

tesi di laurea

# **Applicazione software per la classificazione, guidata da query Xpath, di pagine equivalenti ai fini del testing**

Anno Accademico 2007/2008

## **relatore**

Ch.mo prof. Anna Rita Fasolino

## **correlatore**

Ch.mo prof. Porfirio Tramontana

## **candidato**

Roberto Licciardi

Matr. 534/2091

## Contesto applicativo e problematica

- ✓ La maggior parte delle applicazioni web esistenti sono di tipo *dinamico* e presentano una User Interface (UI) implementata da pagine HTML che possono essere di tipo:
  - *Statico*, il cui contenuto è costante;
  - *Dinamico*, ovvero delle *Built Client Page* (BCP) il cui contenuto è definito a run-time dalla Server page e sulla base dell'input dell'utente e dello stato in cui si trova l'applicazione stessa.
  
- ✓ In molte problematiche è necessario poter classificare le pagine dinamicamente costruite in gruppi semanticamente equivalenti e trovare delle espressioni che consentano di riconoscere automaticamente il gruppo cui una data BCP appartiene.

## Scopi delle classificazione di pagine web

[ La classificazione automatica di pagine web è utile per molteplici scopi legati al Reverse Engineering ma più in generale volti ad incidere nell'ambito di approcci orientati ad uno sviluppo programmatico delle applicazioni web ].

### Vantaggi dal lato utente

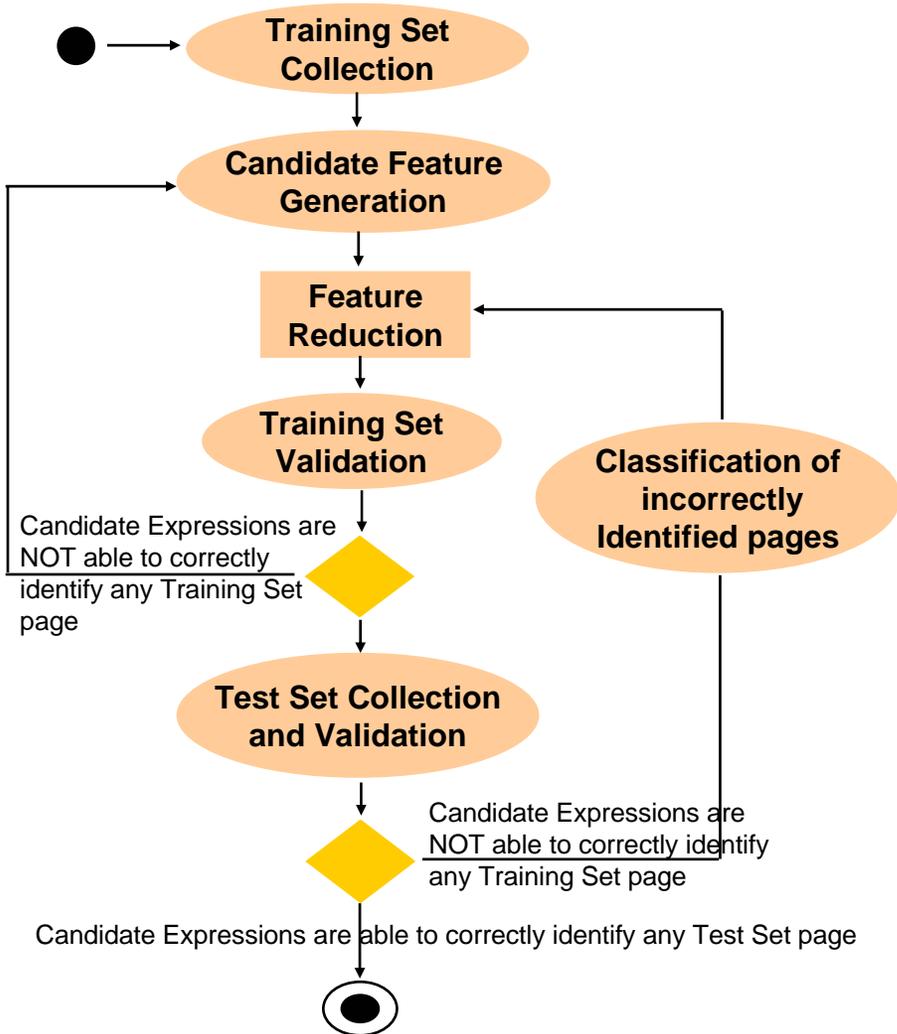
- Reperire ed accedere con più facilità alle risorse sparse nel web;
- Accedere a pagine raggruppate secondo i temi trattati o i servizi offerti.

