

Introduzione al Datamining

Francesco Passantino

francesco@iteam5.net

www.iteam5.net/francesco

Cos'è il datamining

- Processo di selezione, esplorazione e modellazione di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori, e allo scopo di ottenere un risultato di business al proprietario del database.

Categorie di informazioni/1

- **Associazioni:** si ha una associazione quando più eventi o fatti elementari vengono collegati in un unico evento globale.
 - Esempio: l'analisi delle vendite di una vettura di un certo colore può essere associata al fatto che il mese prima la campagna pubblicitaria mostrava l'auto di quel colore guidata da un personaggio famoso. La casa automobilistica può decidere di sfruttare questi dati per pianificare la produzione.
- **Sequenze:** si ha un sequenza se gli eventi elementari vengono collegati nel tempo.
 - Esempio: il 40% delle famiglia che acquistano una casa entro un mese acquistano anche un frigorifero nuovo. Oppure: se una famiglia ha acquistato a Marzo un fornello da campeggio, ad Aprile un sacco a pelo è probabile che a Maggio deciderà di comprare una tenda da campeggio.

Categorie di informazioni/2

- **Classificazione:** una classificazione si ha quando vengono identificati schemi o insiemi di caratteristiche che definiscono il gruppo (classe) a cui appartiene un dato elemento (record). La classificazione parte dall'utilizzo di insiemi esistenti e già classificati e porta alla definizione di alcune regolarità.
 - Esempio: rendita media di un cliente di una banca o la vendita settimanale di un particolare prodotto.
- **Clustering:** è simile alla classificazione ma consente di produrre nuovi gruppi non ancora definiti. Il clustering deriva dal partizionare la base di dati in modo tale che i membri di ogni gruppo siano simili rispetto a qualche criterio.
 - Esempio: segmentare le liste dei clienti in gruppi simili tra loro.

Data Warehouse

- Magazzino di dati a livello di impresa
- Insieme di strumenti per convertire un vasto insieme di dati in informazioni utilizzabili dall'utente
- Strumento di supporto decisionale
- Base informativa per costruire sistemi di analisi e previsione

Definire il modello di dati

- Identificare gli eventi da misurare:
 - Vendite
 - Chiamate al customer-service
 - Interventi di assistenza
 - Produzione
- Mantenere flessibilità per il futuro:
 - Nuovi prodotti
 - Nuovi centri assistenza
 - Nuove linee di produzione

