# Generazione automatica di word cloud dinamiche

## Ingegneria Informatica e dell'Automazione

Università degli Studi di Perugia
Facoltà di Ingegneria

**Candidato**: *Enrico Spataro*
**Relatore**: *Prof.ssa Carla Binucci*
**Relatore**: *Prof. Walter Didimo*

19.02.2016

# Contenuti

# Obiettivi

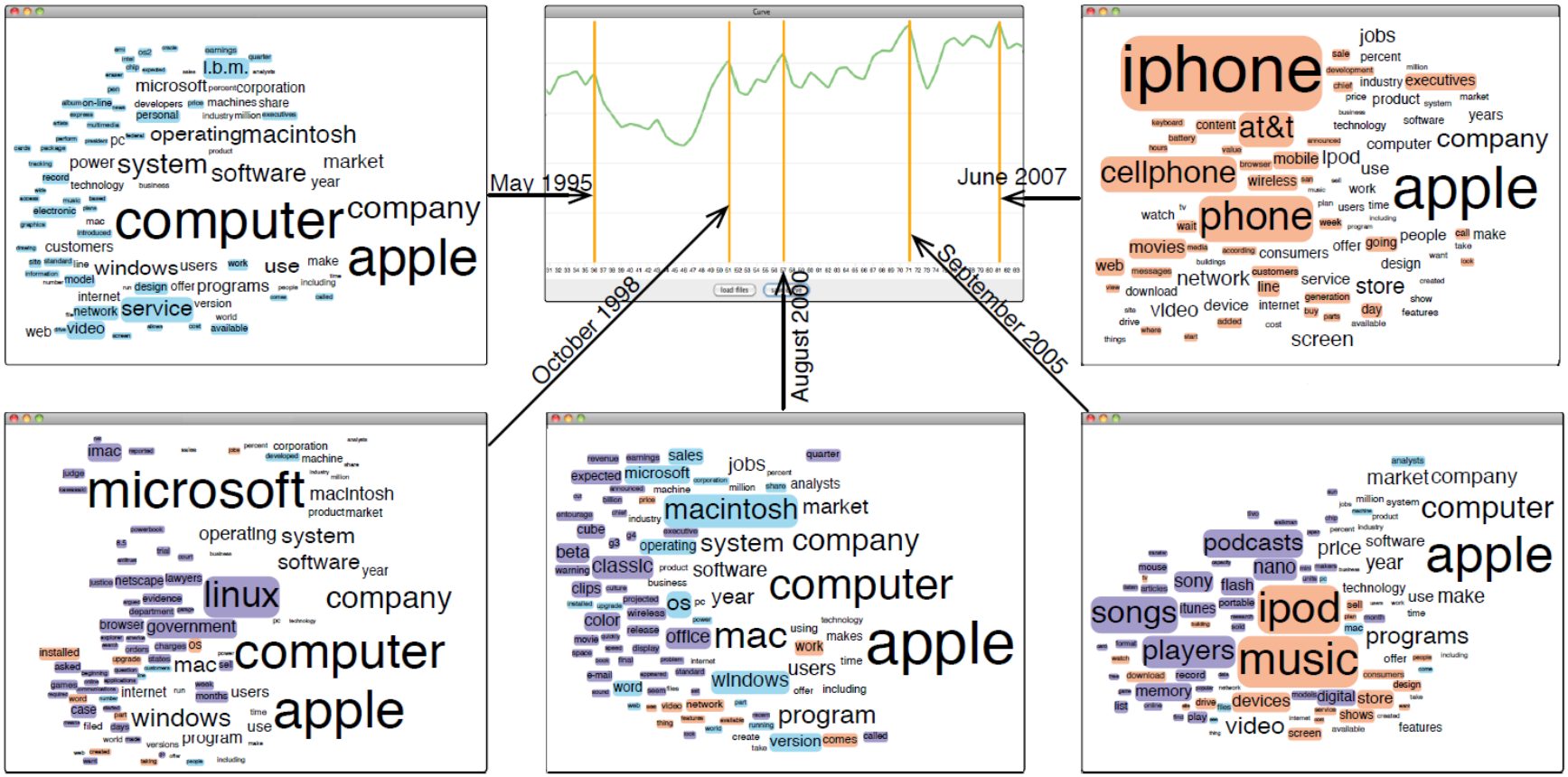
Mostrare l'evoluzione di un testo tramite:

- creazione word cloud semantiche ad intervalli regolari;
- animazioni tramite morphing.

