Università degli Studi Roma Tre

Dipartimento di Ingegneria

Corso di Laurea Triennale in Ingegneria Informatica

XFP: Un Algoritmo per l'inferenza automatica della struttura di siti Web di grandi dimensioni basato sull'Analisi dei Punti Fissi

Relatore:
Dott. Valter Crescenzi

Candidato:
Lukas Canciani Graziani

Estrazione dati siti web

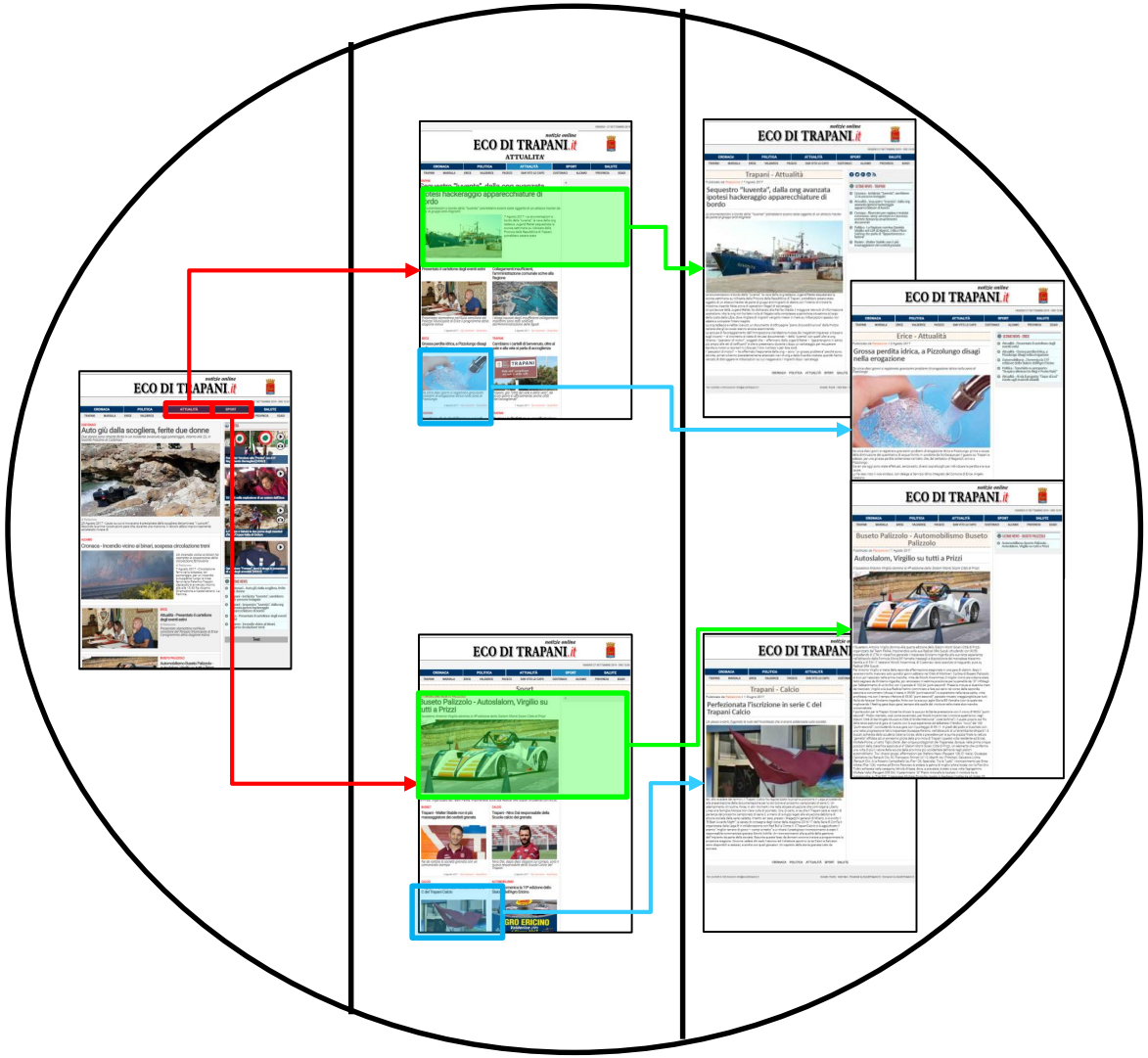
- Internet permette di accedere a enormi quantità di informazioni

