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

tesi di laurea

Uno Strumento per la ricerca di campi Output e Etichetta in pagine Client

Anno Accademico 2006/2007

relatore
Ch.ma prof.ssa Anna Rita Fasolino

correlatore
Ch.mo prof. Porfirio Tramontana

candidato
Pasquale Giacomino
Matr. 534/151

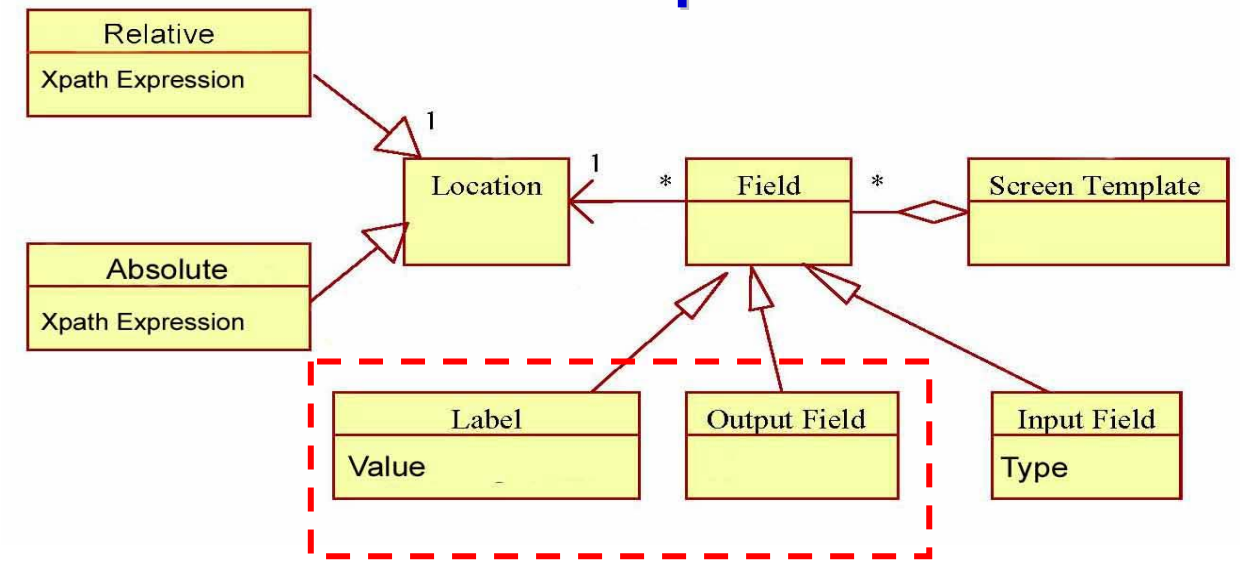
Il contesto

- Reverse Engineering di Applicazioni Web “dinamiche”.
- Estrazioni di informazioni strutturate da pagine client (che per loro natura sono ben poco strutturate).

Il problema

- Analisi delle pagine Html generate dinamicamente ai fini della generazione di un template comune per classi di pagine Equivalenti.
- Per riuscire ad associare un gruppo di pagine equivalenti uno stesso Screen Template, bisogna prima di tutto ottenere una descrizione della schermata di ciascuna pagina del gruppo e successivamente mediante tecniche euristiche di identificazione,estrarre il Template comune.
- La descrizione della schermata è caratterizzata da un insieme di campi di testo che possono essere campi di **input**, campi di **output** e campi di **etichetta** e dalla loro posizione sulla schermata che può essere **assoluta** o **relativa.**

Il Template



Obiettivo: Realizzare un'applicazione che estragga, per un gruppo di pagine equivalenti:

- i field di output con diverso contenuto informativo e la stessa posizione (esprimibile come XpathQuery)
- i field di tipo label con lo stesso contenuto informativo e la stessa posizione (esprimibile come XpathQuery)

