Web Pages Classification using Concept Analysis

Porfirio Tramontana

Anna Rita Fasolino

Dipartimento di Informatica e Sistemistica
University of Naples Federico II, Italy

Giuseppe A. Di Lucca

RCOST – Research Centre on Software Technology

University of Sannio, Benevento, Italy

The Context

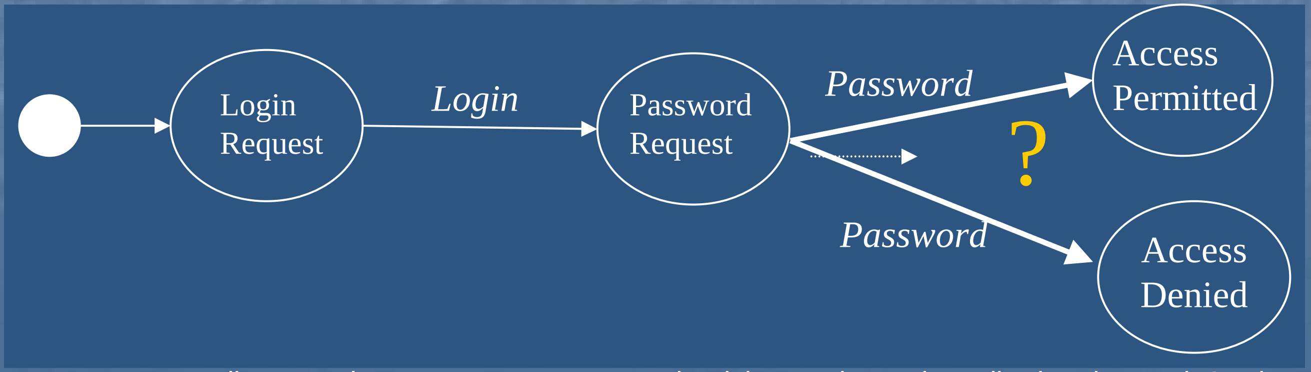
- With the diffusion of new paradigms and technological solutions for the Web (RIA and Ajax, SOA, ...), existing Web Applications are rapidly become legacy
- A strategic objective: integrating that existing applications with the new platforms
- An open issue: defining effective approaches for analysing and classifying the Web Applications User Interface

Automatic Web Pages Classification

- Web Applications interact with users by means of Client Pages that are dynamically generated
- Given a classification of Built Client pages in a set of equivalent classes
  - ... classes may correspond to the different reached Web Application scenarios
  - ... classes may correspond to the different results of the execution of test cases
- The problem:
  - To propose a technique for the automatic identification of the equivalence class a Web page belongs to

A possible application scenario

- Designing wrappers that encapsulate the original UI with the aim of exporting a renewed interface, such as a Web Service one
  - The possible reached scenarios represent the classes to recognise.
  - The Wrapper has to automatically identify the class in order to know the state of the interaction with the UI and the actions to be performed



G. Di Lorenzo, A. R. Fasolino, L. Melcarne, P. Tramontana, V. Vittorini, "*Turning Web Applications into Web Services by Wrapping Techniques*", to appear in the 14th Working Conference on Reverse Engineering WCRE 2007, Vancouver, BC, Canada

