UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

tesi di laurea

Realizzazione di un software a supporto della Classificazione guidata di Pagine Web

Anno Accademico 2007/08

relatore
Ch.mo prof. Porfirio Tramontana

Candidato
Francesco Castiglia
Matr. 534002444

Scopo del progetto

- Lo scopo del progetto è di poter CLASSIFICARE pagine Web.
- E' opportuno fornire esempi di relazione di appartenenza delle pagine alle diverse partizioni definite.
- Si vuole dedurre una regola che consenta la classificazione automatica, ovvero che permetta di associare correttamente le pagine alle diverse partizioni esistenti.

Usi possibili

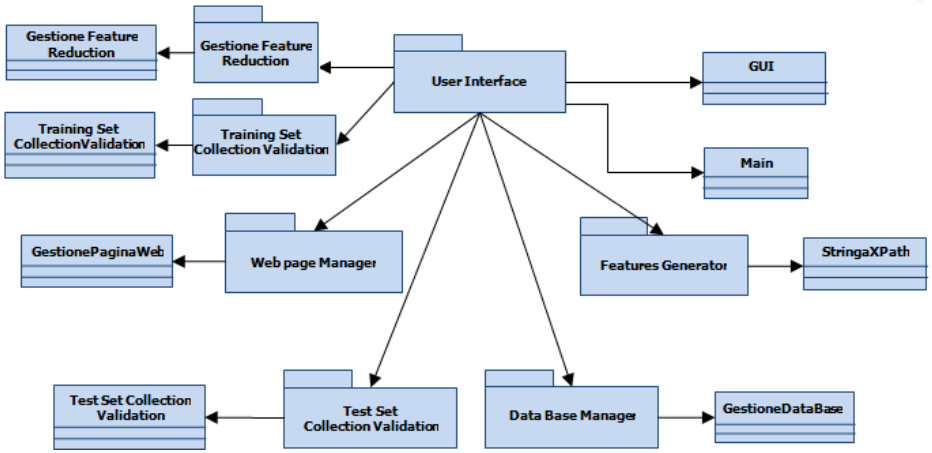
- Implementare Wrappers che incapsulano la User Interface originale al fine di esportare un'interfaccia rinnovata.
- Fornire un supporto automatico di analisi e interpretazione dei risultati di test ottenuti dalle differenti attività di validazione.
- Mettere in atto processi orientati all'ottenimento di un modello della UI con lo scopo di re-ingegnerizzarla secondo le diverse architetture (ad esempio di tipo Model-Driven) o tecnologia (AJAX).

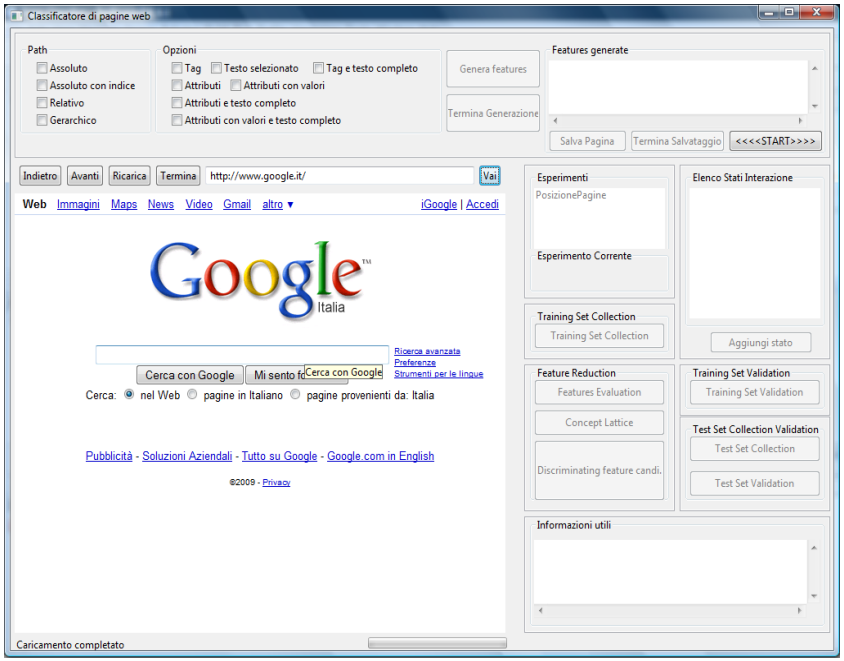
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

