Market Basket Analysis
Regole associative

Relazioni tra prodotti

- Le regole associative e i filtri collaborativi sono metodi di apprendimento non supervisionato molto utilizzati nel marketing
- Nelle regole associative, l'obiettivo è identificare quali gruppi di prodotti tendono a essere acquistati insieme
- Nei filtri collaborativi, l'obiettivo è realizzare una *raccomandazione personalizzata* basata su informazioni a livello di singolo consumatore.

Regole associative

In numerosi ambiti di applicazione la raccolta sistematica di dati si traduce in elenchi di transazioni che si prestano a essere analizzati mediante regole di associazione.

L'uso di queste regole di associazione nel marketing spesso viene indicato come Market Basket Analysis, MBA.

La MBA rappresenta lo studio di "che cosa va con che cosa".

Regole associative

Transazione commerciale: lista di articoli acquistati dallo stesso cliente nell'ambito di una stessa visita.

Tipicamente, per ciascuna transazione che corrisponde a un pagamento vengono registrati:

- elenco di prodotti acquistati e relativa numerosità
- prezzo
- importo complessivo
- modalità di pagamento
- id cliente nel caso di tessere fedeltà

Regole associative

I manager sono interessati a sapere se ci sono gruppi di prodotti che sono spesso acquistati insieme.
Tale informazione può essere utile per

- pianificare la disposizione dei prodotti nei negozi
- il design dei cataloghi
- la messa a punto di promozioni
- l'identificazione di segmenti di consumatori sulla base dei comportamenti di acquisto

Regole associative

Le regole associative hanno quindi lo scopo di identificare regolarità di acquisto.
Tali regolarità di acquisto sono esprimibili tramite regole probabilistiche, del tipo:

- se un cliente acquista il prodotto x, acquista anche il prodotto y con probabilità p_1
- tre prodotti {a,b,c} sono acquistati congiuntamente con probabilità p_2

Il tipico "*frequently bought together*" di Amazon.com si basa su regole di questo tipo.

Regole associative

Per essere utile, una regola associativa deve essere non banale e facilmente comprensibile.
Può trovare giustificazione come conseguenza di:

- azioni esogene: mode, variazioni dei gusti dei consumatori, azioni di concorrenti
- azioni endogene: azioni promozionali svolte nel passato, introduzione di nuovi prodotti che ne hanno cannibalizzati altri

Regole associative

L'idea sottostante alle regole associative è di esaminare tutte le possibili regole tra oggetti (ad es. prodotti) secondo una logica

$$\text{se} \rightarrow \text{allora}$$

$$\text{antecedente} \rightarrow \text{conseguente}$$

L'antecedente e il conseguente sono insiemi di oggetti disgiunti
$\rightarrow$ che quindi non hanno elementi in comune.

Tali insiemi sono chiamati itemset.

Regole associative

Esempio

trans ID	items
1	latte, pane
2	pane, burro
3	birra
4	latte, pane, burro
5	pane, burro

L'insieme di item in questo esempio è $I = \{latte, pane, burro, birra\}$
Un esempio di regola potrebbe essere

$$\{pane, latte\} \Rightarrow \{burro\}$$

se un cliente compra pane e latte, comprerà anche burro.

Regole associative

- Il primo passo nelle regole associative consiste nel generare tutte le possibili regole che potrebbero indicare associazioni tra item.
- Questo significa trovare tutte le possibili combinazioni tra item in un database di p item . . . ma potrebbe essere oneroso e inutile!
- Una soluzione pratica consiste nel considerare solo le combinazioni che compaiono con elevata frequenza nel database, ovvero identificare i cosidetti itemset frequenti.

Regole associative

Che cosa identifica un itemset frequente?

- Dobbiamo definire il concetto di supporto.
- Il supporto di un itemset X è la proporzione delle transazioni nel dataset che contengono quel itemset.
- Nell'esempio di prima, l'itemset $\{latte, pane\}$ ha supporto $2/5 = 0.4$, compare cioè nel 40% delle transazioni.

