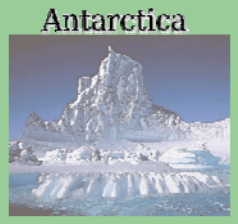
Analisi comparativa dei metodi di stima della similarità semantica di testi in linguaggio naturale non strutturato.

Prova Finale di Laurea Triennale in Ingegneria Informatica

Relatore	**Prof. Giovanni Cantone**
Correlatore	**Dott. Davide Falessi**
Candidato	**Carlo Ieva**

Anno accademico 2007/2008

Sviluppo di uno strumento.
Studio pilota su
requisiti software.
Analisi dei risultati.



UNIVERSITA' degli STUDI di ROMA
TOR VERGATA

FACOLTÀ DI INGEGNERIA

Contenuti

- Introduzione.
- Approcci mutuati dalla Intelligenza artificiale.
- Lo strumento sviluppato.
- Validazione: uno studio pilota.
- Conclusioni.

Introduzione

- ***Requisiti***: descrizione dei vincoli e servizi richiesti al sistema (es. funzionalità, caratteristiche, aspetti non funzionali.)
- ***Software Product Line Engineering***: tecnica ingegneristica per sviluppare un portafoglio di sistemi software traendo vantaggio dalle loro comunanze.
  - Industria automobilistica: gestione delle migliaia di comunalità e varianti di uno stesso modello.

Contesto

- Probabilità elevata numero di requisiti duplicati:
  - Alto numero di progetti,
  - Alto numero di requisiti per progetto,
  - Alto numero di persone a stilare i requisiti.

- Requisiti duplicati:
  - Stesso progetto: errore.
  - Progetti diversi: opportunità di capitalizzazione tramite riutilizzo (“riuso”).

- In entrambi i casi, è importante avere a disposizione degli strumenti efficaci ed efficienti che permettano l'individuazione di coppie di requisiti equivalenti.

Domanda di ricerca

- È possibile e utile, per un ingegnere del software, impiegare tecniche di IA, e quali di queste mutuare, allo scopo di individuare requisiti semanticamente equivalenti?

Obiettivo

- (*GQM*) Realizzare e validare uno strumento allo *scopo* di *caratterizzare,* in maniera oggettiva e quantitativa, le tecniche di IA disponibili, *in termini* di supporto all'identificazione di requisiti semanticamente equivalenti in un *contesto* industriale dal *punto di vista* del ricercatore.

- Fornire informazioni utili per scegliere la tecnica IA più adatta al contesto di riferimento.