Componenti di un modello DW

Tabella delle Dimensioni

Comuni		

Prodotti		

Tempo		

Dimensioni

Tabella dei Fatti

Comune	Prodotto	Tempo	Unità	Fatturato

Misure

Fatti

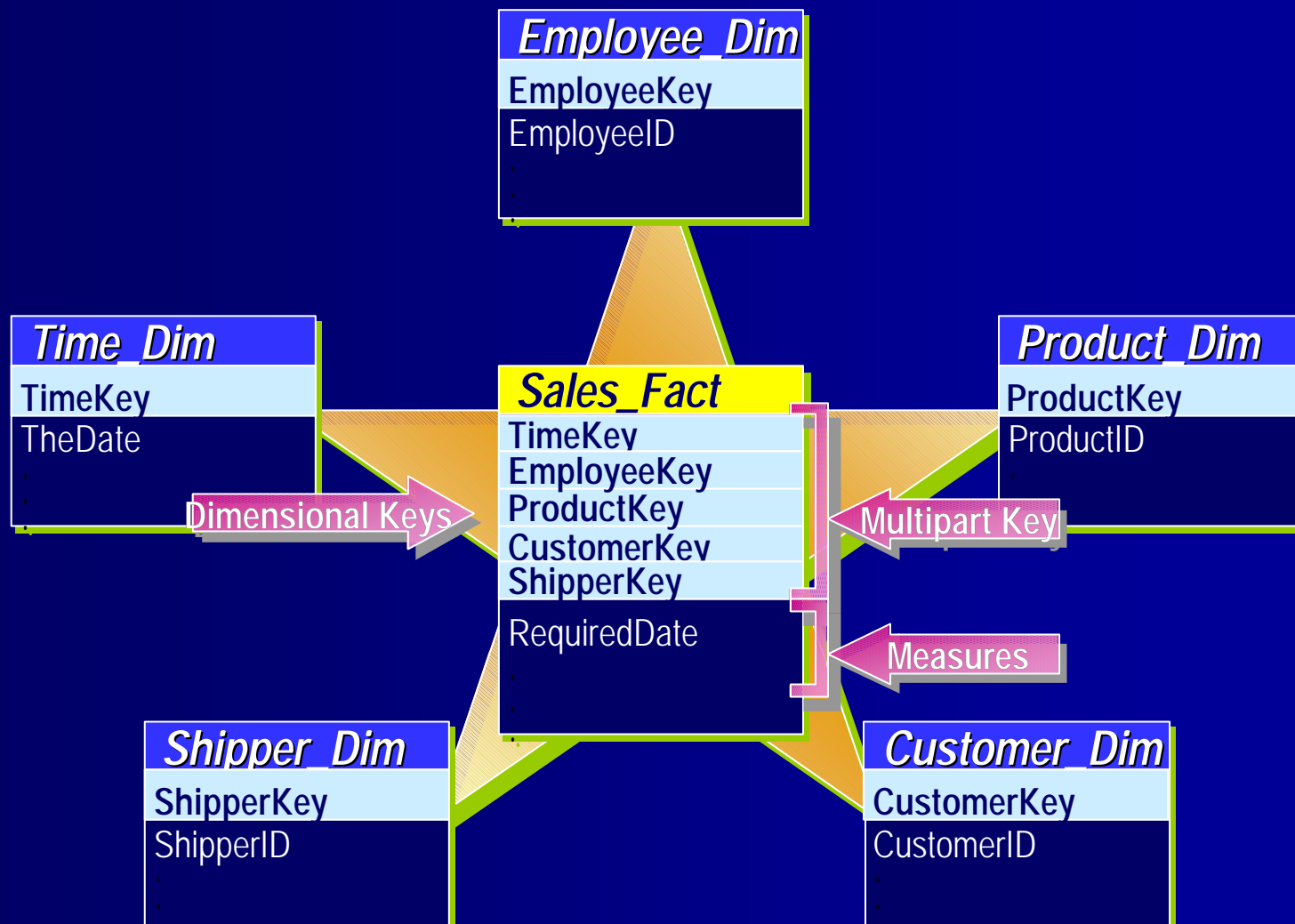
Componenti di un modello DW

- **Tabella dei fatti**
 - Contiene misure numeriche che descrivono un evento di business, come una vendita o una transazione bancaria
- **Fatto**
 - Una riga nella tabella dei fatti; contiene uno o più valori numerici che misurano un evento
- **Misura**
 - Una colonna numerica della tabella dei fatti
- **Dimensione**
 - Una entità di business che descrive il quando, chi, dove, come di un fatto (tempo, prodotto, cliente, ...)

Star Schema

- Lo Star Schema è la modellizzazione più semplice ed efficace dei componenti di un data warehouse
- Ogni tabella dei fatti è associata ad N tabelle dimensionali
- Relazioni gerarchiche all'interno di una dimensione (per es. anno/mese/giorno) vengono mantenute in una sola tabella dimensionale

Star Schema



Base dati multidimensionale



Fatto: Vendite (importo)

