

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).

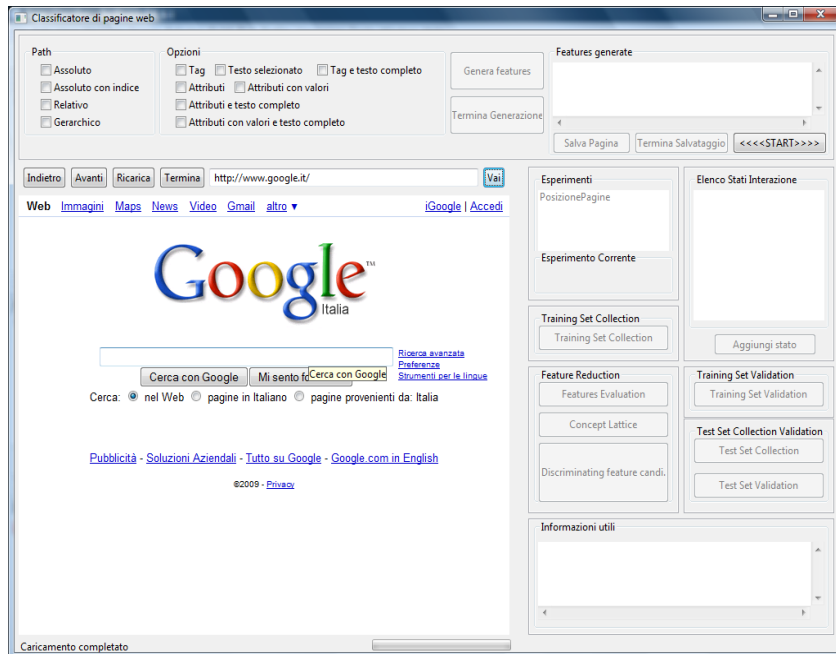
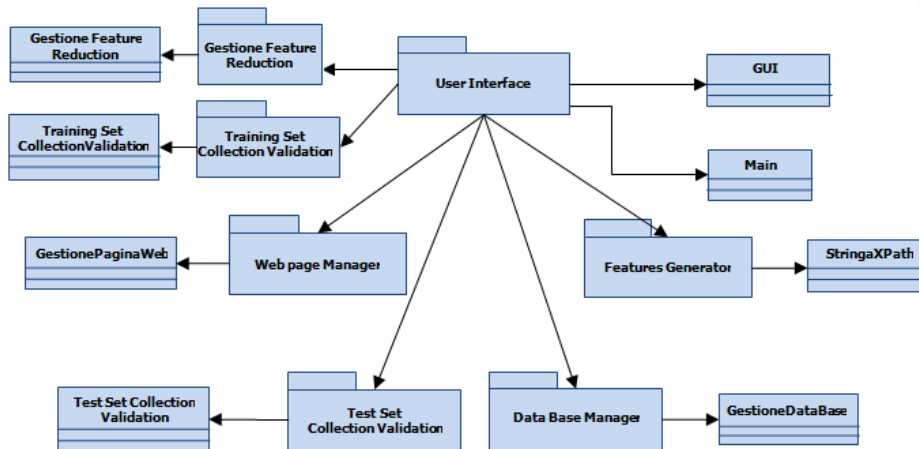
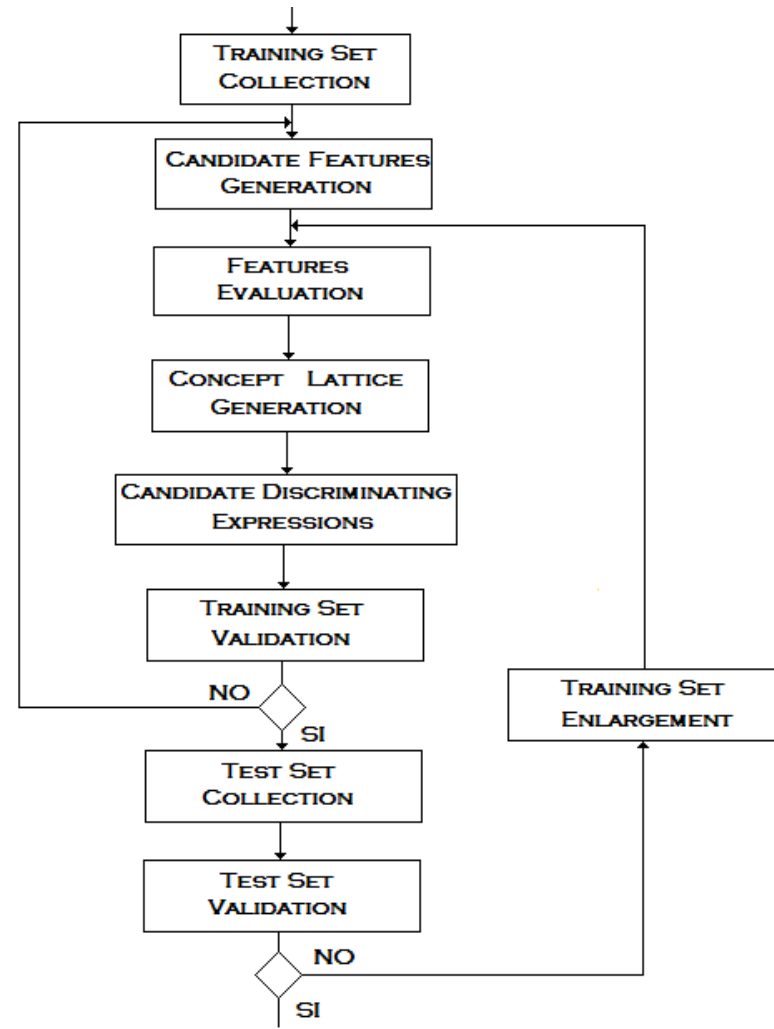
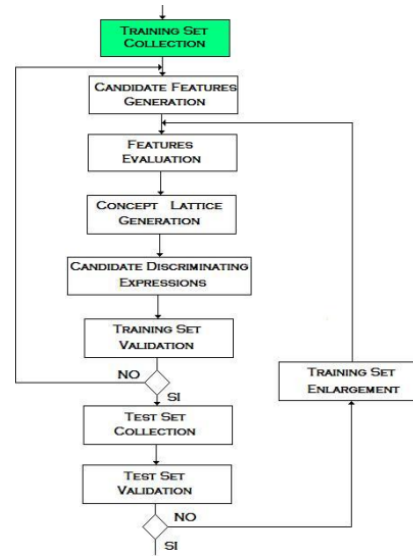