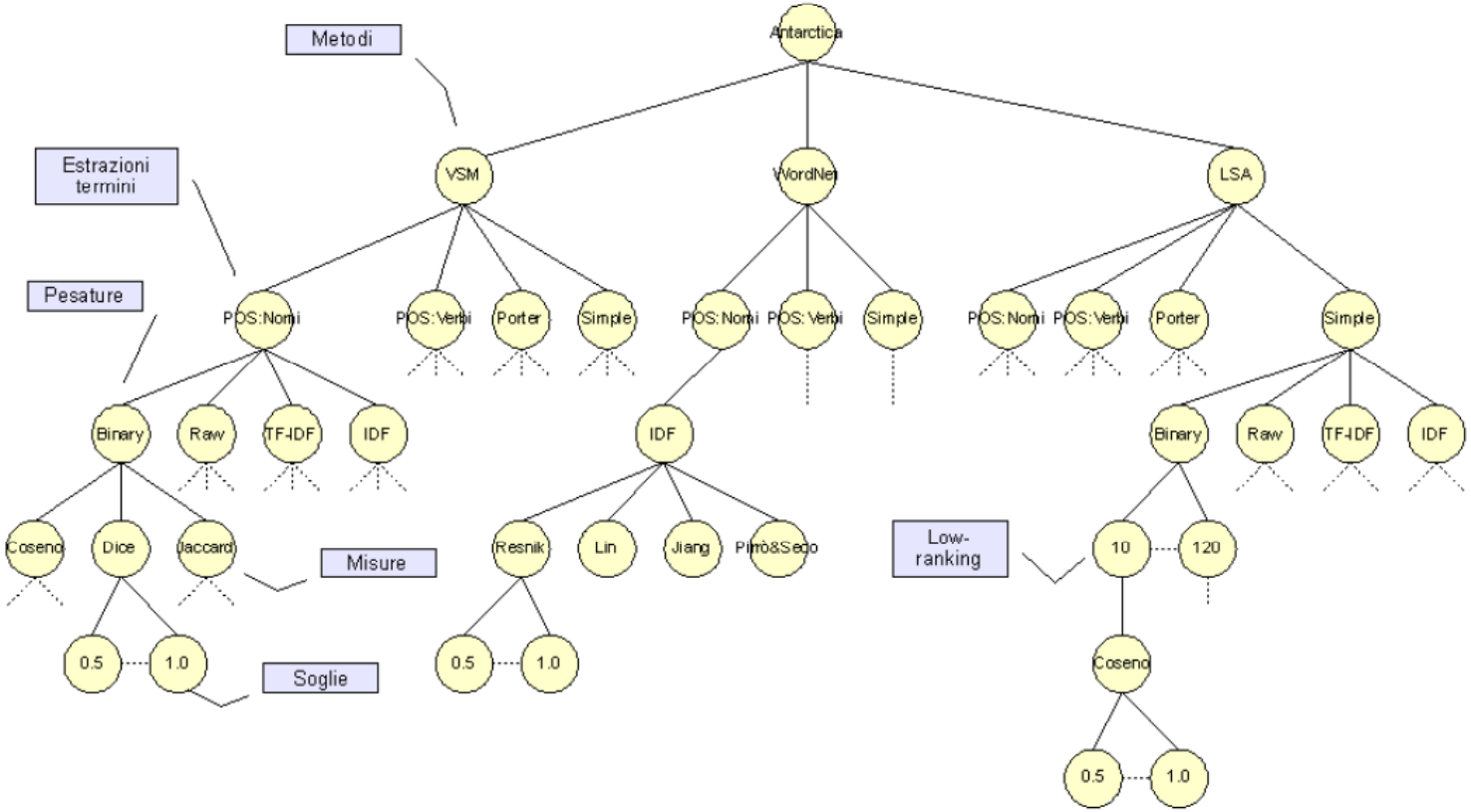
Modalità di valutazione dei testi

- **Modelli algebrici per la misura di similarità semantica**
- **Estrazione dei termini**
- **Schemi di pesatura**
- **Misura di similarità**