# Cos'è una word cloud?

# Word cloud semantiche (Kobourov et al., 2014)

# Word cloud dinamiche (Cui et al., 2010)



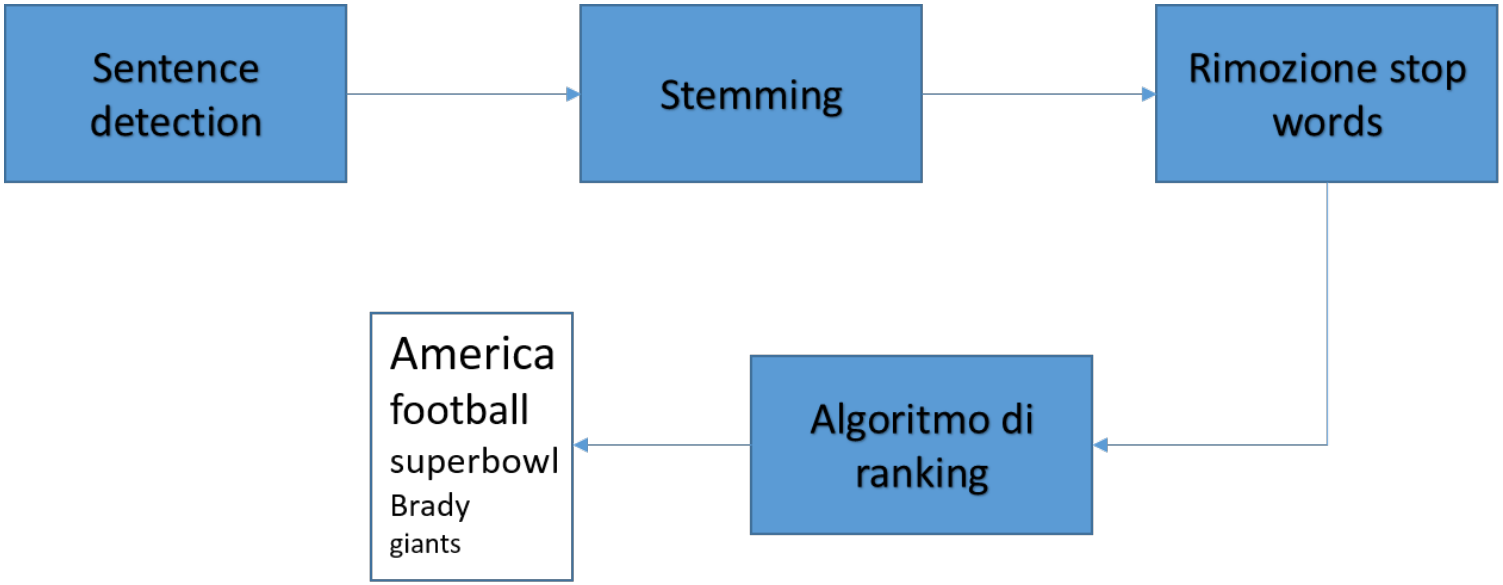
# **Morphing** (Chi et al., 2015)

# Generazione word cloud semantica

1. Estrazione parole chiave
2. Calcolo similarità
3. Clustering
4. Disegno word cloud

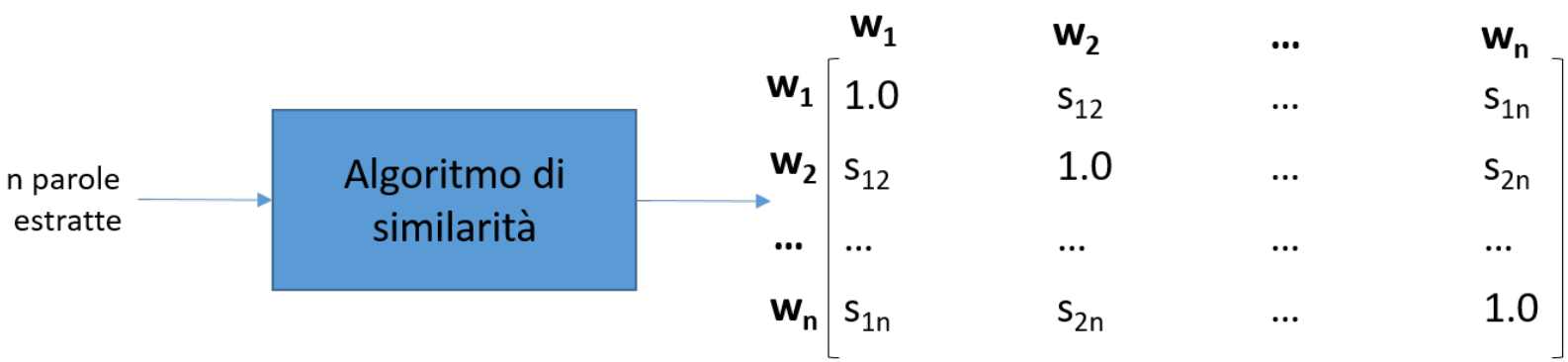
# Generazione word cloud semantica

1. Estrazione parole chiave
2. Calcolo similarità
3. Clustering
4. Disegno word cloud



# Generazione word cloud semantica

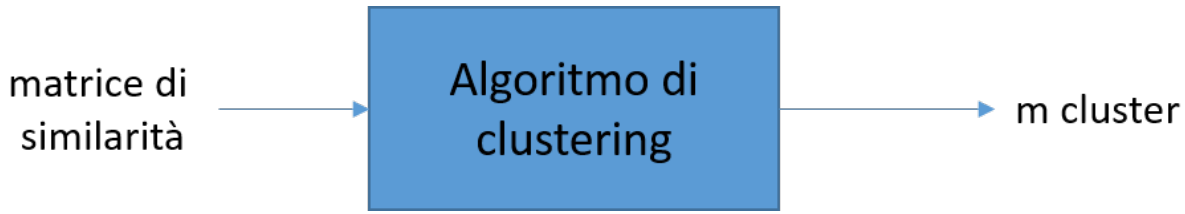
1. Estrazione parole chiave
2. Calcolo similarità
3. Clustering
4. Disegno word cloud




