1.7: Confronto delle percentuali di valori nulli

Dal confronto tra le percentuali di valori nulli nella matrice iniziali e nelle due sottomatrici (Tabella 1.7) si nota come la sottomatrice relativa alle osservazioni anomale si distanzi molto dall'altra, poiché le percentuali di valori nulli sono inferiori. Ciò significa che gli utenti che hanno effettuato visite considerate anomale, non soltanto hanno visitato più pagine rispetto agli altri (Figura 1.5), ma hanno anche visitato più aree diverse durante la loro sessione.

Capitolo 2

Analisi di segmentazione comportamentale

L'analisi si propone di individuare, all'interno di un campione di dati, diversi segmenti comportamentali in cui sia possibile raggruppare gli utenti. Per questo scopo si è fatto uso dell'analisi cluster (Kaufman e Rousseeuw, 1990), tecnica statistica che permette di individuare, nell'insieme dei dati, gruppi di unità con caratteristiche simili tra loro e dissimili rispetto a quelle degli altri gruppi.

Per condurre un'analisi cluster è possibile fare ricorso a tecniche gerarchiche o non gerarchiche. Con le tecniche gerarchiche si costruisce una sequenza di partizioni nidificate: da quella in cui ogni elemento è un gruppo a sé a quella in cui tutti gli elementi appartengono allo stesso gruppo (metodo agglomerativo) o viceversa (metodo divisivo). Con le tecniche non gerarchiche si fissa un numero k di gruppi e si suddividono gli elementi in k gruppi.

Tuttavia, entrambe hanno dei difetti che ne limitano l'utilizzo: l'analisi cluster gerarchica non è adatta per trattare campioni di numerosità elevata, mentre l'analisi cluster non gerarchica non fornisce criteri per decidere il numero di cluster da considerare. Pertanto, nell'analisi si è seguita una procedura che combina entrambi i metodi: si è condotta un'analisi non gerarchica, con il metodo delle k-medie, sull'intero campione, imponendo l'ottenimento di un numero di cluster elevato. Poi si è costruito un nuovo campione contenente le medie di ciascun gruppo e su questo si è condotta un'analisi gerarchica, utilizzando un algoritmo agglomerativo con metodo di Ward per il calcolo delle distanze tra i gruppi. Si è escluso il metodo del legame singolo poiché le osservazioni sembrano molto vicine tra loro, e questo tipo di legame tende a identificare come un unico gruppo unità vicine.

L'indagine è stata condotta sulla matrice “visite2” depurata dalle osservazioni anomale. Ciò è stato suggerito dall'analisi esplorativa dei dati (§ 1.4) ed è stato confermato da un primo tentativo di segmentazione sulla matrice contenente anche le osservazioni anomale: essa ha prodotto un raggruppamento poco rilevante, con l'ottenimento di un gruppo di dimensioni elevate e altri contenenti poche unità.

2.1 Analisi cluster non gerarchica

L'analisi cluster non gerarchica è stata condotta imponendo un numero k di cluster pari a 30. La Figura 2.1 mostra i centroidi (vettori delle medie) di ciascun gruppo utilizzando il grafico bidimensionale delle osservazioni ottenuto mediante le componenti principali[1].

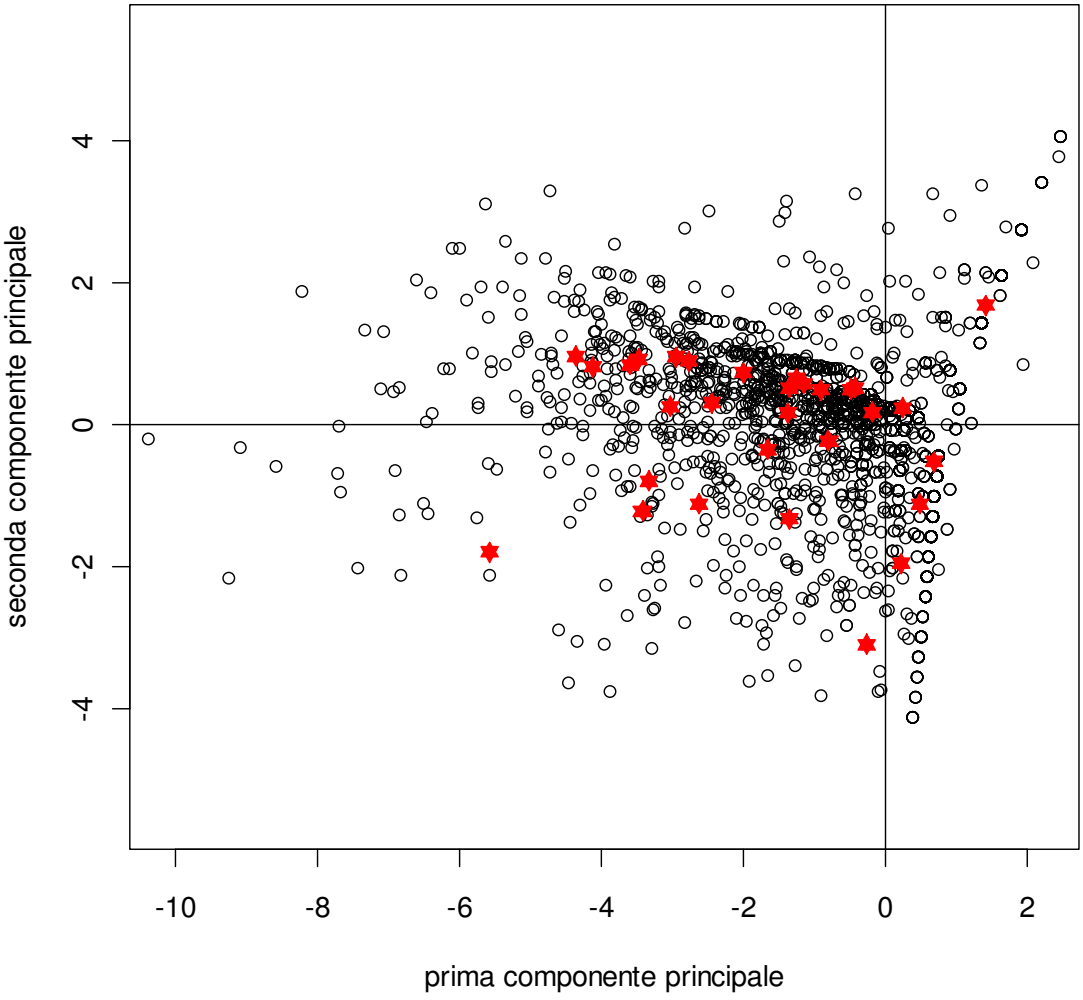

Figura 2.1: Rappresentazione dei centroidi

Per maggior chiarezza, in Figura 2.2 sono riportati in dettaglio solo i centroidi, identificati da un numero progressivo. Tale rappresentazione può essere utile per un confronto nella fase successiva delle analisi.

[1] Si veda la sottosez. 1.4.1

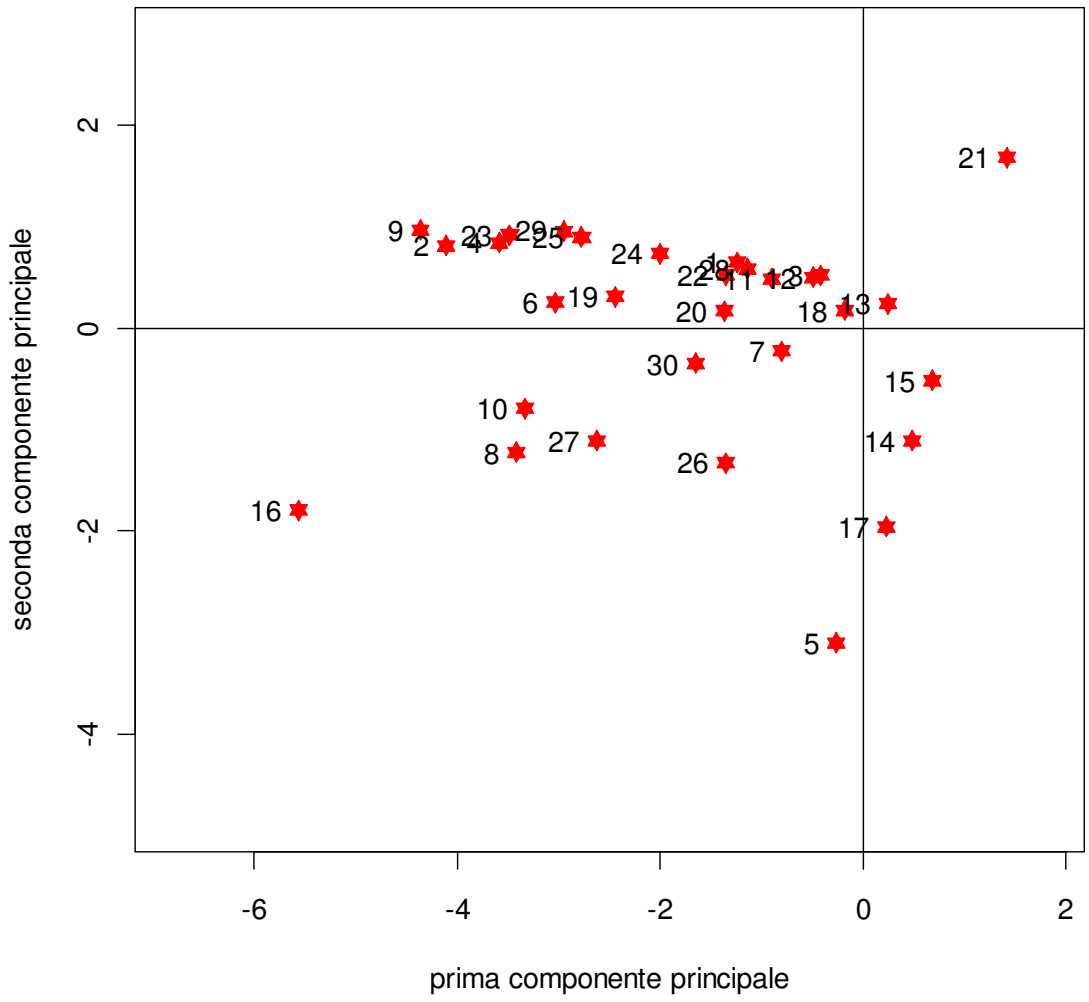

Figura 2.2: Rappresentazione dei centroidi in dettaglio

2.2 Analisi cluster gerarchica

La procedura ha prodotto la sequenza di partizioni rappresentata mediante il dendrogramma di Figura 2.2. Poiché nel dendrogramma l'altezza del segmento che unisce due gruppi rappresenta la distanza tra i gruppi stessi, la partizione avviene in corrispondenza di un "salto" notevole in altezza. Si è scelta la partizione in cinque gruppi perché analisi condotte su altre partizioni hanno evidenziato raggruppamenti non rilevanti.