Approcci alla valutazione dei testi

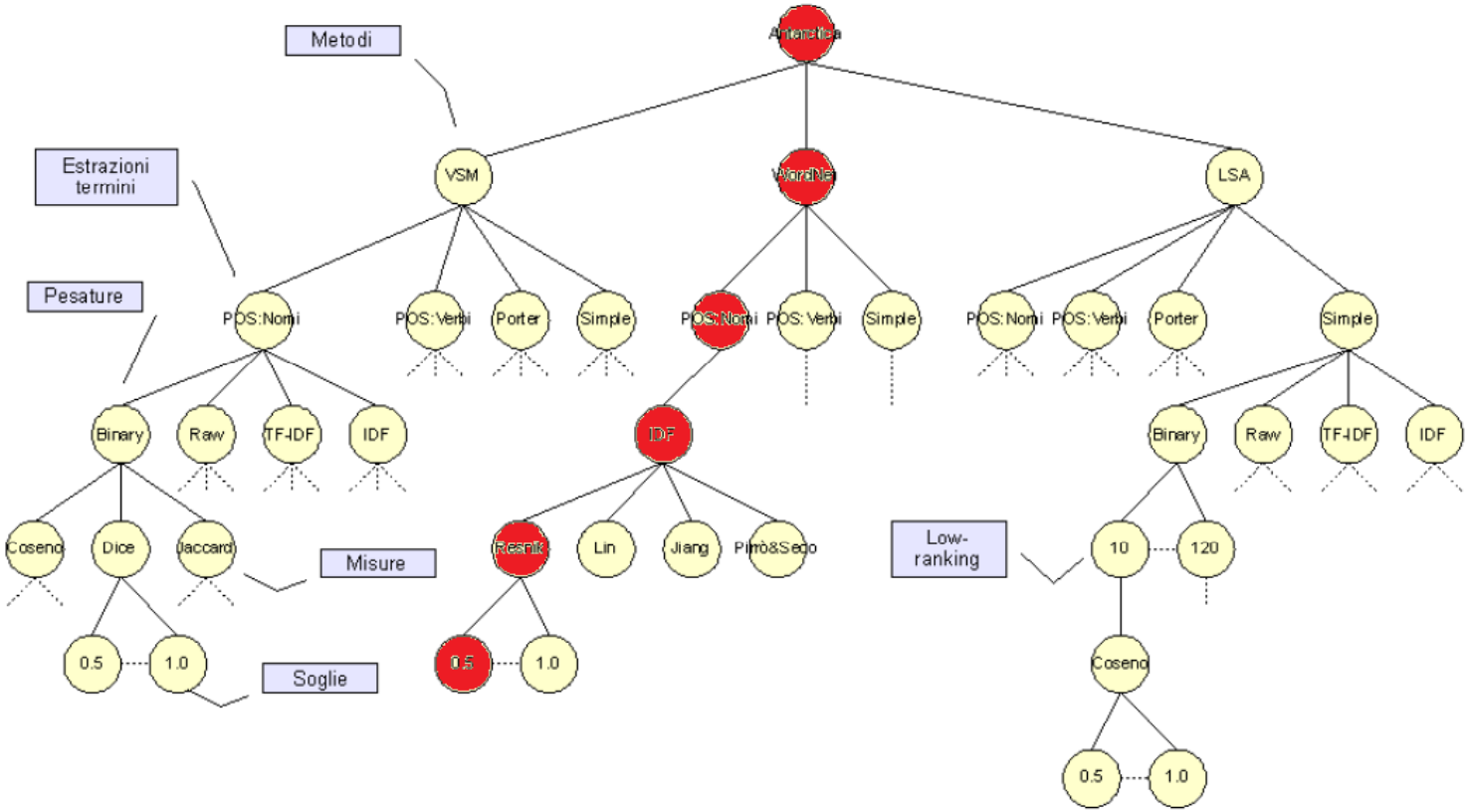
- **Modelli algebrici per la misura di similarità:**
  - VSM: Vector Space Model
  - LSA: Latent Semantic Analysis
  - WordNet Similarity
- **Estrazione dei termini:**
  - Estrazione di tipo "Simple“
  - POS tagging
  - Stemming
- **Schemi di pesatura**
  - Raw
  - Binary
  - Term Frequency (TF)
  - Inverse Document Frequency (IDF)
  - TF-IDF
- **Misura di similarità**
  - Misure di similarità applicate al modello Word-space: Coseno, Dice, Jaccard
  - Misure di similarità applicate alla rete semantica WordNet: Resnik, Lin, Jiang, Pirrò e Seco

Processi di valutazione: visione d'insieme



1300 processi possibili !

