UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

tesi di laurea

Perfezionamento e sperimentazione di uno strumento per la classificazione automatica di pagine Web

Anno Accademico 2008/09

relatore
Ch.mo prof. Porfirio Tramontana

candidato
Ferdinando Celentano
Matr. 534/2848

UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

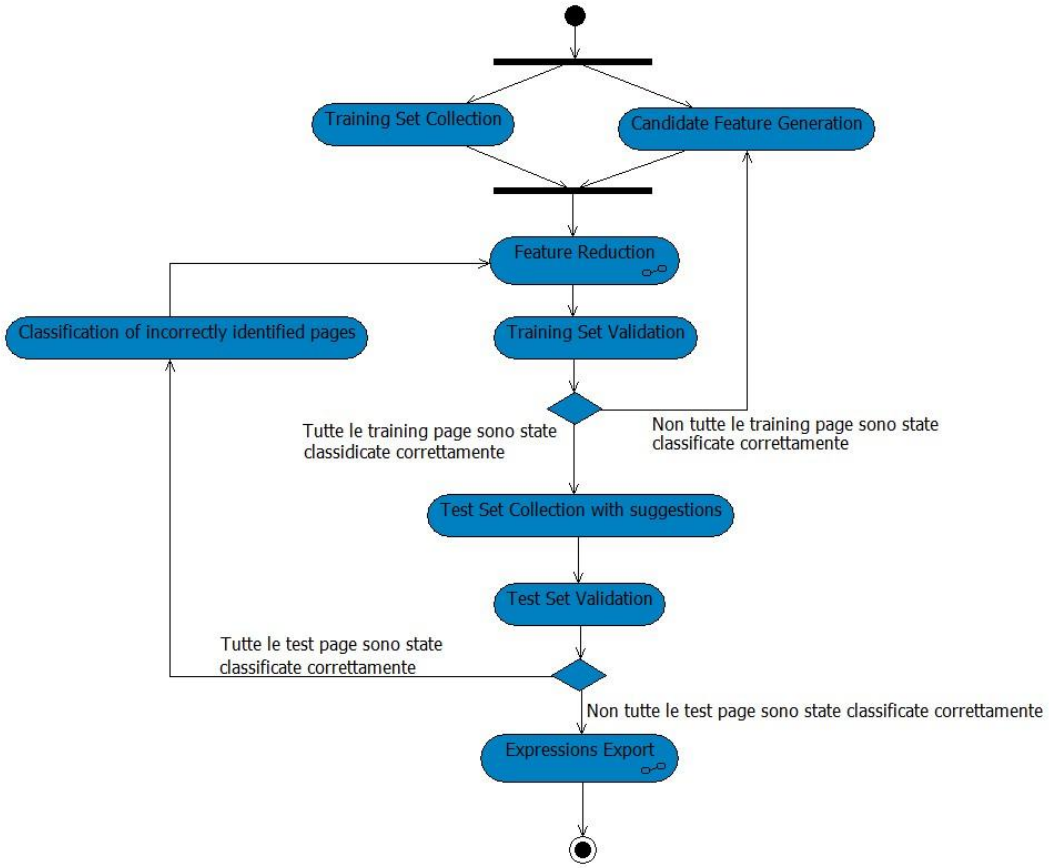
Scopo del progetto:

- **Perfezionamento del tool Page Classifier, che supporta la classificazione semiautomatica di pagine web.**
- **Validazione del funzionamento del tool e dimostrazione del suo funzionamento.**
- **Sperimentazione su problematiche reali.**

Applicazioni:

- **Realizzare una nuova interfaccia, incapsulando quella originale.**
- **Generare casi di test e in particolare di asserzioni.**

Diagramma delle attività del processo di classificazione



Training Set Collection
Consiste nel salvare, le pagine Web, che popoleranno l'insieme Training Set.

Fase manuale

Le due fasi sono svolte in contemporanea.

Infatti mentre si naviga sul browser integrato si può sia salvare la pagina e assegnargli la classe sia selezionare un elemento della stessa per generare una feature.

Feature Generation
Consiste nel proporre le caratteristiche discriminanti per le classi d'equivalenza.

Fase semiautomatica

Le feature, generate in forma di stringa Xpath, vengono valutate, una ad una, sulle pagine del Training Set.

Feature Evaluation
Consiste nel valutare le feature sulle pagine del Training Set.

Fase automatica

Concept Lattice Generation
La Concept Analysis viene adoperata per analizzare le feature che caratterizzano le diverse classi d'equivalenza delle pagine.

Fase automatica

In base alla valutazione delle feature viene generato il Concept lattice.