- Tipicamente il processo estrattore è organizzato in due fasi
  - Crawling
  - Estrazione

Definizione del problema

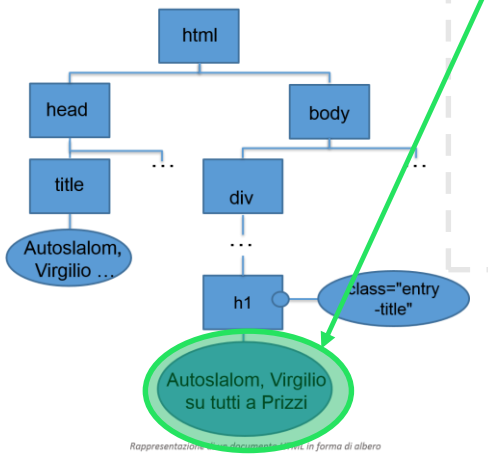
Input

Output

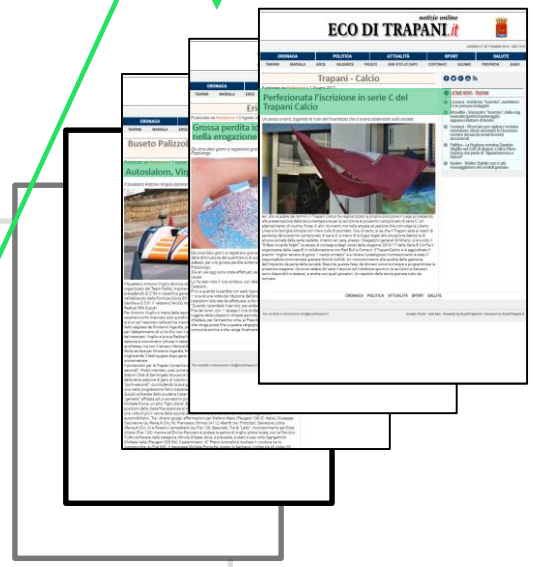


XPath e vettore di valori

- Linguaggio di query per XML
- Ogni espressione XPath identifica uno o più valori
- Valori raggruppati in vettori



//h1[@class="entry-title"]



Vettore di valori

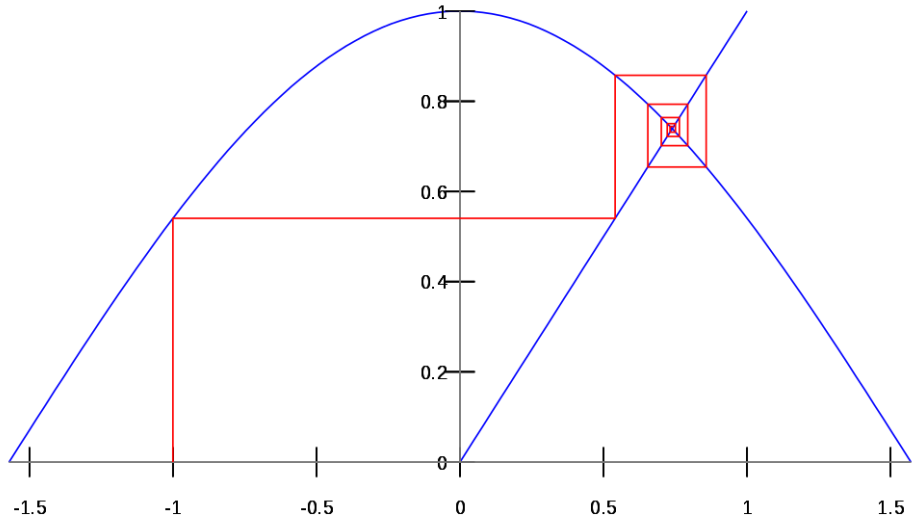
Perfezionata l'iscrizione in serie C del Trapani Calcio
Gossa perdita idrica, a Pizzolungo disagi nella erogazione
Autoslalom, Virgilio su tutti a Prizzi

Si sta cercando un elemento h1 con un attributo «class» con valore «entry-title».