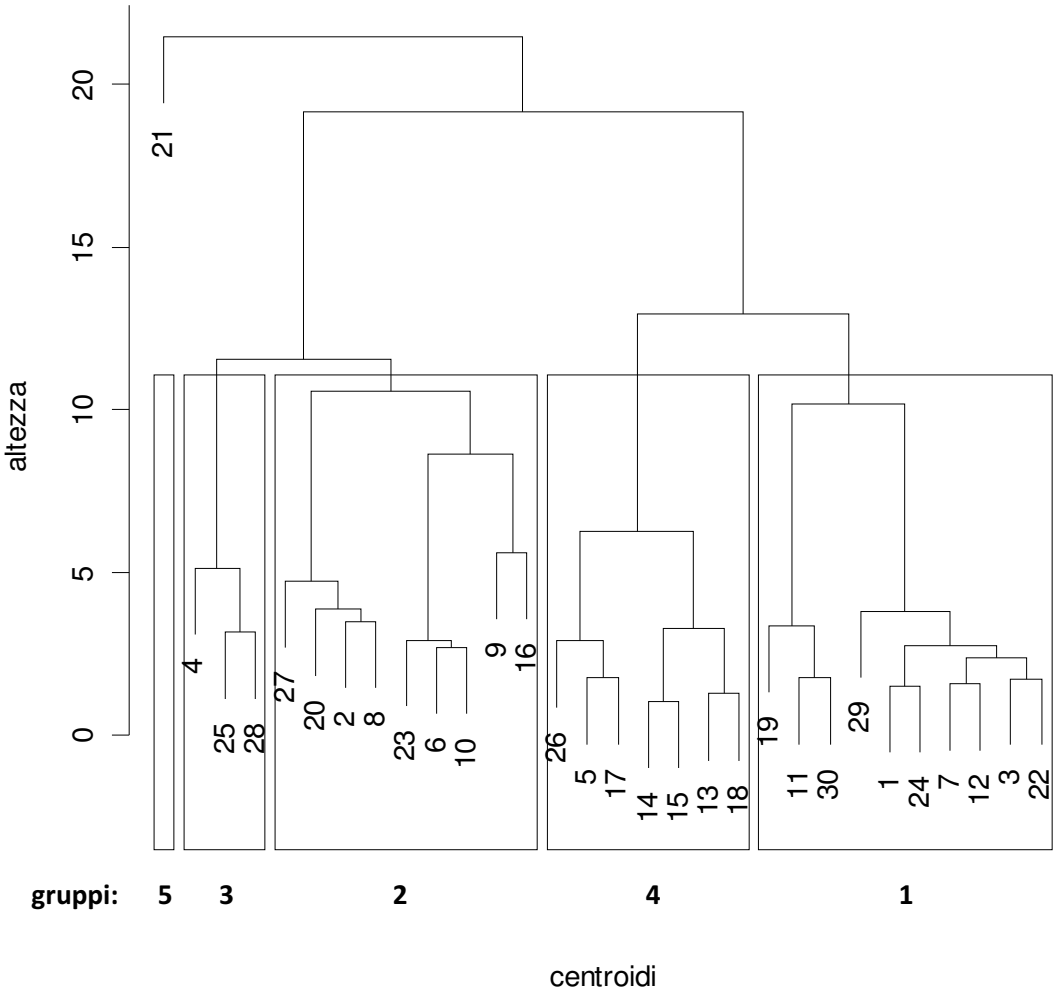

Figura 2.3: Dendrogramma della partizione

Di seguito è riportata distribuzione delle osservazioni nei gruppi mediante un diagramma a torta:

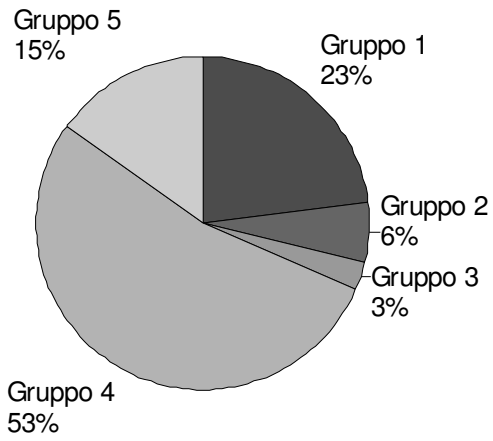

Figura 2.4: Distribuzione percentuale degli utenti nei gruppi

Come si può facilmente osservare, la maggior parte dei visitatori si colloca all'interno del gruppo 4, mentre gli altri visitatori si distribuiscono in modo non uniforme all'interno dei restanti quattro gruppi.

Per poter interpretare più agevolmente il risultato di un metodo di partizione (che consiste semplicemente di una lista dei gruppi e dei loro elementi), è possibile utilizzare uno strumento grafico chiamato "clusplot" (Pison, Struyf e Rousseeuw, 1999). Esso permette di visualizzare gli elementi, la forma e la dimensione di ciascun gruppo e le loro posizioni relative rispetto a un sistema di assi costituito dalle prime due componenti principali calcolate sull'insieme dei dati su cui viene applicato il metodo di partizione.

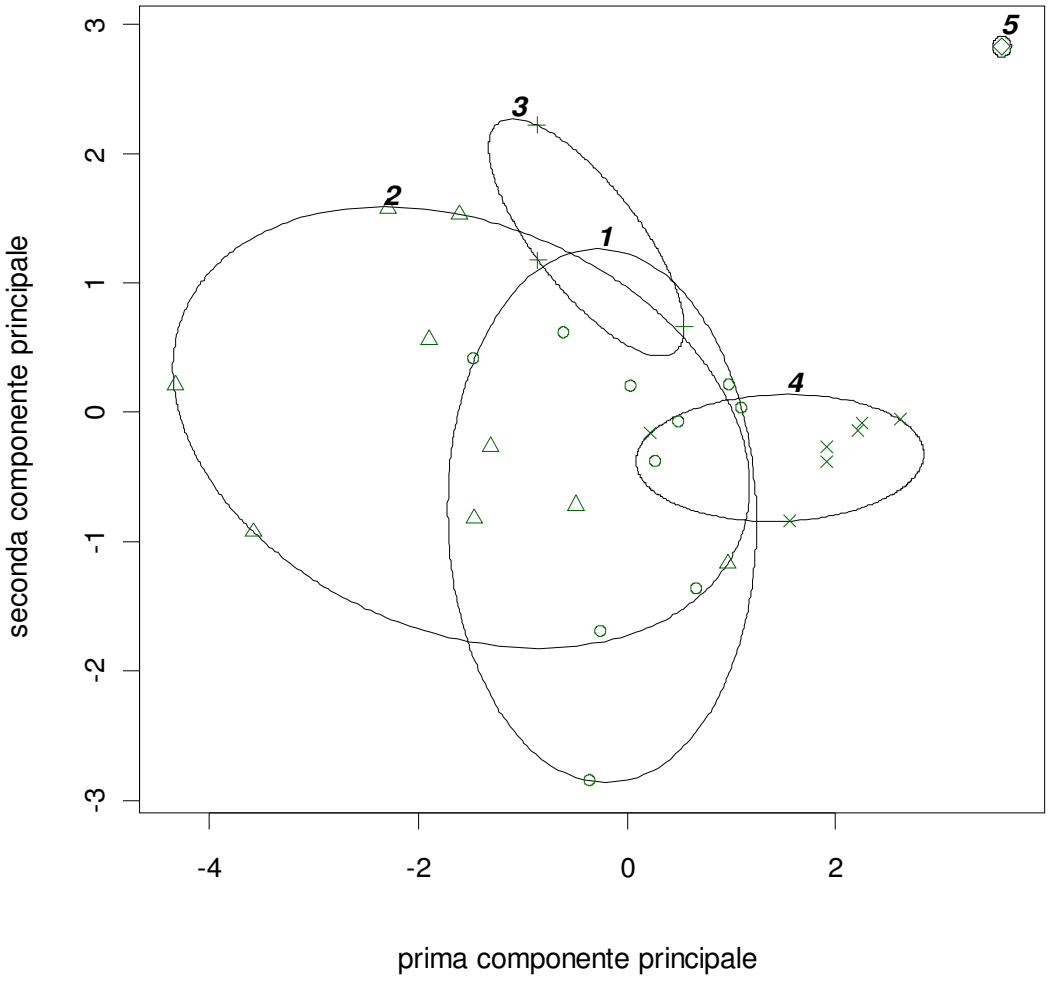

Figura 2.5 : Clusplot dei centroidi ottenuti con il metodo delle k-medie

La Figura 2.5 riporta il clusplot della suddivisione in gruppi: i centroidi sono rappresentati in un grafico a due dimensioni facendo ricorso alle componenti

principali[2]; ogni centroide è contrassegnato da un simbolo secondo il gruppo di appartenenza. Ciascun gruppo è rappresentato da un'elisse la cui area è la minima contenente tutti i punti del gruppo, ciò spiega perché c'è sempre almeno un punto sul contorno dell'ellisse; le ellissi sono costruite basandosi sulla media e sulla matrice di covarianza di ciascun gruppo. Nel caso di gruppi formati da un solo elemento, viene visualizzato un piccolo cerchio attorno al punto. Accanto ad ogni ellisse è indicato il gruppo di riferimento.