Discriminating Expressions Generation
Consiste nel proporre le caratteristiche discriminanti per le classi d'equivalenza.

Fase automatica

Dal lattice dei concetti è possibile risalire al tipo di ogni feature:

- ***Specific**: la feature è soddisfatta da tutte le pagine della classe C.*
- ***Relevant**: la feature è soddisfatta da tutte le pagine della classe C ma anche da altre pagine di classi diverse.*
- ***CSPC(**Conditionally **Specific**):* ***la feature è soddisfatta da un sottoinsieme*** **di pagine della classe C e da nessun'altra delle altre classi.**
- ***Shared**: la feature è soddisfatta da un sottoinsieme di pagine della classe C ma anche da altre pagine di classi diverse.*
- ***Irrelevant**: la feature non è soddisfatta da nessuna pagina della classe C*

Sulla base del tipo delle feature vengono generate le espressioni discriminanti secondo regole ben precise.

Training Set Validation
Verifica per ciascuna classe se l'espressione discriminante ad essa associata riesca ad attribuire correttamente tutte le sue pagine alla classe stessa.

Fase automatica

Se per una classe le metriche recall e precision sono pari ad 1 l'espressione discriminante è corretta, relativamente al Training Set.

$$recall = \frac{\#PagineCorrettamenteAttribuiteAllaClasseC}{\#PagineAppartenentiAllaClasseC}$$

$$precision = \frac{\#PagineCorrettamenteAttribuiteAllaClasseC}{\#PagineAttribuiteAllaClasseC}$$

Se non tutte le espressioni discriminanti sono corrette si torna alla Feature Generation.

Se tutte le espressioni discriminanti sono corrette si può passare alla raccolta di un insieme più grande di pagine su cui testare le espressioni, per valutare l'efficacia delle stesse.

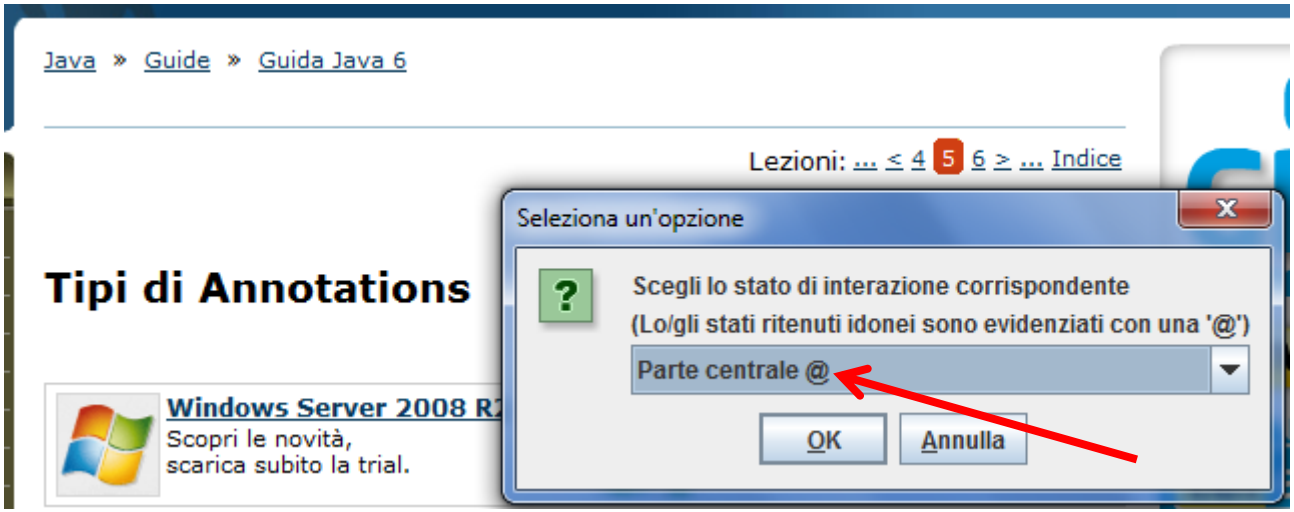
Quando si salva una pagina, si riceve un suggerimento sulla classe di appartenenza.

Se non tutti i suggerimenti sono giusti si aggiungeranno le pagine fallite al Training Test e si tornerà alla Feature Evaluation. Diversamente si passa alla Test Set Validation.

Test Set Collection
Consiste nel salvare, le pagine Web, che popoleranno l'insieme Test Set.

Fase semiautomatica

Tornando sul sistema di suggerimento si ha che la classe d'equivalenza considerata esatta viene evidenziata con una "@" come si può vedere in figura:



Test Set Collection
Consiste nel salvare, le pagine Web, che popoleranno l'insieme Test Set.