$s_{ij} = s_{ji}$ , $s_{ij} \in [0, 1]$

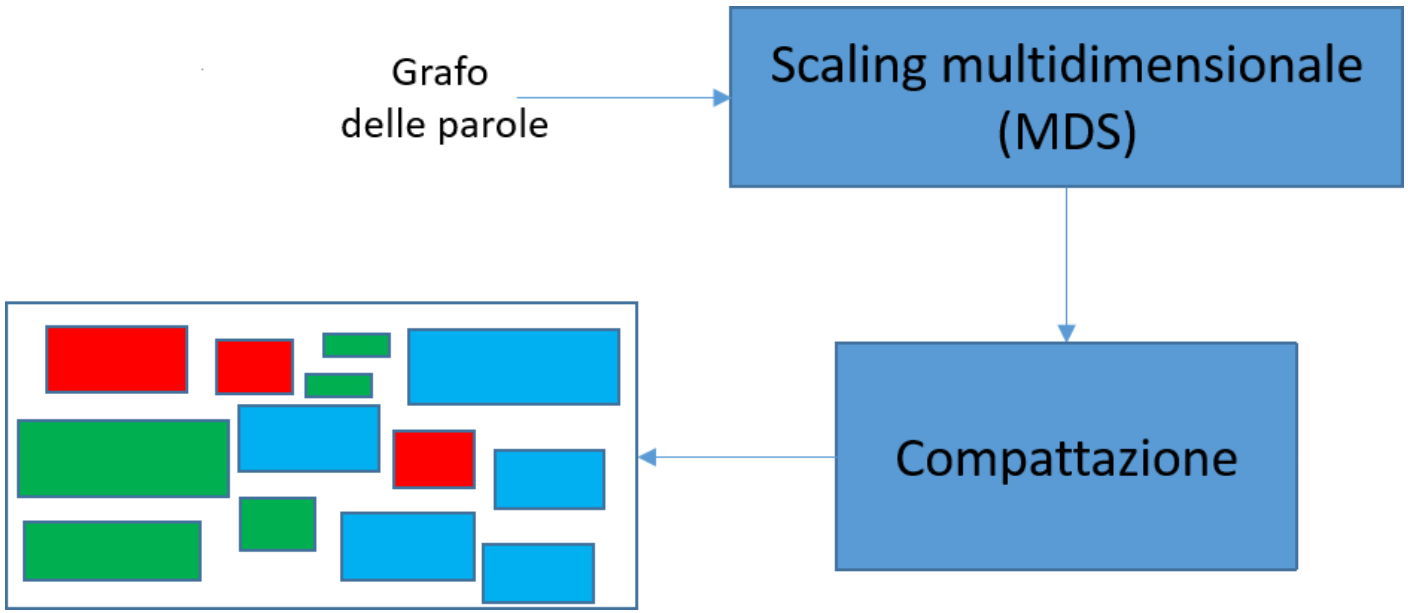
# Generazione word cloud semantica

1. Estrazione parole chiave
2. Calcolo similarità
3. Clustering
4. Disegno word cloud



# Generazione word cloud semantica

1. Estrazione parole chiave
2. Calcolo similarità
3. Clustering
4. Disegno word cloud



# Generazione word cloud dinamica

Obiettivo: create $K$ word cloud dal testo, si vuole passare da una word cloud alla successiva in modo graduale tramite tecniche di morphing.

# Morphing

## Caratteristiche

- Numero di frame
- Gestione stato delle parole

# Stato delle parole

Le parole, tra una word cloud e la successiva, possono:

- scomparire
- apparire
- rimanere nel layout (variando posizione)

# Stato delle parole

Le parole, tra una word cloud e la successiva, possono:

- scomparire
- apparire
- rimanere nel layout (variando posizione)

# Stato delle parole

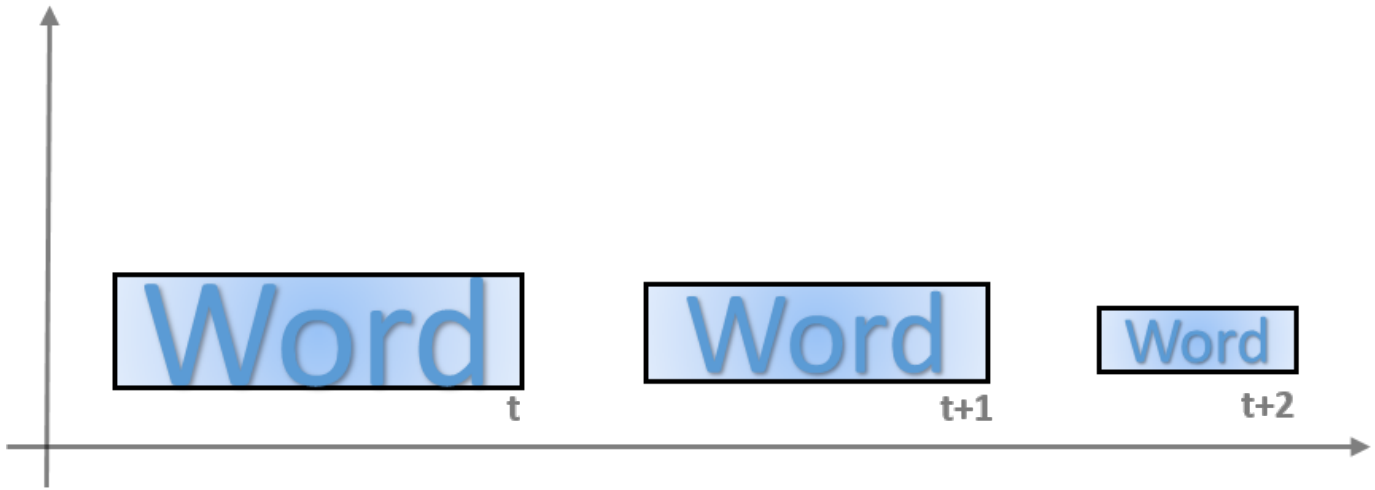
Le parole, tra una word cloud e la successiva, possono:

- scomparire
- apparire
- rimanere nel layout (variando posizione)

# Stato delle parole

Le parole, tra una word cloud e la successiva, possono:

- scomparire
- apparire
- rimanere nel layout (variando posizione)