Soluzione proposta

Strumenti e tecnologie utilizzati

- **JAVA**

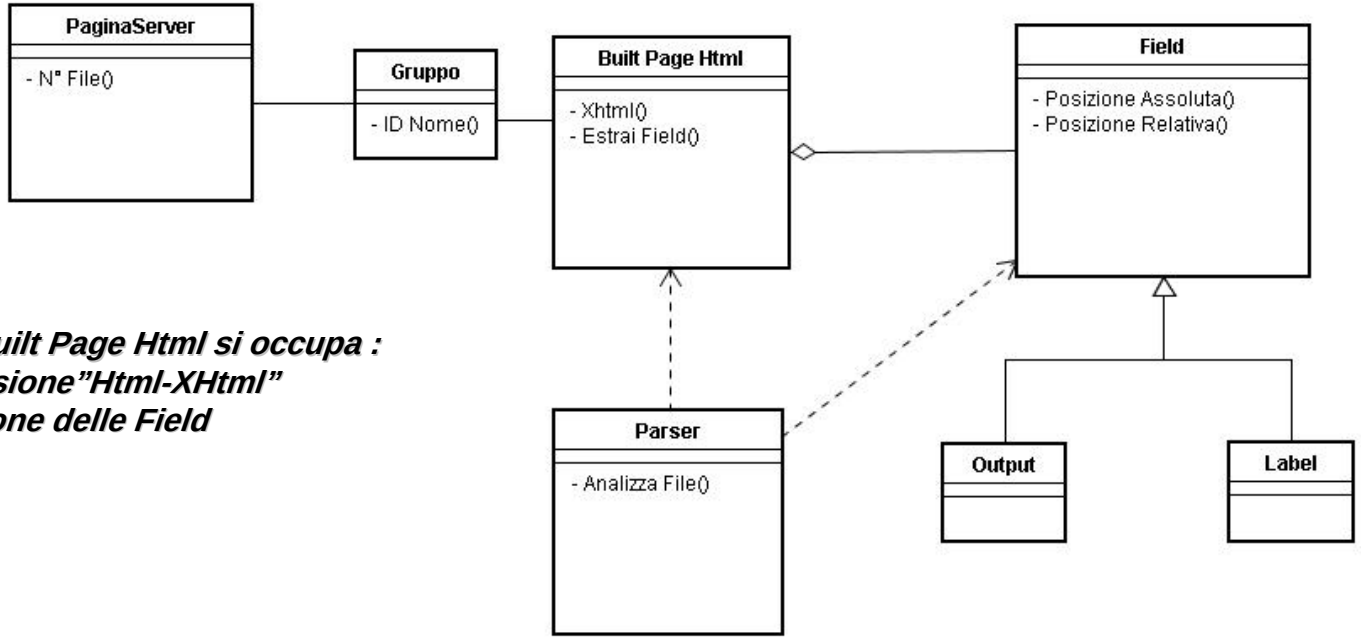
- **XML**
- **XPATH**
- **HTTPUNIT**

- **JTIDY**

L' applicazione realizza le seguenti funzionalità:

- Ricerca campi di output o etichetta
- Popola un DB
- Genera un file di output esportabile

Diagramma di Progetto



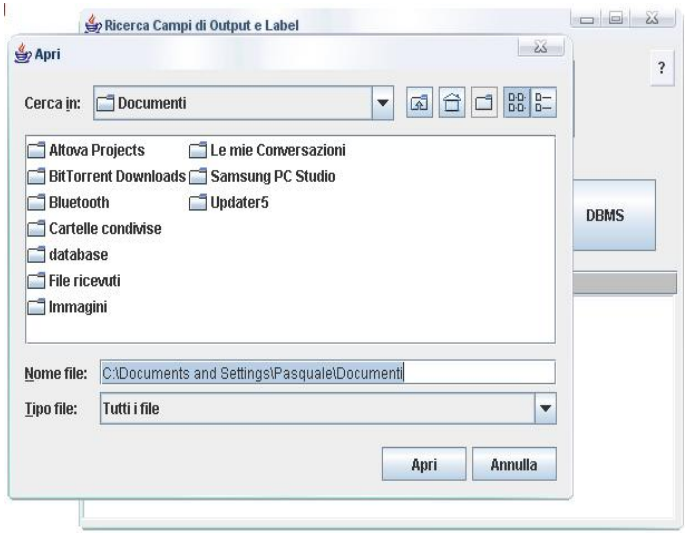
- ***La classe Built Page Html si occupa :***
  - ***conversione"Html-XHtml"***
  - ***estrazione delle Field***

- ***La Classe Parser si occupa:***
  - ***confronto field estratte***
  - ***ricava field di output***
  - ***ricava field di tipo label***

- ***Classe Field si occupa:***
  - ***genera posizione XpathQuery relativa***
  - ***genera posizione XpathQuery assoluta***

