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

tesi di laurea

Realizzazione di Web Service per l'estrazione di informazioni da siti web enciclopedici

Anno Accademico 2008/2009

relatore
Ch.mo prof. Porfirio Tramontana
Ch.mo prof. Annarita Fasolino

candidato
Cirillo Giovanni
Matr. 534/2881

Cos'è un Web service?

Secondo la definizione data dal World Wide Web Consortium (W3C) un Web Service (servizio web) è un sistema software progettato per supportare l'interoperabilità tra diversi elaboratori su di una medesima rete; caratteristica fondamentale di un Web Service è quella di offrire un'interfaccia software (descritta in un formato automaticamente elaborabile quale, ad esempio, il Web Services Description Language) utilizzando la quale altri sistemi possono interagire con il Web Service stesso attivando le operazioni descritte nell'interfaccia tramite appositi "messaggi" inclusi in una "busta" (la più famosa è SOAP): tali messaggi sono, solitamente, trasportati tramite il protocollo HTTP e formattati secondo lo standard XML.

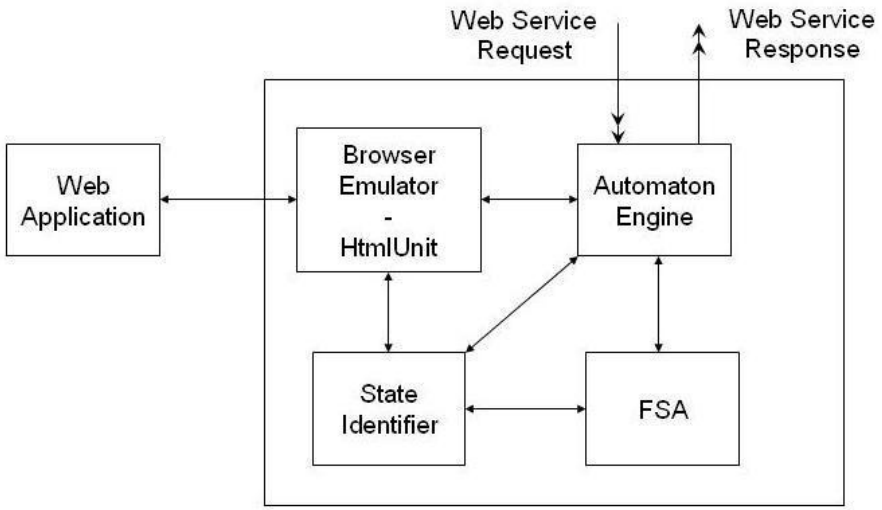
- ***Quali sono i vantaggi? E gli svantaggi?***
- ***Perché creare un Web Service?***
- ***Quale tecnologie sono utilizzate nel Web Service?***

- ***Perché trasformare Web application in Web Service?***

XML and XPATH

WRAPPER

- Lo scopo del Wrapper è quello di interagire con l'applicazione web per ottenere dalla sua esecuzione le funzionalità che devono essere esportate come web service.
- Il wrapper diviso in due componenti principali: un interprete e un automa che deve essere interpretato.



I passi fondamentali da seguire per la costruzione di un Wrapper saranno:

- Selezionare la funzionalità di interesse che si vuole migrare verso web service;
- Effettuare un reverse enginerring dell'interfaccia utente della Web Application;
- Creare il Modello di Iterazione descritto tramite l'FSA;
- Convalidare il Wrapper;
- Esportare la funzionalità creata verso web service.