XFP: Un algoritmo basato sui punti fissi

- Invariante

- Convergenza:
  - x_0
  - $x_{i+1} = f(x_i) \; \forall \; i \geq 0$
  - $f(x') = x'$

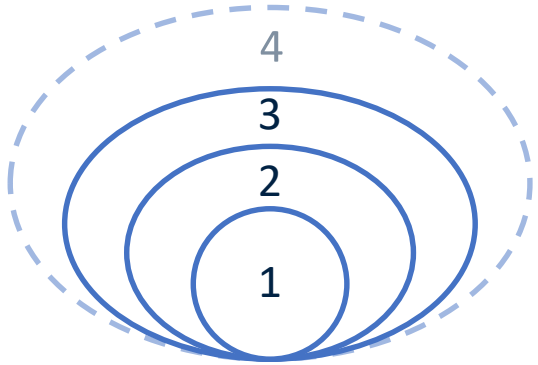


Interazione di punto fisso sulla funzione coseno

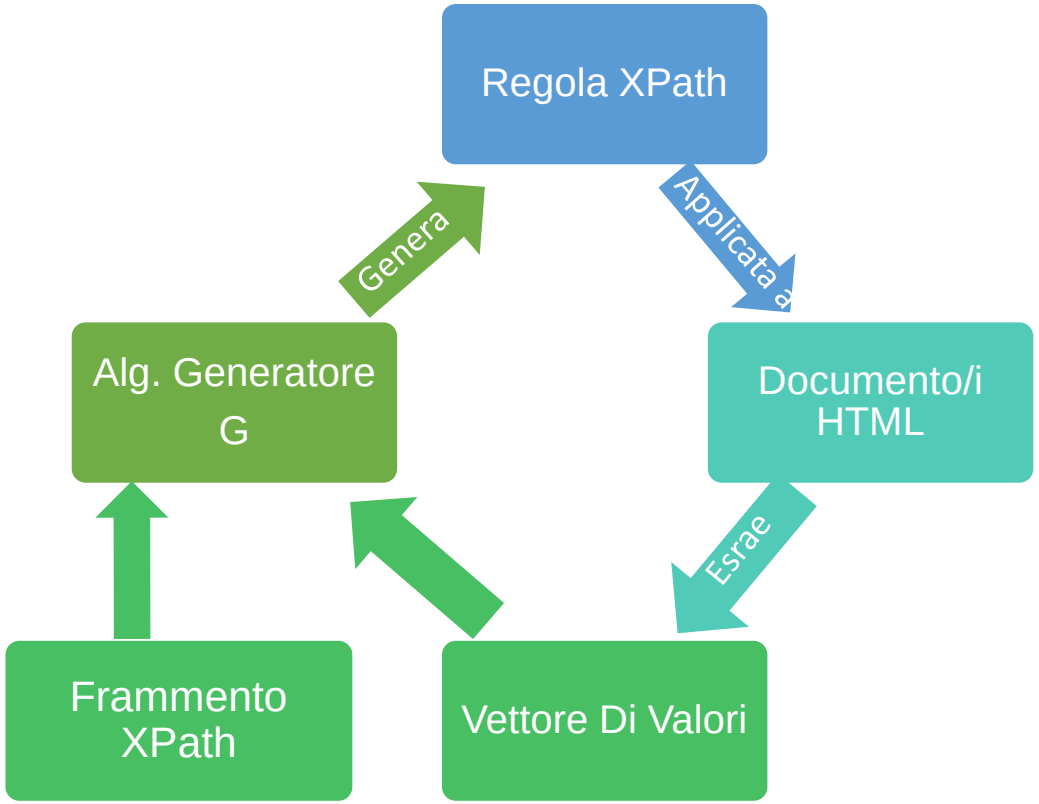
Generazione di regole e frammento XPath

- L'algoritmo Generatore (G) a partire da un vettore

- L'algoritmo utilizza un frammento XPath configurabile



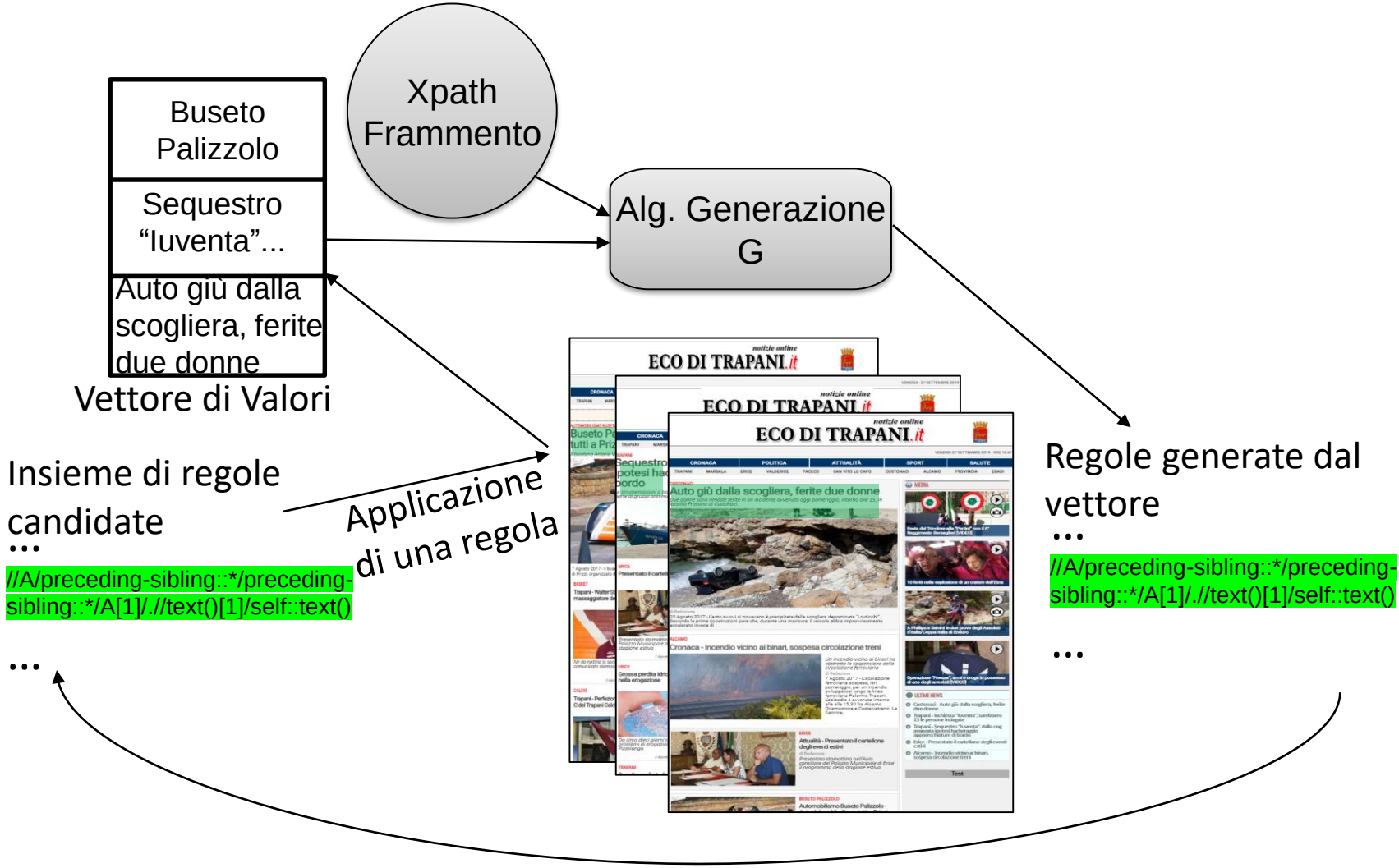
Frammento XPath