Si può notare come il gruppo 5, composto da utenti che hanno visitato solo l'area "pubblicazioni", sia nettamente separato dagli altri ed essendo costituito da un unico elemento, si può affermare che esso era già stato identificato nella prima fase dell'analisi cluster, relativamente al raggruppamento tramite l'algoritmo delle k-medie.

Inoltre, confrontando la distribuzione delle osservazioni nei gruppi (Figura 2.3) con la rappresentazione dei gruppi nel clusplot (Figura 2.4), si nota come non vi sia una relazione tra dimensione e ampiezza dei cluster. Per esempio, il gruppo 4, pur contenendo più della metà delle osservazioni, non è il più esteso: ciò significa che le osservazioni appartenenti al gruppo sono meno disperse rispetto alle altre.

2.3 Interpretazione dei risultati

Per ogni cluster è riportato un grafico con il confronto tra la media del gruppo (rappresentata dal punto pieno) e la media dell'intero campione (rappresentata dal punto vuoto) corredato da una descrizione del comportamento di visita degli utenti.

[2] È importante sottolineare che le componenti principali utilizzate nel clusplot vengono calcolate sull'insieme dei dati su cui viene applicato il metodo di partizione. Pertanto esse vengono calcolate sui centroidi, non sull'insieme di tutte le osservazioni. Questo spiega perché la rappresentazione dei centroidi nel clusplot differisce da quella della Figura 2.1.

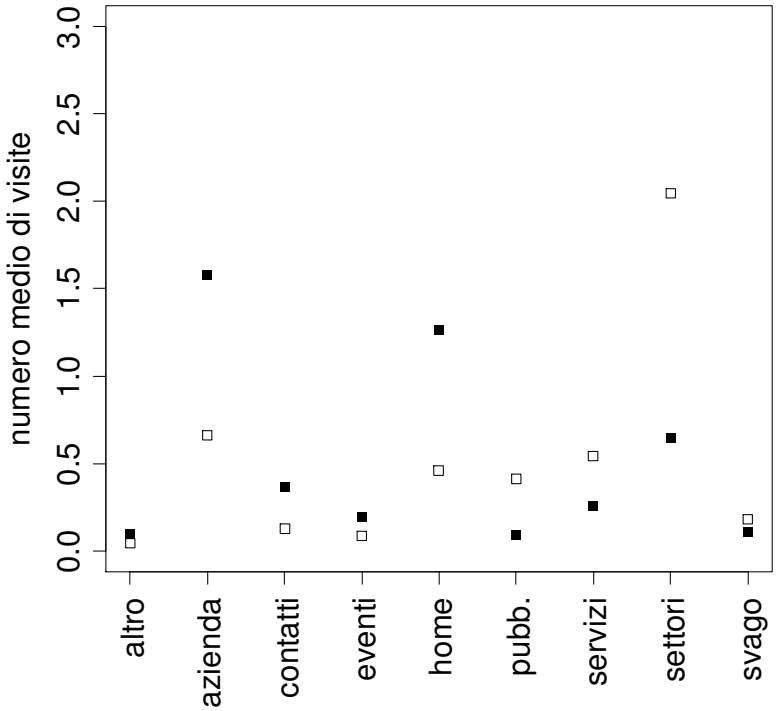

Figura 2.6: Cluster 1

Segmento 1: è costituito da utenti che visitano congiuntamente l'home page e la sezione del sito dedicata alla presentazione dell'azienda: Per queste aree, e anche per l'area "contatti", le visite sono superiori alla media. Si può ipotizzare che gli utenti di questo segmento non conoscano l'azienda e accedano a queste aree del sito per reperire delle informazioni generali sull'azienda stessa.

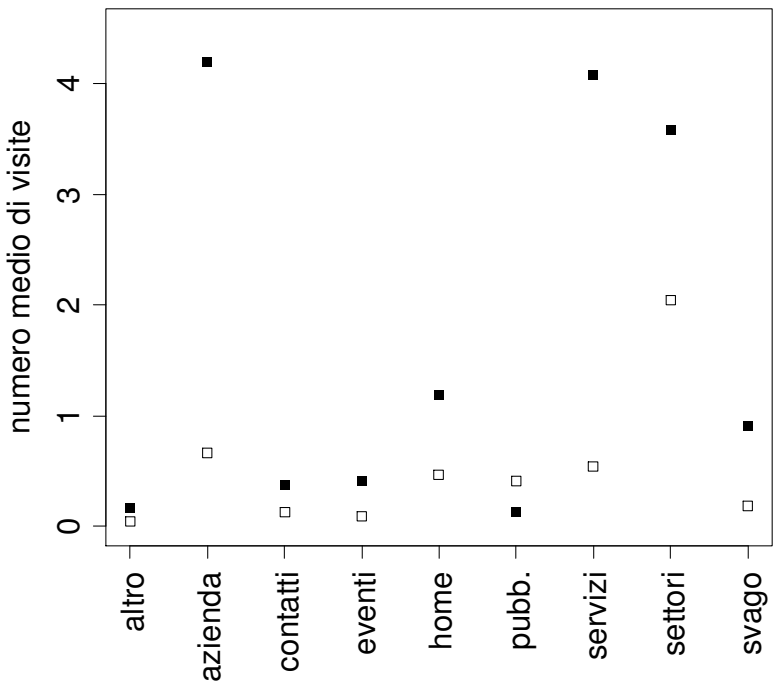

Figura 2.7: Cluster 2

Segmento 2: i visitatori appartenenti a questo gruppo sono caratterizzati da un utilizzo del sito rivolto soprattutto all'accesso ad aree relative all'azienda, ai servizi che essa offre e ai settori nei quali lavora. Le visite alle altre aree sono comunque superiori alla media (anche se di poco), tranne che per l'area "pubblicazioni". Probabilmente, gli utenti di questo segmento desiderano approfondire la conoscenza dell'azienda.

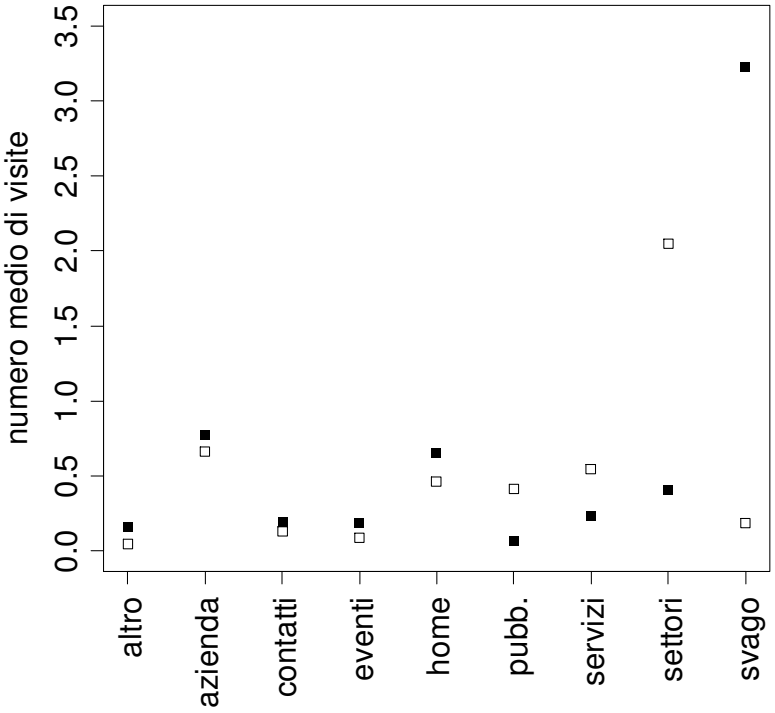

Figura 2.8: Cluster 3

Segmento 3: è caratterizzato da un elevato accesso all'area "svago" e da un numero di visite pari alla media del campione per le aree "home" e "azienda". Gli utenti di questo segmento, dopo aver raccolto informazioni sull'azienda, probabilmente sono attratti dall'area "relax".

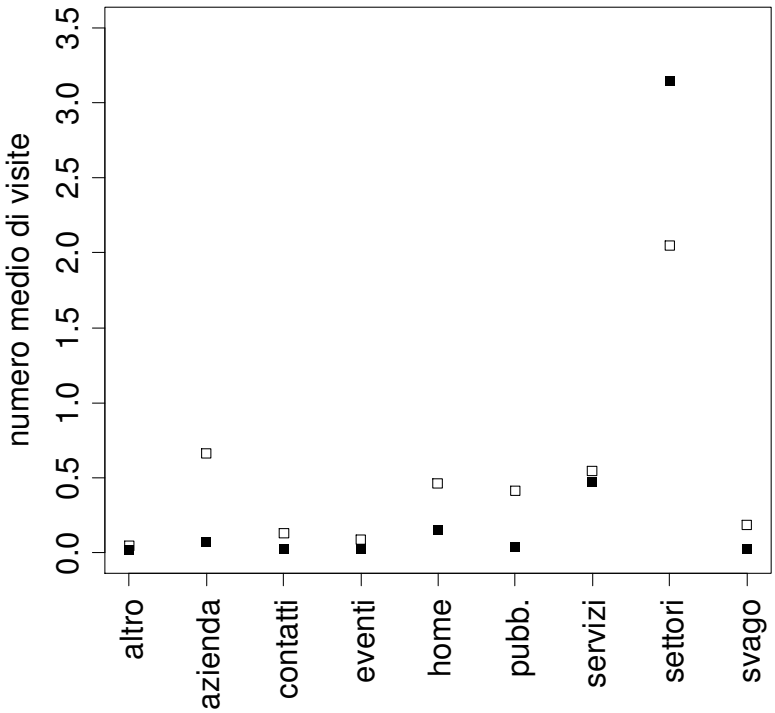

Figura 2.9: Cluster 4

Segmento 4: i visitatori appartenenti a questo gruppo accedono quasi esclusivamente all'area "settori" infatti sono molto vicini allo zero i valori delle medie per le altre aree, tranne che per "servizi". Inoltre, il numero medio di visite all'area "settori" risulta superiore alla media dell'intero insieme. Ciò fa supporre che gli utenti appartenenti a questo gruppo abbiano avuto accesso all'area da link presenti in altri siti oppure che siano stati indirizzati verso il sito dell'azienda da un motore di ricerca e, successivamente, spinti da curiosità, abbiano visitato anche altre pagine della stessa area.