Regole associative

- Definiamo inoltre la fiducia di una regola come

$$\text{fiducia} = \frac{(\text{transazioni con antecente e conseguente})}{(\text{transazioni con antecedente})}$$

- La fiducia può essere vista come una probabilità condizionata

$$\text{fiducia} = \frac{p(\text{conseguente} \cap \text{antecedente})}{p(\text{antecedente})} = p(\text{cons}|\text{ant})$$

Regole associative

Esempio

- Supponiamo che in un supermercato ci siano 100000 transazioni.
- Di queste, 2000 includono sia succo d'arancia che aspirina, e 800 includono anche l'acquisto di sapone.
- La regola $\{succo, aspirina\} \Rightarrow \{sapone\}$ ha
- supporto = $800/100000 = 0.8\%$
- fiducia = $800/2000 = 40\%$

Regole associative

- Un elevato livello di fiducia può suggerire una regola associativa forte.
- Tuttavia bisogna fare attenzione perchè se tutti i consumatori comprano banane e quasi tutti comprano gelato, il livello di fiducia nella regola "se banane, allora gelato" sarà elevato *a prescindere dalla reale associazione tra gli item*.

Regole associative

- Un modo più affidabile per valutare la forza di una regola associativa consiste quindi nel confrontare la fiducia di una regola con un benchmark, nel quale assumiamo che antecedente e conseguente siano indipendenti.
- In ipotesi di indipendenza,

$$\text{fiducia} = p(\text{conseguente})$$

Regole associative

- Possiamo quindi confrontare la fiducia di una regola con il suo benchmark, attraverso il lift

$$\text{lift} = \frac{p(\text{conseguente}|\text{antecedente})}{p(\text{conseguente})}$$

- Se il lift è maggiore di 1, allora la regola è utile.
- Valori di lift maggiori di 1 indicano una buona regola, con capacità predittiva superiore alla semplice conoscenza della probabilità del conseguente.

Regole associative

Esempio

- In un negozio di informatica un addetto alle vendite nota una forte relazione tra l'acquisto di macchine fotografiche digitali e stampanti a colori.
- Immaginiamo 1000 transazioni. Di queste:
  - 600 includono fotocamere digitali
  - 750 includono stampanti a colori
  - 400 includono entrambe.
- La regola fotocamera $\rightarrow$ stampante ha una fiducia di $p = 400/600 = 0.66$ e un supporto $s = 400/1000 = 0.4$.

Regole associative

- Se si stabiliscono dei valori soglia di $s_{min} = 0.3$ e $p_{min} = 0.6$ la regola viene selezionata e suggerisce che l'acquisto di una fotocamera induce l'acquisto di una stampante.
- Sembra ragionevole incentivare questo tipo di comportamento con promozioni o pacchetti commerciali?
- La probabilità di acquistare una stampante è 0.75, ben più alta della probabilità di acquistare una stampante dato l'acquisto di una fotocamera!

Regole associative

In questo caso si ha che il lift vale $0.66/0.75 = 0.88$.
Se il lift è minore di 1 la regola che nega il conseguente risulta più efficace della regola originaria.

Regole associative

Esempio

Consideriamo gli acquisti di t-shirt in vari colori

transazione	rosso	bianco	blu	arancio	verde	giallo
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

Cerchiamo regole associative con un supporto del 20%, ovvero regole basate su item che sono stati acquistati insieme in almeno il 20% delle transazioni.

Regole associative

Primo step:
selezionamo gli itemset *frequenti* (con supporto minimo del 20%)

Itemset	Supporto $\geq 20\%$
{rosso}	0.6
{bianco}	0.7
{blu}	0.6
{arancio}	0.2
{verde}	0.2
{rosso, bianco}	0.4
{rosso, blu}	0.4
{rosso, verde}	0.2
{bianco, blu}	0.4
{bianco, arancio}	0.2
{bianco, verde}	0.2
{rosso, bianco, blu}	0.2
{rosso, bianco, verde}	0.2

Regole associative

Secondo step: selezioniamo solo regole con un dato livello di fiducia.
Consideriamo alcune regole associative e i rispettivi livelli di fiducia e lift

Itemset	Fiducia	Lift
{rosso, bianco} $\rightarrow$ {verde}	2/4= 50%	50%/20%= 2.5
{verde} $\rightarrow$ {rosso}	2/2= 100%	100%/60%= 1.67
{bianco, verde} $\rightarrow$ {rosso}	2/2=100%	100%/60%=1.67

Se il livello di fiducia minimo richiesto è del 70% allora utilizzeremo solo la seconda e la terza regola.

Regole associative

E' importante osservare che:

- la fiducia di una regola serve a determinare l'utilità commerciale e operativa di una regola: una regola con un livello di fiducia troppo basso potrebbe non giustificare il costo di promuovere il conseguente in tutte le transazioni in cui è presente l'antecedente
- il lift indica l'efficienza previsiva di una regola, rispetto a una selezione casuale. Ovviamente una regola efficiente è sempre preferibile, ma deve comunque avere un supporto adeguato.