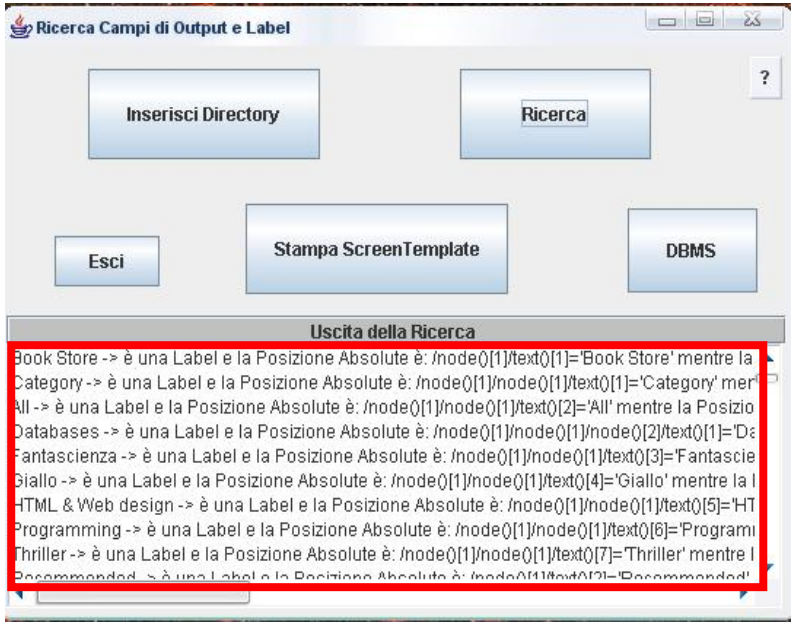
Ricerca campi di Output e Etichetta

- **Prima di effettuare la ricerca bisogna, inserire un gruppo di file che abbiano caratteristiche equivalenti.**



- **Il tool realizzato non ricerca i campi di input perché sono lavoro di una precedente tesi, a differenza dei campi di output e etichetta non presentano particolari problemi di estrazione.**

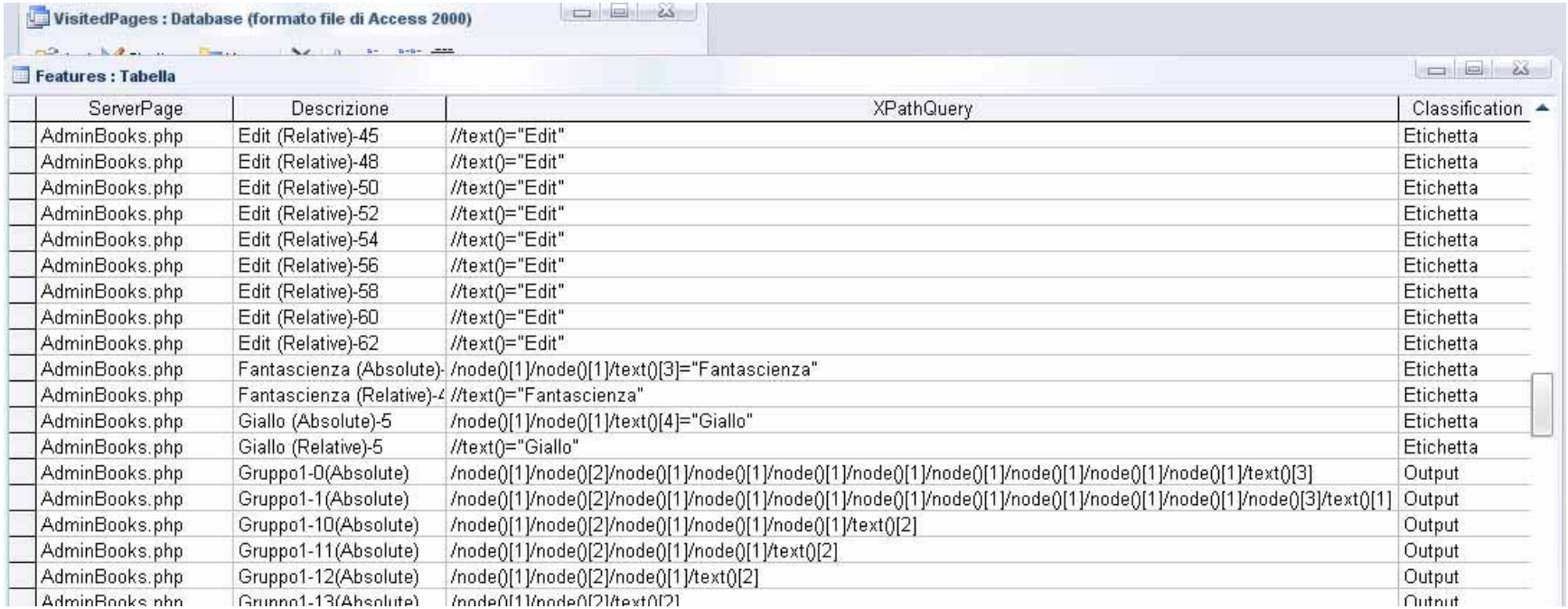
- **I risultati vengono mostrati nell' apposito riquadro.**