Fase semiautomatica

Test Set Validation
Convalida le espressioni discriminanti.

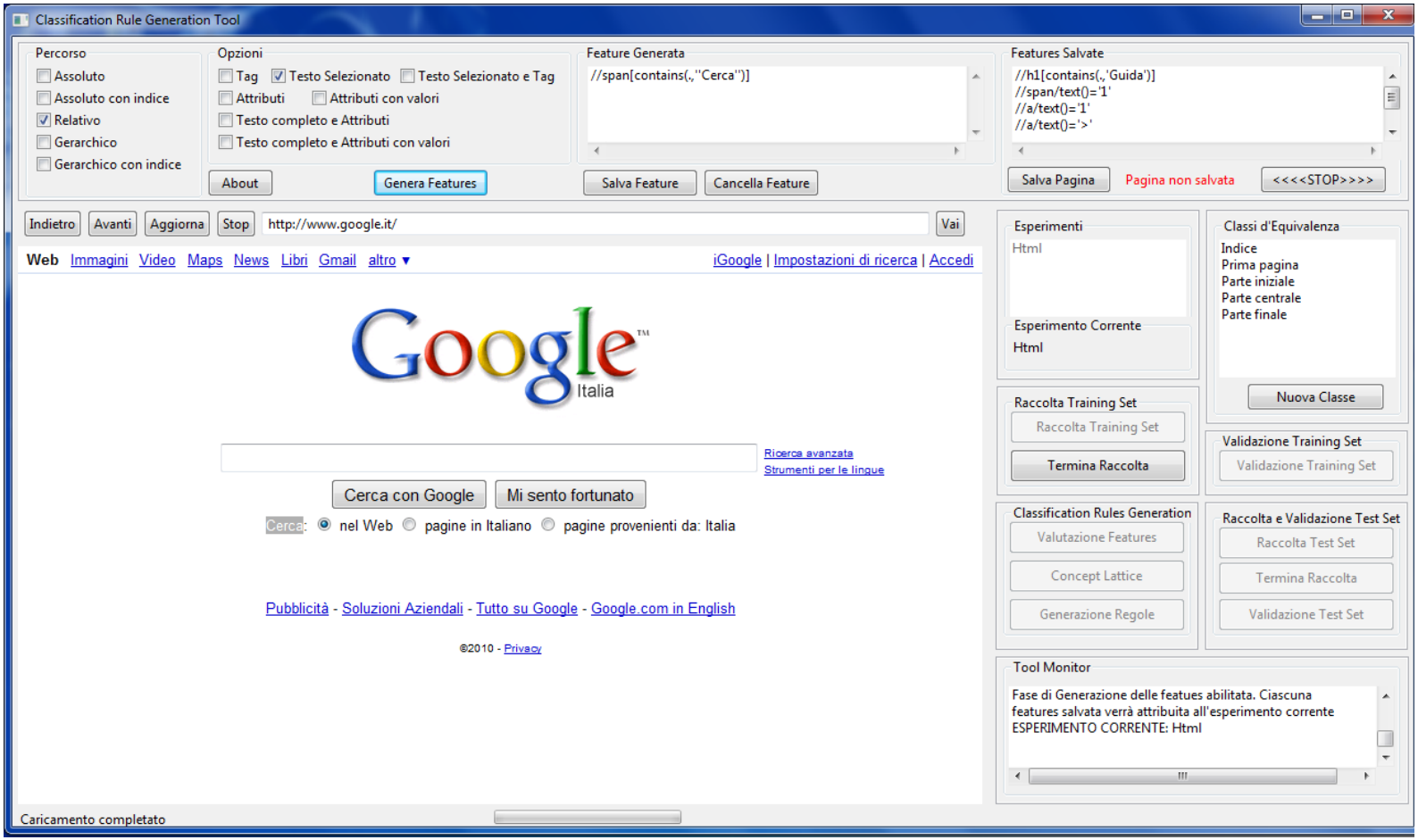
Fase automatica

Una volta convalidate le espressioni si ha la possibilità di esportarle sia in forma XPath che in forma di Test Case.

Expressions Export
Esporta le espressioni discriminanti per un uso esterno al tool.