Dimensioni: Prodotto, Regione, Tempo

Percorsi **gerarchici** di sintesi

Prodotto

Industria

Categoria

Prodotto

Regione

Paese

Regione

Città

Ufficio

Tempo

Anno

Trimestre

Mese

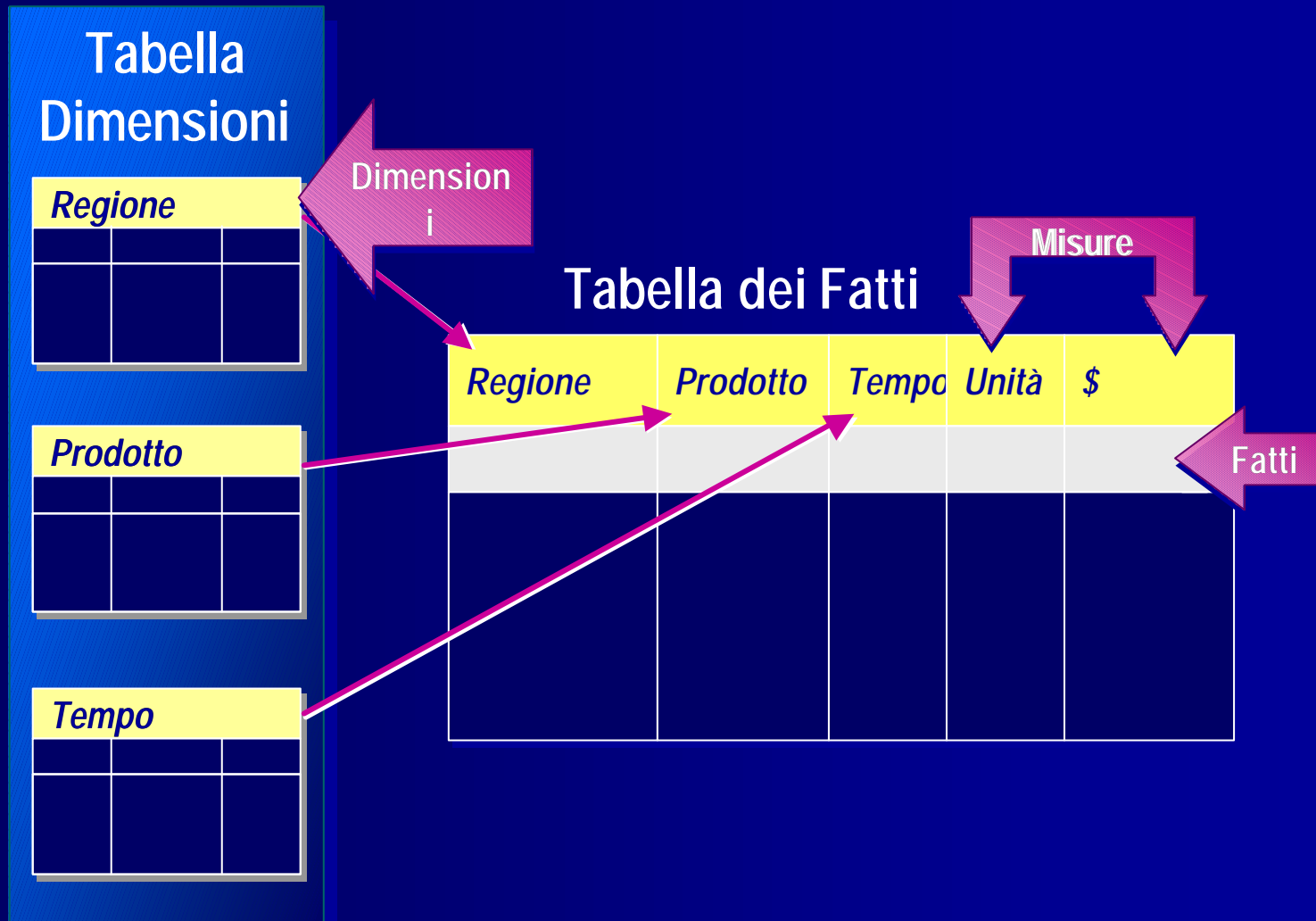
Settimana

Giorno

Fatti, Misure, Dimensioni, Gerarchie

- L'analisi multidimensionale dei dati analizza uno o più **fatti misurabili** al variare di una o più **dimensioni** organizzate in uno o più **livelli gerarchici**
- Nell'esempio precedente, si individua un **cubo delle vendite**, dove:
 - Le vendite sono il **fatto** oggetto di analisi
 - Importo e Volume sono **misure** del fatto
 - Regione è una **dimensione**
 - Paese, Regione, Città, Ufficio sono i **livelli gerarchici** della dimensione Regione