Algoritmo generatore

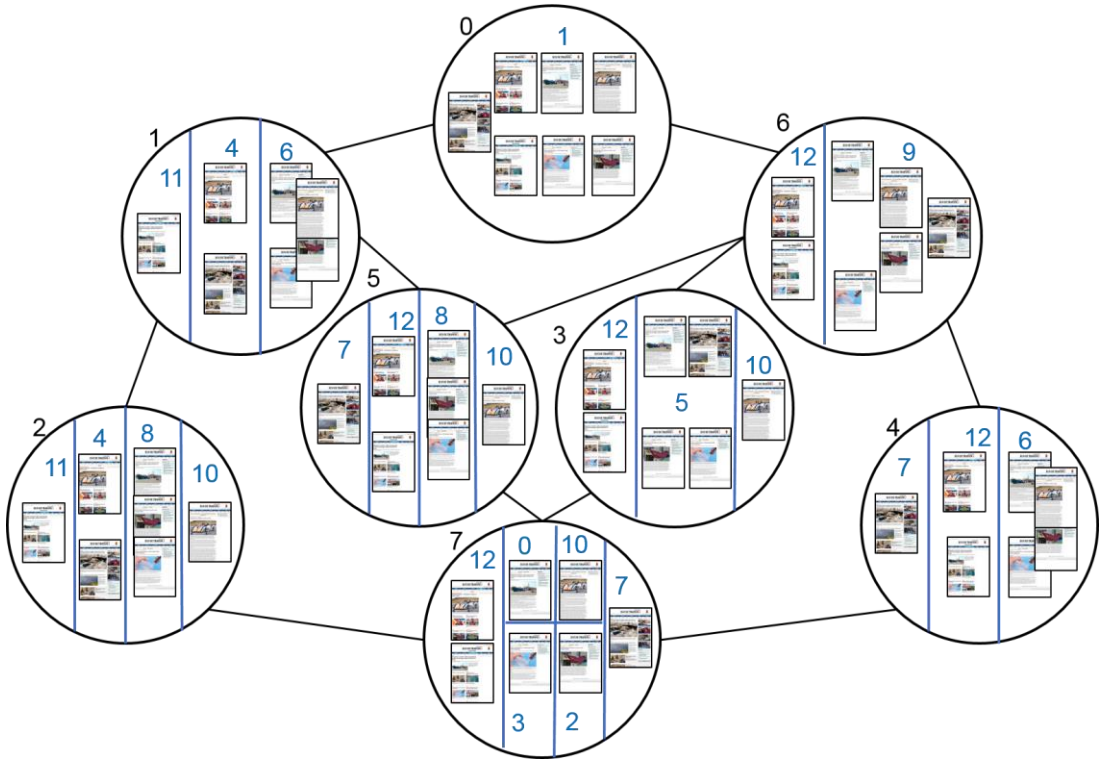
Punto fisso negli XPath

- Una regola XPath risulta punto fisso di un insieme di documenti se applicando l'algoritmo al vettore di valori che estrae si ottiene nuovamente la regola iniziale.