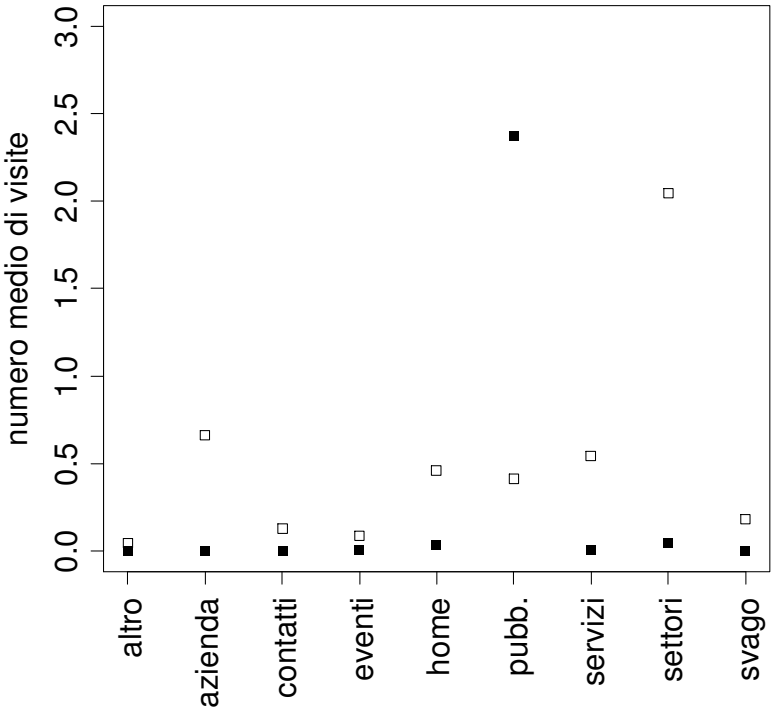

Figura 2.10: Cluster 5

Segmento 5: è caratterizzato da un numero medio di visite all'area “pubblicazioni” molto superiore alla media dell'intero insieme, mentre sono praticamente nulli gli accessi alle altre aree. Anche in questo caso, come per il segmento 4, si può ipotizzare che gli utenti appartenenti a questo gruppo abbiano avuto accesso all'area da link in altri siti oppure che siano stati indirizzati verso il sito dell’azienda da un motore di ricerca e, successivamente, spinti da curiosità, abbiano visitato anche altre pagine della stessa area.

Capitolo 3

Analisi delle sequenze di visita

3.1 Regole di associazione

La possibilità di individuare delle regolarità nelle dinamiche di navigazione all'interno di un sito web rappresenta un'opportunità per capire i comportamenti di visita degli utenti al sito e quindi anche per adeguare il *layout* del sito stesso alle loro esigenze di navigazione.

Una delle metodologie più utilizzate per lo studio delle sequenze di visita si basa sull'analisi delle regole di associazione (Agrawal, Imielinski e Swami, 1993). Le *regole associative* rappresentano una tecnica di data mining per l'apprendimento non supervisionato, attraverso cui è possibile individuare le sequenze di pagine più ricorrenti e le relazioni significative tra insiemi di pagine.

Solitamente, per questo tipo di analisi, i dati sono organizzati in una matrice in cui ogni colonna rappresenta una variabile binaria che indica se la corrispondente pagina del sito è stata visitata almeno una volta e ogni riga rappresenta una sessione utente.

Nell'ambito dell'analisi delle sequenze di visita, si usa l'espressione A→B per indicare che, se è stata visitata la pagina A, allora è stata visitata anche la pagina B, all'interno della stessa sessione utente. L'espressione A→B viene chiamata "regola associativa"; il termine di sinistra è chiamato "antecedente", quello di destra "conseguente". Antecedente e conseguente possono essere costituiti non solo da un'unica pagina, ma anche da insiemi di più pagine, chiamati *itemset*[1].

A una regola possono essere associati i seguenti indici: *supporto*, *confidenza*, *lift.*

[1] Viene chiamato *itemset* ciascun sottoinsieme ricavabile dall'insieme di tutte le pagine del sito web. Di seguito un generico *itemset* A sarà indicato con il simbolo I_A

Il *supporto* (o *support*) per la regola $I_A \rightarrow I_B$ è il rapporto tra il numero di sessioni in cui gli itemset A e B sono contemporaneamente presenti e il numero totale di sessioni del data set:

$$support(I_A \rightarrow I_B) = \frac{\mathrm{N}(I_A \cup I_B)}{\mathrm{N}}$$

dove $\mathrm{N}(I_A \cup I_B)$ è il numero di sessioni utente in cui compare la regola $I_A \rightarrow I_B$ e N è il numero totale delle sessioni utente.

Dalla definizione di supporto si evince che tale rapporto non dipende dalla direzione della regola, pertanto esso è simmetrico; inoltre, se il numero complessivo di sessioni utente è elevato, l'indice di support esprime la probabilità che una sessione utente contenga la regola $I_A \rightarrow I_B$, cioè $P(I_A \cup I_B)$.

La *confidenza* (o *confidence*) per la regola $I_A \rightarrow I_B$ è il rapporto tra il numero di sessioni utente in cui gli itemset A e B sono contemporaneamente presenti e il numero di sessioni utente in cui è presente l'itemset A:

$$confidence\ (I_A \rightarrow I_B) = \frac{N(I_A \cup I_B)}{N(I_A)} = \frac{\frac{N(I_A \cup I_B)}{N}}{\frac{N(I_A)}{N}} = \frac{support\ (I_A \rightarrow I_B)}{support\ (I_A)}$$

L'indice di confidence può essere interpretato come la probabilità condizionata $P(I_B|I_A)$, ossia come la probabilità che un utente visiti l'*itemset* B, avendo visualizzato nella stessa sessione di visita l'*itemset* A. Tale indice, pertanto, è di tipo asimmetrico.

"Se il *supporto* di una regola associativa serve a valutarne la rilevanza statistica, la *confidenza* rappresenta una misura della significatività statistica dell'implicazione logica che lega antecedente e conseguente." (Milanato, 2008)

Molto spesso, tuttavia, utilizzando solamente questi due indici, si ottengono degli insiemi di regole con numerosità elevata. In queste situazioni si fa ricorso ad un altro indice, chiamato *lift* (Brine et al., 1997).

Esso è definito come segue:

$$lift\ (I_A \rightarrow I_B) = \frac{confidence\ (I_A \rightarrow I_B)}{support\ (I_B)} = \frac{support\ (I_A \rightarrow I_B)}{support\ (I_A)\ support\ (I_B)}$$

L'indice di lift essere interpretato come una stima di $\frac{P(I_A \cup I_B)}{P(I_A)\,P(I_B)}$

È importante notare che l'indice di lift è dato dal rapporto tra la confidenza di una regola e la confidenza attesa per la stessa regola in condizione di indipendenza tra antecedente e conseguente.

L'indice di lift assume un valore:

- superiore all'unità, se $confidence\ (I_A \rightarrow I_B) > support\ (I_B)$
  In questo caso esiste una correlazione positiva fra antecedente e conseguente della regola, e quindi la regola esprime più efficacemente la probabilità che in una sessione utente sia presente la pagina B piuttosto che il solo *support* (I_B)
- inferiore all'unità, se $confidence\ (I_A \rightarrow I_B) < support\ (I_B)$
  In questo caso la regola non deve considerarsi statisticamente attendibile, poiché è meno efficace nel prevedere che la sessione utente contenga il conseguente B piuttosto che il solo *support* (I_B)

Una regola associativa è utile per l'analisi se fornisce informazioni rilevanti, per questo si richiede che una regola soddisfi dei valori minimi per gli indici di support, confidence e lift. Se ciò accade, essa viene detta *regola associativa forte*.

Per estrarre le regole più frequenti e significative tra tutte quelle possibili si utilizza un algoritmo di ricerca chiamato "Apriori" (Agrawal, Imielinski, and Swami, 1994). Esso si basa sull'omonimo principio secondo cui, se un itemset composto da k elementi è frequente, allora un qualsiasi suo sottoinsieme è anch'esso frequente.

Si ha a disposizione, quindi, una tecnica che permette di estrarre dall'insieme di tutte le regole possibili solo le regole "forti", eliminando automaticamente ("a priori", appunto) tutti gli insiemi di elementi di cardinalità superiore[2] a quella di ciascun itemset non frequente, senza dover individuarli e poi scartarli, dopo averne calcolato il supporto. Esso si compone di 2 fasi:
prima fase: individuazione degli itemset frequenti
seconda fase: generazione delle regole associative forti

[2] Un itemset A ha cardinalità maggiore di un itemset B se A contiene tutti gli elementi di B e altri elementi non contenuti in B

A queste due fasi può esserne aggiunta una terza: la fase di valutazione dell'importanza delle regole forti individuate.