### Vantaggi per l'ingegneria del software

- Implementare wrappers che incapsulano la UI originale della web application, al fine di esportare un'interfaccia rinnovata;
- Mettere in atto processi orientati all'ottenimento di un modello della UI con lo scopo di re-ingegnerizzarlo secondo le diverse architetture (Model-Driven) o tecnologie (AJAX).
- Fornire un supporto nel campo del testing.

Tra questi vantaggi quello particolarmente interessante è l'implementazione di wrappers in modo da consentire la migrazione di applicazioni web di tipo *legacy* verso un paradigma Service Oriented Architecture (SOA) basato sulla tecnologia dei *Web Service*, che mediante la UI permette l'interoperabilità di applicazioni software scritte in diversi linguaggi di programmazione e implementate su diverse piattaforma hardware.

## Processo di classificazione



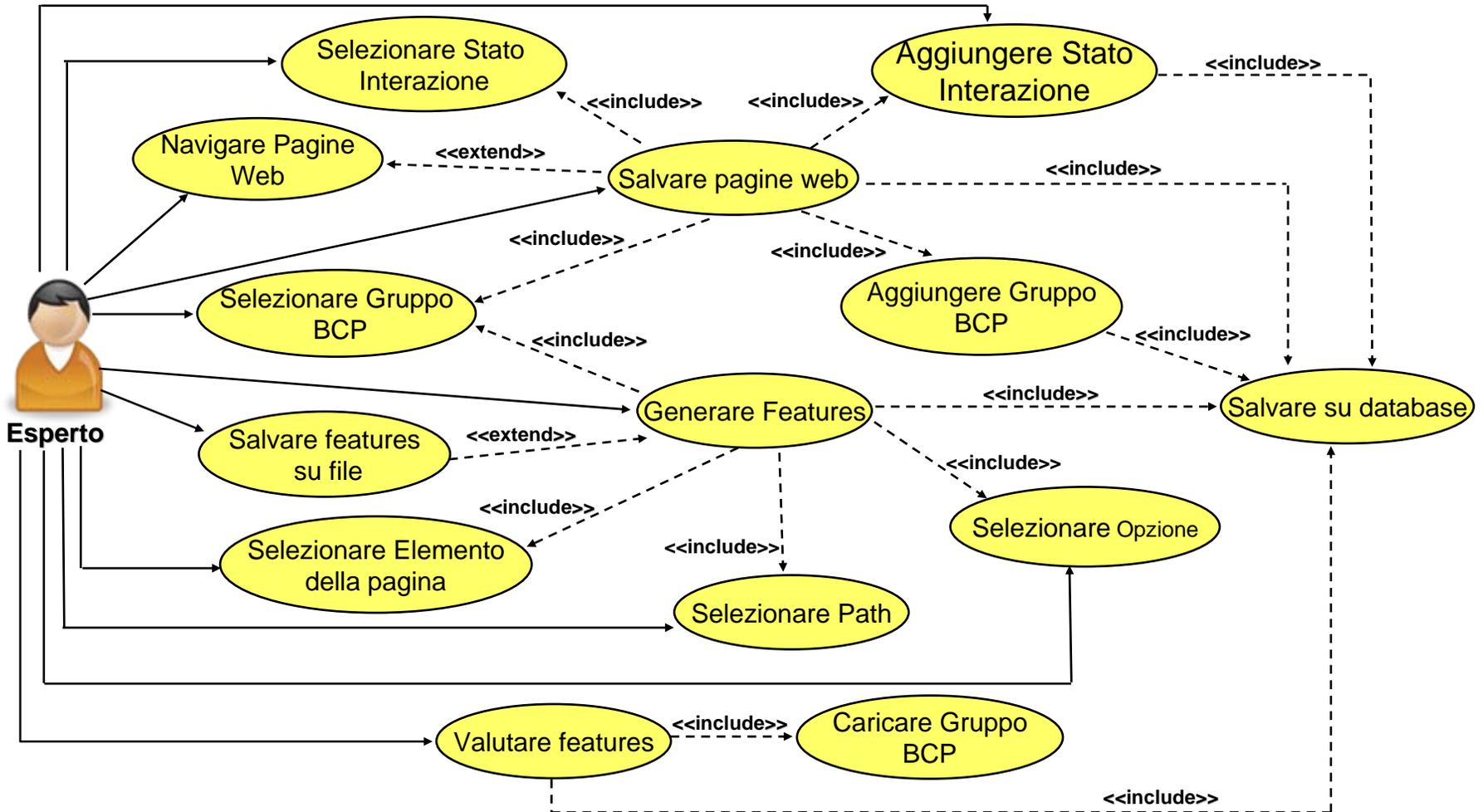
Il processo è caratterizzato da tre punti chiave:

- ❖ vi è una fase di learning del sistema;
- ❖ è iterativo;
- ❖ classifica le pagine in modo deterministico.

Una parte del processo in esame è stata realizzata mediante l'implementazione del tool denominato **Page Classifier** che in particolare ricopre le seguenti fasi:

- Training Set Collection;
- Candidate Feature Generation;
- Feature Evaluation, la prima sottofase di Feature Reduction.

# Il Tool sviluppato a supporto del processo



## Background tecnologico

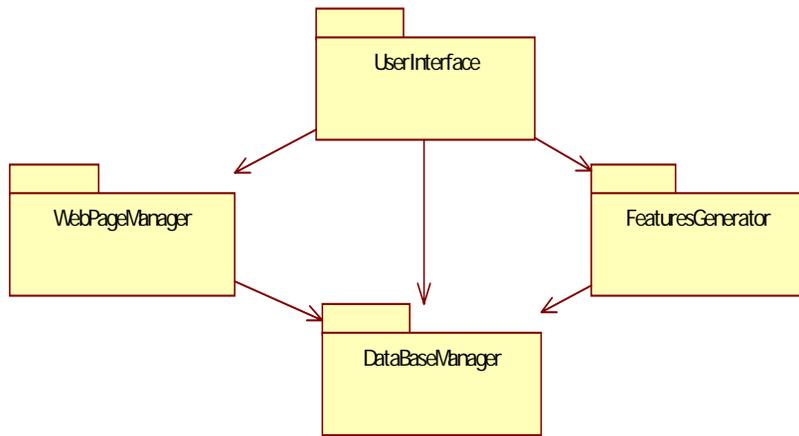
Il tool a supporto del processo è stato implementato utilizzando le seguenti tecnologie:

- ✓ **Java**: per garantire la massima portabilità su qualsiasi piattaforma che abbia installato la JVM.
- ✓ **SWT**: toolkit utilizzato per lo sviluppo della GUI. È un toolkit Heavy-Weight i quali si integrano fortemente con il sistema operativo, utilizzando i Widget che esso mette a disposizione.
- ✓ **JavaXPCOM**: consente la comunicazione tra Java e XPCOM, tale che un'applicazione Java può accedere agli oggetti XPCOM e quest'ultimo può accedere a qualsiasi classe Java che implementa un'interfaccia XPCOM.
- ✓ **XULRunner**: è un pacchetto a run-time che consente di installare, avviare, aggiornare, disinstallare applicazioni XUL + XPCOM. Fornisce delle librerie dette *libxul* che incorporano le tecnologie Mozilla attraverso le quali vengono salvate le pagine web in formato HTML.
- ✓ **HtmlUnit**: è un framework Java per il test di applicazioni web-based. Consente la creazione di un WebClient utilizzato per la valutazione delle features sulle BCP, in quanto *sincrono, meno avido di risorse di memoria e di tempo*, rispetto al browser SWT utilizzato per l'interfaccia grafica.
- ✓ **JDBC**: è un'API che mette a disposizione le funzionalità per poter lavorare con database relazionali.