Regole associative

Da un punto di vista pratico, abbiamo bisogno di algoritmi che riescano a trovare le regole più interessanti all'interno di un database.

- L'algoritmo Apriori rappresenta un metodo efficiente per estrarre le regole forti contenute in un insieme di transazioni, in base all'identificazione di itemset frequenti.
- Principio Apriori: Se un insieme di oggetti è frequente, allora anche tutti i suoi sottoinsiemi sono frequenti.
- Corollario: Ogni sovrainsieme di un itemset non frequente non può essere frequente

Regole associative

- E' semplice generare itemset di dimensione 1 frequenti: basta contare, per ogni item quante transazioni lo includono.
  Se un itemset ha un supporto minore di un livello minimo desiderato, viene eliminato.
- A questo punto, per generare un itemset frequente di dimensione 2, basta utilizzare gli itemset frequenti di dimensione 1.
  Se un certo itemset di dimensione 1 non aveva un supporto minimo, nessun itemset che lo contiene potrà avere quel supporto minimo ...

Regole associative

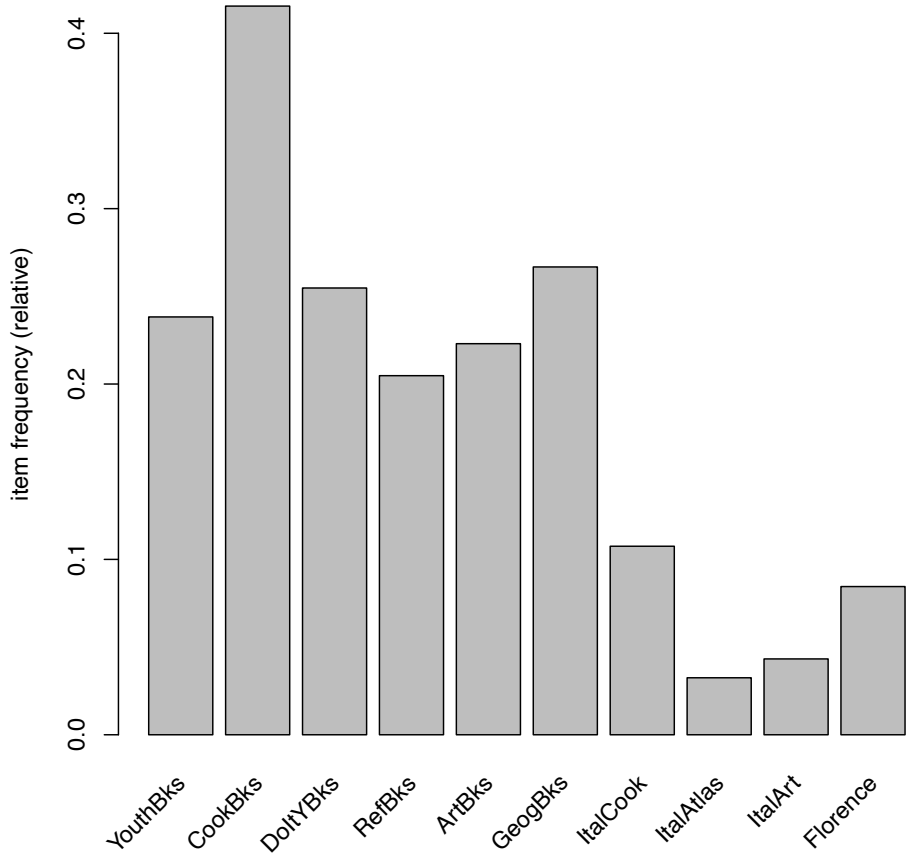
Algoritmo Apriori

1. Assegnare delle soglie sul supporto s e la fiducia p
2. Trovare gli itemset di dimensione 1 con supporto maggiore di s;
3. Per $k = 2; 3; \ldots$
   - generare gli itemset candidati con lunghezza $k+1$ dagli itemset di lunghezza k frequenti;
   - ottenere il supporto degli itemset di lunghezza $k+1$
   - eliminare i candidati non frequenti, tenendo solamente quelli con supporto maggiore di s;
4. Per ogni sottoinsieme non vuoto calcolare le regole e scegliere quelle con fiducia maggiore di p

Regole associative

Esempio

Il direttore di una libreria è interessato a promuovere alcune categorie di libri e per farlo vorrebbe studiare la possibilità di proporre dei 'pacchetti'.
Decide quindi di realizzare una MBA a questo scopo.