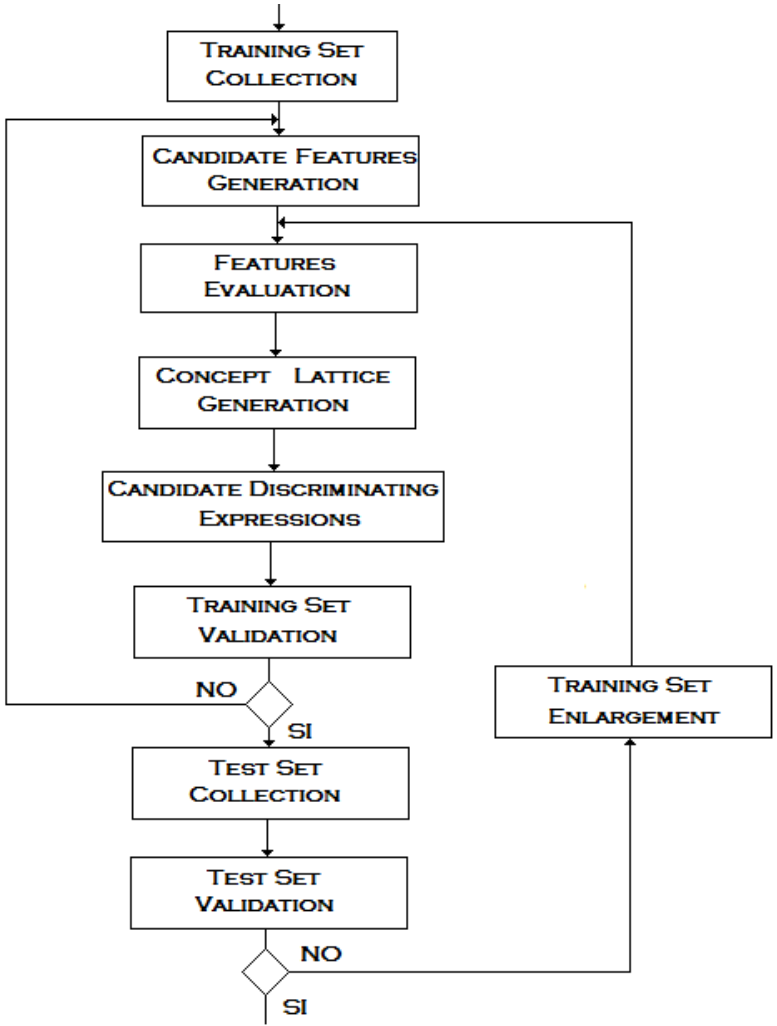
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

Realizzazione di un software a supporto della classificazione guidata di pagine web

