



# METODI SEMANTICI PER TEXT MINING

*Alessio Leoncini - Fabio Sangiacomo*

*Relatore: Chiar.<sup>mo</sup> Prof. Ing. Rodolfo Zunino*  
*Correlatore: Dott. Ing. Simone Tacconi*

25 settembre 2009



# Sommario

- Obiettivi
- Il text mining ed il suo ruolo nelle investigazioni digitali
- Clustering di documenti
- WordNet e le relazioni semantiche
- Risultati sperimentali
- Conclusioni e sviluppi futuri

# Obiettivi

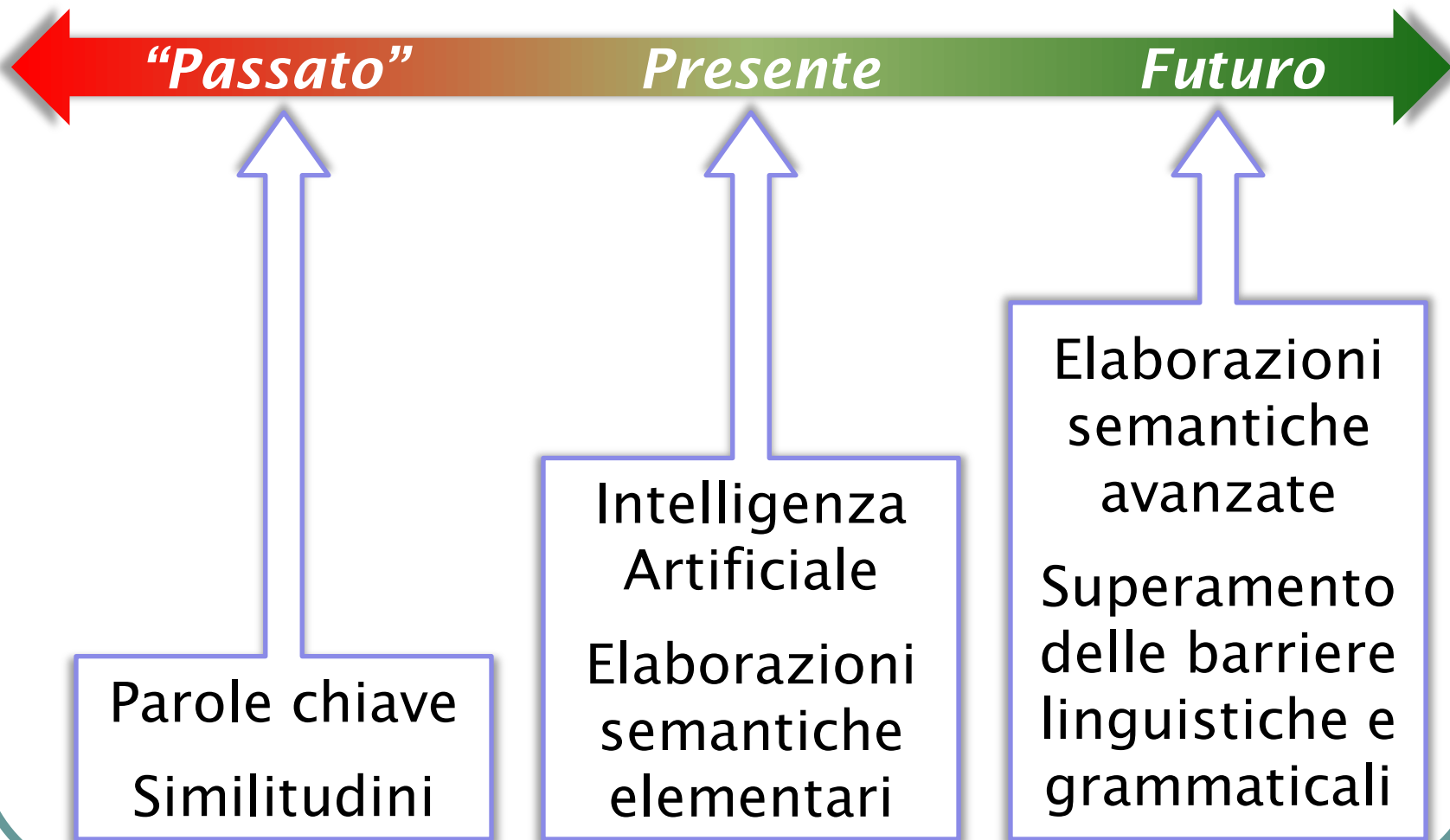
- Una panoramica sul mondo del text mining e sulla sua rilevanza nella computer forensics
- Realizzare un software capace di estrarre informazioni semantiche da un documento