Regole associative

Transazioni= 4000, supporto minimo $s = 0.05$, fiducia minima $p = 0.5$

```
     lhs                      rhs         support confidence lift
[1]  {DoItYBks,GeogBks}    => {YouthBks} 0.05450 0.5396040  2.264864
[2]  {CookBks,GeogBks}     => {YouthBks} 0.08025 0.5136000  2.155719
[3]  {CookBks,RefBks}      => {DoItYBks} 0.07450 0.5330948  2.092619
[4]  {YouthBks,GeogBks}    => {DoItYBks} 0.05450 0.5215311  2.047227
[5]  {YouthBks,CookBks}    => {DoItYBks} 0.08375 0.5201863  2.041948
[6]  {YouthBks,RefBks}     => {CookBks}  0.06825 0.8400000  2.021661
[7]  {YouthBks,DoItYBks}   => {GeogBks}  0.05450 0.5278450  1.978801
[8]  {YouthBks,DoItYBks}   => {CookBks}  0.08375 0.8111380  1.952197
[9]  {DoItYBks,RefBks}     => {CookBks}  0.07450 0.8054054  1.938400
[10] {RefBks,GeogBks}      => {CookBks}  0.06450 0.7889908  1.898895
[11] {YouthBks,GeogBks}    => {CookBks}  0.08025 0.7679426  1.848237
[12] {DoItYBks,GeogBks}    => {CookBks}  0.07750 0.7673267  1.846755
[13] {YouthBks,ArtBks}     => {CookBks}  0.05150 0.7410072  1.783411
[14] {DoItYBks,ArtBks}     => {CookBks}  0.05300 0.7114094  1.712177
[15] {RefBks}              => {CookBks}  0.13975 0.6825397  1.642695
[16] {ArtBks,GeogBks}      => {CookBks}  0.05525 0.6800000  1.636582
[17] {YouthBks}            => {CookBks}  0.16100 0.6757608  1.626380
[18] {DoItYBks}            => {CookBks}  0.16875 0.6624141  1.594258
[19] {ItalCook}            => {CookBks}  0.06875 0.6395349  1.539193
[20] {GeogBks}             => {CookBks}  0.15625 0.5857545  1.409758
[21] {ArtBks}              => {CookBks}  0.11300 0.5067265  1.219558
```

Misura di lift decrescente: quali regole sembrano più interessanti?

Regole associative

- L'esito finale delle regole associative non è un modello globale, quanto piuttosto un insieme di risultati interessanti;
- Può essere prodotta una grande quantità di regole . . . necessaria operazione di selezione!
- Le regole associative non sono sottoposte ad alcuna procedura statistico-inferenziale;
- Le regole associative nascono in ambito commerciale ma possono essere usate in altri contesti: analisi di testi, analisi click-stream, analisi mediche.

Sistemi di raccomandazione
Filtri collaborativi

Filtri collaborativi

- Negli ultimi anni la personalizzazione dell'offerta è diventata un fattore cruciale per molte aziende, specie se orientate alla vendita online.
- Grande interesse è stato rivolto all'analisi e alla modellazione dei gusti individuali, con l'obiettivo di suggerire nuovi prodotti.
- Questo meccanismo è alla base dei cosiddetti *recommendation systems*, sistemi di raccomandazione, ormai adottati da diverse piattaforme tipo Amazon, Netflix, Spotify, Facebook.

Filtri collaborativi

- Il sistema è costruito in modo da fornire una raccomandazione personalizzata basata sulle informazioni dell'utente e su quelle di utenti simili.
- I filtri collaborativi sono una tecnica normalmente utilizzata da questi sistemi di raccomandazione
- Idea: identificare degli item rilevanti per uno specifico utente da un ampio insieme di item (filtro) a partire dalle preferenze di molti utenti (collaborativo).

Filtri collaborativi

- Il filtro collaborativo richiede che per ogni combinazione "item-utente" ci sia una qualche misura della preferenza dell'utente per quell'item.
- La preferenza può essere un rating numerico o un comportamento binario come un acquisto, un "like", un "click".