Nella prima fase, l'algoritmo Apriori individua tutti gli itemset frequenti, cioè quelli con supporto superiore alla soglia minima (definita secondo il problema trattato), partendo da quelli costituiti da un solo elemento. Successivamente, sulla base del principio Apriori, individua gli itemset frequenti composti da due elementi e itera il procedimento fino a raggiungere il numero massimo di elementi in un itemset, pari al numero di pagine del sito.

Nella seconda fase vengono individuate le regole associative forti, partendo dagli itemset frequenti estratti nella fase 1. Per ciascun itemset frequente si costruiscono tutte le possibili regole, date dalla combinazione degli elementi dell'itemset come antecedente e conseguente. Per ogni regola viene calcolata la confidenza e vengono eliminate tutte le regole che non soddisfano la soglia minima prefissata. Il risultato è un insieme di regole forti.

Al termine di questa fase, viene calcolato l'indice di lift per ciascuna regola forte, in modo da trovare le regole maggiormente esplicative rispetto alla sola presenza del conseguente. Tali regole presentano un valore per l'indice di lift superiore all'unita.

3.2 Applicazione delle regole associative ai dati

La matrice di dati utilizzata nella analisi è stata ricavata dalla matrice "utenti" (§ 1.1), sostituendo le variabili discrete (indicanti il numero di visite effettuato a ciascuna pagina) con delle variabili dicotomiche che assumono valori 1 e 0, indicanti rispettivamente se la pagina è stata visitata almeno una volta oppure no.

Le regole di associazione sono state calcolate mediante l'algoritmo "Apriori" imponendo un valore minimo per l'indice di support pari a 0.01 e pari a 0.80 per l'indice di confidence.

Il set di regole ottenute è riportato nella Tabella 3.1.

	antecedente		conseguente	supporto	confidenza	lift
1	{col139}	=>	{col133}	0,010	0,953	35,013
2	{col142}	=>	{col145}	0,011	0,841	47,640
3	{col142}	=>	{col133}	0,012	0,899	33,016
4	{col145}	=>	{col133}	0,016	0,903	33,161
5	{col142, col145}	=>	{col133}	0,011	0,962	35,343
6	{col133, col142}	=>	{col145}	0,011	0,900	50,997
7	{col136, col145}	=>	{col133}	0,011	0,930	34,149

Tabella 3.1: Regole associative

col139: "http://[...]/azienda/metodologia_di_lavoro" [3]

col133: "http://[...]/azienda"

col136: "http://[...]/azienda/qualifica_personale"

col142: "http://[...]/azienda/collaborazioni"

col145: "http://[...]/azienda/personale_dirigente"

Tutte le regole individuate fanno riferimento a pagine dell'area "azienda" e sono rappresentate nel grafo di Figura 3.1. Per le regole il cui antecedente è composto da un'unica pagina (linea tratteggiata), il conseguente di una regola è individuato dalla punta della freccia, l'antecedente dall'estremità opposta. Per le regole il cui antecedente è composto da due pagine (linea continua), tali pagine sono unite da un segmento non orientato, dal quale parte la freccia in direzione del conseguente.

[3] L'indirizzo dei ciascuna pagina è stato modificato per evitare riferimenti all'azienda e, nello stesso tempo, per dare un'indicazione sul contenuto della pagina.

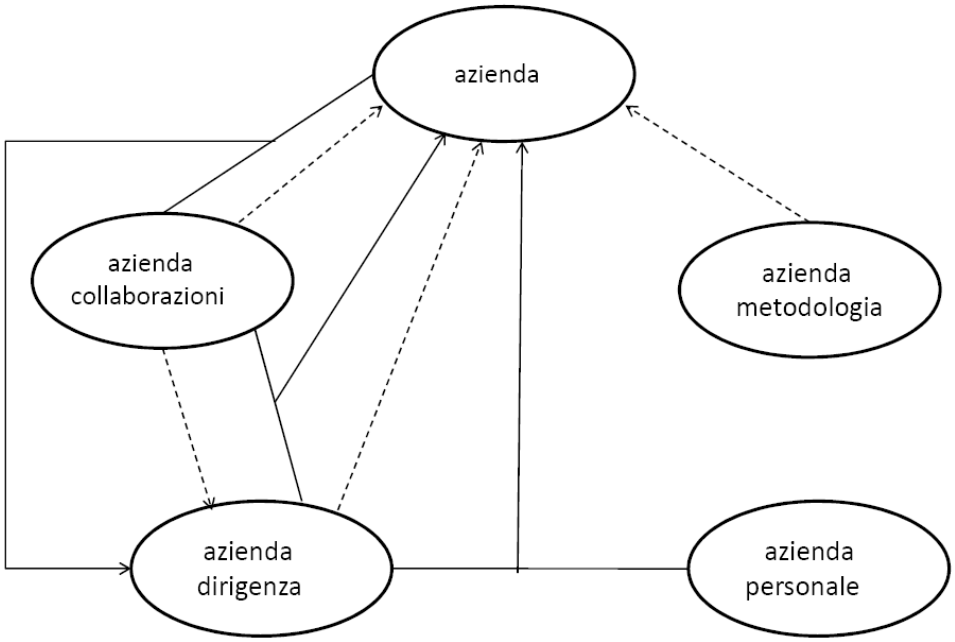

Figura 3.1: Grafo delle regole associative individuate nella matrice di dati completa

Poiché l'insieme di dati a disposizione è caratterizzato dalla presenza di un numero molto elevato di sessioni utente contenenti un'unica visita, si conduce l'analisi solo sulle sessioni di lunghezza maggiore di uno.

Imponendo come valori minimi 0,02 per il supporto, 0,9 per la confidenza e 10 per l'indice di lift, sono state individuate delle regole associative[4] tra pagine all'interno di ciascuna delle aree "azienda", "servizi" e "svago" e trasversalmente tra le aree "azienda", "servizi" e "settori", come mostrato nel grafo di Figura 3.1.

Nel caso in cui le pagine dell'itemset antecedente appartengano alla stessa area (linea tratteggiata), il conseguente di una regola è individuato dalla punta della freccia, l'antecedente dall'estremità opposta. Per itemset antecedenti composti da pagine appartenenti a due o tre aree diverse (rispettivamente linea continua e punto e linea), tali aree sono state unite da un segmento non orientato, dal quale parte la freccia in direzione del conseguente. Lo spessore dei tratti indica a quali aree appartengono le associazioni più frequenti. Se non diversamente specificato, l'indice di lift per le regole assume un valore compreso tra 11 e 12.

[4] Si veda l'allegato 2

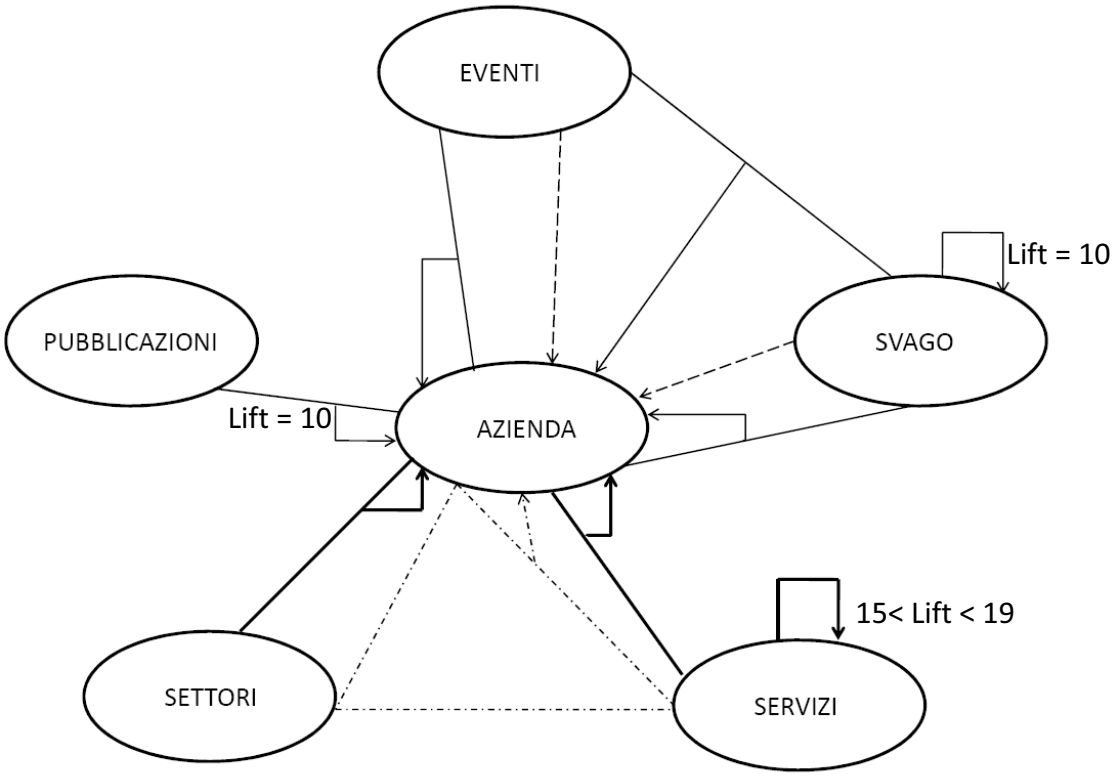

Figura 3.2: Grafo delle regole associative individuate nella matrice di dati ridotta

Dall'analisi delle regole individuate in entrambi i casi, risulta che le aree "azienda" e "settori" sono caratterizzate da una correlazione maggiore al loro interno, in quanto la probabilità di che compaia una pagina dell'area "azienda" (o "settori") in una sessione che contiene già due pagine appartenenti alla medesima area è circa del 90%. Oltre alle associazioni all'interno di queste due aree, emergono regole associative anche *tra* le aree "azienda", "settori" e "servizi", poiché, nonostante siano presenti in un numero ridotto di sessioni utenti (supporto pari al 2%), la probabilità di avere un pagina dell'area "azienda" in una sessione che contiene pagine delle aree "azienda", "settori" e "servizi" è, anche in questo caso, pari al 90%.