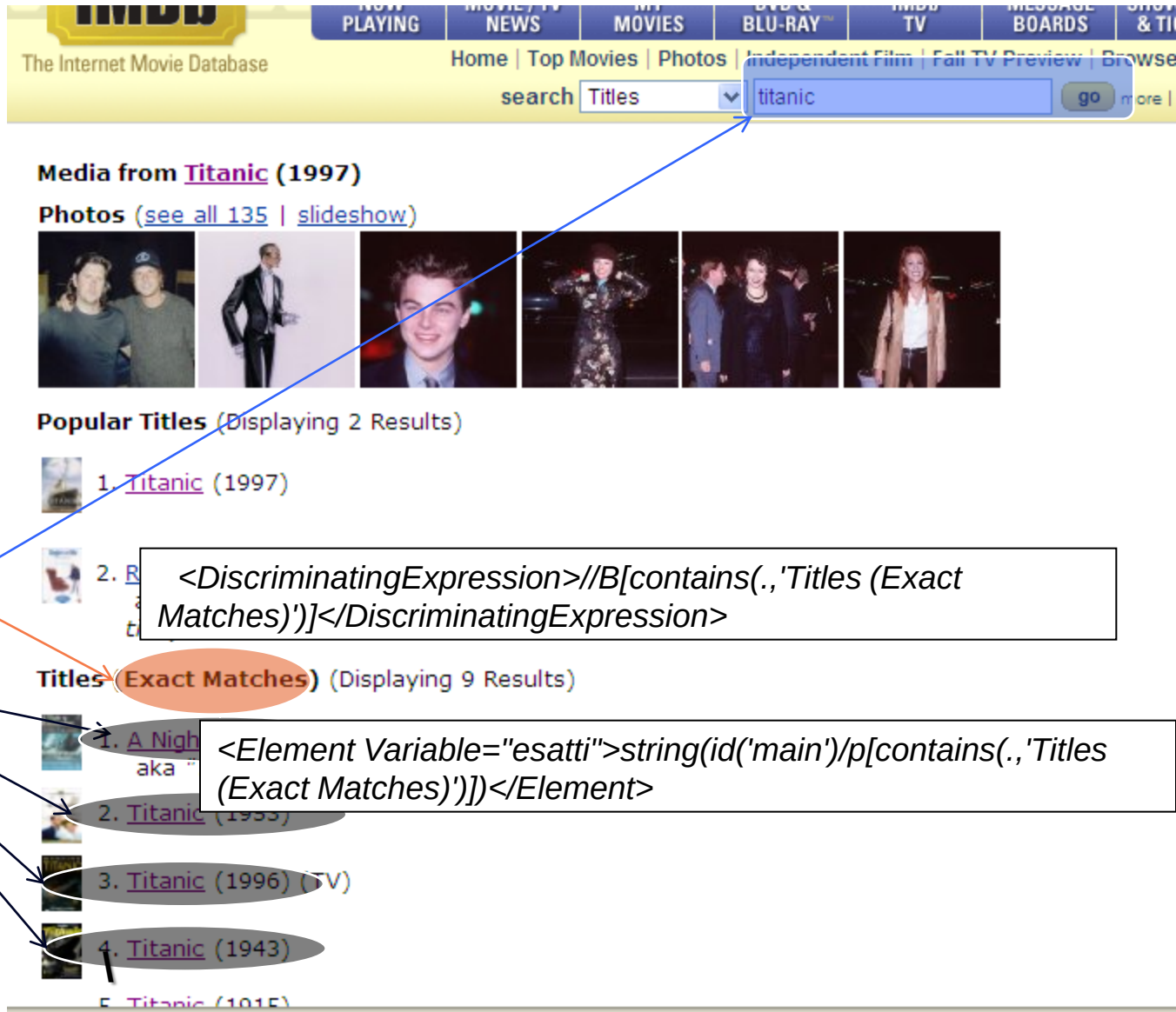
UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II
Facoltà di Ingegneria
Corso di Studi in Ingegneria Informatica

Struttura dell'automa

L'automa è costituito da sequenza di stati con:

- Nome
- Numero identificativo
- Discriminante
- Stati successivi
- Variabili di I/O

```
<Input Type="Click">
   <Element>//INPUT[@type="image" and @title="go"]</Element>
 </Input>
```