Filtri collaborativi

Per n utenti $(u_1, u_2, \ldots, u_n)$ e p item $(i_1, i_2, \ldots, i_p)$ possiamo pensare a una matrice $n \times p$ contenente le preferenze di ciascun utente.

	i_1	i_2	$\ldots$	i_p
u_1	$r_{1,1}$	$r_{1,2}$	$\ldots$	$r_{1,p}$
u_2	$r_{2,1}$	$r_{2,2}$	$\ldots$	$r_{2,p}$
$\vdots$				
u_n	$r_{n,1}$	$r_{n,2}$	$\ldots$	$r_{n,p}$

Dato che non ogni utente compra o valuta ogni item . . . la matrice sarà sparsa.

Filtri collaborativi

Sostanzialmente, un filtro collaborativo si basa su due principi

- Identificare utenti simili a quello di interesse (vicini)
- Considerare solo gli item che ovviamente l'utente non ha già comprato, e che risultano i preferiti dai vicini

Naturalmente questo implica di cercare una misura di prossimità tra utenti . . . e l'approccio dei k-nearest neighbors si rivela spesso il più adatto.

Filtri collaborativi

Una volta identificati i k utenti più vicini, è necessario scegliere che cosa raccomandare all'utente che ci interessa

Come?
Ovviamente vale un principio di prevalenza –il prodotto più acquistato, il più valutato, o quello valutato meglio–
Se l'approccio k-nearest neighbors risulta dispendioso sul piano computazionale, un'altra possibilità è di usare metodi di clustering ovvero identificare gruppi omogenei di utenti e misurare la distanza dell'utente da ognuno di questi gruppi.

Filtri collaborativi

Limite dei filtri collaborativi

Approccio che soffre del cosidetto *cold start* . . . cioè non può essere usato per creare raccomandazioni per nuovi utenti

Case Study

Introduzione al caso

- Una società di consulenza è interessata a conoscere meglio i clienti che visitano la sua pagina web
- Scopo: identificare azioni di marketing mirate per cercare di personalizzare la relazione con ciascun cliente
- Data mining applicato al web $\rightarrow$ web mining
- Web mining può essere diviso in tre categorie:
  - web usage mining: uso di 'click-stream data'
  - web content mining: estrazione di informazione utile da contenuti di pagine web (testi, immagini, video, audio ...)
  - web structure mining: estrazione di informazione utile da hypelinks e strutture dei documenti

Introduzione al caso

Per conoscere meglio i clienti, utilizziamo due strategie:

- realizziamo un'analisi di segmentazione, classificando i clienti in gruppi omogenei sulla base delle pagine visitate
- analizziamo le sequenze di pagine visitate, identificando i percorsi di navigazione più probabili

Introduzione al caso

Prima di procedere con le analisi specifiche, descriviamo il dataset

- 26157 'hits' anonimi sulla pagina web della società di consulenza
- per ogni 'hit' si conosce il numero di pagine visitate in un intervallo temporale fissato
- la pagina web contiene 231 pagine
- pagine visualizzate in totale: 47387
- quindi ogni pagina è stata visitata in media 205 volte
- ogni 'hit' di un cliente ha una media di 1.81 pagine