Diagramma delle attività del: **PROCESSO DI CLASSIFICAZIONE DELLE PAGINE WEB**



Single Page:



TRAINING SET COLLECTION:

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

- Fase completamente manuale

Category: Giallo Recommended: No Search

Books					
Edit	Title	Author	Price	Category	Recommended
Edit	10 piccoli indiani	Agatha Christie	15	Giallo	No
Edit	Sipario Nero	C. Woolrich	10	Giallo	No
Edit	Il caso Saint facre	G. Simonon	5	Giallo	No
Edit	I delitti della rue Morgue	E. A. Poe	2	Giallo	No
Edit	Delitto sull'Orient Express	A. Christie	4	Giallo	No
Edit	Il meglio di Edgar Wallace	E. Wallace	15	Giallo	No
Edit	Shining	Stephen King	15	Giallo	No
Edit	Io uccido	G. Faletti	12	Giallo	No
Edit	Il giorno della civetta	L. Sciascia	7	Giallo	No
Edit	Il sole nudo	I. Asimov	6	Giallo	No
Edit	I delitti dei vedovi neri	I. Asimov	14	Giallo	No
Edit	A ciascuno il suo	L. Sciascia	5	Giallo	No
Edit	Il cane di terracotta	A. Camilleri	5	Giallo	No
Edit	Nero Wolfe	Rex Stout	10	Giallo	No
Edit	Il padrino	M. puzo	20	Giallo	No
Edit	Il gattopardo	Tomasi di lampedusa	15	Giallo	No

Add New

Central Page:

Category: All Recommended: All Search

Books					
Edit	Title	Author	Price	Category	Recommended
Edit	Io uccido	G. Faletti	12	Giallo	No
Edit	Delitto sull'Orient Express	A. Christie	4	Giallo	No
Edit	Sipario Nero	C. Woolrich	10	Giallo	No
Edit	HTML 4 for the World Wide Web Visual Quickstart Guide	Elizabeth Castro	15.99	HTML & Web design	No
Edit	Oracle8i Web Development	Bradley D. Brown, Brad Brown	41.99	Databases	No
Edit	MySQL (OTHER NEW RIDERS)	Paul DuBois	39.99	Databases	No
Edit	Beginning Active Server Pages 3.0	David Buser, Chris Ullman, Jon Duckett?	31.99	Programming	No
Edit	Mastering ColdFusion 4.5	Arman Danesh, Kristin Aileen Motlagh, Kristin Motlagh	39.99	Programming	No
Edit	Il gattopardo	Tomasi di lampedusa	15	Giallo	No
Edit	Il cane di terracotta	A. Camilleri	5	Giallo	No
Edit	PHP and MySQL Web Development	Luke Welling, Laura Thomson	39.99	Programming	No
Edit	Il sole nudo	I. Asimov	6	Giallo	No
Edit	Shining	Stephen King	15	Giallo	No
Edit	Poirot a Styles Court	Agatha Christie	-5	Giallo	Yes
Edit	Il Ladro di merendine	Andrea Camilleri	10	Giallo	Yes
Edit	Beginning ASP Databases	John Kaufman, Thearon Willis, David Buser, Kevin Spencer, kauffman, John Kauffman	39.99	Databases	Yes
Edit	Web Database Development : Step by Step	Jim Buysens	39.99	Databases	Yes
Edit	MySQL and mSQL	Randy Jay Yarger, George Reese, Tim King	27.98	Databases	Yes
Edit	MySQL & PHP From Scratch	Wade Maxfield	23.99	Programming	Yes
Edit	Il cane di terracotta	Andrea Camilleri	10	Giallo	Yes

Add New Previous [2] Next

First Page:

Category: All Recommended: All Search

Books					
Edit	Title	Author	Price	Category	Recommended
Edit	Poirot a Styles Court	Agatha Christie	-5	Giallo	Yes
Edit	I delitti della rue Morgue	E. A. Poe	2	Giallo	No
Edit	Delitto sull'Orient Express	A. Christie	4	Giallo	No
Edit	A ciascuno il suo	L. Sciascia	5	Giallo	No
Edit	Il caso Saint facre	G. Simonon	5	Giallo	No
Edit	Il cane di terracotta	A. Camilleri	5	Giallo	No
Edit	Il sole nudo	I. Asimov	6	Giallo	No
Edit	Il giorno della civetta	L. Sciascia	7	Giallo	No
Edit	Il Ladro di merendine	Andrea Camilleri	10	Giallo	Yes
Edit	Nero Wolfe	Rex Stout	10	Giallo	No
Edit	Sipario Nero	C. Woolrich	10	Giallo	No
Edit	Il cane di terracotta	Andrea Camilleri	10	Giallo	Yes
Edit	Io uccido	G. Faletti	12	Giallo	No
Edit	I delitti dei vedovi neri	I. Asimov	14	Giallo	No
Edit	Il meglio di Edgar Wallace	E. Wallace	15	Giallo	No
Edit	10 piccoli indiani	Agatha Christie	15	Giallo	No
Edit	Il gattopardo	Tomasi di lampedusa	15	Giallo	No
Edit	Shining	Stephen King	15	Giallo	No
Edit	Perl and CGI for the World Wide Web: Visual QuickStart Guide	Elizabeth Castro	15.19	Programming	No
Edit	HTML 4 for the World Wide Web Visual Quickstart Guide	Elizabeth Castro	15.99	HTML & Web design	No

Add New Previous [1] Next

Empty Page:

Category: Fantascienza Recommended: No Search

Books					
Edit	Title	Author	Price	Category	Recommended
No records					

Add New

Final Page:

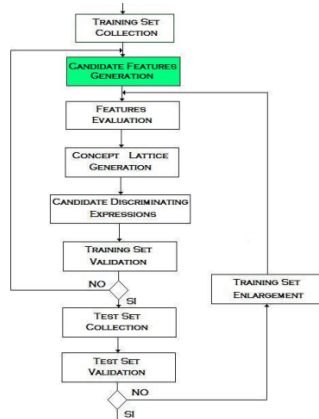
Category: All Recommended: All Search

Books					
Edit	Title	Author	Price	Category	Recommended
Edit	Black Belt Web Programming Methods: Servers, Security, Databases and Sites		30	Programming	Yes