# Generazione word cloud dinamica

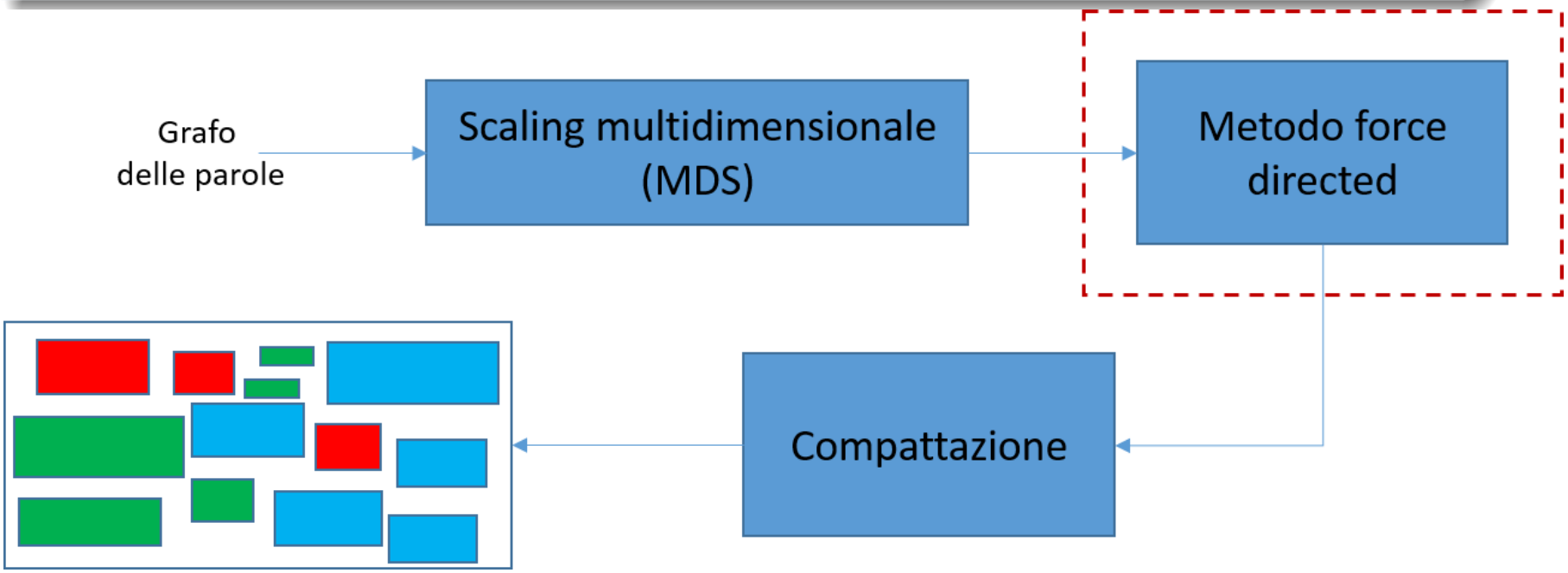
## Possibili problematiche

1. Variazione posizione delle parole
2. Variazione dei cluster

# Generazione word cloud dinamica

## Possibili problematiche

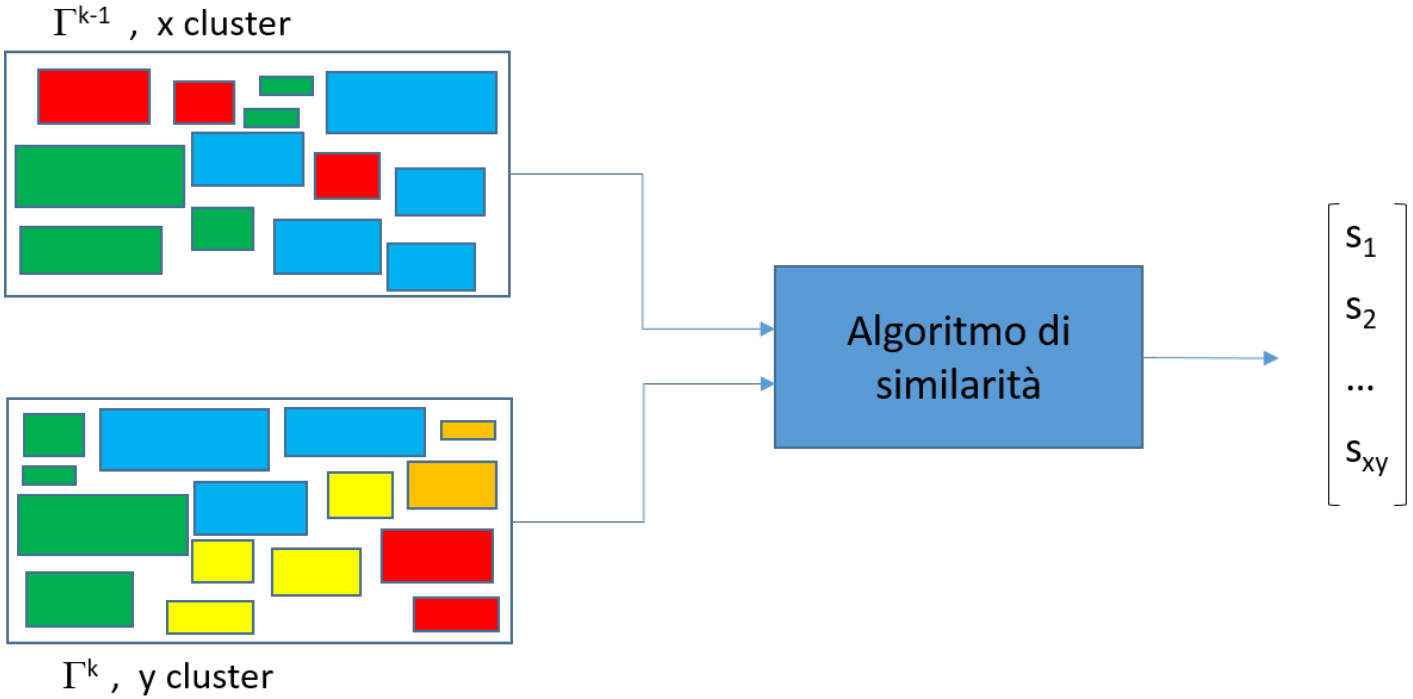
1. Variazione delle parole
2. Variazione cluster