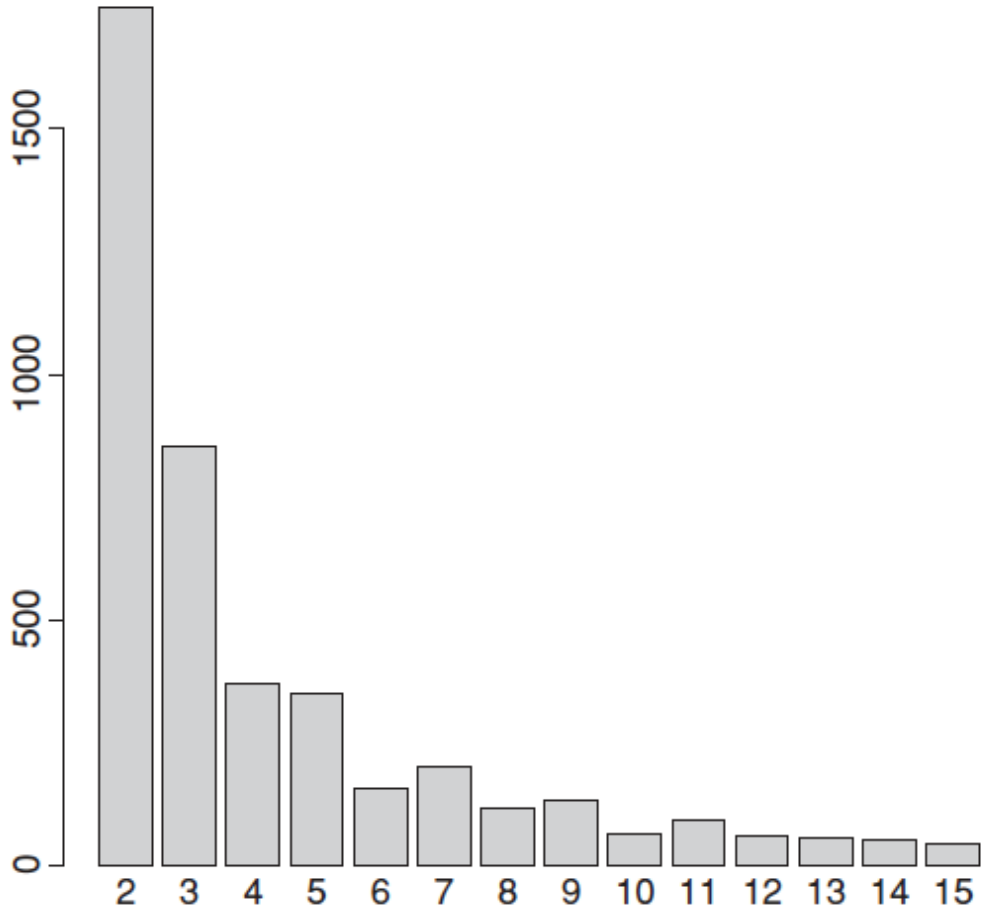
Introduzione al caso

Le pagine sono state raggruppate in 8 categorie in funzione del loro contenuto:

1. home
2. contacts
3. communications
4. events
5. company
6. white papers
7. business units
8. consulting

Introduzione al caso

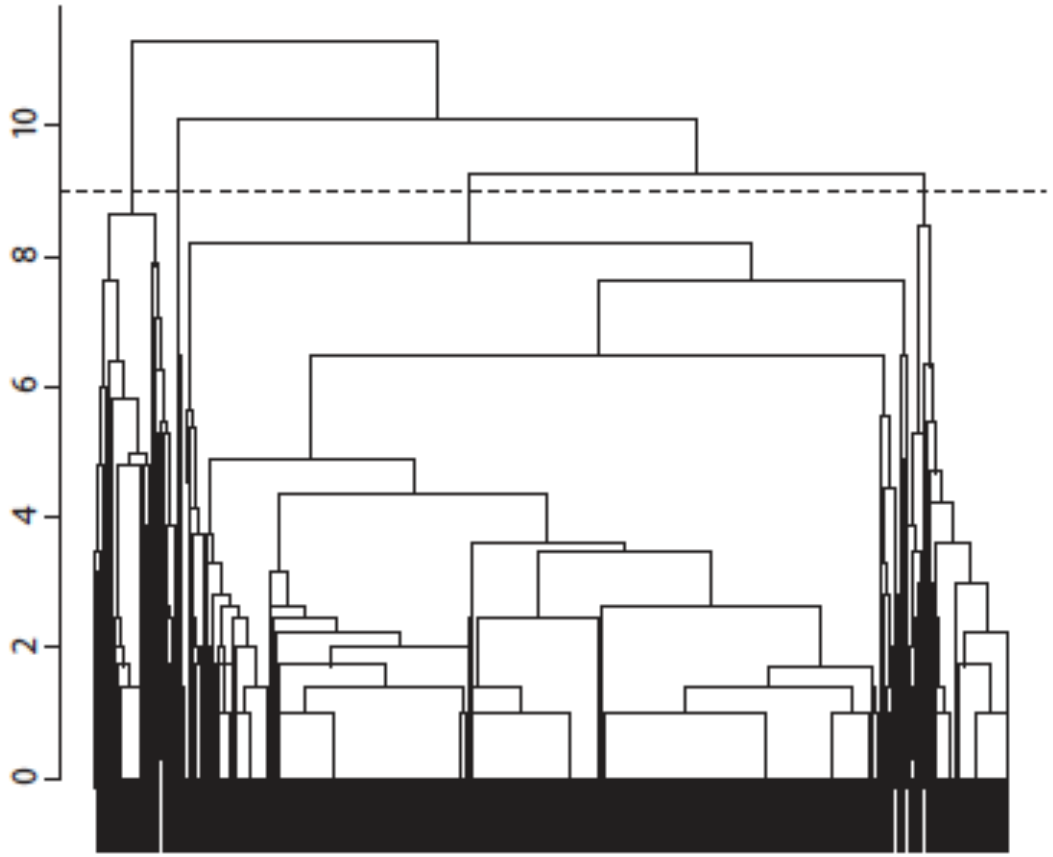
Siamo interessati ad analizzare i clienti che hanno visitato almeno due pagine: 4572