Diagramma delle attività del:
PROCESSO DI CLASSIFICAZIONE DELLE PAGINE WEB





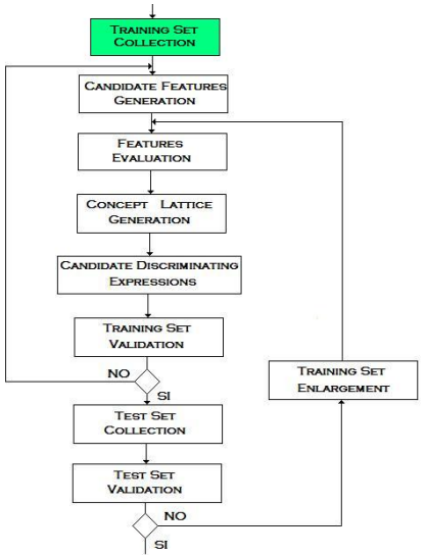


UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

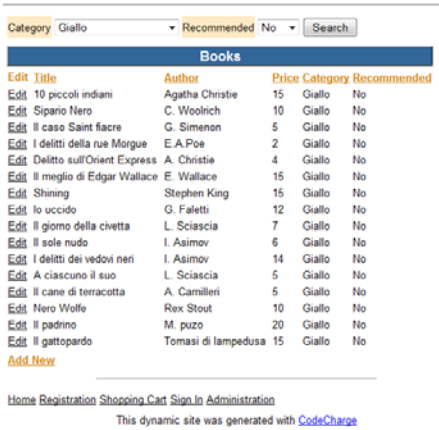
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

TRAINING SET COLLECTION:

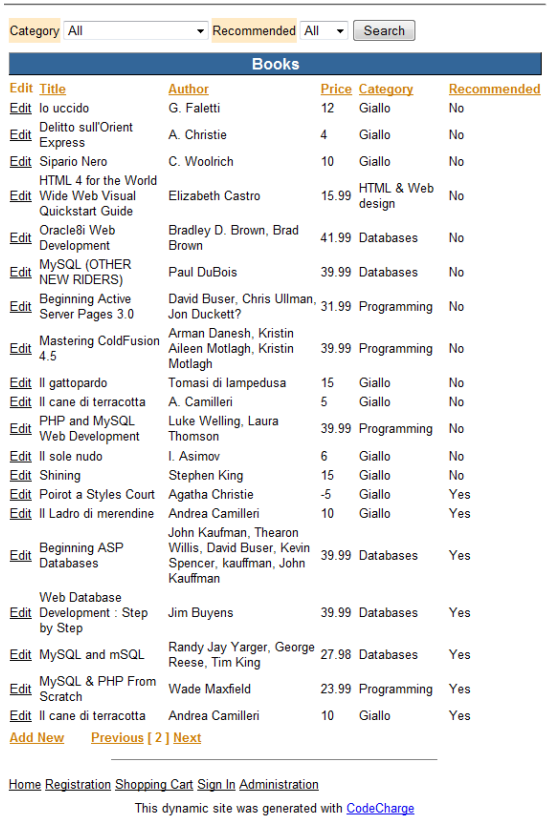
- **Tale fase consiste nel collezionare, e quindi salvare, le diverse pagine Web che popoleranno l'insieme Training Set.**
- **Fase *completamente manuale***



Single Page:

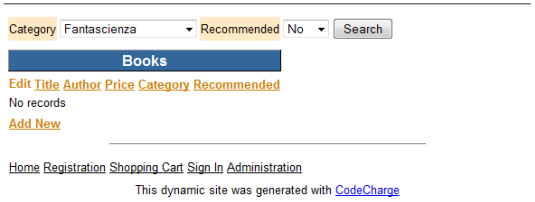


Central Page:

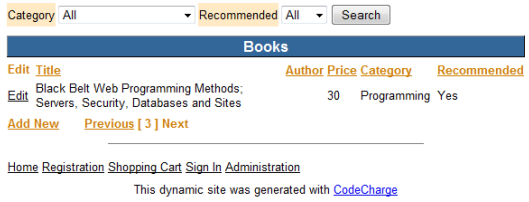


First Page:

Empty Page:

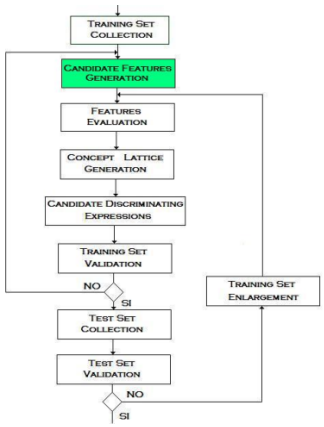


Final Page:



CANDIDATE FEATURES GENERATION:

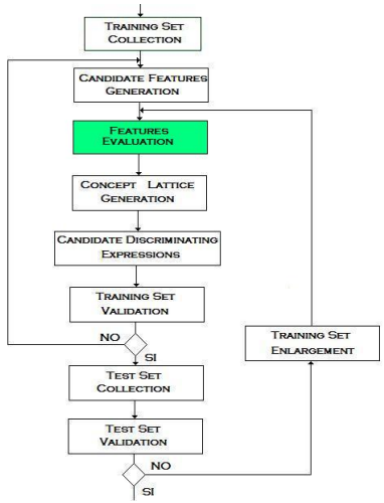
- Elemento caratterizzante di una pagina Web. ⟹ Codificato come Query Xpath.



Fase semiautomatica

FEATURES EVALUATION:

- Questo step valuta ogni singola feature, precedentemente generata, su tutte le pagine del training set.



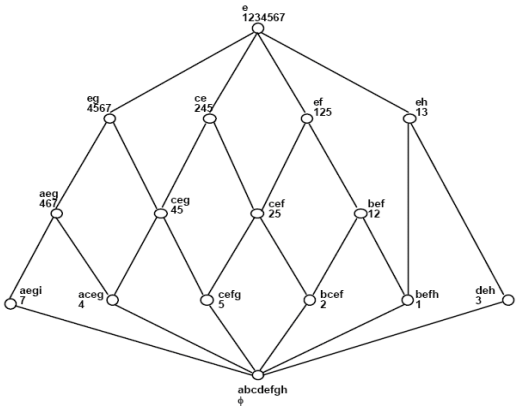
Fase automatica

Features candidate

Pagine Training	F1	F2	F3	F4	F5
P1	X				X
P2		X		X	X
P3	X				X
P4		X			X
P5	X		X		X
P6	X		X		X
P7	X		X		X

CONCEPT LATTICE:

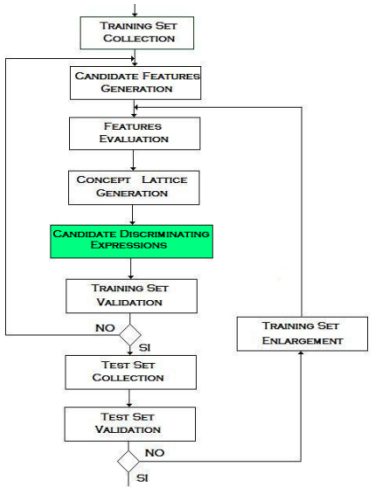
- La Concept Analysis viene adoperata per analizzare le feature che caratterizzano le diverse classi di equivalenza delle pagine.



CANDIDATE DISCRIMINATING EXPRESSIONS

Da lattice dei concetti è possibile risalire alla tipologia di ogni singola feature:

- ***Specific***: La feature è soddisfatta da tutte le pagine della partizione C
- ***Relevant:*** La feature è soddisfatta da tutte le pagine della partizione C ma anche da altre pagine di partizioni diverse.
- ***CSPC:*** (Condizionatamente specifiche) E' soddisfatta da un sottoinsieme di pagine della partizione C e da nessun'altra delle altre classe.
- ***Shared:*** La feature è soddisfatta da un sottoinsieme di pagine della partizione C ma anche da altre pagine di partizioni diverse
- ***Irrelevant:*** Non è soddisfatta da nessuna pagina della partizione C



Sulla base della tipologia delle features vengono proposte le espressioni discriminanti secondo regole ben precise