# Il Text Mining - Applicazioni

- Quantità sempre crescenti di informazioni sono memorizzate in testi non strutturati
- Il text mining consente l'analisi ed l'organizzazione automatica dei dati
- Applicazioni di maggior interesse:
  - business intelligence
  - gestione di grandi database
  - ricerche in ambito biomedico
  - investigazioni
  - ricerche di marketing



# Il Text Mining - Evoluzione



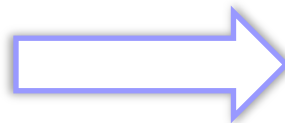
# Computer forensics (1)

- Scienza che studia ogni forma di trattamento del dato informatico per essere valutato in un processo giuridico
- Requisiti:
  - raccolta evidenze senza alterare il sistema sorgente
  - completa congruenza delle copie con l'originale
  - l'analisi non deve alterare le copie
- Importanza della catena di custodia e di un metodo di lavoro collaudato per i forenser



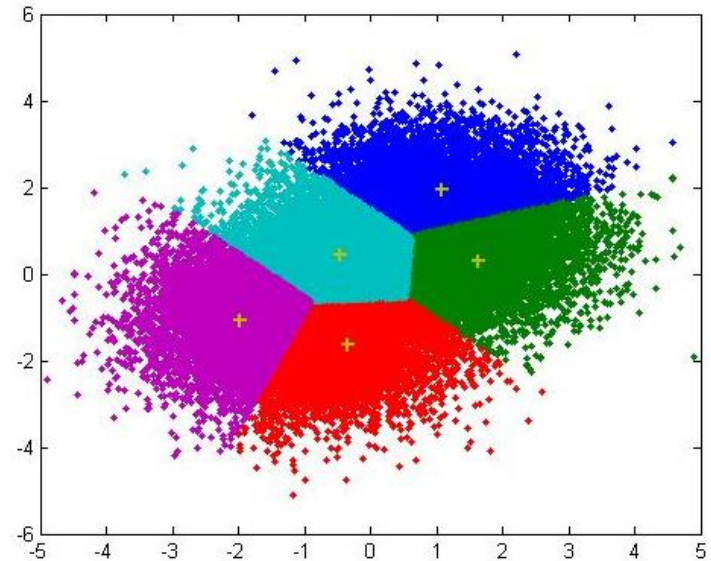
# Computer forensics (2)

- Un text miner permette di estrarre velocemente evidenze da grandi quantità di documenti testuali, come messaggi di posta elettronica



# Clustering di documenti

- K-means: algoritmo diffuso per il grouping non supervisionato
- Step fondamentali:
  - rappresentazione vettoriale dei documenti
  - applicazione di una metrica
  - passaggio allo spazio del kernel
  - individuazione dei cluster