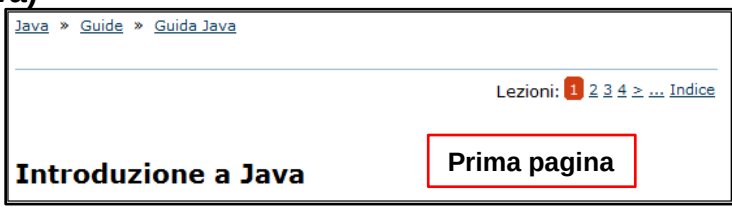
Fase automatica

Interfaccia del Tool



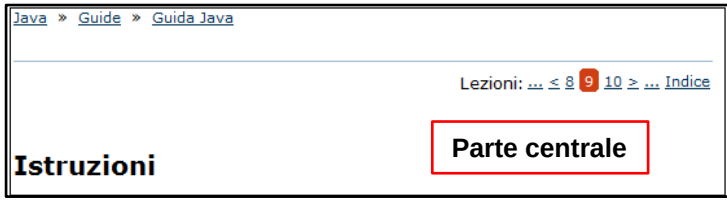
Esperimento Applicazione Html.it

- Numero di classi: 5 (Indice, Prima pagina, Parte iniziale, Parte centrale, Parte finale)
- Numero di Training page: 5 (una per classe, collezionate da una guida su java)
- Feature proposte: 5



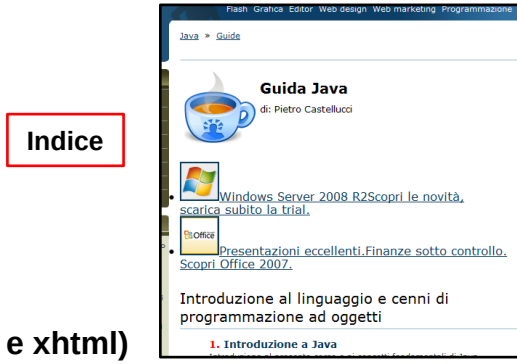
- Tempo di valutazione delle 5 feature sulle 5 pagine: 23.56 secondi
- (tempo medio per valutare 1 feature su 1 pagina: 0,9424 secondi)

- Tempo di generazione del concept lattice: 0.104 secondi
- Tempo di generazione delle espressioni discriminanti: 0.16 secondi



- Espressioni discriminanti:

Cluster	Espressione discriminante
Indice	//h1[contains(.,'Guida')]
Prima pagina	//span/text()='1'
Parte iniziale	//a/text()='1'
Parte centrale	//a/text()='<' and //a/text()='>'
Parte finale	//a/text()='<' and not //a/text()='>'



- Tempo di validazione sul Training Set: 0.3 secondi
- Numero di Test page: 25 (5 per classe, collezionate da cinque guide su php, xml, css, ajax e xhtml)

- Tempo di valutazione delle 5 espressioni sulle 25 pagine: 85.067 secondi
- (tempo medio per valutare 5 espressioni su 1 pagina: 3,40268 secondi)

- Espressioni esportate con successo

Esperimento Applicazione Bed&Breakfast

- Numero di classi: 8 (Home Page, Camere, Servizi, Prezzi, Offerte, Prenotazioni, Dove siamo, Contatti)

- Vengono raccolte 2 pagine e proposte 2 feature per ciascuna classe, per distinguere le classi a prescindere dalle due localizzazioni dell'applicazione (italiano e inglese)

- Per verificare la correttezza dell'intero processo ho commesso di proposito degli errori:

  - Fase 1: mancano le feature per 2 classi (Home Page e Camere) e le training page per 1 classe (Servizi); di conseguenza è fallita la Training Set Validation, per la mancanza delle feature

  - Fase 2: essendo tornato alla Feature Generation ed avendo proposto le feature mancanti, la Training Set Validation è andata a buon fine. E' fallita però la Test Set Validation, perché non sono state riconosciute le pagine della classe Servizi, in quanto non è stata generata l'espressione per tale classe, dato che manca la relativa training page.

  - Fase 3: le pagine fallite della classe Servizi sono state aggiunte automaticamente al Training Set. A questo punto, seguendo nuovamente il processo di classificazione sono state generate correttamente le espressioni discriminanti per tutte le classi.

Fase	# Training Page	# Feature	Tempo medio valutazione feature	Tempo medio di validazione del Test Set
1	14	12	0.235 secondi	/
2	14	16	0.138 secondi	6.215 secondi (senza successo)
3	16	16	0.131 secondi	6.056 secondi

Conclusioni:

Le fasi più onerose del processo sono quelle, che per forza di cose, sono manuali: le due fasi di raccolta delle pagine.

Tuttavia il vantaggio nell'utilizzo del tool è notevole per vari motivi:

- **Il tool consente di salvare le pagine e assegnarvi la classe d'equivalenza con pochi click.**
- **In caso di errata validazione può generare nuove feature basandosi solo sulle pagine fallite.**
- **La raccolta del Test Set è adiuvata dal sistema di suggerimento.**
- **Il guadagno di tempo nelle fasi semiautomatiche e automatiche è notevole e rende applicabile il processo anche a esperimenti più complessi.**