Genera File e Popola Dbms

- **Una volta effettuata la ricerca dei campi, posso scegliere di generare lo Screen Template e di popolare il DB.**
- **La stampa dello ScrenTemplate produrrà la creazione di un file xml.**

- **Mentre per il popolamento del DB, il tool progettato andrà a popolare la tabella Features del database "VisitedPages", il database in questione serve a mantenere informazioni sulle pagine visitate.**


VisitedPages : Database (formato file di Access 2000)

Features : Tabella

ServerPage	Descrizione	XPathQuery	Classification
AdminBooks.php	Edit (Relative)-45	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-48	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-50	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-52	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-54	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-56	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-58	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-60	//text()="Edit"	Etichetta
AdminBooks.php	Edit (Relative)-62	//text()="Edit"	Etichetta
AdminBooks.php	Fantascienza (Absolute)-	/node()[1]/node()[1]/text()[3]="Fantascienza"	Etichetta
AdminBooks.php	Fantascienza (Relative)-4	//text()="Fantascienza"	Etichetta
AdminBooks.php	Giallo (Absolute)-5	/node()[1]/node()[1]/text()[4]="Giallo"	Etichetta
AdminBooks.php	Giallo (Relative)-5	//text()="Giallo"	Etichetta
AdminBooks.php	Gruppo1-0(Absolute)	/node()[1]/node()[2]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/text()[3]	Output
AdminBooks.php	Gruppo1-1(Absolute)	/node()[1]/node()[2]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[1]/node()[3]/text()[1]	Output
AdminBooks.php	Gruppo1-10(Absolute)	/node()[1]/node()[2]/node()[1]/node()[1]/node()[1]/text()[2]	Output
AdminBooks.php	Gruppo1-11(Absolute)	/node()[1]/node()[2]/node()[1]/node()[1]/text()[2]	Output
AdminBooks.php	Gruppo1-12(Absolute)	/node()[1]/node()[2]/node()[1]/text()[2]	Output
AdminBooks.php	Gruppo1-13(Absolute)	/node()[1]/node()[2]/text()[2]	Output