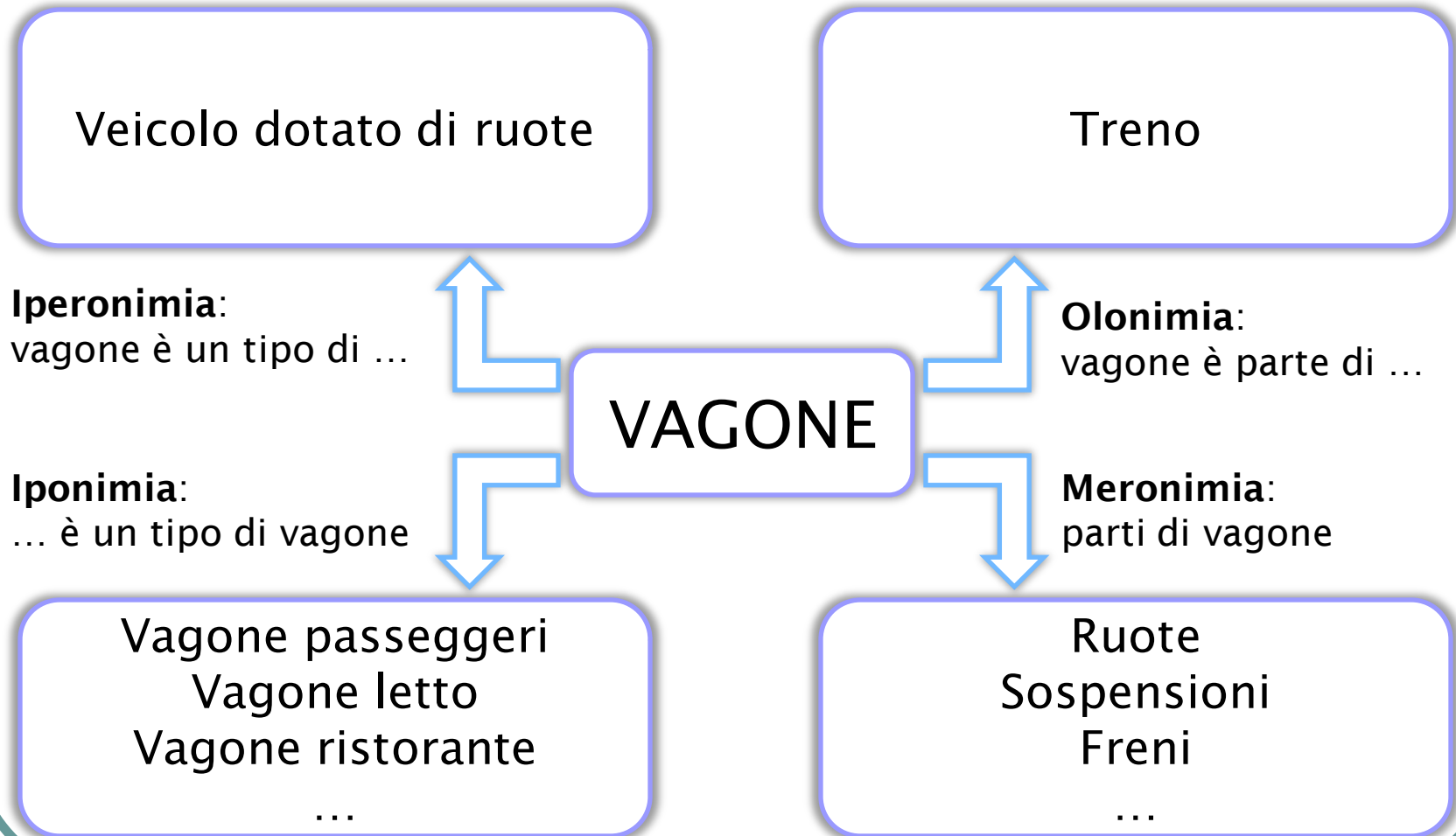


# WordNet English 2.1

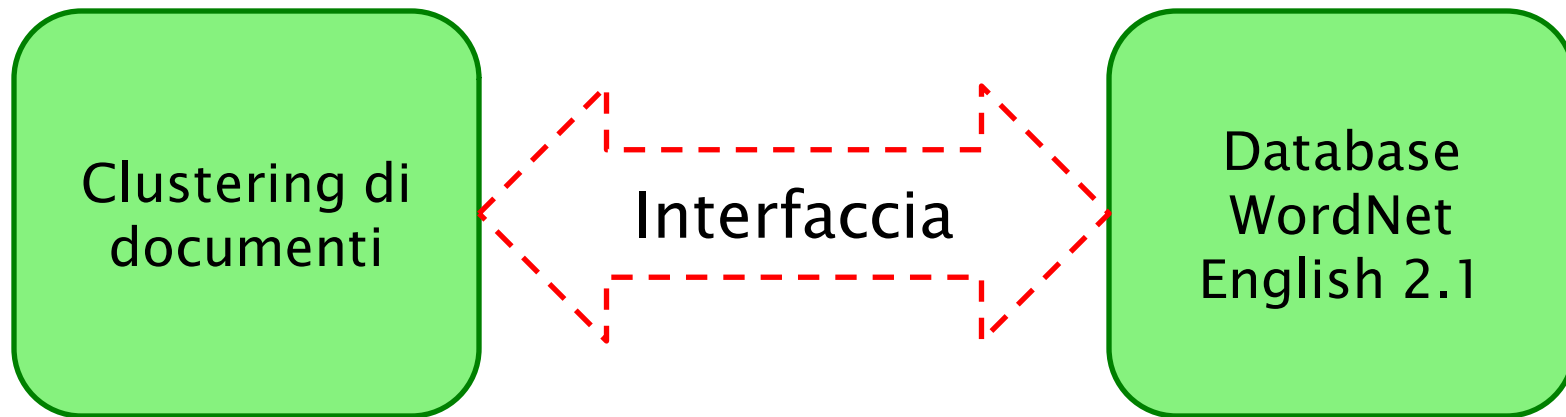
- Database semantico-lessicale
- Struttura basata sul **synset**
- Il text miner sfrutta le relazioni:
  - lessicali, che legano i lemmi in un singolo synset
  - semantiche, che legano più synset
- Attenzione all'importanza dello “stemmer”