Reticolo di Partizioni

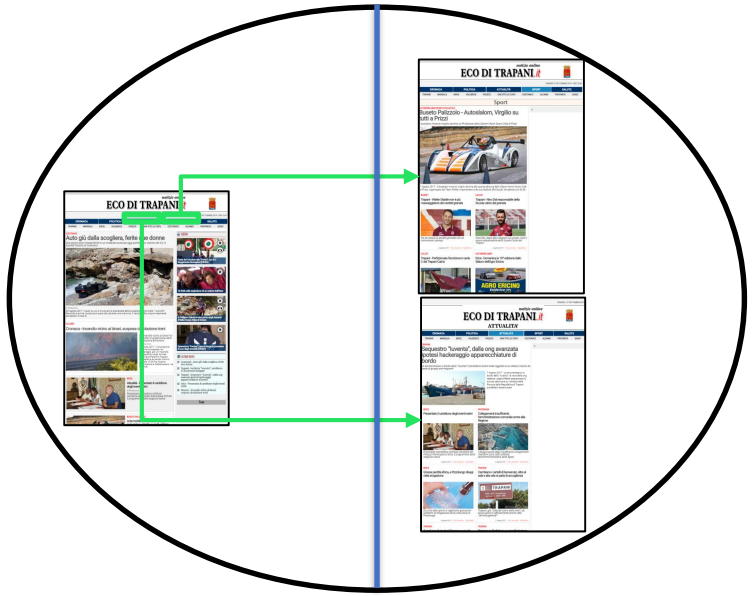
- Una partizione è un insieme di classi di pagine disgiunte che ricoprono l'intero input iniziale
- Spazio delle possibile soluzioni



Compatibilità Navigazionale dei Punti fissi

- Un punto fisso che estrae collegamenti ad altre pagine HTML è detto navigazionale
- Un punto fisso navigazione è compatibile con una partizione se è “interno” ad essa

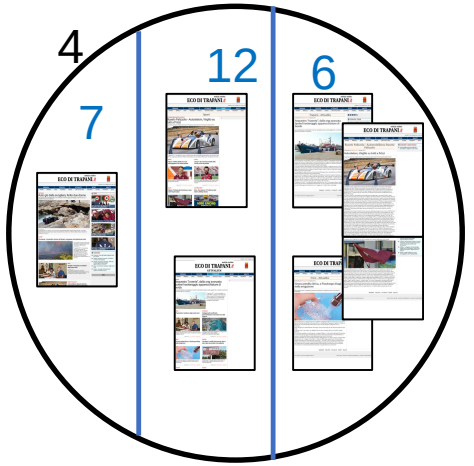
```
//DIV[@id='footer']/.//A/self::*[@href]
```



Punteggio per partizione

- Sulla base dei punti fissi, dando peso maggiore a quelli navigazionali compatibili

- Punteggi in base alle caratteristiche :
  - Costante/Variabile
  - Opzionale/Non Opzionale



Partizione 4
Navigazionali: Constanti: 7 Variabili: 18(Totali: 25)
Nivigazionali Opzionali: Constanti: 8 Variabili: 4(Totali: 12)
Dati: Constanti: 14 Variabili: 34(Totali: 48)
Dati Opzionali: Constanti: 12 Variabili: 15(Totali: 27)

Rank = (((18 + 7) / 3) * 10) + ((34*2 + 14)/3) = 110,67

$$Rank = (((NFPVariabili + NFPCostanti)/NumeroClassiDiPagine) * 10) + ((DFPVariabili * 2 + DFPCostanti)/NumeroClassiDiPagine)$$

