Un'applicazione **di Text Mining**

Knowledge Discovery in Text (KDT)

Problema

- Un’azienda erogatrice di servizi intende analizzare il testo delle telefonate in arrivo al proprio numero verde al fine di migliorare il servizio e agevolare l’utente nell’utilizzo dello stesso.

Soluzione

Applicazione del text mining nelle richieste e reclami testuali rivolti al **call center.**

Estrarre parole chiave dalle telefonate pervenute al call center.

Parole chiave

- Rappresentative dei testi oggetti di studio
- Dotate di un valore semantico che ha il potere di richiamare i principali temi contenuti nelle telefonate

L’analisi

Segue le quattro fasi del KDT:

- **Comprensione del problema**
- **Raffinamento del testo**
- **Text Mining**
- **Valutazione dei risultati**

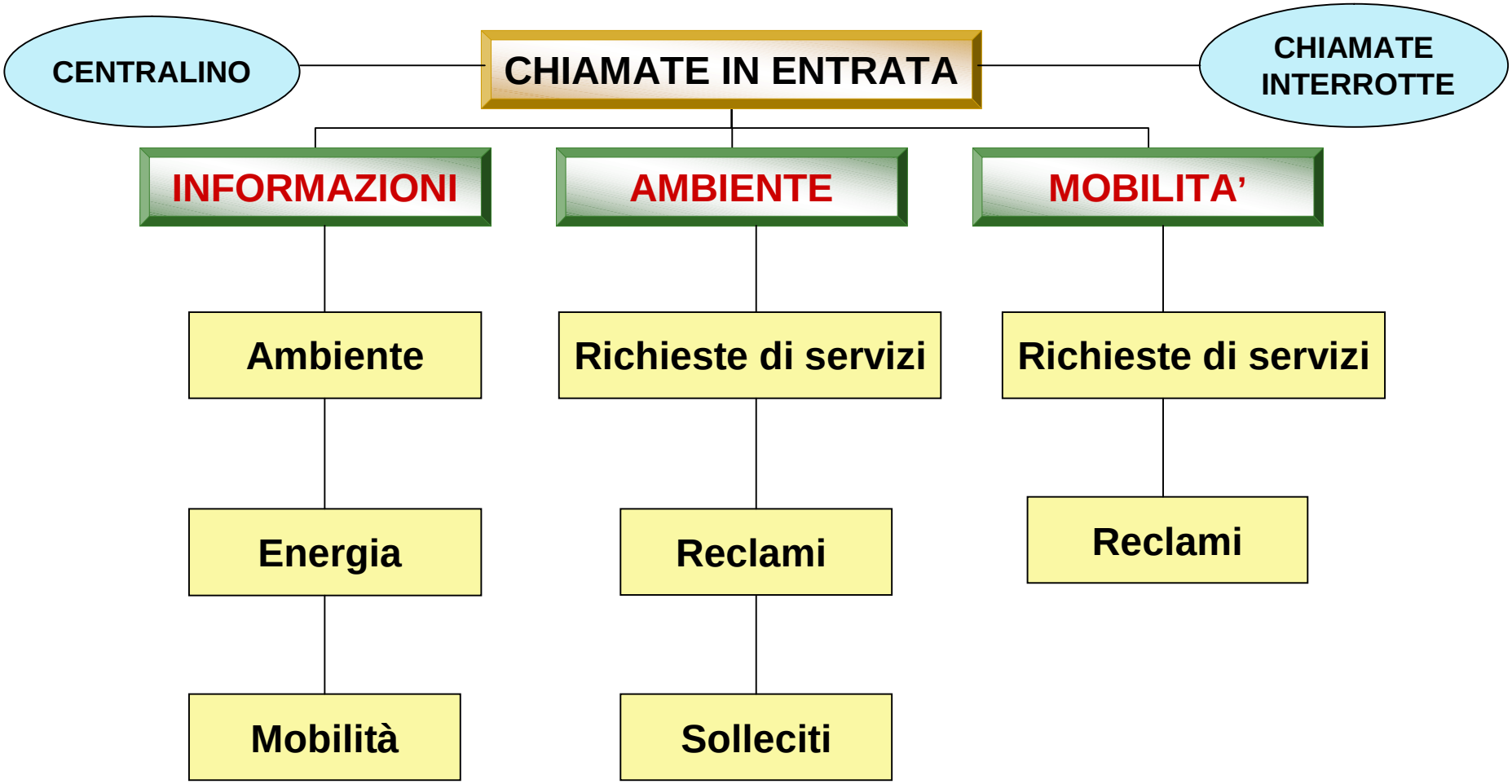
Prima Fase: **Comprensione del Problema**

1. Comprensione del dominio di applicazione
2. Comprensione degli obiettivi dell'utilizzatore finale
3. Creazione del campione

Prima Fase: **Comprensione del Problema**

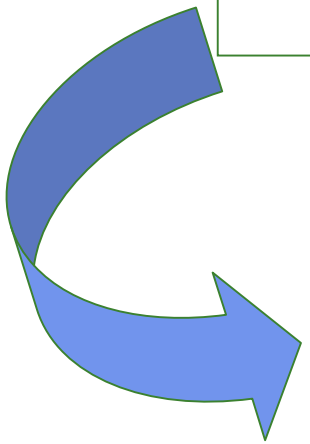
- **Memorizzazione delle telefonate** (schede)
- **Database** (record)
- Al fine di identificare gruppi di schede omogenee per contenuto **si è eseguita una classificazione preliminare** delle stesse