# Relazioni semantiche



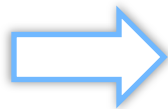
# Interfaccia per WordNet



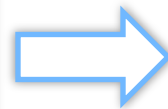
L'interfaccia conferisce funzionalità semantiche al clusterizzatore

- La creazione dei cluster non è più solo basata sulle parole, ma sui significati
- I significati sono estratti grazie alle relazioni semantiche

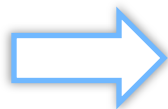
# Analisi semantica



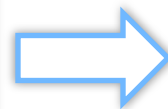
pesca  
acqua  
cattura  
...



**Sport**  
**Attività**  
**commerciale**



pesca  
polpa  
pianta  
...



**Frutta**  
**Alimentazione**

# Risultati sperimentali (1)

- Test condotti sul database di e-mail “Enron”: dati diffusi online durante le indagini sul fallimento dell’azienda
- 61 9449 messaggi di posta elettronica, di proprietà di 1 58 impiegati della Enron, in gran parte dirigenti
- Consentono di simulare il caso di un contesto investigativo

# Risultati sperimentali (2)

- Scelti a caso alcuni dipendenti, si vuole capire qualcosa di essi grazie alle loro e-mail

Dipendente: Smith M.	
Cluster ID	Parole più rilevanti
1	employe, businnes, hotel, houston, company
2	pipeline, social, database, report, link, data
3	ect, enronxg
4	coal, oil, gas, nuke, west, test, happy, business
5	yahoo, compubank, ngcorp, dynegi, night, plan
6	shank, trade
7	travel, hotel, continent, airport, flight, sheraton
8	questar, paso, pice, gas
9	schedule, london, server, sun, contact, report
10	trip, weekend, plan, ski

# Risultati sperimentali (3)

- ferc: Federal Energy Regulatory Commission

Dipendente: Steffes J.	
Cluster ID	Parole più rilevanti
1	ferc, rto, epsa, nerc
2	market, ferc, edison, contract, credit, order, rto
3	ferc, report, approve, task, image, attach
4	market, ee, meet, november, october
5	california, protect, attach, testimoni, washington
6	stock, billion, financial, market, trade, investor
7	market, credit, ee, energy, util
8	attach, gov, energy, sce
9	affair, meet, report, market
10	gov, meet, november, imbal, pge, usbr

# Conclusioni

- Il text miner analizza un elevato numero di documenti, senza bisogno di criteri di ricerca preimpostati
- L'organizzazione in cluster permette una ricerca più approfondita, limitata ai gruppi di maggior interesse

# Sviluppi futuri (1)

- L'uomo comunica in linguaggio naturale: sarà sempre più importante avere strumenti capaci di analizzare parole, interpretare frasi, comprendere concetti
- Prevalenza dell'analisi semantica sull'analisi lessicale
- Semantic web: motori di ricerca semantici (esempio: “aspetti nutrizionali della pesca”)

# Sviluppi futuri (2)

- Text miner sempre più indipendenti dalle lingue e dalle regole grammaticali dei documenti
- Futuri progressi delle prestazioni di hardware e software porteranno i text miner a:
  - elaborazioni più avanzate a parità di tempo
  - più ambiti di utilizzo