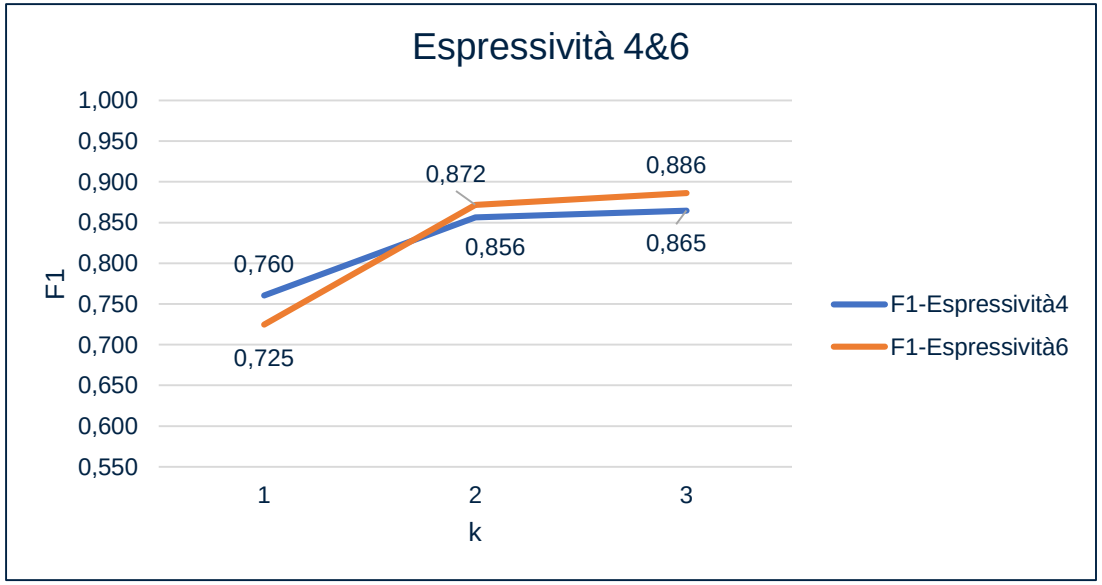
Esperimenti

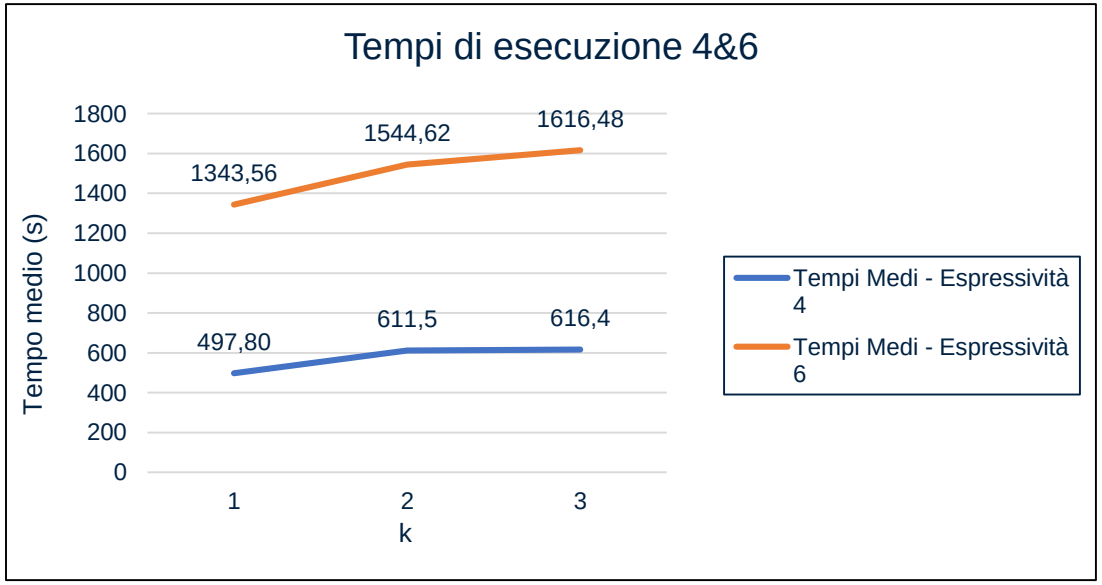
- Eseguiti su 50 siti web reali con una media di circa 45 pagine a sito
- Eseguiti a due livelli di espressività e sui top *k* punteggi
- Una possibile soluzione selezionata preventivamente
- Viene calcolata una F1 score basata su precision e recall

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

Grafici Esperimenti

- Corrispondenza tra:
  - Espressività
  - F1 score
  - Tempi di esecuzione





Sviluppi futuri

- Ottimizzazione
- Compromesso tra espressività e tempi di esecuzione