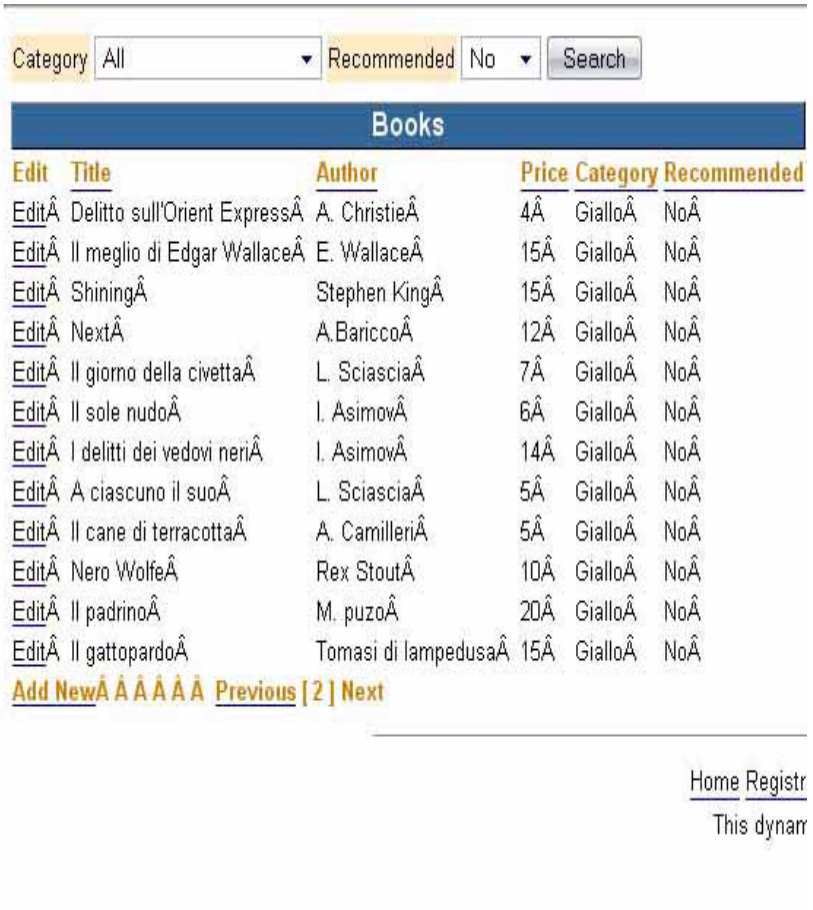
La Sperimentazione 1° Fase

Nell'esperimento sono stati valutati due distinti obiettivi:

- *se i field di output e label proposti dal tool sono correttamente riconosciuti*
- *se i field di output e label proposti dal tool sono in grado di formare espressioni discriminanti*

L'esempio considerato ha coinvolto un'applicazione Web open source chiamata Online Bookstore. Il processo di validazione è stato eseguito su un sottoinsieme di queste pagine.

BCP (pagine client costruite)	Risultati ottenuti
First Page Books	Il tool riesce ad individuare tutti i campi di Output, mentre per le Label: il tag con la scritta "Search" non riesce ad individuarlo, in quanto la scritta "Search" è un attributo del tag; il tag "EditA" lo considera un'etichetta invece è un collegamento ipertestuale, questo errore è dovuto alla conversione Html in XHtml.
Central Page Books	Il tool individua tutti i campi di output e label.
Last Page Books	Il tool riesce ad individuare tutti i campi di output e etichetta fino a che ha un riscontro da tutte le pagine analizzate, nei rimanenti campi i risultati ottenuti possono essere sbagliati o non esserci proprio. Per migliorare il tool realizzato, si potrebbe aggiungere un ulteriore controllo su i file che hanno un numero di campi testo diversi.
Single Page Books	Lo stesso del Last Page Books.
Empty Books List	Il tool individua solo le label, in quanto i campi di output non sono presenti.



La Sperimentazione 2° Fase

La seconda parte della sperimentazione si occuperà di vedere se le features trovate dal tool, sono discriminanti cioè una singola feature deve essere vera in tutte le pagine del gruppo e falsa altrimenti . Per verificare se i campi di output o label trovati sono veramente discriminanti ho utilizzato il Concept Analyzer.

La Candidatura delle features

Nell'ambito della sperimentazione effettuata su un *Training Set* di BCP (pagine client costruite) per ogni classe di equivalenza è proposta una Espressione Discriminante (cioè una combinazione logica di espressioni riguardanti la presenza di un field) e due indici di prestazione quali *Recall* e *Precision*.

BCP	Espressione Discriminante	Recall	Precision
Central Page Books List	Previous (Relative)-72	5/5	5/15
Empty Books List	No records (Relative)-20	5/5	5/5
First Page Books List	Previous (Relative)-72	5/5	5/15
Last Page Books List	Previous (Relative)-72	5/5	5/15
Single Page Books List	!Previous (Relative)-72 AND !No records (Relative)-20	5/5	5/5

RECALL = (n° di BCPs rilevanti ritornate) / (n° totale di BCPs rilevanti nella EBCP)

PRECISION = (n° di BCPs rilevanti ritornate) / (n° totale di BCPs restituite dalla Discriminant Expression)

Conclusioni e sviluppi futuri

Il tool realizzato ci permette di identificare tutti i campi di output e etichetta, in un gruppo di file appartenenti allo stesso scenario di esecuzione e di ottenere infine un Modello di Schermata. Inoltre il tool permette di popolare una tabella di un database, che ci servirà in un secondo momento per verificare se i campi trovati sono discriminanti.

L'applicazione implementata costituisce un tassello fondamentale del progetto di cui essa fa parte che è la realizzazione del Wrapper, il componente che permette la migrazione da Web Application a Web Service.

Per quel che riguarda gli sviluppi futuri, il suddetto tool potrà essere impiegato in diversi ambiti:

- ***sviluppare ulteriori metodi di ricerca, dei campi di output e di etichetta, che controlli le singole parti della stringa di testo del tag analizzato, inoltre si potrebbero aggiungere all'analisi del tag attributi e aggregati.***
- ***di trovare in automatico la query xpath, che individua i campi di output e di etichetta nel file xml, che sia una via di mezzo tra la query xpath relativa e assoluta già trovata nel tool realizzato.***