### Tecnologie scartate:

- x **JTidy**: libreria per salvare pagine web e conversione in formato XHTML. Introduce modifiche alle pagine per cui il DOM che esce dal parsing XHTML non è identico a quello originale, ciò comporta che le features generate dalla pagina HTML in esecuzione nel browser possano essere diverse da quelle generate dalla pagina XHTML e di conseguenza i risultati delle valutazioni inaffidabili.

## Architettura a package del sistema

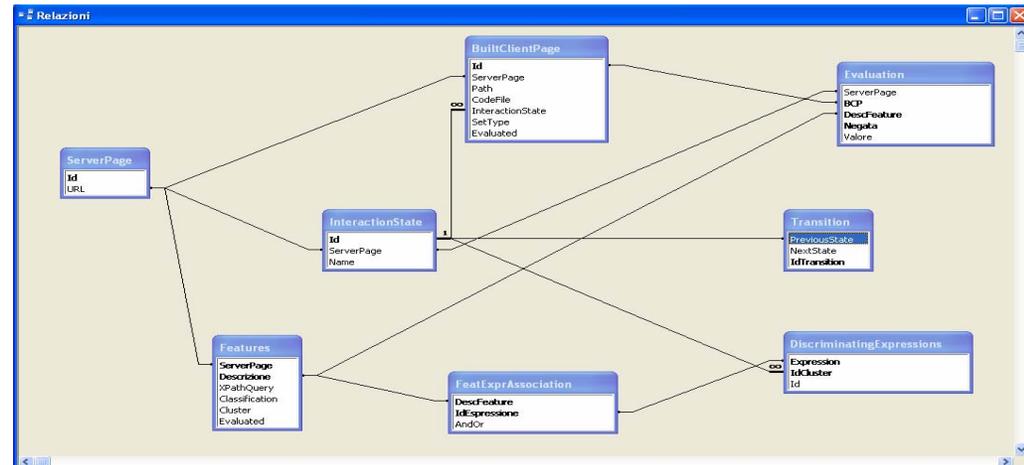


L'ossatura del sistema è rappresentata da quattro packages:

- **UserInterface**: contiene la classe Main che avvia il tool e la fase parametrica; la classe GUI che definisce l'interfaccia grafica.
- **WebPageManager**: contiene la classe GestionePaginaWeb che implementa le funzioni di business per la raccolta e ripristino delle pagine web, nonché la valutazione delle features.
- **FeaturesGenerator**: presenta la classe StringaXPath utilizzata per la generazione automatica delle features.
- **DataBaseManager**: include la classe GestioneDataBase che gestisce tutte le transazioni con il database.

## Architettura del database

Il sistema, per garantire la persistenza delle informazioni raccolte dall'esperto, è dotato di un database Access. Ma è possibile utilizzare qualsiasi sistema DBMS, poiché il tool presenta una fase di parametrizzazione iniziale che consente di specificare il path assoluto del database locale ed il relativo driver di gestione.



# Esempio d'uso del tool Page Classifier

The screenshot displays the 'Classificatore di pagine web' application window. The main interface shows a browser view of the Posteitaliane website. Several dialog boxes and tool actions are highlighted with red circles and arrows, numbered 1, 2, and 3:

- 1:** An 'Input' dialog box titled 'Aggiungi un gruppo' with the text 'posteHome' in the input field. An arrow points from this dialog to the 'Aggiungi gruppo' button in the 'Elenco Gruppi BCP' panel.
- 2:** An 'Input' dialog box titled 'Aggiungi uno stato di interazione' with the text 'iniziale' in the input field. An arrow points from this dialog to the 'Aggiungi stato' button in the 'Elenco Stati Interazione' panel.
- 3:** A red text box with the instruction 'Cliccare due volte sullo stato di interazione' and an arrow pointing to the 'Aggiungi stato' button.