Capitolo 4

Previsione dei comportamenti di visita

4.1 Descrizione dei dati

Per l'azienda può essere interessante valutare se esiste una relazione tra il percorso di visita effettuato e l'area a cui appartiene la pagina che conclude la sessione utente, soprattutto se si considera l'area "contatti".

I dati sulle sessioni utente sono stati organizzati in una matrice con 26226 righe, una per ogni sessione considerata, e 113 colonne: la prima corrisponde alla prima pagina visitata nella sessione, la seconda colonna alla seconda pagina e così via, fino all'ultima colonna, che rappresenta l'ultima pagina visitata dall'utente che ha riportato la sessione di visita più lunga. In questo modo ogni riga riporta la successione delle pagine visitate, mantenendo l'ordinamento temporale della visita. Per maggior chiarezza, di seguito è riportato uno stralcio della matrice.

	pagina 1	pagina 2	pagina 3	pagina 4	pagina 5	pagina 6	pagina 7	pagina 8
1	pubblicazioni	-	-	-	-	-	-	-
2	pubblicazioni	-	-	-	-	-	-	-
3	pubblicazioni	-	-	-	-	-	-	-
4	home	servizi	servizi	servizi	servizi	servizi	home	settori
5	home	-	-	-	-	-	-	-
6	azienda	azienda	azienda	svago	svago	svago	svago	-
7	home	-	-	-	-	-	-	-
8	home	contatti	-	-	-	-	-	-
9	home	-	-	-	-	-	-	-
10	settori	settori	-	-	-	-	-	-
11	azienda	-	-	-	-	-	-	-
12	contatti	altro	servizi	-	-	-	-	-
13	home	-	-	-	-	-	-	-
14	settori	settori	-	-	-	-	-	-
15	settori	-	-	-	-	-	-	-
16	settori	-	-	-	-	-	-	-

Tabella 4.1: Stralcio della matrice dei dati

L'insieme di dati considerato nelle analisi è stato costruito a partire dalla matrice precedente, considerando le ultime cinque pagine visitate per gli utenti che hanno visitato almeno cinque pagine nella loro sessione[1]. La matrice così ottenuta contiene 1603 osservazioni e cinque variabili qualitative, una per ognuna delle ultime cinque pagine visitate, che indicano l'area di appartenenza della pagina secondo la classificazione utilizzata nel § 1.1

Per poter valutare la bontà della previsione in seguito alla stima del modello, l'insieme di dati è stato suddiviso in due campioni: campione di apprendimento (1303 unità) e campione di verifica (300 unità).

Di seguito sono riportate le frequenze percentuali di visita a ciascuna area del sito, limitatamente all'ultima pagina visitata per il campione di apprendimento:

area	frequenza percentuale
settori	38,9
azienda	16,0
servizi	11,7
home	9,7
svago	7,0
contatti	6,3
pubblicazioni	5,6
eventi	3,2
altro	1,5

Tabella 4.2: Distribuzione percentuale delle frequenze di visita per l'ultima pagina visitata

Dall'analisi della tabella 4.2 si nota che solo il 6,3% delle sessioni utenti si conclude con l'accesso all'area "contatti"; tale percentuale risulta inferiore a un ragionevole tasso di errore accettabile per un modello di previsione. Infatti, prevedendo tutti gli ultimi accessi come non appartenenti all'area "contatti" si avrebbe un errore inferiore al 7%, quindi accettabile, anche se una previsione di questo tipo sarebbe totalmente inutile.

Una soluzione solitamente adottata in queste situazioni consiste nell'utilizzare per la stima del modello un insieme di dati che contenga tutte le unità che hanno come ultimo accesso l'area "contatti" e un campione casuale di numerosità fissata (in questo caso pari al doppio del numero di eventi favorevoli) delle rimanenti

[1] Si veda l'allegato 1

osservazioni. In questo modo si ha a disposizione un campione bilanciato per la stima del modello.

Il campione bilanciato contiene 246 osservazioni: tutte le 82 sessioni per le quali l'ultima pagina visitata appartiene all'area contatti e 164 osservazioni estratte casualmente tra le rimanenti. Le variabili considerate sono le stesse del campione di apprendimento, tranne l'ultima: la variabile qualitativa indicante l'area di appartenenza dell'ultima pagina visitata è stata sostituita da una variabile dicotomica indicante se la sessione utente si è conclusa con la visita all'area "contatti" o meno.

4.2 Previsione per l'area "contatti"

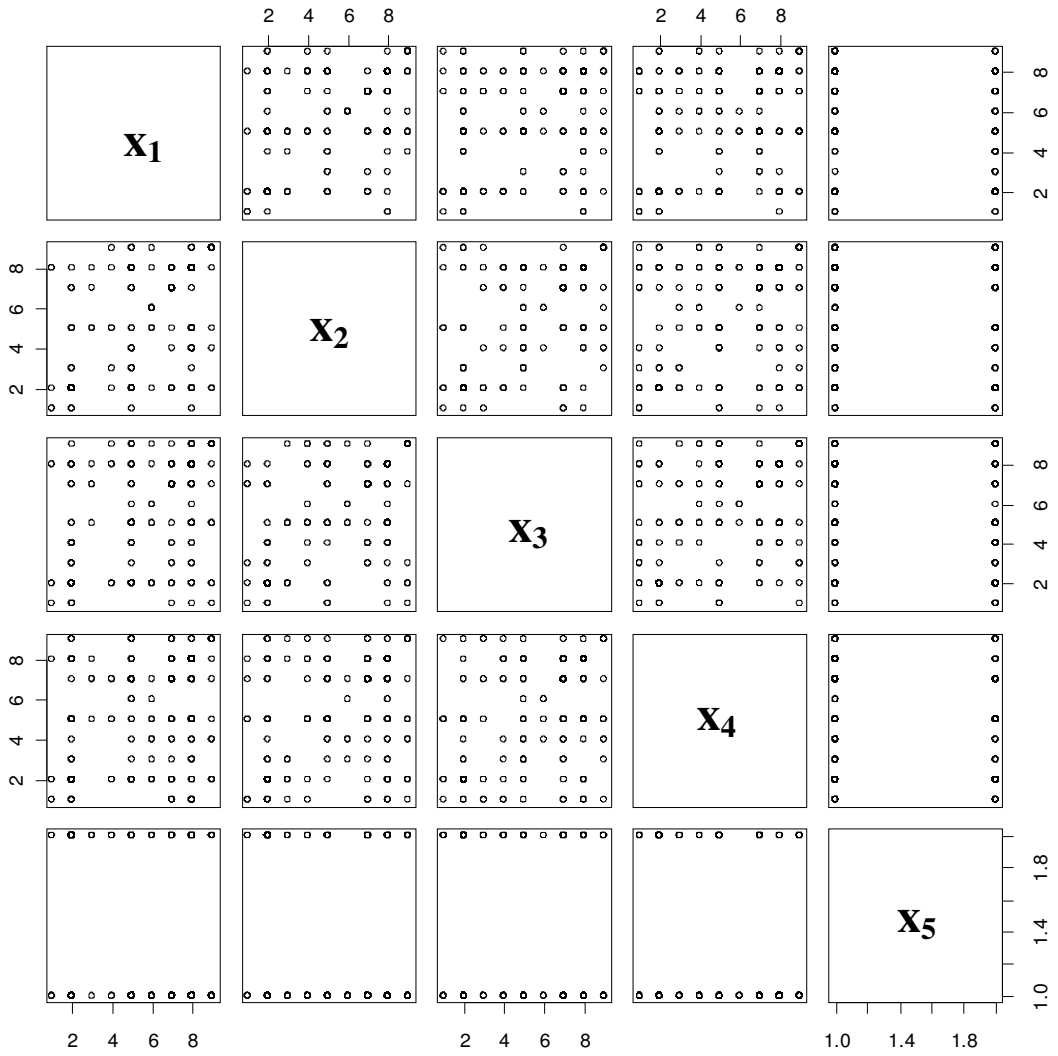

Figura 4.1: Matrice di dispersione per il campione di apprendimento bilanciato

Dall'analisi della matrice di dispersione (Figura 4.1) si può notare che sono presenti valori zero e uno per la variabile dicotomica in corrispondenza di quasi ogni valore delle altre variabili: ciò suggerisce che probabilmente non esiste una relazione significativa tra l'area di appartenenza dell'ultima pagina visitata e le altre variabili prese singolarmente.