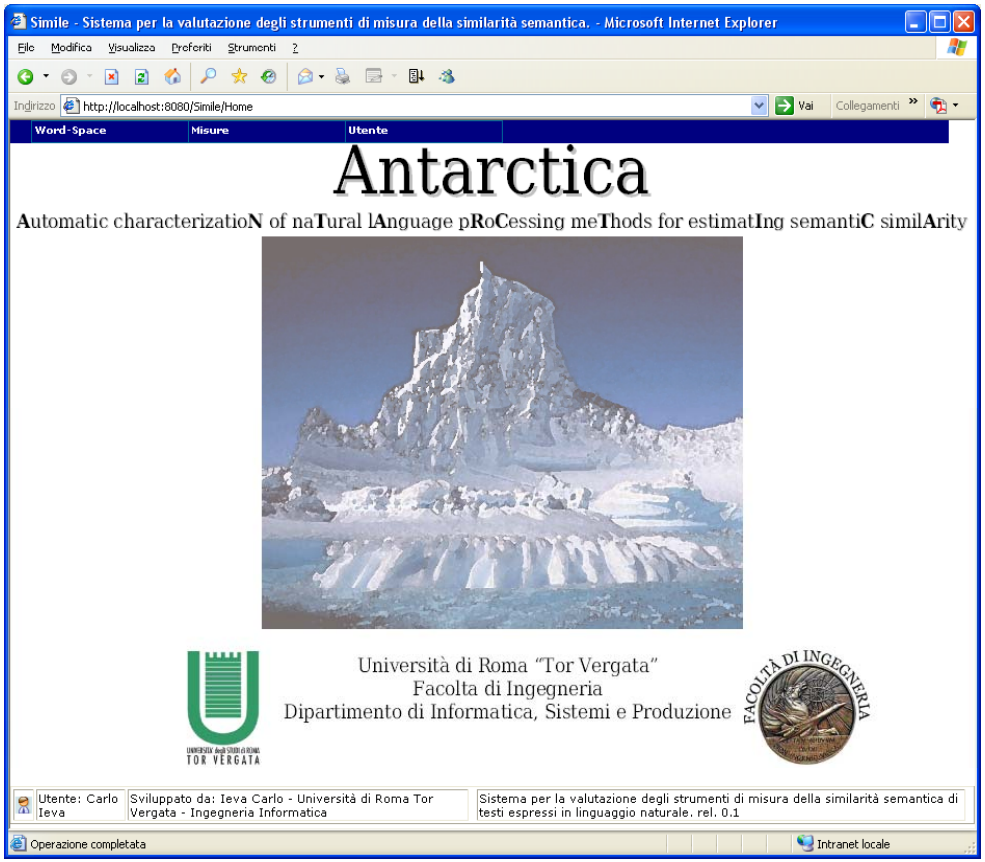
Processo di valutazione



Domande di ricerca

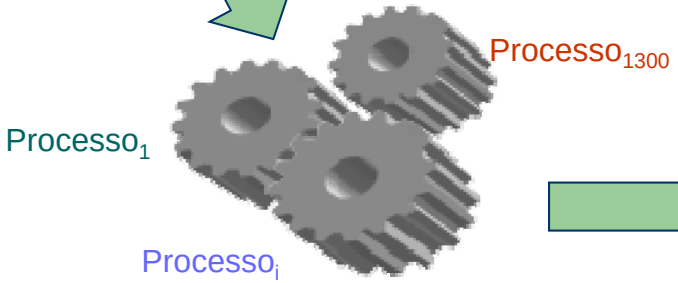
- Come stabilire quale tra questi processi è il più adatto per un certo contesto?

Il sistema “ANTARCTICA”