Web usage mining: analisi dei gruppi

Analisi dei gruppi gerarchica: legame completo

la linea tratteggiata indica il punto in cui si sceglie di 'tagliare' il dendrogramma $\rightarrow$ si identificano 4 gruppi



Web usage mining

Per ogni gruppo, si riportano medie e deviazioni standard di numero di visite divise per categoria

Area	Cluster A	Cluster B	Cluster C	Cluster D	Overall mean
Business area	9.43 (5.83)	0.58 (2.02)	1.76 (1.92)	0.30 (0.73)	2.27 (3.36)
Communications	0.67 (1.59)	0.02 (0.14)	0.27 (1.24)	0.16 (0.62)	0.29 (1.22)
Company	3.15 (6.40)	0.07 (0.43)	0.69 (2.03)	0.53 (1.90)	0.89 (2.79)
Consulting	2.23 (3.80)	0.41 (1.81)	0.13 (0.46)	3.73 (3.76)	0.72 (2.13)
Contacts	0.37 (0.78)	0.02 (0.14)	0.12 (0.39)	0.21 (0.59)	0.15 (0.47)
Events	0.43 (0.79)	0.11 (0.37)	0.09 (0.35)	0.13 (0.58)	0.12 (0.45)
Home	1.67 (2.14)	0.62 (1.35)	0.43 (0.88)	1.10 (2.41)	0.62 (1.36)
White papers	0.40 (1.83)	14.75 (14.18)	0.46 (0.94)	0.02 (0.14)	0.57 (2.38)
Number of visitors	409	53	3603	507	4572

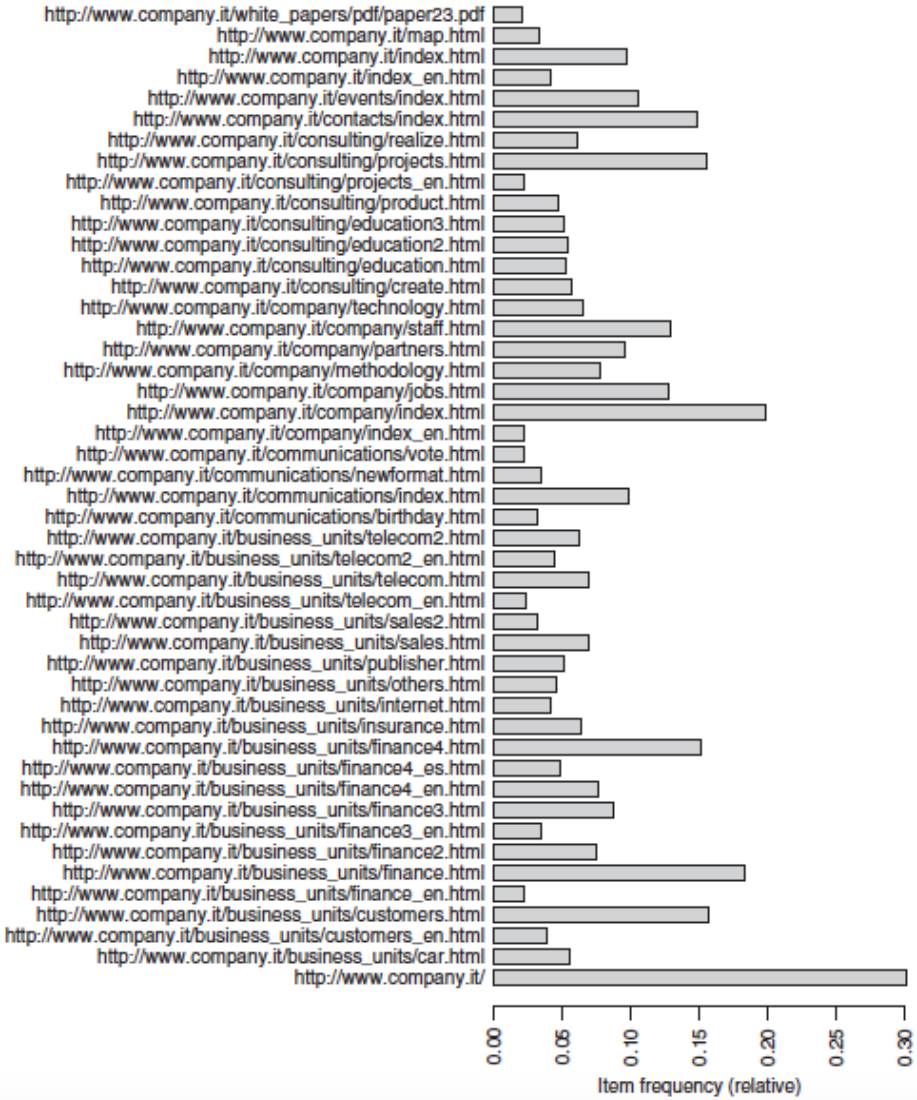
Che caratteristiche presentano questi 4 gruppi?