The proposed solution

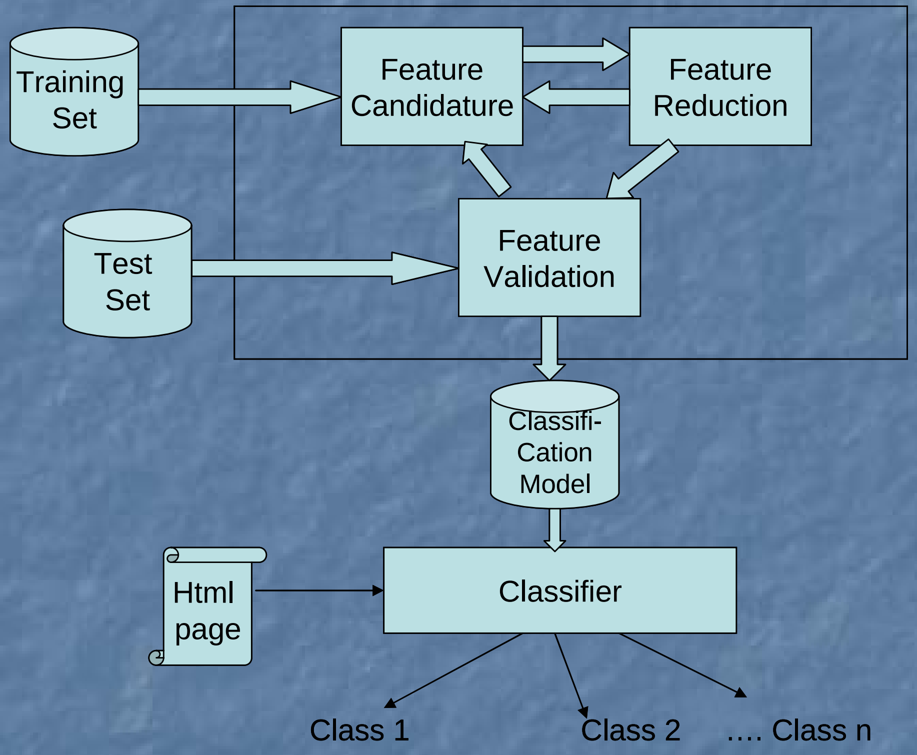
- To find, for any class, a combination of Web page features for any class that are satisfied just by pages belonging to that class (Classification model)

- We are interested to a solution that is:
  - Accurate
    - The classification approach should be very reliable, in order to support automatic tasks
  - Efficient
    - The classification approach should support a fast classification of dynamic built client pages

The proposed approach

Three step iterative process:

1. Feature Candidature
   - Generation of a set of features, representing page characteristics
2. Feature Reduction
   - Proposition of a combination of the candidate features allowing the identification of the class of any Web page
3. Feature Validation
   - Evaluation of the accurateness of the proposed classification model (i.e. the set of discriminating expressions)



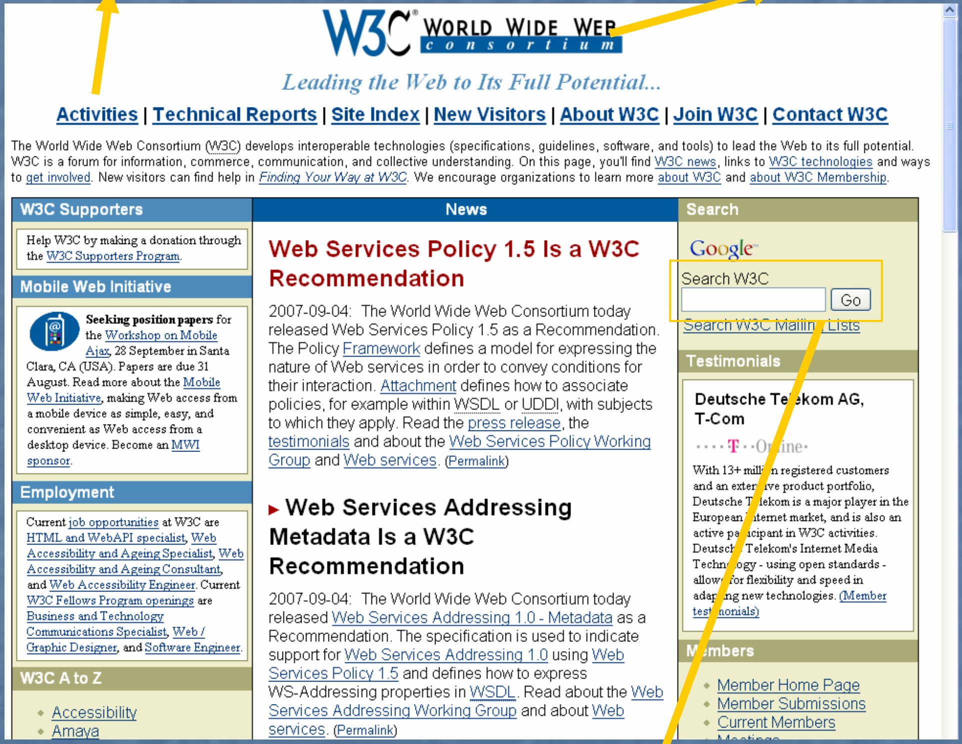
The resulting classification model allows a Classifier module to automatically determine the class any Web page belongs to

1 - Feature Candidature

- A feature is a characteristic retrievable in a Web page interface
  - It is retrievable by analysing its HTML code
    - A feature can be associated to a XPath query (if the HTML code is XHTML-compatible)
- A set of features that could be useful for discriminating a class are proposed
  - Ideally speaking, we search for features that are retrievable just in the pages belonging to a class

//a/text()=“Activities”

//h1/img



Leading the Web to Its Full Potential...