Fatti, Misure, Dimensioni



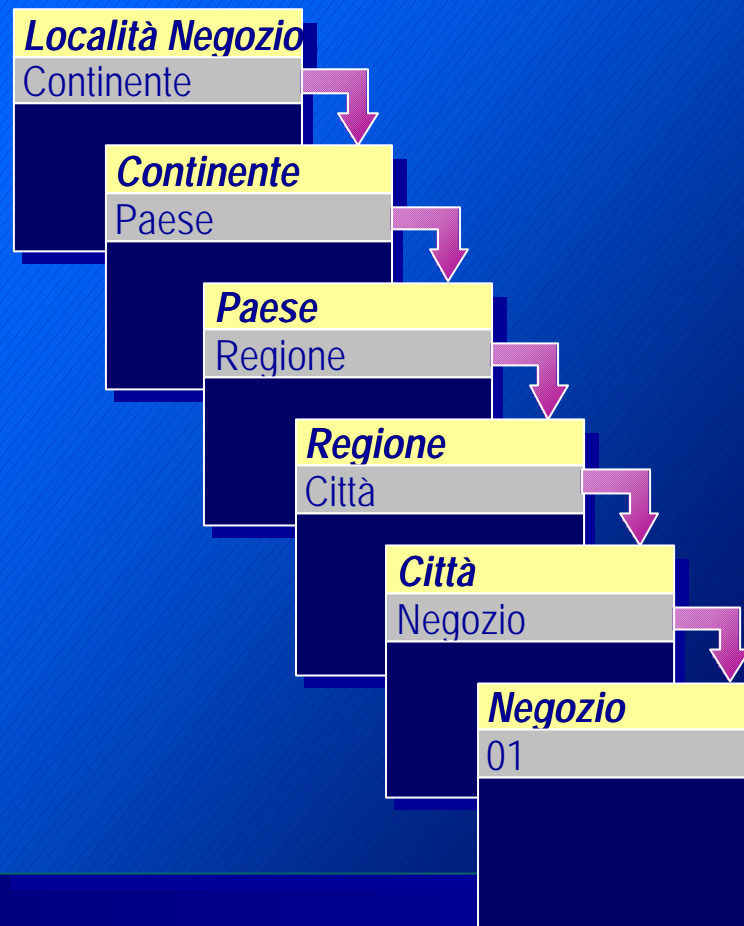
Gerarchia delle Dimensioni

Gerarchia Consolidata

Località Negozio

<i>Continente</i>	<i>Paese</i>	<i>Regione</i>	<i>Città</i>	<i>Negozio</i>

Gerarchia Separata



Cubi e Ipercubi

- Il cubo consente di rappresentare in modo intuitivo e maneggevole la dipendenza di un fatto da 3 dimensioni
- L'ipercubo è una generalizzazione del cubo su n dimensioni, con $1 \leq n \leq \infty$
- Per semplicità, si usa fare riferimento al "cubo" indipendentemente dal numero di dimensioni

Aggregazioni

Fatturato	Hardware	Software	Totale
Home	100	80	180
Business	70	30	100
Totale	170	110	280

- Celle di dettaglio: **4 (A)**
- Celle di sintesi: **5 (B)**
- Celle complessive: **9 (C)**
- Rapporto (C)/(A): **2.25**

Aggregazioni

Fatturato	Hardware				Software			Totale
	Desktop	Laptop	Server	Totale	Italiano	Inglese	Totale	
Home	70	30		100	80		80	180
Business	50	15	5	70	25	5	30	100
Totale	120	45	5	170	105	5	110	280

- Celle di dettaglio: **10 (di cui 2 vuote) (A)**
- Celle di sintesi: **14 (B)**
- Celle complessive: **24 (C)**
- Rapporto (C)/(A): **2.4**

Caso 1: CRM

- Customer Relationship Management
- Individuazione di gruppi omogenei di clienti in termini di comportamenti di acquisto di caratteristiche socio-demografiche
- L'individuazione delle diverse tipologie di clienti permette di effettuare campagne di marketing diretto personalizzate e di valutarne gli effetti, nonché di ottenere indicazioni su come modificare la propria offerta per fidelizzare il cliente.