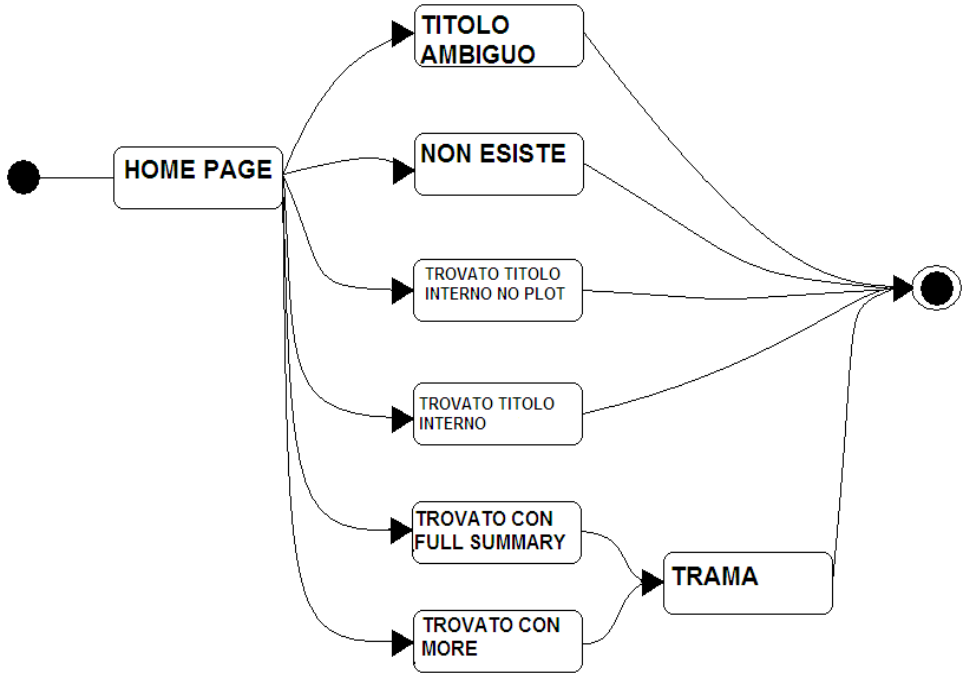

```
<DiscriminatingExpression>//B[contains(.,'Titles (Exact Matches)')]</DiscriminatingExpression>
```

```
<Element Variable="esatti">string(id('main')/p[contains(.,'Titles (Exact Matches)')])</Element>
```

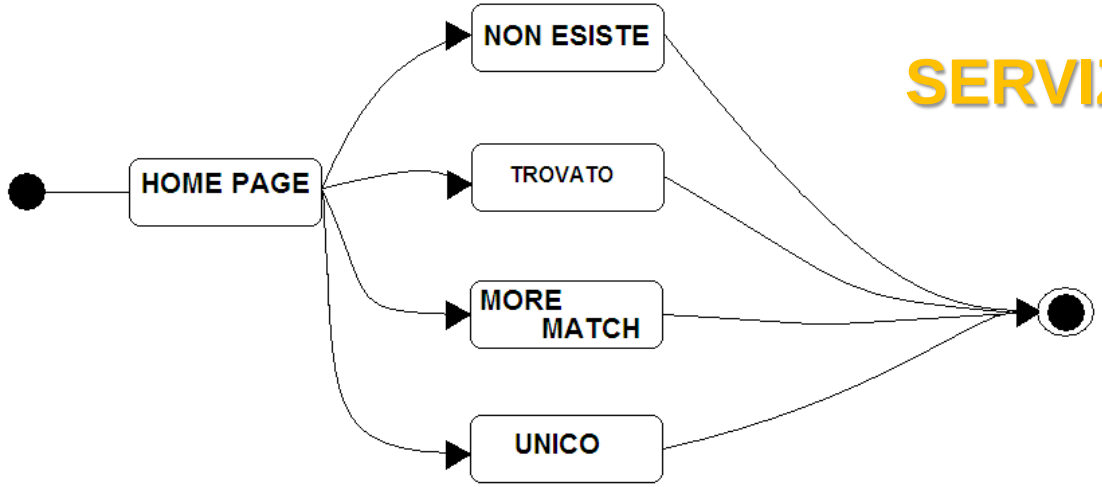
SERVIZIO IMDB AUTOMA UNO

Il sito Internet Movie Database è un database online di informazioni su film, attori, registi, DVD, programmi televisivi, spot pubblicitari e videogiochi.