Activities | Technical Reports | Site Index | New Visitors | About W3C | Join W3C | Contact W3C

The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential. W3C is a forum for information, commerce, communication, and collective understanding. On this page, you'll find W3C news, links to W3C technologies and ways to get involved. New visitors can find help in *Finding Your Way at W3C*. We encourage organizations to learn more about W3C and about W3C Membership.

W3C Supporters

Help W3C by making a donation through the W3C Supporters Program.

Mobile Web Initiative

Seeking position papers for the Workshop on Mobile Ajax, 28 September in Santa Clara, CA (USA). Papers are due 31 August. Read more about the Mobile Web Initiative, making Web access from a mobile device as simple, easy, and convenient as Web access from a desktop device. Become an MWI sponsor.

Employment

Current job opportunities at W3C are HTML and WebAPI specialist, Web Accessibility and Ageing Specialist, Web Accessibility and Ageing Consultant, and Web Accessibility Engineer. Current W3C Fellows Program openings are Business and Technology Communications Specialist, Web / Graphic Designer, and Software Engineer.

W3C A to Z

- Accessibility
- Amaya

News

Web Services Policy 1.5 Is a W3C Recommendation

2007-09-04: The World Wide Web Consortium today released Web Services Policy 1.5 as a Recommendation. The Policy Framework defines a model for expressing the nature of Web services in order to convey conditions for their interaction. Attachment defines how to associate policies, for example within WSDL or UDDI, with subjects to which they apply. Read the press release, the testimonials and about the Web Services Policy Working Group and Web services. (Permalink)

▸ Web Services Addressing Metadata Is a W3C Recommendation

2007-09-04: The World Wide Web Consortium today released Web Services Addressing 1.0 - Metadata as a Recommendation. The specification is used to indicate support for Web Services Addressing 1.0 using Web Services Policy 1.5 and defines how to express WS-Addressing properties in WSDL. Read about the Web Services Addressing Working Group and about Web services. (Permalink)

Search

Search W3C

Go

Search W3C Mailing Lists

Testimonials

Deutsche Telekom AG, T-Com

With 13+ million registered customers and an extensive product portfolio, Deutsche Telekom is a major player in the European Internet market, and is also an active participant in W3C activities. Deutsche Telekom's Internet Media Technology - using open standards - allows for flexibility and speed in adapting new technologies. (Member testimonials)

Members

- Member Home Page
- Member Submissions
- Current Members

/html/body/div[4]/form

2 - Feature Reduction

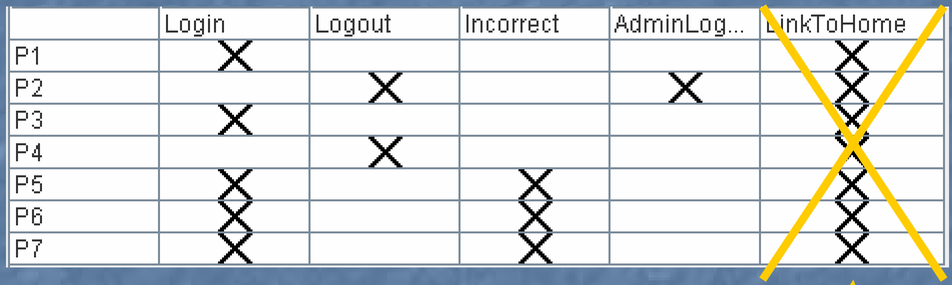
- The goal of the Feature Reduction step is to obtain a Discriminating Expression for each class, i.e. a combination of features that is true for each Web page belonging to the class, false elsewhere
- The Feature Reduction is executed on the set of candidate features
  - A training set of pre-classified pages is adopted