Add New Previous [3] Next

CANDIDATE FEATURES GENERATION:

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



Fase semiautomatica

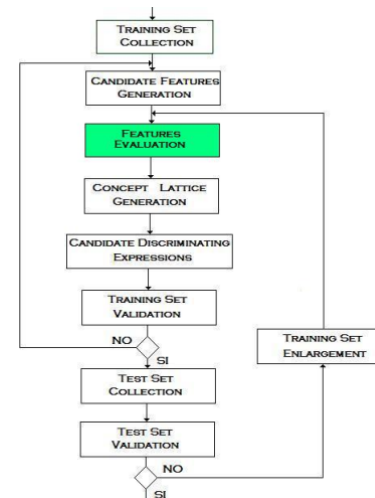
FEATURES EVALUATION:

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

Pagine
Training

Features candidate

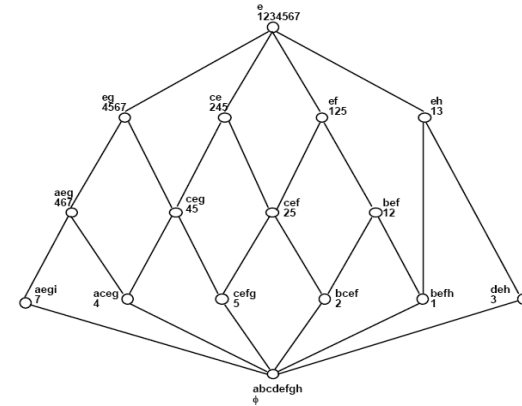
	F1	F2	F3	F4	F5
P1	X				X
P2		X		X	X
P3	X				X
P4		X			X
P5	X		X		X
P6	X		X		X
P7	X		X		X



Fase automatica

CONCEPT LATTICE:

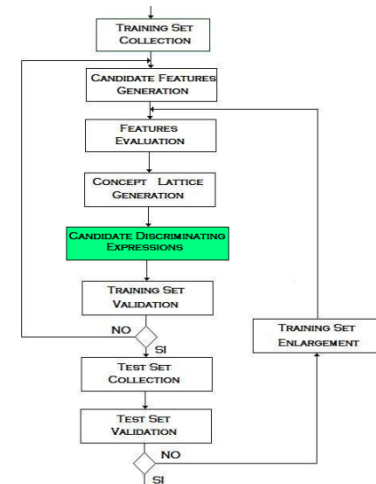
- La Concept Analysis viene adoperata per analizzare le feature che caratterizzano le diverse classi di equivalenza delle pagine.



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

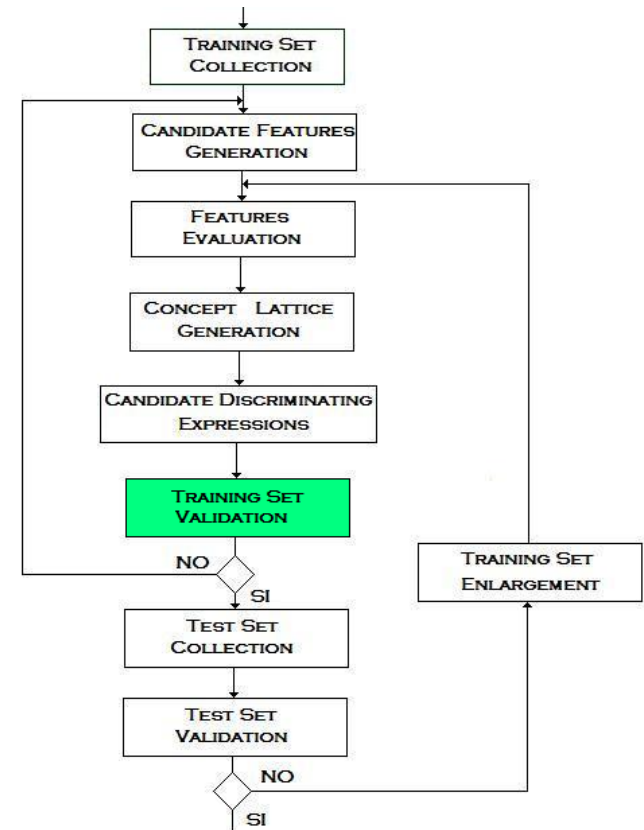
■ TRAINING SET VALIDATION

- ❑ 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: $\frac{\# \text{Pagine correttamente attribuite alla classe C}}{\# \text{Pagine che soddisfano l'espressione Expr}(C)}$