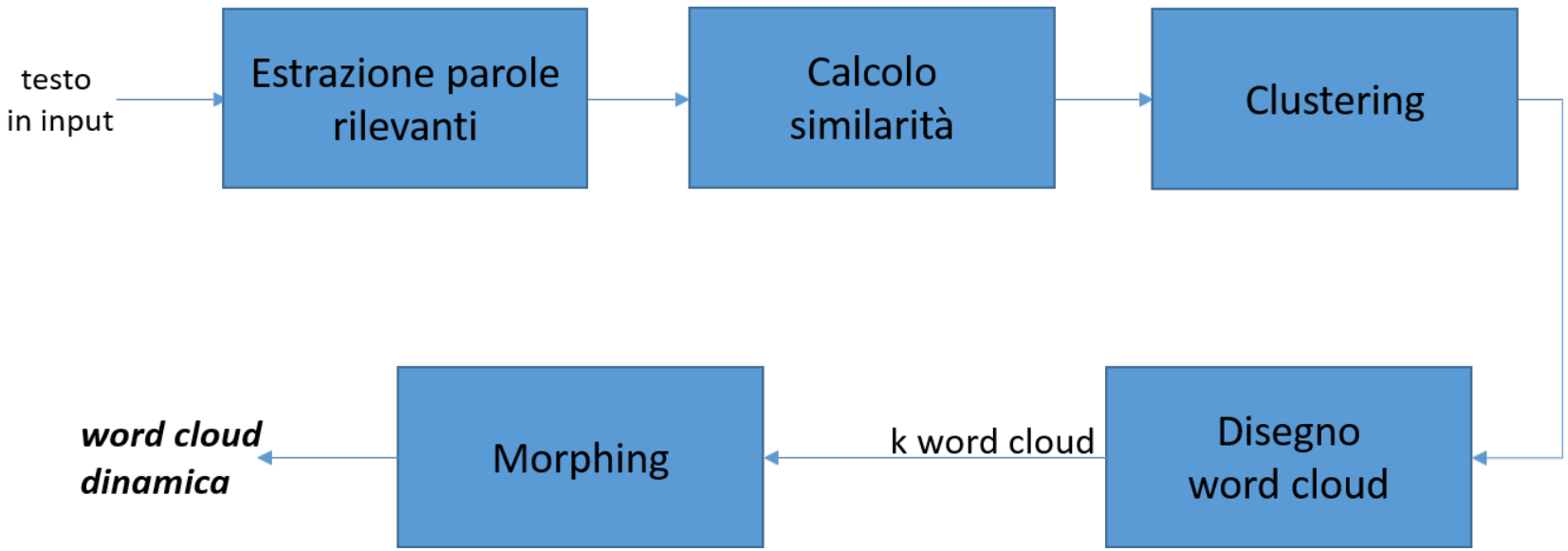
# Generazione word cloud dinamica

## Possibili problematiche

1. Variazione delle parole
2. Variazione cluster



# Architettura



## Word cloud dinamica

# Risultati sperimentali

Test suite:

- 200 discorsi estratti (file *.txt*) dal set di conferenze annuali TED (Technology Entertainment Design).
- Lunghezza media 17/18 minuti.
- 4 campionamenti per ogni testo.
- Parole estratte: 20,40,60