Evaluation Matrix

M(i,j) is the value of the Feature j for the Web Page i

Candidate Features

Training Set Pages



	Login	Logout	Incorrect	AdminLog...	~~LinkToHome~~
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~~

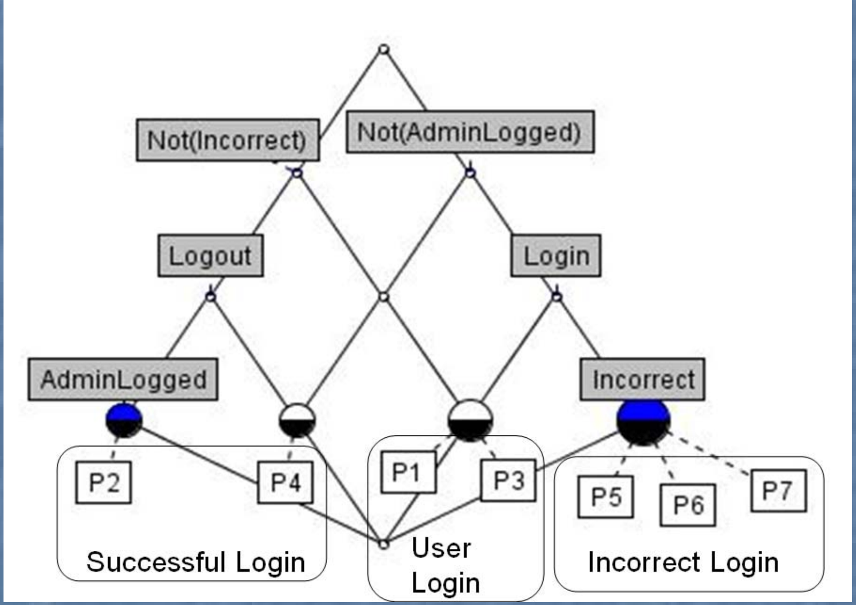
Class C1={P1, P3}
Class C2={P2, P4}
Class C3={P5, P6, P7}

The features with all true (or false) values are deleted: they surely are not useful to discriminate Classes

Concept Analysis

- Concept Analysis is adopted to find features that characterise classes of equivalent pages

- A Concept Lattice may be automatically built on the basis of the Evaluation Matrix values (evaluated on the Training Set)