❖**Il nostro scopo è quello di: dato in input un titolo di un film, restituirne in output il direttore (regista), il genere del film (romantico, commedia, horror etc..) e il plot (trama del film).**



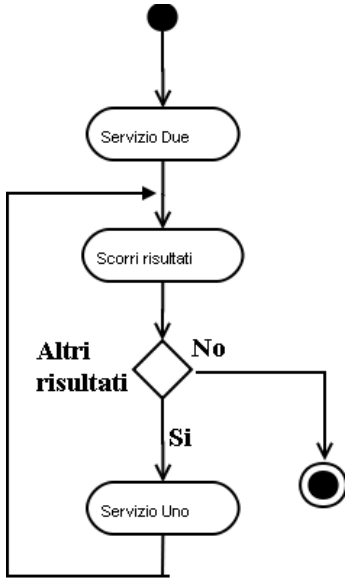
Precondizioni	**Presente nel database**
Input	*Film da cercare*
Output	*Direttore, Genere e Trama del film*
Descrizione	*L'automa interroga l'applicazione che ne restituisce direttore genere e trama del film cercato.*



SERVIZIO IMDB AUTOMA DUE

Il risultato finale sarà un automa del tipo:

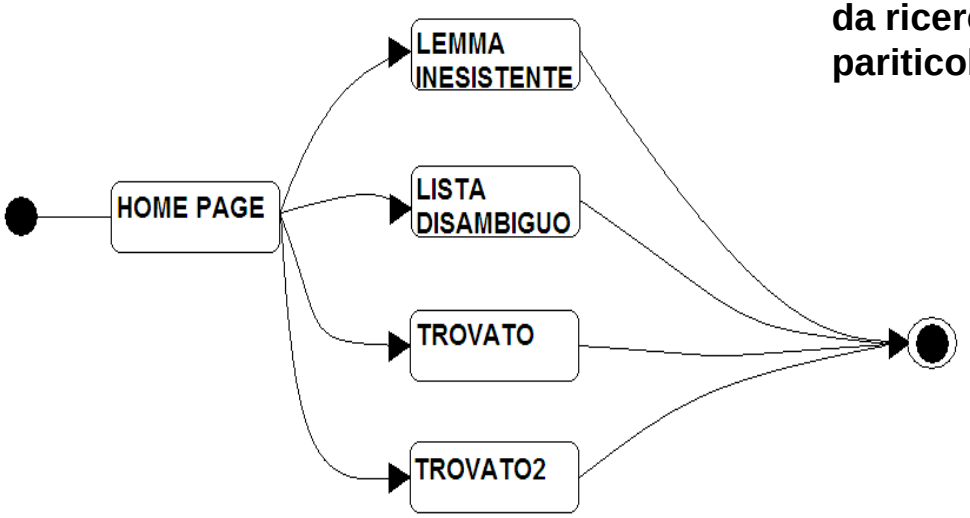
Precondizioni	**Presente nel database**
Input	*Lemma da cercare*
Output	*Lista dei film*
Descrizione	*L'automa interroga l'applicazione che ne restituisce la lista di tutti i film che hanno come titolo il nome esatto del film cercato.*



SERVIZIO WIKIPEDIA AUTOMA UNO

Wikipedia è un'enciclopedia online, multilingue, a contenuto libero, redatta in modo collaborativo da volontari e sostenuta dalla Wikimedia Foundation, un'organizzazione senza fine di lucro.

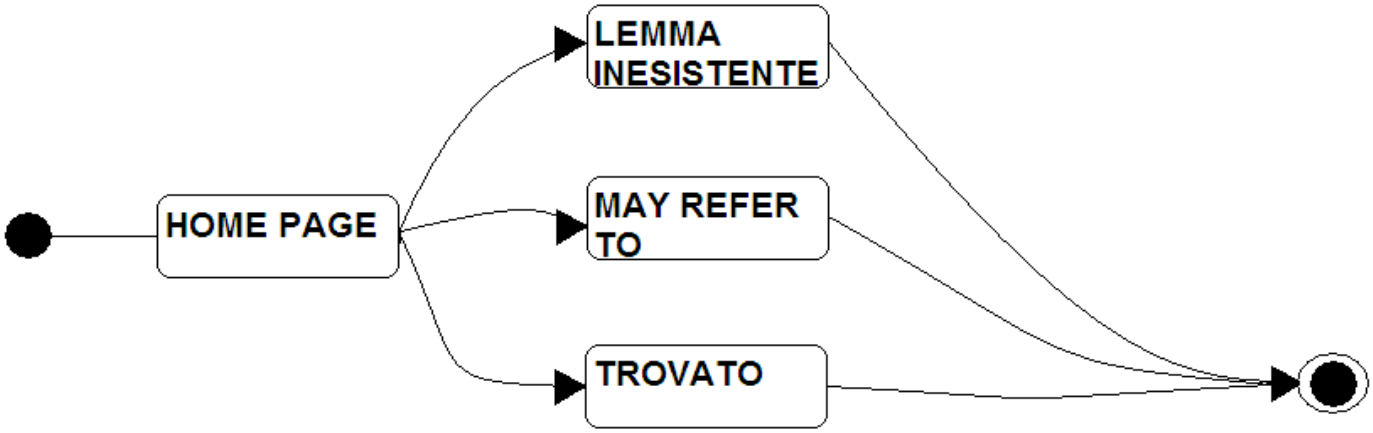
❖**Lo scopo di questo servizio è: dato in input un lemma da ricercare, ci restituisca il risultato della ricerca, in pariticolar modo l'incipit.**