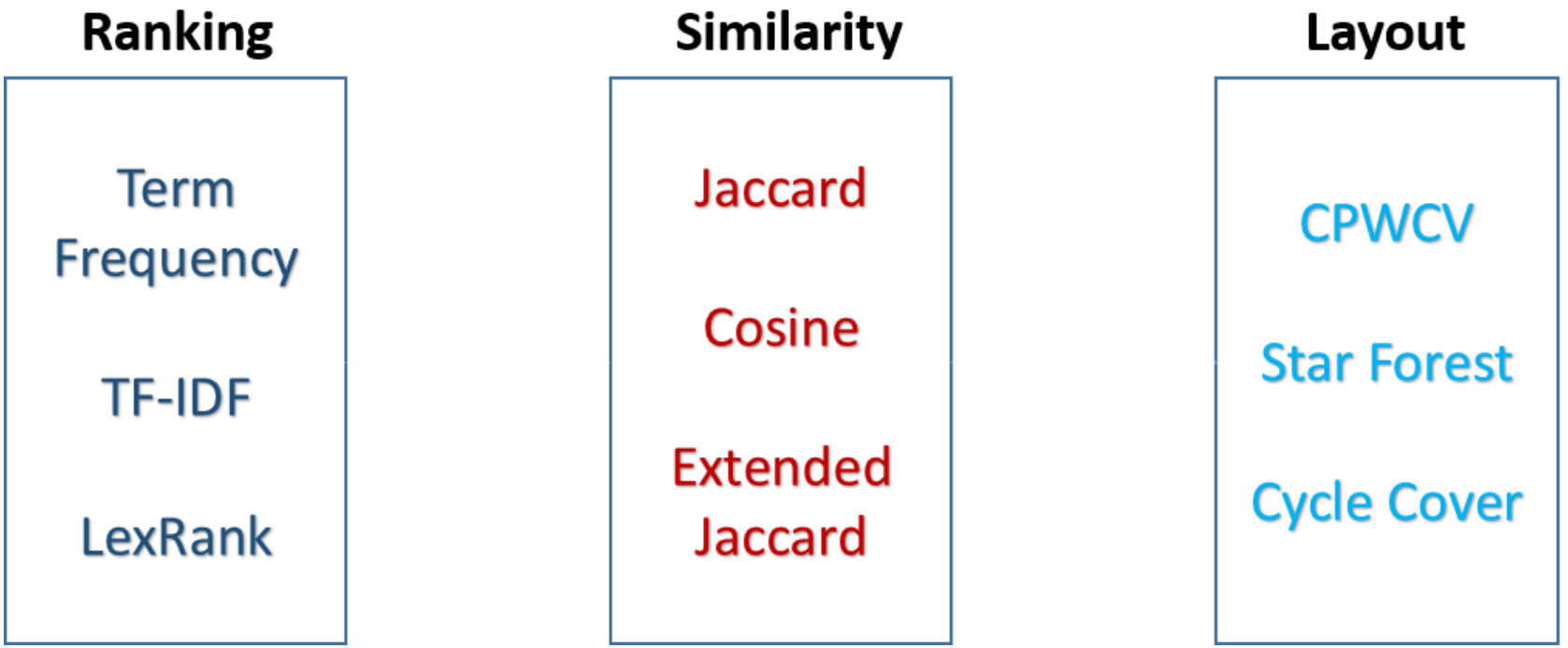
# Risultati sperimentali

Metriche adottate:

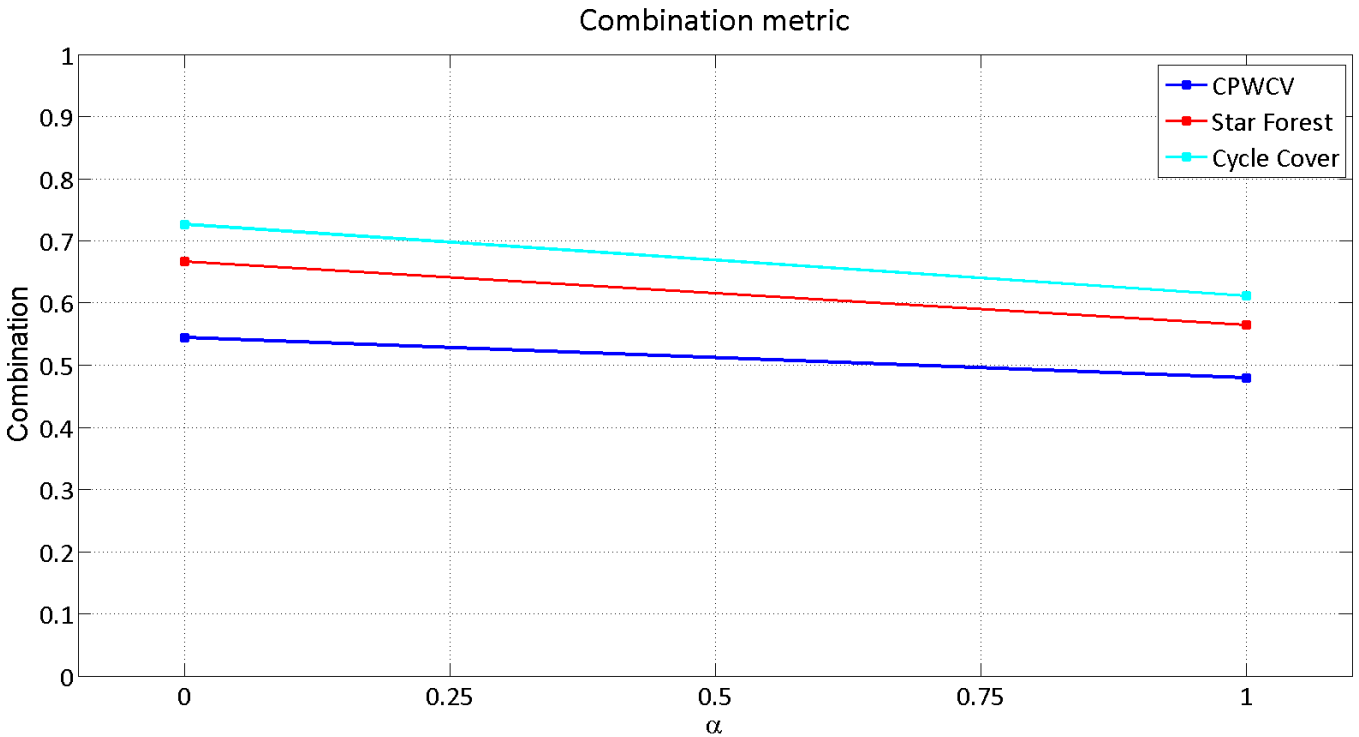
- Combination metric $\Rightarrow \nu = \frac{1}{K}\sum_{k=1}^{K}(\alpha S^k + \beta \vartheta^{k,k-1})$
  - Distortion metric $\Rightarrow S^k = \frac{\sum_{ij} c_{ij} s_{ij}}{\sum_{ij} s_{ij}}$
  - Coherence metric $\Rightarrow \vartheta^{k,k-1} = 1 - \frac{\sum_{i=1}^{p} \sigma(w_i^k, w_i^{k-1})}{pD}$
- Space metric $\Rightarrow \gamma = 1 - \frac{\mu}{\varphi}$
- Tempo d'esecuzione