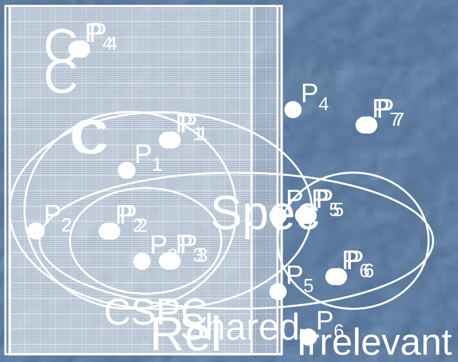
Attributes ≡ Candidate Features

Objects ≡ Training Set Pages

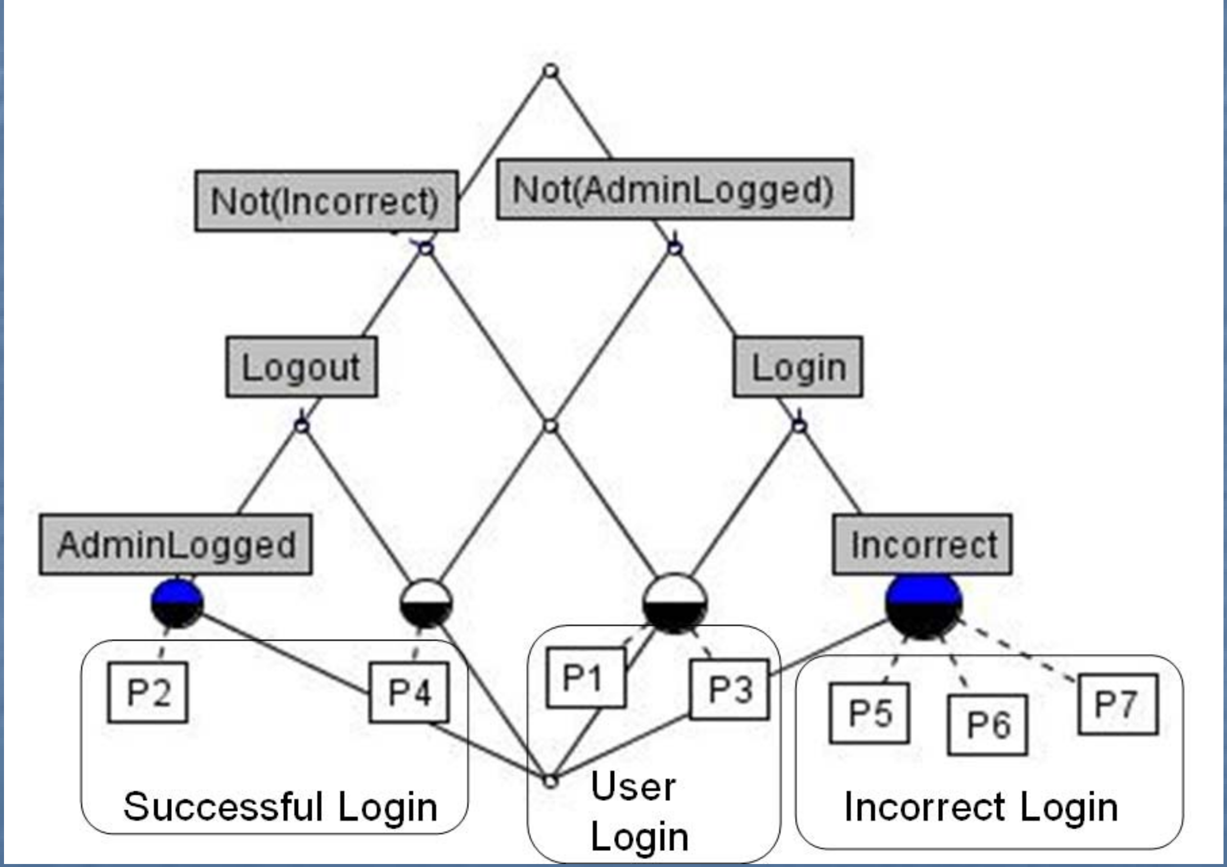
Feature Classification

According to Eisenbarth, Koschke and Simon*, features are classified, with respect to classes as:

- **Specific,** if it is satisfied by all and only the pages of the class C;
- **Relevant:** if it is satisfied by all the pages of the class C, but also by other pages from other classes;
- **CSPC** (i.e., Conditionally Specific): if it is satisfied by a subset of pages of the class C and by no other page from other classes;
- **Shared**: if it is satisfied by a subset of pages of the class C and by other pages from other classes too;
- **Irrelevant**: if it is not satisfied by any page of the classes C.



* T. Eisenbarth, R. Koschke, D. Simon, “Locating features in source code”, IEEE Transactions on Software Engineering, Volume 29, Issue 3, IEEE CS Press, March 2003, pp.210 – 224



For the class Successful Login

- Logout is Specific
- Not(Incorrect) is Relevant
- AdminLogged is CSPC
- Not(AdminLogged) is Shared
- Login is Irrelevant

Feature Classification Evaluation

- The Classification of the features may be evaluated on the basis of the Concept Analysis results:

$$Specific(C) = \bigcap_{p \in C} Intent(OC(p)) - \bigcup_{p \in TS - C} Intent(OC(p))$$

$$\mathrm{Re}\,levant(C) = \bigcap_{p \in C} Intent(OC(p)) - Specific(C)$$

$$CSPC(C) = \bigcup_{p \in C} Intent(OC(p)) - \bigcup_{p \in TS - C} Intent(OC(p)) - Specific(C)$$

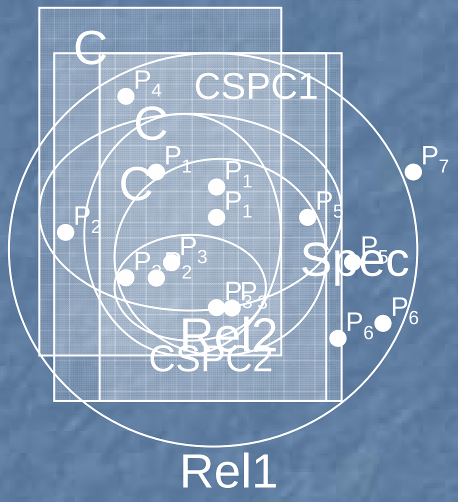
$$Shared(C) = \bigcup_{p \in C} Intent(OC(p)) - Specific(C) - \mathrm{Re}\,levant(C) - CSPC(C)$$

$$Irrelevant(C) = F - Specific(C) - \mathrm{Re}\,levant(C) - CSPC(C) - Shared(C)$$