Web usage mining: regole associative

- Utile completamento dell'analisi: considerare ogni singola pagina e analizzare tutti i 'percorsi fatti'
- Le regole associative possono essere usate per analizzare il percorso di navigazione più probabile e vedere quali pagine saranno visitate . . . a partire da quelle già viste
- Viene usata una estensione dell'algoritmo Apriori: sequential pattern discovery using equivalence classes (SPADE)
- In questo caso l'ordine delle pagine visitate è fondamentale!

Web usage mining: regole associative

Mostriamo item con supporto almeno del 20%



Web usage mining: regole associative

- L'algoritmo ha trovato 186 sequenze con supporto di almeno 0.5%.
- Selezionando regole con almeno 60% di fiducia si ottengono 15 sequenze.
- Supporto: indica che la percentuale di utenti che hanno visitato le due pagine era in sequenza
- Fiducia: indica la probabilità che la seconda pagina della sequenza sia stata vista da utenti interessati alla prima pagina (o al gruppo di prime pagine)

Web usage mining: regole associative

Regole elencate con livello di fiducia decrescente

	Rule	Support	Confidence	Lift
1	<http://www.company.it/business_units/finance4.html, http://www.company.it/business_units/customers.html> => <http://www.company.it/business_units/finance4.html>	0.0056	0.7228	10.04
2	<http://www.company.it/company/index.html, http://www.company.it/company/staff.html, http://www.company.it/company/partners.html> => <http://www.company.it/company/jobs.html>	0.0069	0.7054	35.35
3	<http://www.company.it/company/staff.html, http://www.company.it/company/partners.html> => <http://www.company.it/company/jobs.html>	0.0073	0.6931	34.73
4	<http://www.company.it/, http://www.company.it/company/index.html, http://www.company.it/company/staff.html> => <http://www.company.it/company/partners.html>	0.0065	0.6552	49.53
5	<http://www.company.it/, http://www.company.it/company/index.html> => <http://www.company.it/company/staff.html>	0.0100	0.6525	36.86
6	<http://www.company.it/, http://www.company.it/company/staff.html> => <http://www.company.it/company/partners.html>	0.0068	0.6520	49.29
7	<http://www.company.it/company/index.html, http://www.company.it/company/technology.html> => <http://www.company.it/company/partners.html>	0.0054	0.6498	49.12
8	<http://www.company.it/company/index.html, http://www.company.it/company/partners.html> => <http://www.company.it/company/jobs.html>	0.0074	0.6467	32.40
9	<http://www.company.it/company/index.html, http://www.company.it/company/staff.html> => <http://www.company.it/company/partners.html>	0.0099	0.6434	48.64
10	<http://www.company.it/consulting/create.html> => <http://www.company.it/consulting/realize.html>	0.0052	0.6398	75.38
11	<http://www.company.it/company/index.html, http://www.company.it/company/technology.html> => <http://www.company.it/company/staff.html>	0.0053	0.6359	35.93
12	<http://www.company.it/company/partners.html> => <http://www.company.it/company/jobs.html>	0.0082	0.6185	30.99
13	<http://www.company.it/company/technology.html> => <http://www.company.it/company/partners.html>	0.0055	0.6128	46.32
14	<http://www.company.it/company/index.html, http://www.company.it/company/staff.html> => <http://www.company.it/company/jobs.html>	0.0093	0.6060	30.36
15	<http://www.company.it/company/technology.html> => <http://www.company.it/company/staff.html>	0.0054	0.6000	33.90