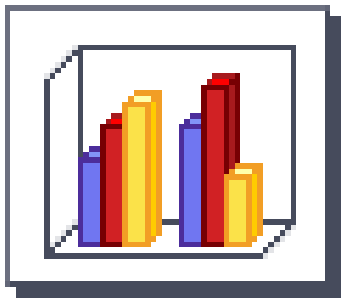


Sistema: vista d'insieme

Oracolo



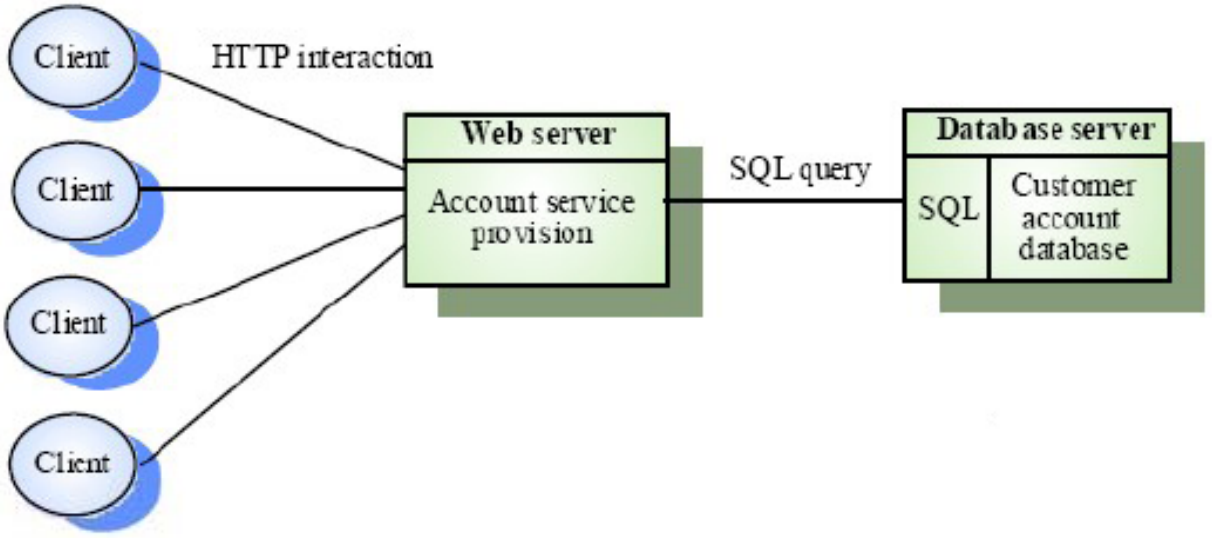
Strumento



Caratterizzazione dei metodi

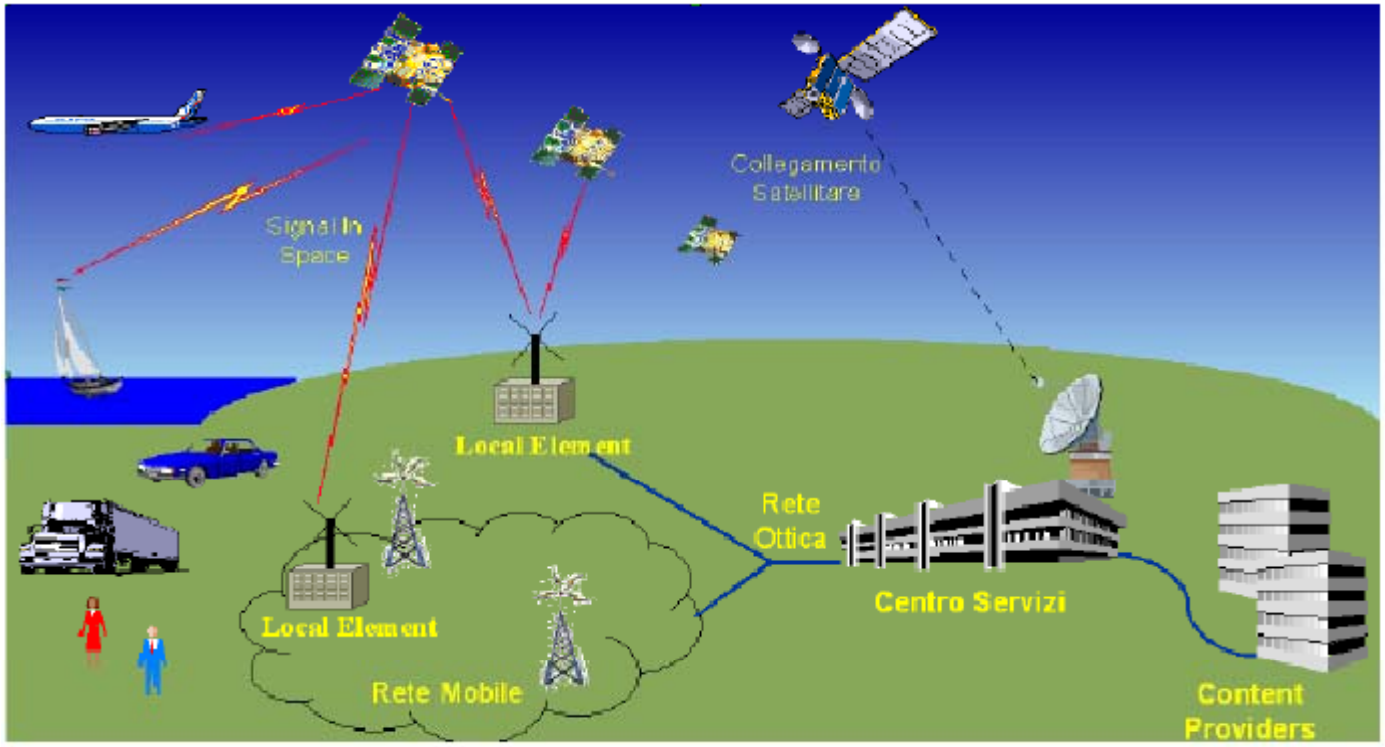
Sistema: architettura

Three-tier: Client, Application server e Data server.