Rules for generating Discriminating Expressions

- Our goal is to obtain combinations of features (*Discriminating Expressions*) that are *specific* for a single class, i.e. that are true for any page belonging to the class and false elsewhere

1. If the class C includes at least one specific feature, any of these specific features can be a candidate discriminating expression of the class.
2. If there are not specific features, then a candidate discriminating expression can be obtained by considering the logic and of the relevant features
   - If there is a couple of features f1 and f2 so that f1==> f2 (i.e. Extent(AC(f1))$\subseteq$ Extent(AC(f2))), then it is possible to simplify the expression by discarding the feature f2 from it .



3. If there are neither specific nor relevant features, then a candidate discriminating expression can be obtained by considering the logic or of the CSPC features
   - If there is a couple of features f1 and f2 so that f1==>f2, it is possible to simplify the expression by discarding the feature f1 from it .

3 - Discriminating Expression Validation

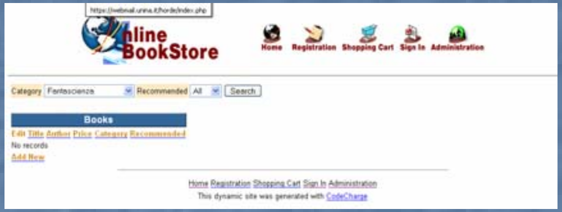
- Recall And Precision:

$$Recall = \frac{\#(pagesAttributedToClassC)}{\#(AnalysedPagesBelongingToClassC)} \qquad Precision = \frac{\#(pagesCorrectlyAttributedToClassC)}{\#(pagesAttributedToClassC)}$$

- *Failed Classification*: the page class has not been identified, because no discriminating feature (or more than one) has been satisfied by the page;
- *Incorrect Classification*: the page has been attributed to an incorrect class;
- *Correct Classification*: the page has been attributed to the correct class

A Classification Example

- Web Pages related to the AdminBooks page of the open source Web application Bookstore has been considered
- Five classes of Web pages
- 25 Web pages training set
- 65 Web pages test set



Category Fantascienza ▾ Recommended All ▾ Search

Books

Edit Title Author Price Category Recommended

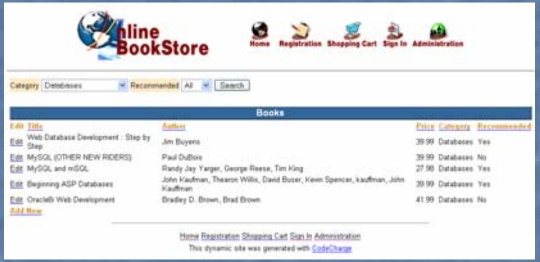
No records

Add New

Home Registration Shopping Cart Sign In Administration

This dynamic site was generated with CodeCharge

Empty List



Category Databases ▾ Recommended All ▾ Search

Books

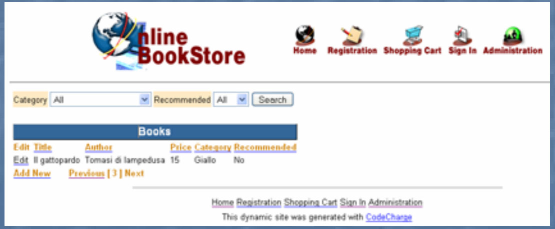
Edit	Title	Author	Price	Category	Recommended
Edit	Web Database Development : Step by Step	Jim Buyens	39.99	Databases	Yes
Edit	MySQL (OTHER NEW RIDERS)	Paul DuBois	39.99	Databases	No
Edit	MySQL and mSQL	Randy Jay Yarger, George Reese, Tim King	27.98	Databases	Yes
Edit	Beginning ASP Databases	John Kaufman, Thearon Willis, David Buser, Kevin Spencer, kauffman, John Kauffman	39.99	Databases	Yes
Edit	Oracle8i Web Development	Bradley D. Brown, Brad Brown	41.99	Databases	No