È stato stimato un modello di regressione logistica sui dati del campione di apprendimento bilanciato considerando come variabile risposta la variabile dicotomica "x_5" e come regressori le quattro variabili qualitative "x_1", "x_2", "x_3", "x_4". I parametri stimati sono riportati nella Tabella 4.3.

parametri	stima	Standar Error	z value	p-value	
intercetta	0,70	1,39	0,50	0,615	
x1azienda	0,11	1,18	0,09	0,926	
x1contatti	2,71	1,77	1,53	0,126	
x1eventi	1,19	1,52	0,78	0,433	
x1home	-0,46	1,23	-0,37	0,709	
x1pubblicazioni	1,93	1,67	1,16	0,247	
x1servizi	0,34	1,25	0,27	0,784	
x1settori	-0,56	1,17	-0,48	0,631	
x1svago	-1,16	1,40	-0,83	0,407	
x2azienda	0,37	1,16	0,32	0,753	
x2contatti	-0,93	1,39	-0,67	0,504	
x2eventi	-2,14	1,43	-1,50	0,133	
x2home	-0,92	1,20	-0,76	0,445	
x2pubblicazioni	-34,40	1897,98	-0,02	0,986	
x2servizi	-1,65	1,20	-1,37	0,170	
x2settori	-0,58	1,17	-0,50	0,618	
x2svago	-0,73	1,32	-0,55	0,583	
x3azienda	-0,72	0,97	-0,74	0,458	
x3contatti	0,28	1,20	0,23	0,817	
x3eventi	1,32	1,20	1,10	0,272	
x3home	1,46	1,12	1,30	0,193	
x3pubblicazioni	16,87	1275,26	0,01	0,989	
x3servizi	-0,45	1,07	-0,42	0,673	
x3settori	-0,62	0,99	-0,63	0,531	
x3svago	1,13	1,12	1,00	0,316	
x4azienda	-0,15	0,79	-0,19	0,846	
x4contatti	-0,99	1,11	-0,89	0,376	
x4eventi	-0,59	1,08	-0,55	0,583	
x4home	0,44	0,79	0,56	0,576	
x4pubblicazioni	-32,54	2219,92	-0,02	0,988	
x4servizi	-0,34	0,83	-0,41	0,684	
x4settori	-2,44	0,88	-2,78	0,005	**
x4svago	-0,35	0,91	-0,38	0,702	

Tabella 4.3: Stime di massima verosimiglianza dei parametri del modello di regressione logistica

I parametri sono tutti non significativi, tranne uno: sembra che esista una relazione solo tra aver visitato l'area settori nella penultima pagina e visitare per ultima l'area "contatti". Provando a considerare come regressore solo la penultima pagina (variabile "x_4") si ottiene lo stesso risultato (Tabella 4.4): risulta significativo solo il parametro relativo alla modalità "settori" della variabile con un p-value inferiore a 0.1% .

parametri	stima	Standar Error	z value	p-value	
intercetta	0,0000	0,5345	0,0000	1,0000	
x4azienda	0,0715	0,5977	0,1200	0,9048	
x4contatti	-0,6931	0,8864	-0,7820	0,4342	
x4eventi	-0,8473	0,8729	-0,9710	0,3317	
x4home	0,2683	0,6492	0,4130	0,6794	
x4pubblicazioni	-15,5700	840,3000	-0,0190	0,9852	
x4servizi	-0,8210	0,6455	-1,2720	0,2034	
x4settori	-2,8180	0,7421	-3,7980	0,0001	***
x4svago	-0,1178	0,7224	-0,1630	0,8705	

Tabella 4.4: Stime di massima verosimiglianza dei parametri del modello di regressione logistica con regressore x_4

I risultati ottenuti suggeriscono di utilizzare come regressore solamente una variabile dicotomica (chiamata "$x_4$settori") che indichi se la penultima pagina visitata appartiene all'area "settori" o meno (Tabella 4.5).

parametri	stima	Standar Error	z value	p-value	
intercetta	-0,2180	0,1521	-1,433	0,152	
x4settori	-2,6004	0,5367	-4,845	$1.27*10^{-06}$	***

Tabella 4.5: Stime di massima verosimiglianza dei parametri del modello di regressione logistica con regressore $x_4$settori

Per valutare la bontà del modello vengono calcolate le previsioni sul campione di validazione. La Tabella 4.6 riporta i risultati ottenuti:

		previsione		
		0	**1**	**totale**
$\mathbf{x_5}$	**0**	284	0	284
	1	16	0	16
	totale	300	0	300

Tabella 4.6: Confronto tra previsioni corrette ed errate

Il modello non risulta utile per prevedere la probabilità che una sessione si concluda con la visita all'area "contatti", perché sembra non esistere nessuna relazione tra il percorso di visita e la sua conclusione, come già si intuiva dall'analisi della matrice di correlazione dei dati.

4.3 Previsione per l'area "settori"

Poiché la previsione considerando l'area "contatti" non è emersa alcuna relazione tra la visita a questa area e il percorso precedente, si è ripetuta l'analisi effettuata precedentemente considerando come area d'interesse per l'ultima pagina l'area "settori", che ha frequenza relativa maggiore nel campione.

Il campione di riferimento per le analisi è l'intero campione di apprendimento; la variabile qualitativa indicante l'area di appartenenza dell'ultima pagina visitata è stata sostituita da una variabile dicotomica indicante se la sessione utente si è conclusa con la visita all'area "settori" o meno. Per la stima del modello di regressione logistica è stata considerata come variabile risposta la variabile "x_5", e come regressori le quattro variabili qualitative "x_1", "x_2", "x_3", "x_4".

parametri	stima	standard error	z value	p-value	
intercetta	-5,375	1,521	-3,535	0,000	***
x1azienda	-0,401	0,906	-0,442	0,659	
x1contatti	0,216	1,062	0,203	0,839	
x1eventi	0,763	1,101	0,693	0,489	
x1home	0,266	0,912	0,291	0,771	
x1pubblicazioni	-0,665	1,350	-0,493	0,622	
x1servizi	0,510	0,931	0,548	0,584	
x1settori	1,878	0,891	2,107	0,035	*
x1svago	0,638	0,988	0,646	0,518	
x2azienda	1,581	0,995	1,589	0,112	
x2contatti	0,252	1,151	0,219	0,827	
x2eventi	0,893	1,081	0,826	0,409	
x2home	0,080	1,005	0,079	0,937	
x2pubblicazioni	3,150	1,595	1,976	0,048	*
x2servizi	1,072	1,001	1,071	0,284	
x2settori	0,336	0,959	0,350	0,726	
x2svago	1,000	1,050	0,953	0,341	
x3azienda	1,222	0,939	1,302	0,193	
x3contatti	1,000	1,087	0,920	0,358	

x3eventi	0,935	1,122	0,834	0,404	
x3home	0,936	0,927	1,009	0,313	
x3pubblicazioni	-0,256	1,613	-0,159	0,874	
x3servizi	1,127	0,917	1,229	0,219	
x3settori	2,363	0,882	2,680	0,007	**
x3svago	1,213	0,980	1,237	0,216	
x4azienda	-0,127	0,865	-0,147	0,883	
x4contatti	0,092	0,954	0,096	0,923	
x4eventi	0,904	0,937	0,965	0,334	
x4home	1,313	0,852	1,541	0,123	
x4pubblicazioni	-2,062	1,693	-1,218	0,223	
x4servizi	0,330	0,858	0,384	0,701	
x4settori	3,073	0,818	3,757	0,000	***
x4svago	-0,761	0,978	-0,778	0,437	

Tabella 4.7: Stime di massima verosimiglianza dei parametri del modello di regressione logistica

Il modello stimato ha come parametri significativi l'intercetta (con p-value < 0.1%), e i parametri relativi alla modalità "settori" per le variabili "x_3" e "x_4" (con p-value < 5%). Ciò significa che esiste una relazione tra la l'accesso nelle ultime pagine dell'area settori e la conclusione della sessione utente con la visita a una pagina della stessa area.

Considerando come regressore le quattro variabili dicotomiche indicanti se la corrispondente pagina visitata appartiene all'area "settori" o meno, si è ottenuto il seguente risultato (Tabella 4.8):

parametri	stima	standard error	z value	Pr(>|z|)	
intercetta	-2,806	0,149	-18,816	$< 2*10^{-16}$	***
x1settori	1,561	0,215	7,254	$4,04*10^{-13}$	***
x2settori	-0,501	0,262	-1,916	0,0554	.
x3settori	1,372	0,233	5,881	$4,08*10^{-13}$	***
x4settori	2,625	0,202	13,018	$< 2*10^{-16}$	***

Tabella 4.8: Stime di massima verosimiglianza dei parametri del modello di regressione logistica con regressore $x_4$settori

I parametri sono tutti altamente significativi, (tranne quello relativo alla variabile "$x_2$settori" che risulta essere al limite della significatività), perciò si può affermare che esiste una relazione tra aver visitato nelle ultime pagine l'area settori e aver concluso con una visita alla medesima area la sessione utente.

Le previsioni calcolate sul campione di validazione hanno mostrato un tasso di corretta classificazione pari a 85%. (Tabella 4.9)

		previsione		
		0	**1**	**totale**
x_4 settori	**0**	158	17	175
	1	28	97	125
	totale	186	114	300

Tabella 4.9: Confronto tra previsioni corrette e errate

Capitolo 5

Conclusioni

Sin dall'analisi esplorativa si è potuta operare una prima suddivisione degli utenti in due gruppi in base al comportamento di visita, separando gli utenti che ha visitato un'unica pagina nella propria sessione (circa l'83%) dagli altri.

Considerando le sessioni utente di lunghezza maggiore di uno, l'analisi delle sequenze di visita ha confermato quanto emerso dall'analisi cluster: gli utenti che visitano l'area "pubblicazioni" generalmente non visitano altre aree del sito. Questo risultato è coerente con ciò che era stato rilevato dall'analisi della matrice di correlazione dei dati, in quanto la variabile "pubblicazioni" risultava correlata negativamente con tutte le altre variabili.

I risultati ottenuti mediante l'applicazione analisi cluster hanno confermato e approfondito quanto già emerso in un primo tentativo di interpretazione delle osservazioni compiuto nella sottosez. 1.4.1: infatti erano già stati evidenziati due possibili gruppi di osservazioni: quello caratterizzato da valori elevati per le visite all'area "pubblicazioni" (corrispondente al cluster 5) e quello caratterizzato da valori elevati per le visite all'area "settori" (corrispondente al cluster 4).

L'analisi delle sequenze di visita ha evidenziato che, a differenza delle pagine dell'area "pubblicazioni", che si contraddistinguono per essere visitate soprattutto singolarmente (nelle sessioni utente composte da un'unica visita), e solo in misura minore a gruppi, le aree "azienda" e "settori" sono caratterizzate da una correlazione maggiore al loro interno (in particolare, per l'area "settori", anche mediante l'analisi cluster è emerso un gruppo, il più numeroso, caratterizzato da un numero di visite a quest'area superiore alla media). Oltre alle associazioni all'interno di queste due aree, emergono regole associative anche *tra* le aree "azienda", "settori" e "servizi", come evidenziato anche in uno dei gruppi individuati mediante l'analisi cluster.