TRAINING SET VALIDATION

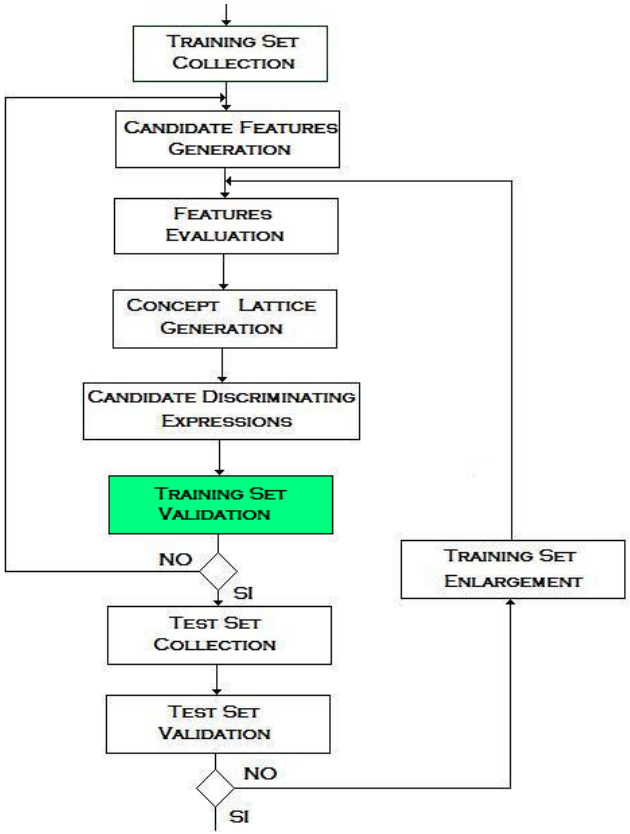
- **Verifica per ciascuna partizione se l'espressione discriminante ad essa associata riesca ad attribuire correttamente tutte le sue pagine alla partizione stessa.**

- ***Fase completamente automatica***

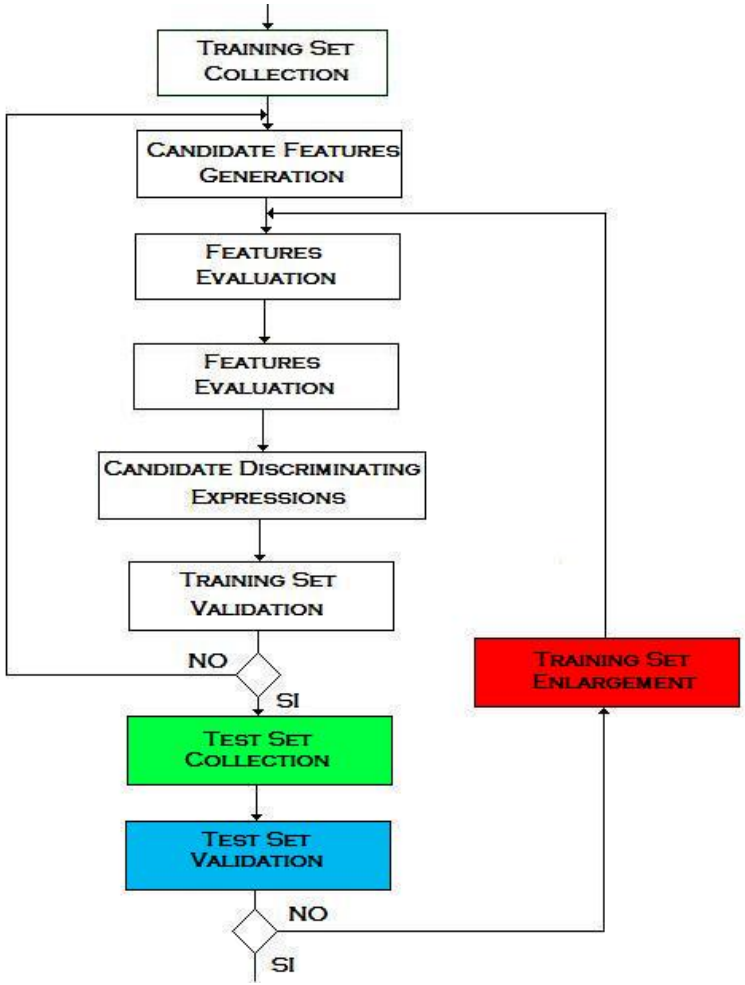
Recall: $\frac{\#\text{Pagine correttamente attribuite alla classe C}}{\#\text{pagine della classe C}}$