Validazione: uno studio pilota

Selex SI e l'Ingegneria dei Sistemi di sistemi

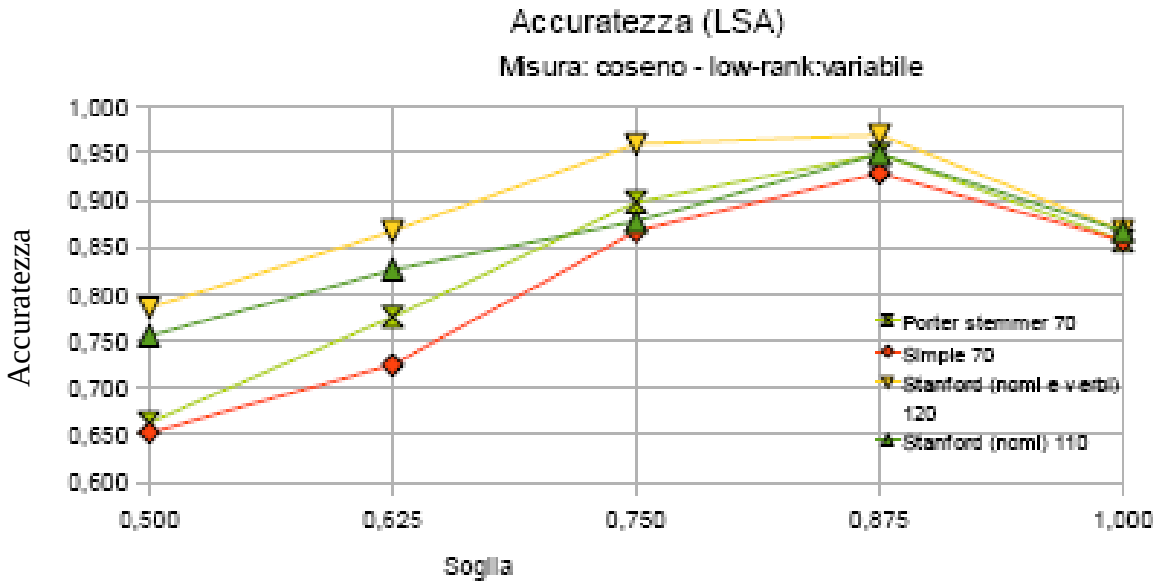


Analisi supportate.

1/2 Modalità non supervisionata → Accuratezza

	Similarità < Soglia	*Similarità >= Soglia*	*Totali*
Non equivalenti	A Veri negativi	B Falsi positivi	A+B
Equivalenti	C Falsi negativi	D Veri positivi	C+D
Totali	A+C	B+D	A+B+C+D