Other visible elements include the 'Salva su file' button, a 'Messaggio per l'utente' dialog box with the message 'Pagina Web salvata!', and various navigation and search options in the browser interface.

Aggiunta Gruppo BCP, Stato Interazione e salvataggio pagina web

The screenshot shows the 'Classificatore di pagine web' application interface. The main window displays the URL 'http://www.poste.it/' and the page content of 'Posteitaliane'. A red oval labeled '2' highlights the 'Prodotti postali' link in the navigation menu. A red oval labeled '3' encompasses the 'Opzioni' section, which includes checkboxes for 'Tag', 'Testo selezionato', 'Tag e testo completo', 'Attributi', 'Attributi con valori', 'Attributi e testo completo', and 'Attributi con valori e testo completo'. A red oval labeled '4' highlights the 'Genera features' button. A red oval labeled '1' highlights the 'Elenco Gruppi BCP' dropdown menu, which currently shows 'posteHome'. A red oval labeled '5' highlights the 'Valuta features' button. A 'Messaggio per l'utente' dialog box is open, displaying Xpath queries and their corresponding generated features. The 'Features generate' section shows the following queries and features:

```

//a[contains(., "Prodotti postali")]
//a/text()="Prodotti postali"
//a[@title="Prodotti postali"][@href="/postali/"]
//a[@title="Prodotti postali"][@href="/postali/"]/text()="Prodott
  
```

The 'Input' dialog box prompts the user to 'Scegli un gruppo BCP sul quale valutare le features' and has 'posteHome' entered in the text field.

## Generazione e valutazione features

Output - Porting (run)

```
init:  
deps-jar:  
compile:  
run:  
Aggiornamento Features  
Fatto  
Numero di pagine web:1  
Numero di features:8  
Processo di valutazione delle features:  
Pagina caricata: 266... tempo di caricamento:2969  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 0 Risultato Xpath: false  
La feature non è stata inserita nella Evaluation poichè il risultato XPath è falso  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query
```

```
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 0 Risultato Xpath: false  
La feature non è stata inserita nella Evaluation poichè il risultato XPath è falso  
  
Singola feature caricata  
Feature già valutata? false  
  
Risutati Trovati: 1 Risultato Xpath: true  
Inserita nella Evaluation il risultato della query  
Inserita nella Evaluation il risultato negato della query  
  
TUTTE LE features del gruppo POSTERHOME sono state valutate su tutte le pagine del medesimo gruppo. ...Tempo impiegato: 4250  
  
BUILD SUCCESSFUL (total time: 1 minute 39 seconds)
```

Output della valutazione delle features e relativo aggiornamento del database



## Conclusioni

- ❑ Il tool realizzato **assiste** l'ingegnere del software in alcune fasi del processo di classificazione delle BCP generate dalla web application.
- ❑ **Supera i limiti** legati sia alle diffuse incompatibilità delle pagine web con lo standard XHTML, sia ai software per la conversione on-the-fly XHTML - HTML.
- ❑ Il processo di classificazione è stato proposto con lo scopo di definire tecniche e tools sia per consentire la **migrazione** di Web Application Legacy in Architetture Service Oriented, sia per fornire supporto nel campo del **testing**.

## Sviluppi futuri

Per completare il framework a supporto del processo è necessario sviluppare nuovi componenti software da integrare con quelli progettati nel lavoro di tesi:

- **Feature Reduction:** per creare il Concept Lattice sulla base delle features candidate e valutate; per classificare le features in: *Specific, Relevant, CSPC, Shared e Irrelevant*, per produrre le features discriminanti.
- **Training Set Validation:** tool per valutare in maniera automatica l'efficacia delle features discriminanti relative alle classi di equivalenza del Training Set.
- **Test Set Collection and Validation:** tool per validare l'efficacia dell'espressioni Xpath su un insieme di pagine più ampio.