Precision: $\frac{\#\text{Pagine correttamente attribuite alla classe C}}{\#\text{Pagine che soddisfano l'espressioneExpr(C)}}$

- **Le espressioni sono discriminanti se entrambe le metriche Recall e Precision assumono valore unitario**



- **TEST SET COLLECTION:**
  - Tale fase provvede a popolare l'insieme, detto di test, sul quale verrà valutata l'efficacia delle espressioni discriminanti.
  - L'insieme potrebbe essere collezionato da una persona diversa dall'esperto.
- **TEST SET VALIDATION:**
  - Per ogni pagina test salvata propone la classe di equivalenza di appartenenza.
- **TRAINING SET ENLARGEMENT:**
  - Tale step è attivo solo se la partizione suggerita dal tool non coincide con quella voluta dall'esperto.
  - La pagina test viene aggiunta al training set.



- **Esperimento Applicazione Web Bookstore (parte 1):**

  - **N° classi di equivalenza: 5 (Single Page, First Page, Central Page, Final Page, Empty Page)**

  - **Numero di pagine del training set: 7 (2 – 2 – 1 – 1 – 1)**

  - **Feature proposte: 11**
    Tempo di valutazione delle 11 features sulle 7 pagine: 5.131 sec

  - **Espressioni discriminanti:**

CLUSTER	EXPR	RECALL	PRECISION
CentralPage	(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')	2/2	2/2
EmptyPage	NOT(//a[@href][@style]/text()='Next') AND NOT(//a[@href][@style]/text()='Previous')	1/1	1/2
FinalPage	NOT(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')	1/1	1/1
FirstPage	(//a[@href][@style]/text()='Next') AND NOT(//a[@href][@style]/text()='Previous')	2/2	2/2
SinglePage	NOT(//a[@href][@style]/text()='Next') AND NOT(//a[@href][@style]/text()='Previous')	1/1	1/2

Tempo di generazione espressioni discriminanti: 0,13 sec
Tempo di validazione sul Training Set: 0,25 sec

- **Esperimento Applicazione Web Bookstore (parte 2):**

  - **Nuove features proposte: 5**

    Tempo di valutazione delle 5 features sulle 7 pagine: 2.237 sec

  - **Espressioni discriminanti:**

CLUSTER	EXPR	RECALL	PRECISION
CentralPage	(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')	2/2	2/2
EmptyPage	//td/text()='No records'	1/1	1/1
FinalPage	NOT(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')	1/1	1/1
FirstPage	(//a[@href][@style]/text()='Next') AND NOT (//a[@href][@style]/text()='Previous')	2/2	2/2
SinglePage	NOT(//a[@href][@style]/text()='Next') AND NOT (//a[@href][@style]/text()='Previous') AND NOT (//td/text()='No records')	1/1	1/1

    Tempo di generazione espressioni discriminanti: 0,097 sec

    Tempo di validazione sul Training Set: 0,203 sec

  - **N° pagine Test Set: 35 (FirstPage, CentralPage, FinalPage, EmptyPage, SinglePage) : (9, 5, 10, 3, 8)**

    Tempo di valutazione sul Test Set: 7,145 sec

    Tempo di validazione sul Test Set: 0,562 sec ***ESITO POSITIVO !!!***

Conclusioni:

- L'utilizzo del tool è utile

1. Quando le espressioni discriminanti sono difficilmente deducibili.
2. Per esperimenti di classificazione molto complessi.

- L'applicazione offre vantaggi rispetto ad una procedura manuale per due motivi:

1. Dai risultati mostrati durante la validazione l' esperto intuisce facilmente come e dove agire per evitare un eventuale nuovo fallimento.
2. In caso di errata validazione può generare nuove features osservando le sole pagine fallite.

- La fase più onerosa del processo è la valutazione delle features sulle pagine.
- L'esperto generalmente impiega molto tempo per popolare l'insieme di training.