$$\text{Accuracy} = \frac{(A+D)}{(A+B+C+D)}$$



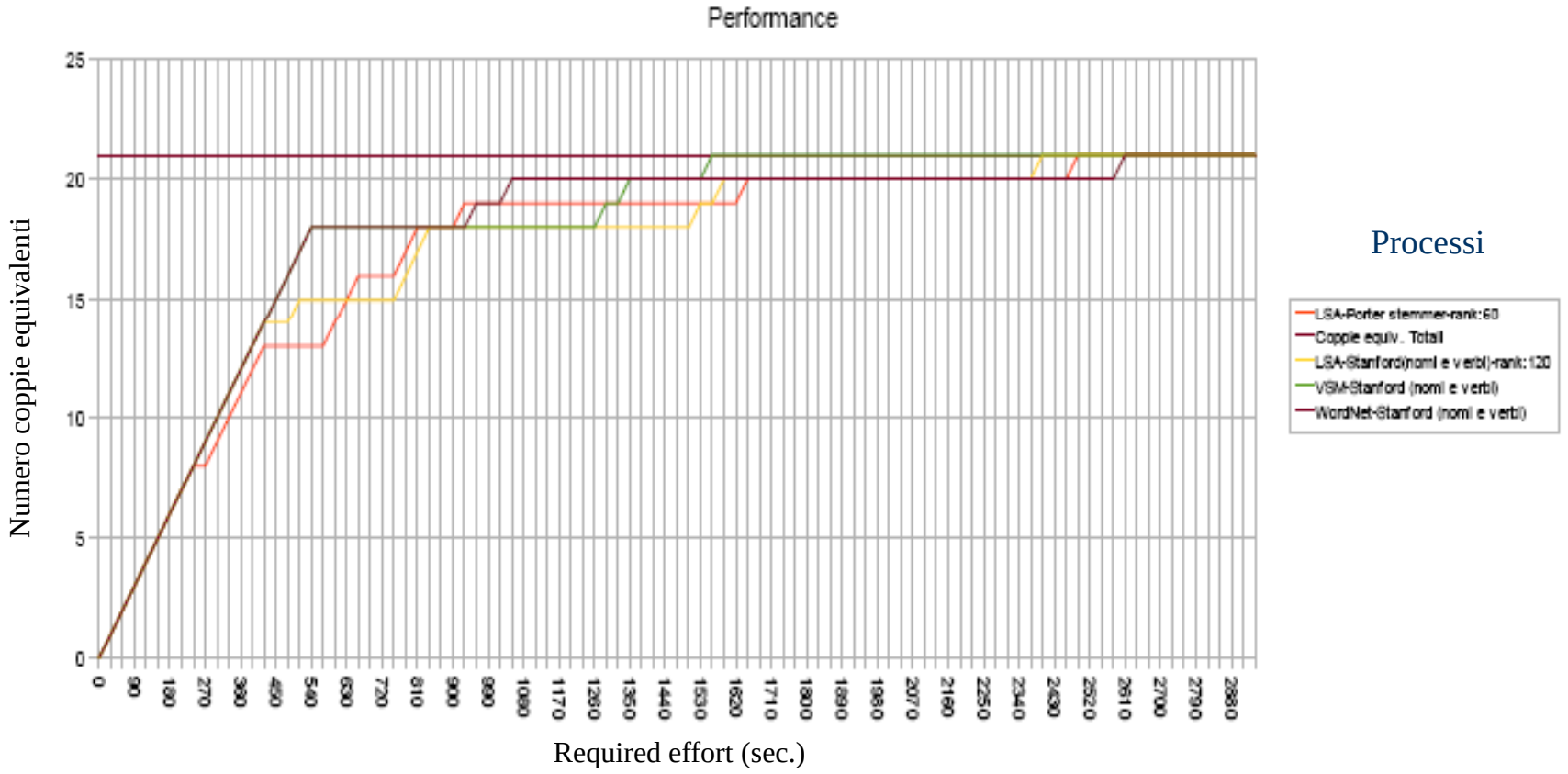
Analisi supportate:
2/2 Modalità supervisionata

In un approccio supervisionato:

1) le coppie di requisiti vengono ordinate in maniera decrescente in base al loro grado di similarità stimato da un certo processo di valutazione.

2) L'operatore decide coppia per coppia se si tratta di un duplicato o meno.

Analisi supportate:
Modalità supervisionata → Effort richiesto



Analisi supportate:
Modalità supervisionata→Effort risparmiato

L'analisi dell'effort risparmiato stima il tempo che un operatore risparmierebbe, al fine di individuare tutte le coppie di requisiti equivalenti, utilizzando l'ordinamento proposto da un dato metodo di stima della similarità semantica, rispetto ad un ordinamento casuale.

Metodo	Misura	Estrazione termini	Pesatura	Pr	Effort risp. (sec.)
WordNet	Resnik	POS tag (nomi e verbi)	IDF	87	253.64
LSA	Coseno	POS tag (nomi e verbi)	TF-IDF	76	583.64
LSA	Coseno	Porter stemmer	TF-IDF	55	1213.64
VSM	Coseno	POS tag (nomi e verbi)	TF-IDF	40	1663.64

Conclusioni

- Realizzazione di uno strumento software in grado di caratterizzare, in maniera oggettiva e quantitativa, le tecniche di IA in termini di supporto nell'identificazione di requisiti semanticamente equivalenti.
- Validazione del sistema mediante l'utilizzo di requisiti facenti riferimento a uno studio pilota svolto in un reale ambiente industriale, la Selex SI.