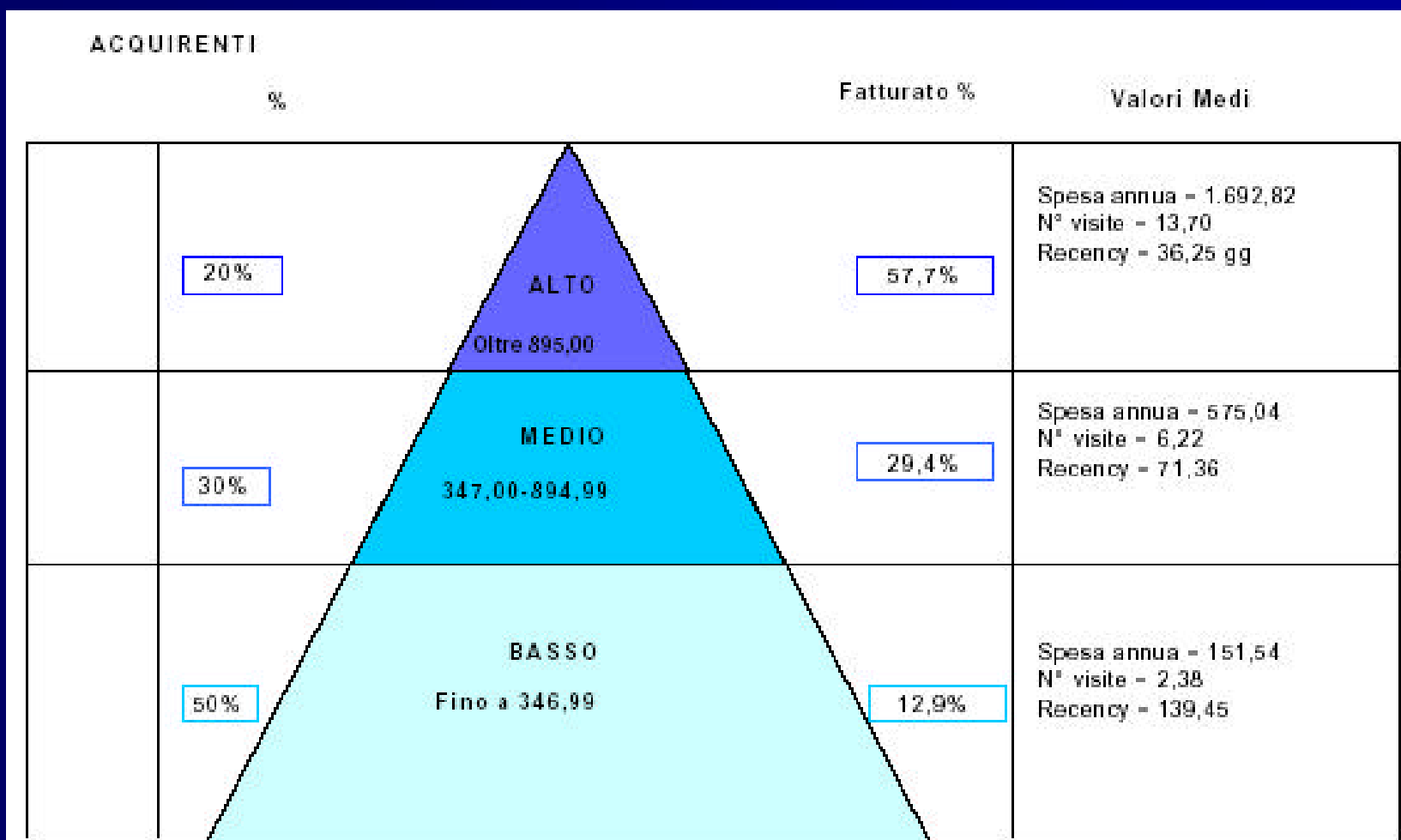
Classificazione

- Il caso delle fidelity card di COIN
- Scopo dell'analisi è stato quello di qualificare il consumatore secondo i parametri di :
 - Recency
(“recentezza” dell’ultimo atto d’acquisto)
 - Frequency
(frequenza degli atti d’acquisto)
 - Monetary
(valore monetario)

Risultati dell'analisi

- Le matrici di migrazione indicano gli spostamenti dei clienti tra le varie classi di valore nell' arco di 2 periodi di osservazione
- Nel biennio 2000-2001 hanno visitato almeno una volta un punto vendita Coin 185.000 clienti "coincardisti". Di questi:
 - Il 40% è rimasto nella stessa classe di valore
 - Il 35% è migrato verso una fascia di valore più alta
 - Il 25% è migrato verso una fascia di valore più bassa o è diventato inattivo
- I clienti sono stati classificati anche per numero di categorie acquistate (1,2,multi)