Classificazione delle telefonate



2. Comprensione degli obiettivi dell'utente finale

L'obiettivo dell'Azienda è **creare una struttura standard per la memorizzazione delle chiamate**



Utilizzare menù a scorrimento contenenti le opzioni che si presentano con maggior frequenza in una telefonata

3. Creazione del campione

- Analisi della “parte in chiaro” delle telefonate

- Tipologia di schede analizzate:
  - **Ø Mobilità**
  - **Ø Ambiente**

- Periodo oggetto di studio

Seconda Fase: **Raffinamento del testo (Text Refining)**

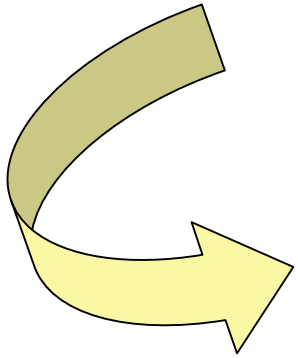
1. Eventuali trasformazioni sul testo

2. La rimozione di parole vuote (stopword) e parole con bassa capacità discriminante

Ottenere una forma intermedia "strutturata" per la successiva fase di mining senza perdere informazioni utili

1.Eventuali trasformazioni sul testo

Sostituzione delle Abbreviazioni



Come ad esempio:

bus **à** autobus

xchè **à** perché

cass.tto **à** cassonetto

2.Eliminazione delle “*Stopword*”

Lista di Stopword eliminata

(Numeri, Articoli, Preposizioni Semplici e Articolate, Pronomi)

Terza Fase: Text Mining

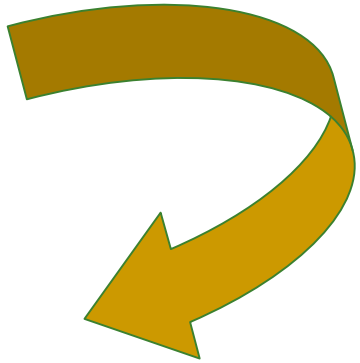
Scelta del metodo di elaborazione dati

Scoperta di informazioni significative attraverso la scoperta di pattern interessanti

Studio delle associazioni testuali

La scoperta di associazioni testuali tramite *Association Node*

Utilizzato per identificare regole tra parole



Regole che si presentano contemporaneamente, indipendentemente dall'ordine nel corso della telefonata

La scoperta di associazioni testuali tramite *Association Node*

Le regole di associazione assumono la seguente forma:

A à B

Dove A e B sono una o più parole distinte

COUNT	SUPPORT (%)	CONFIDENCE (%)	LIFT	RULE
31	12,25	40,79	1,15	autista => segnala
27	10,67	24,55	1,73	non =>autista & autobus
29	11,46	60,42	2,01	fermata & non => autista

Indici di Associazione

Frequenza della Regola (SUPPORT)

E' il rapporto tra il numero di documenti che contengono contemporaneamente le parole **A** e **B,** costituenti la regola $\mathbf{n_{AB}}$ e il numero di documenti totali

$$Support = \frac{n_{AB}}{n} \times 100$$

Indici di Associazione

Indice di compresenza (CONFIDENCE)

Misura la forza della regola di associazione fornendo una **probabilità condizionata**, rilevando la percentuale dei documenti in cui si presenta il termine **B** data la presenza del termine **A**

$$Confidence = \frac{n_{AB}}{n_A} \times 100$$

Indici di Associazione

Indice di compresenza relativa (LIFT)

Misura la dipendenza tra gli elementi della regola:

un Lift pari a 1 indica indipendenza tra gli elementi della regola

$$\text{Lift} = \frac{\frac{n_{AB}}{n_A}}{n^*_{AB}}$$

Dove:

$$n^*_{AB} = n^*_A \times n^*_B$$

Estrazione delle associazioni

Definita una **soglia minima** di documenti in cui le parole devono comparire insieme per generare la regola

Selezione delle associazioni al fine di ottenere **sequenze chiave** **che descrivano il contenuto delle telefonate**

Le associazioni significative sono state rappresentate mediante **struttura grafica**

Quarta Fase: **Valutazione dei risultati (Evaluation)**

1. Analisi mediante pareri “esperti” di dipendenti e consulenti dell’azienda

2. Lemmatizzazione e sostituzione delle parole con significato affine

3. Reiterazione dell’intero processo

Conclusioni

- E’ possibile identificare una **struttura nelle telefonate sulla base dell’analisi testuale**
- La **struttura** ottenuta si è dimostrata **direttamente operativa**, quindi è utile coinvolgere gli esperti dell’azienda
- E’ riproponibile, visto **l’entusiasmo con cui è stata accolta la soluzione ottenuta**
- **Manca l’analisi longitudinale** (stesse persone che chiamano ripetutamente) e la valutazione statistica dei parametri è essenziale