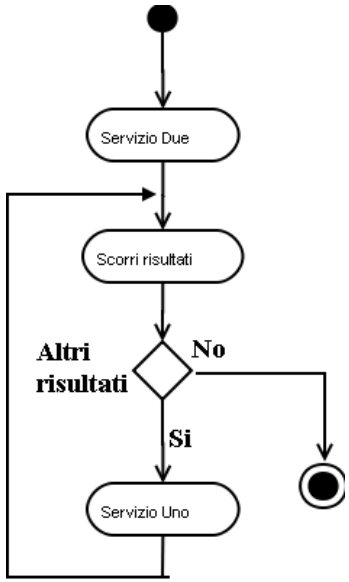


Precondizioni	**Presente nel database**
Input	*Lemma da cercare*
Output	*Incipit della pagina*
Descrizione	*L'automa interroga l'applicazione che ne restituisce l'incipit della pagina con il suo relativo URL*

SERVIZIO WIKIPEDIA AUTOMA DUE

Precondizioni	**Presente nel database**
Input	*Lemma da cercare*
Output	*Lista delle disambiguazioni*
Descrizione	*L'automa interroga l'applicazione che ne restituisce la lista delle disambiguazioni o, se unico, il risultato del lemma cercato.*





SPERIMENTAZIONE

Affidabilità E Performance

Analizziamo l'affidabilità e le prestazione di ogni servizio composto dai due automi:

➢**Nel caso di Wikipedia:**
In base ai risultati ottenuti, su un campione di 528 ricerche, possiamo dire che la percentuale di successo è del 97,7%. In totale sul calcolo dei tempi di 2640 ricerche la media per l'estrazione delle informazioni è stato di 3976,563 ms, con una deviazione standard dei valori di 954 ms.

➢**Nel caso di IMDB:**
la percentuale di successo è stata ancora maggiore 98,41% dei casi su un campione di 563 ricerche.
La media, fatta su un calcolo di 2815 ricerche, è 7021,825ms con una deviazione standard di 1581,073ms.

Lemma	Prova 1	Prova 2	Prova 3	Prova 4	Prova 5
English	5078	4202	7249	7906	6770
Colonial	4547	4453	4485	4844	4250
North America	4359	5578	6218	5922	6000
Settlement	3813	3422	3427	3391	3343
location	3312	3218	3375	3578	3157
Captain John Smith	4172	4218	3323	3828	3266

Lemma	Prova 1	Prova 2	Prova 3	Prova 4	Prova 5
Titanic	5563	9063	9062	5781	6406
Batman	8875	5875	9016	9625	6235
Superman	6562	5937	6078	6141	8718
UP	5719	5610	6062	5782	5656
Blow	9938	5656	5875	5890	9765
Dark	6406	6890	6281	5907	6047

CONCLUSIONI E SVILUPPI

- In questo lavoro di Tesi è stata presentata una metodologia per l'estrazione di informazioni da siti enciclopedici attraverso un Web Service:
  - ✓ L'estrazione avviene tramite un Wrapper che, in maniera trasparente all'utente, interagisce con le Web Application;
  - ✓Abbiamo creato due tipi di servizi complessi a loro volta formati da servizi più semplici in modo da rendere modulabile la nostra applicazione.
- Eventuali sviluppi degli automi possono portare a servizi diversi:
  - ✓ Raccogliere informazioni diverse o di diverso tipo;
  - ✓ Implementazioni di automi per raccogliere le informazioni da altre fonti.