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



TEST SET COLLECTION:

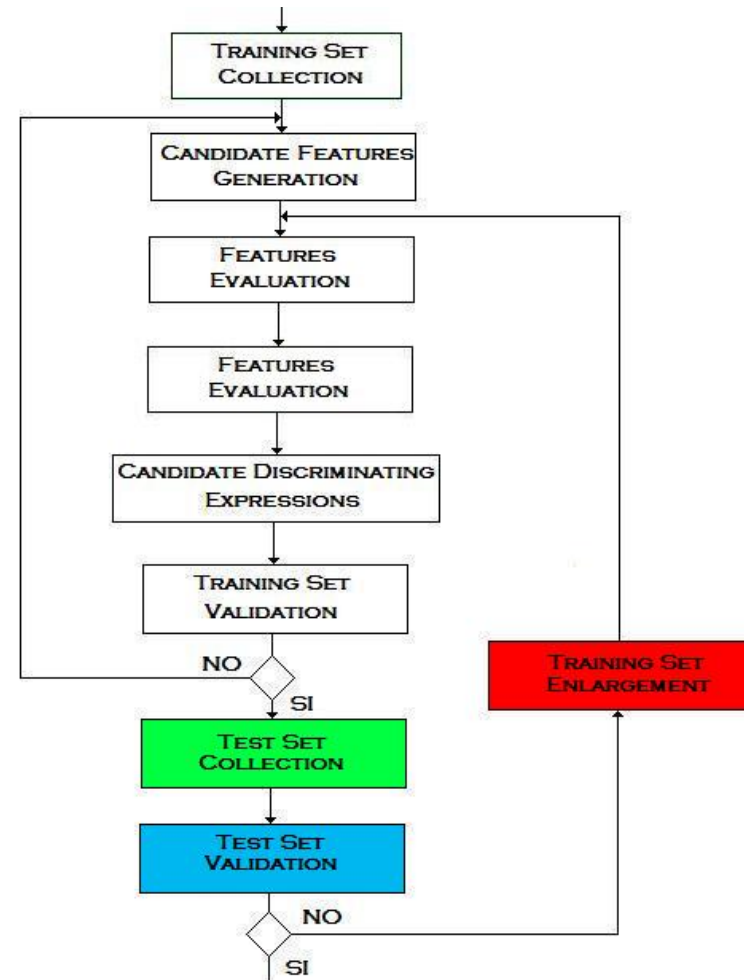
- Tale fase provvede a popolare l'insieme, detto di test, sul quale verrà valutata l'efficacia delle espressioni discriminanti.
- L'insieme potrebbe essere collezionato da una persona diversa dall'esperto.

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	<code>(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')</code>	2/2	2/2
EmptyPage	<code>NOT(//a[@href][@style]/text()='Next') AND NOT(//a[@href][@style]/text()='Previous')</code>	1/1	1/2
FinalPage	<code>NOT(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')</code>	1/1	1/1
FirstPage	<code>(//a[@href][@style]/text()='Next') AND NOT(//a[@href][@style]/text()='Previous')</code>	2/2	2/2
SinglePage	<code>NOT(//a[@href][@style]/text()='Next') AND NOT(//a[@href][@style]/text()='Previous')</code>	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	<code>(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')</code>	2/2	2/2
EmptyPage	<code>//td/text()='No records'</code>	1/1	1/1
FinalPage	<code>NOT(//a[@href][@style]/text()='Next') AND (//a[@href][@style]/text()='Previous')</code>	1/1	1/1
FirstPage	<code>(//a[@href][@style]/text()='Next') AND NOT (//a[@href][@style]/text()='Previous')</code>	2/2	2/2
SinglePage	<code>NOT(//a[@href][@style]/text()='Next') AND NOT (//a[@href][@style]/text()='Previous') AND NOT (//td/text()='No records')</code>	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.**