Tuttavia, per quanto riguarda la previsione dei comportamenti di visita, non è stato possibile individuare uno schema ricorrente che permettesse di prevedere la

conclusione del percorso con l'accesso all'area "contatti". L'unica relazione emersa tra il percorso di visita e la conclusione dello stesso si è avuta con la visita all'area "settori", per la quale si può affermare che esiste una relazione tra la visita nelle ultime pagine l'area settori e aver concluso la sessione utente con una visita alla medesima area.

ALLEGATO 1

Per le analisi del capitolo 2 è stato utilizzato il pacchetto "cluster"

Per le analisi del capitolo 3 è stato utilizzato il pacchetto "arules"

Capitolo 1

```
> # il data set è stato importato in R con il nome di "dati"
> # per ordinare le righe secondo l'identificativo utente:
> dati<-dati[order(dati$sessionID),]
> # per eliminare le duplicazioni di righe:
> dati<-dati[!duplicated(dati$ID),]
> # "url" è la variabile d'interesse contenente gli indirizzi delle
pagine visitate. È stata ripulita
> dati<-dati[dati$url!=levels(dati$url)[1],]
> dati<-dati[dati$url!=levels(dati$url)[2],]
> dati$url<-factor(dati$url)
> url<-dati$url
> # è stata costruita la variabile "area" che sostituisce, per ogni
pagina visitata nella variabile "url", l'area di appartenenza della
pagina stessa

> # costruzione del data set "utenti"
> utenti<-matrix(NA,nsess, 308)
> for (i in 1:nsess) {
+ utenti[i,] <- table(dati$url[dati$sessionID==id.sess[i]])
+ }
> utenti<-data.frame(utenti)

> # costruzione dei data set "visitatori"

> ni<-table(dati$sessionID)
> nsess<-length(ni)
> id.sess<-as.numeric(names(ni))

> visitatori<-matrix(NA,nsess,9)
> for (i in 1:nsess) {
+ visitatori[i,] <- table(area[dati$sessionID==id.sess[i]])
+ }
> dimnames(visitatori)[[2]]<-levels(area)
> visitatori<-data.frame(visitatori)

> # costruzione dei data set "visite1" e "visite2"
> somma<-apply(visitatori,2,sum)
> visite1<-visitatori[somma=1,]
> visite2<-visitatori[somma>1,]

> # eliminazione degli outlier dal data set "visite2"

> quantile<-apply(visite,2,quantile,probs=0.99)
> quantile
altro azienda contatti eventi home
3.00 15.00 3.00 2.39 7.00
pubblicazioni servizi settori svago
8.00 12.00 18.00 9.00
```

```
> visite2<-visite2[visite2[,1]<=3,]
> visite2<-visite2[visite2[,2]<=15,]
> visite2<-visite2[visite2[,3]<=3,]
> visite2<-visite2[visite2[,4]<=2,]
> visite2<-visite2[visite2[,5]<=7,]
> visite2<-visite2[visite2[,6]<=8,]
> visite2<-visite2[visite2[,7]<=12,]
> visite2<-visite2[visite2[,8]<=18,]
> visite2<-visite2[visite2[,9]<=9,]
```

Capitolo 3

```
> # data set contenente tutte le unità
> dtab<-matrix(0,nsess, 308)
> for (i in 1:nrow(utenti))
+ for (j in 1:ncol(utenti)) {
+ if (utenti[i,j]>=1) dtab[i,j]<-1
+ }
> adtab<-as(dtab,"itemMatrix")

> # data set contenente solo le unità che hanno visitato più di una
pagina
> # rimozione delle unità che hanno effettuato una sola visita
> somma1<-apply(dtab,1,sum)
> dtab<-dtab[somma1>1,]

> # rimozione delle variabili per cui non è stata effettuata alcuna
visita
> somma2<-apply(dtab,2,sum)
> dtab<-dtab[,somma2>0]

> adtab<-as(dtab,"itemMatrix")
```

Capitolo 4

```
> ni<-table(dati$sessionID)
> nsess<-length(ni)
> id.sess<-as.numeric(names(ni))

> sequenze<-matrix(NA,nsess, 113)
> for (i in 1:nsess) {
+ sequenze[i,1:ni[i]] <- area[dati$sessionID==id.sess[i]]
+ }
> sequenze<-data.frame(sequenze)

> # costruzione del data set "matrice"
> uni<-114-ni
> matrice<-matrix(NA,nsess,113)
> for (i in 1:nrow(sequenze)) {
+ matrice[i,uni[i]:113]<-sequenze[i,1:ni[i]]
+ }
> dimnames(matrice)[[2]]=dimnames(sequenze)[[2]]
> matrice<-matrice[,109:113]
> matrice<-data.frame(matrice)
> completo<-complete.cases(matrice)
> matrice<- matrice[completo,]
> row.names(matrice)<-c(1:nrow(matrice))
> names(matrice)<-c("x1","x2","x3","x4","x5")
> matrice$settori<-as.factor(matrice$settori)
```

ALLEGATO 2

Regole associative per la matrice di dati senza sessioni utente di lunghezza unitaria. Supporto > 0.02, confidenza > 0.9 e lift > 10

	antecedente		conseguente	supporto	confidenza	lift
1	{col112,					
	col133}	=>	{col104}	0,021	0,986	10,059
2	{col218,					
	col230}	=>	{col236}	0,030	0,938	15,454
3	{col218,					
	col221}	=>	{col236}	0,029	0,919	15,138
4	{col104,					
	col148}	=>	{col139}	0,023	0,911	11,754
5	{col91,					
	col148}	=>	{col139}	0,026	0,901	11,623
6	{col148,					
	col233}	=>	{col139}	0,032	0,918	11,843
7	{col148,					
	col271}	=>	{col139}	0,031	0,909	11,728
8	{col218,					
	col221,					
	col230}	=>	{col236}	0,026	0,968	15,954
9	{col218,					
	col221,					
	col236}	=>	{col230}	0,026	0,902	19,182
10	{col221,					
	col230,					
	col233}	=>	{col236}	0,026	0,920	15,156
11	{col218,					
	col230,					
	col233}	=>	{col236}	0,027	0,932	15,355
12	{col218,					
	col221,					
	col233}	=>	{col236}	0,027	0,921	15,169
13	{col104,					
	col142,					
	col148}	=>	{col139}	0,021	0,925	11,933
14	{col104,					
	col145,					
	col148}	=>	{col139}	0,022	0,917	11,826
15	{col104,					
	col133,					
	col148}	=>	{col139}	0,023	0,921	11,886

16	{col91,					
	col142,					
	col148}	=>	{col139}	0,024	0,923	11,908
17	{col91,					
	col136,					
	col148}	=>	{col139}	0,024	0,922	11,897
18	{col91,					
	col145,					
	col148}	=>	{col139}	0,025	0,927	11,960
19	{col142,					
	col148,					
	col233}	=>	{col139}	0,028	0,925	11,936
20	{col142,					
	col148,					
	col271}	=>	{col139}	0,028	0,917	11,826
21	{col136,					
	col148,					
	col233}	=>	{col139}	0,027	0,930	11,998
22	{col136,					
	col148,					
	col271}	=>	{col139}	0,027	0,913	11,773
23	{col145,					
	col148,					
	col233}	=>	{col139}	0,030	0,929	11,987
24	{col145,					
	col148,					
	col271}	=>	{col139}	0,030	0,920	11,873
25	{col148,					
	col233,					
	col271}	=>	{col139}	0,024	0,913	11,779
26	{col133,					
	col148,					
	col233}	=>	{col139}	0,032	0,925	11,933
27	{col133,					
	col148,					
	col271}	=>	{col139}	0,031	0,909	11,728

Classificazione delle pagine in aree:

azienda: col132 – col150
contatti: col83 - col86
eventi: col90 - col 97
home: col79 - col82 e col126 - col128
pubblicazioni: col151 – col217
servizi: col218 - col238
settori: col239 - col308
svago: col98- col125
altro: col87 - col89 e col129 - col131

Bibliografia

Agrawal, R., Imielinski, T. e Swami, A.: Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pagine 207-216. ACM Press, 1993. URL http://doi.acm.org/10.1145/170035.170072.

Azzalini, A. e Scarpa, B. (2004): *Analisi dei dati e data mining*, Springer

Berry, M. J. A. e Linoff, G. (2001): *Data mining*, Apogeo

S. Brin, R., Motwani, J., D. Ullman, e S. Tsur. Dynamic itemset counting and implication rulesfor market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pagine 255-264, Tucson, Arizona, USA, Maggio 1997.

Giudici, P. (2001): *Data Mining. Metodi statistici per le applicazioni aziendali*, McGraw-Hill

Hahsler, M., Gruen, B. e Hornik, K. (2005), arules -- A Computational Environment for Mining Association Rules and Frequent Item Sets, *Journal of Statistical Software* 14/15. URL: http://www.jstatsoft.org/v14/i15/.

Hastie, T., Tibshirani, R. e Friedman, J. (2001): *The Elements of Statistical Learning: Data mining, Inference and Prediction*, Spinger

Iacus, S. M. e Masarotto, G. (2003): *Laboratorio di statistica con R*, McGraw-Hill

Kaufman, l. e Rousseeuw, P. J. (1990): *Finding Groups in Data. An Introduction to Cluster Analysis*, New York: Jhon Wiley and Sons.

Mardia, K., Kent, J. e Bibby, J. (1979): *Multivariate Analysis*, Academic Press.

Milanato, D. (2008): *Demand Planning – Processi, metodologie e modelli matematici per la gestione della domanda commerciale*, Springer- Verlag Italia, Milano.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Pison, G., Struyf, A. and Rousseeuw, P.J. (1999), Displaying a clustering with CLUSPLOT, Computational Statistics and Data Analysis, 30, 381-392.

Venables, W. N. e Ripley, B. D. (1999): Modern Applied Statistics with S-PLUS, terza edizione, Springer