Analisi di concentrazione



Caso 2: Market Basket Analysis

- Individuazione di associazioni fra dati di vendita, al fine di conoscere quali prodotti sono acquistati congiuntamente ad altri
- Questo tipo d'informazione consente di migliorare l'offerta dei prodotti (disposizione sugli scaffali) e di incrementare le vendite di alcuni prodotti tramite prodotti ad essi associati.
- Scoperta di regole di associazione

Esempio basket analysis

Scontrino	Item
1	A B C
2	A C
3	A D
4	B E F

- La regola $A \Rightarrow C$ ha
 - un supporto pari al 50%, perché $\{A,C\}$ compare in 2 transazioni su 4
 - una confidenza pari al 66.6%, perché su 3 transazioni in cui compare A in 2 compare anche C
- La regola $C \Rightarrow A$ ha
 - ancora un supporto pari al 50%
 - una confidenza pari al 100%

Restando al supermercato

- **Trova tutte le regole che hanno “noccioline” nel conseguente:**
le regole trovate possono essere usate per capire quali prodotti il supermercato deve comprare per favorire la vendita di noccioline
- **Trova tutte le regole che hanno “noccioline” nell’antecedente:**
l’analisi può prevedere quali prodotti possono subire una riduzione delle vendite se il supermercato decide di non vendere più noccioline
- **Trova tutte le regole che hanno “noccioline” nell’antecedente e birra” nel conseguente:**
può servire per capire quali altri prodotti oltre alle noccioline servono per favorire la vendita di birra
- **Trova tutte le regole che riguardano item delle corsie 10 e 11:**
regole di questo tipo possono essere usate ai fini di una migliore organizzazione dei prodotti nelle corsie
- **Trova le regole più “interessanti” per le “noccioline”:**
ad esempio le regole con maggiore confidenza e/o supporto

Caso 3: Web Profiling

- Analisi dei dati di visita ad un sito web, con l'obiettivo di estrarre informazioni sui comportamenti di visita, classificando i visitatori in base ai profili comportamentali
- La segmentazione comportamentale potrà essere impiegata nelle successive decisioni di marketing relazionale
- Individuazione di cluster

Clusterizzazione

Individuazione dei cluster con caratteristiche omogenee

- **Gold:** clienti con alto fatturato, poco speculativo, distribuito su un ampio assortimento, con alta frequenza
- **Silver:** frequenza e fatturato più bassi rispetto ai Gold
- **Cacciatori di promozioni:** fatturato e visite medi, alta incidenza degli acquisti promozionali
- **Occasionalisti:** frequenza molto bassa, fatturato basso, poco sensibili alle promozioni

Analisi / controlli all'interno dei singoli cluster, con particolare attenzione e cura ai clienti più importanti per l'azienda (Gold e Silver)

esempi: Alert - Migration fra cluster - basket analysis...

Customizzazione



Permettere al cliente di configurare i servizi offerti secondo le esigenze.
Es. My Yahoo

Personalizzazione



The screenshot shows the Amazon.com website interface. The main content area displays the book 'Data Mining: Building Competitive Advantage' by Robert Goeth. The page includes a search bar, navigation tabs, and a 'Customers who bought this book also bought' section. The 'Customers who bought this book also bought' section lists several related books:

- [Building Data Mining Applications for CRM](#); Alex Berson, et al
- [Mastering Data Mining: The Art and Science of Customer Relationship Management](#); Michael J. A. Berry, Gordon
- [Data Mining Your Website](#); Jesus Mena
- [The Data Webhouse Toolkit : Building the W](#)
Ralph Kimball, Richard Merz

Offrire al cliente servizi personalizzati creati analizzando il suo comportamento online.
Es. Amazon

Bibliografia

Cerca su Google:

- PDF "Metodi e Modelli di Data Mining"
- PDF "Il Caso Coin"
- PDF "Paolo Ciaccia" "Data Mining"



Paolo Giudici
Data mining
Metodi statistici per le applicazioni aziendali
Mcgraw-hill
2001

www.ateneonline.it/giudici

www.hiknow.com