# Risultati sperimentali

Algoritmi utilizzati:



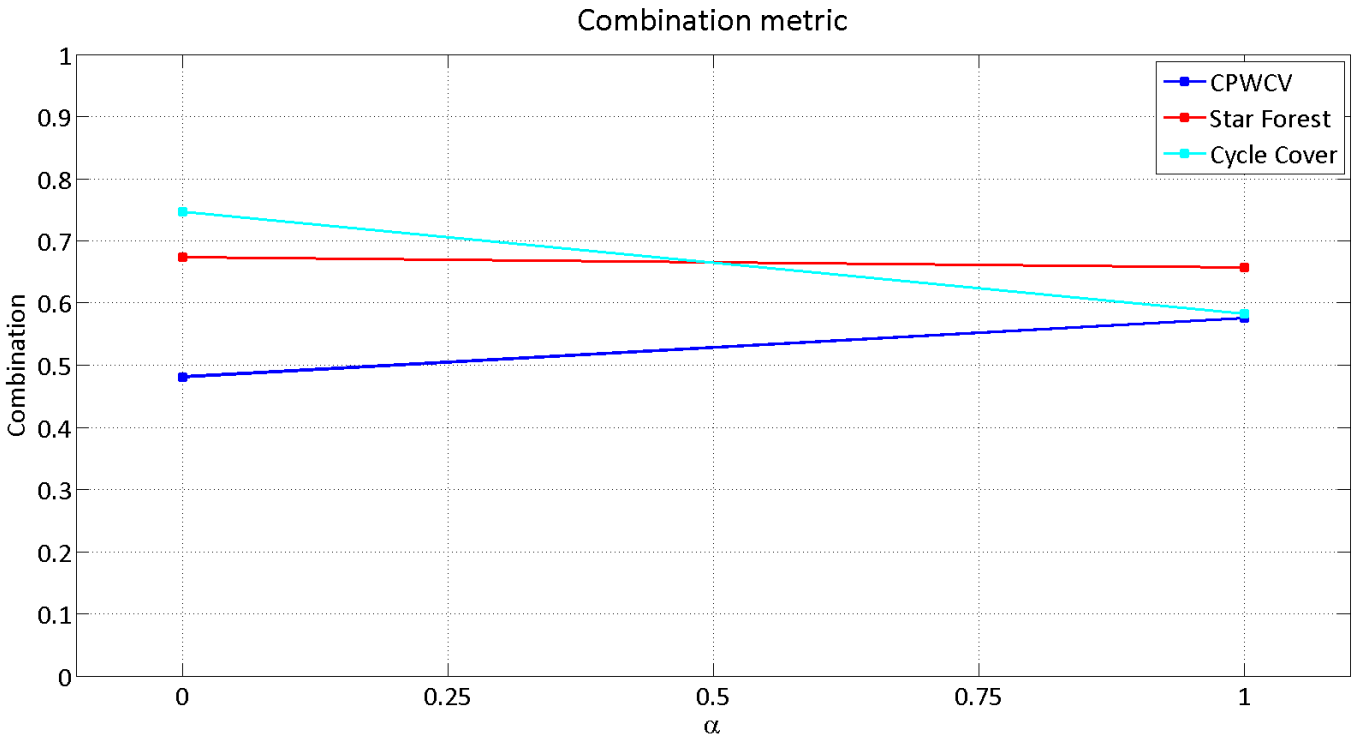
# Combination metric




Parole estratte: 20

Algoritmi utilizzati: TF-IDF Ranking, Jaccard Similarity

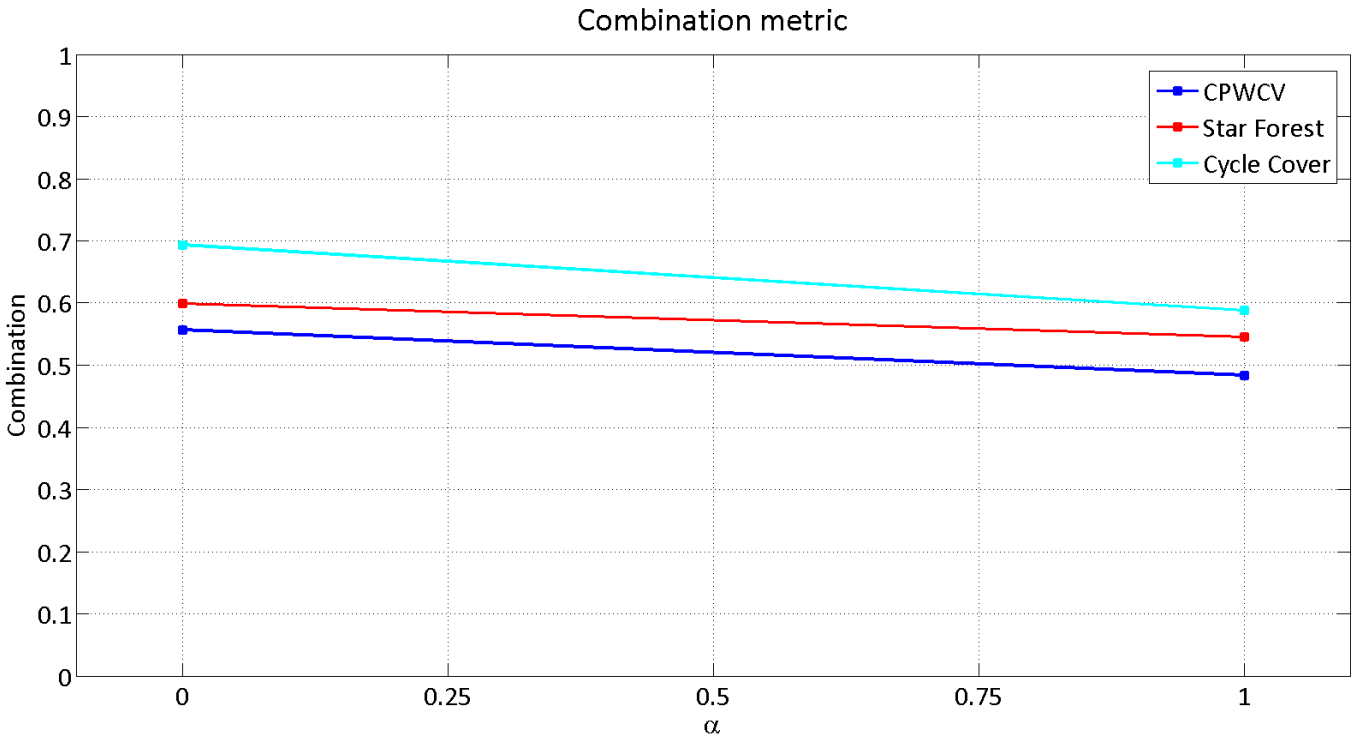
# Combination metric




Parole estratte: 40

Algoritmi utilizzati: TF-IDF Ranking, Jaccard Similarity

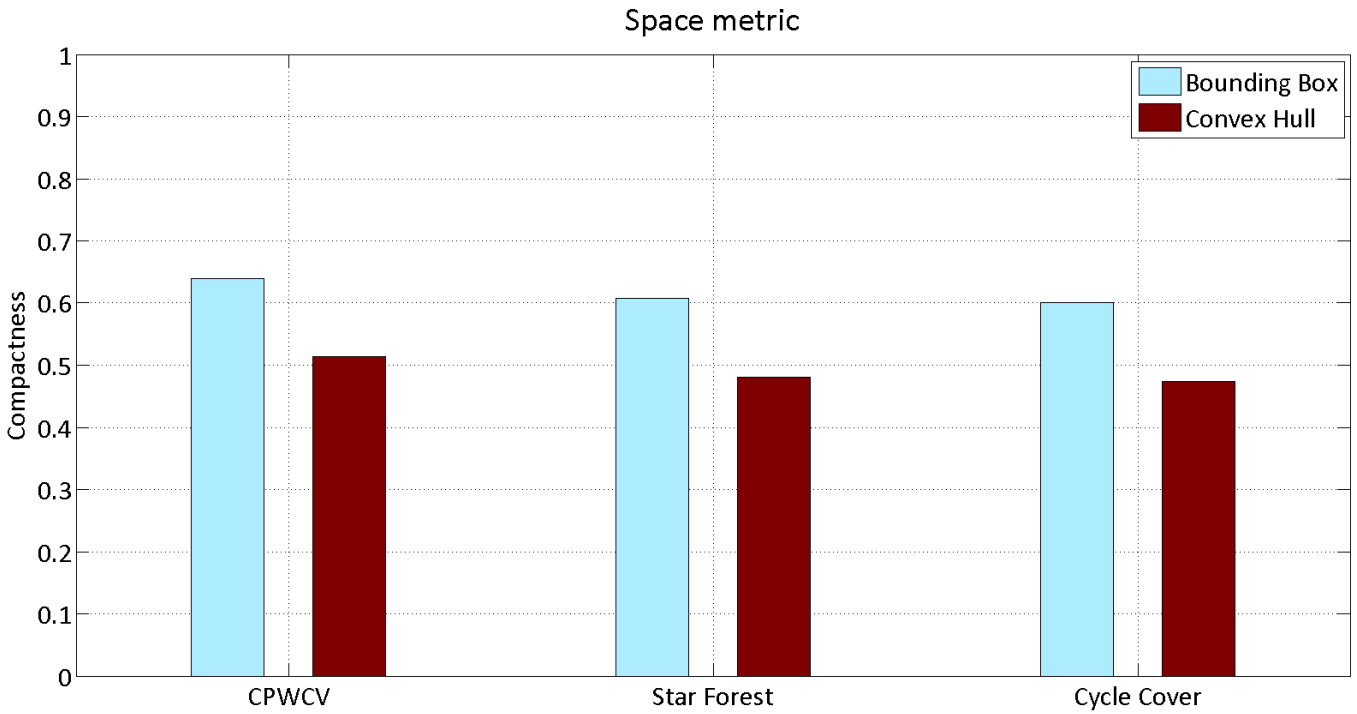
# Combination metric




Parole estratte: 60

Algoritmi utilizzati: TF-IDF Ranking, Jaccard Similarity

# Space metric




Parole estratte: 40

Algoritmi utilizzati: TF-IDF Ranking, Jaccard Similarity

# Tempo d'esecuzione

- Configurazione caso peggiore: TFIDF, Jaccard Similarity e Star Forest
- Configurazione caso migliore: Term Frequency, Cosine Similarity e CPWCV

| Parole estratte | Caso peggiore | Caso migliore |
|---|---|---|
| 20 | 2.65$s$ | 0.77$s$ |
| 40 | 3.01$s$ | 0.94$s$ |
| 60 | 3.66$s$ | 1.25$s$ |

# Conclusioni

- Nuovo approccio per la generazione di word cloud dinamiche.
- Buoni risultati dagli algoritmi di layout utilizzati.
- Tempo d'esecuzione soddisfacente.

**Sviluppi futuri:**

- Utilizzo di tecniche più sofisticate di Text Mining (e.g. *POS Tagging*).
- Definizione di metriche che tengano conto dell'interfaccia grafica + user study.

# Grazie per l'attenzione