Add New

Home Registration Shopping Cart Sign In Administration

This dynamic site was generated with CodeCharge

Single Page



Category All ▾ Recommended All ▾ Search

Books

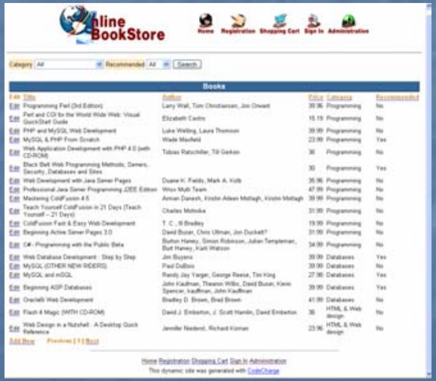
Edit	Title	Author	Price	Category	Recommended
Edit	Il gattopardo	Tomasi di lampedusa	15	Giallo	No

Add New Previous | 3 | Next

Home Registration Shopping Cart Sign In Administration

This dynamic site was generated with CodeCharge

Last Page



Category All ▾ Recommended All ▾ Search

Books

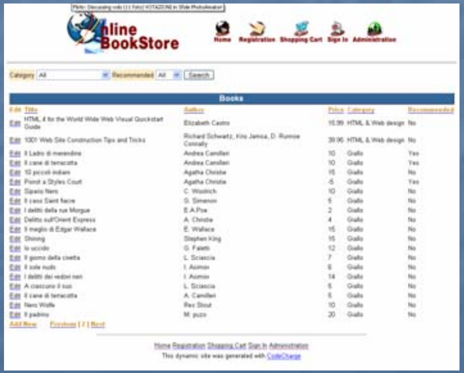
Edit	Title	Author	Price	Category	Recommended
Edit	Programming Perl (3rd Edition)	Larry Wall, Tom Christiansen, Jon Orwant	39.96	Programming	No
Edit	Perl and CGI for the World Wide Web: Visual QuickStart Guide	Elizabeth Castro	15.19	Programming	No
Edit	PHP and MySQL Web Development	Luke Welling, Laura Thomson	39.99	Programming	No
Edit	MySQL & PHP From Scratch	Wade Maxfield	23.99	Programming	Yes
Edit	Web Application Development with PHP 4.0 (with CD-ROM)	Tobias Ratschiller, Till Gerken	36	Programming	No
Edit	Black Belt Web Programming Methods: Servers, Security, Databases and Sites		30	Programming	Yes
Edit	Web Development with Java Server Pages	Duane K. Fields, Mark A. Kolb	35.96	Programming	No
Edit	Professional Java Server Programming J2EE Edition	Wrox Multi Team	47.99	Programming	No
Edit	Mastering ColdFusion 4.5	Arman Danesh, Kristin Aileen Motlagh, Kristin Motlagh	39.99	Programming	No
Edit	Teach Yourself ColdFusion in 21 Days (Teach Yourself -- 21 Days)	Charles Mohnike	31.99	Programming	No
Edit	ColdFusion Fast & Easy Web Development	T. C., III Bradley	19.99	Programming	No
Edit	Beginning Active Server Pages 3.0	David Buser, Chris Ullman, Jon Duckett?	31.99	Programming	No
Edit	C# - Programming with the Public Beta	Burton Harvey, Simon Robinson, Julian Templeman, Burt Harvey, Karli Watson	34.99	Programming	No
Edit	Web Database Development : Step by Step	Jim Buyens	39.99	Databases	Yes
Edit	MySQL (OTHER NEW RIDERS)	Paul DuBois	39.99	Databases	No
Edit	MySQL and mSQL	Randy Jay Yarger, George Reese, Tim King	27.98	Databases	Yes
Edit	Beginning ASP Databases	John Kaufman, Thearon Willis, David Buser, Kevin Spencer, kauffman, John Kauffman	39.99	Databases	Yes
Edit	Oracle8i Web Development	Bradley D. Brown, Brad Brown	41.99	Databases	No
Edit	Flash 4 Magic (WITH CD-ROM)	David J. Emberton, J. Scott Hamlin, David Emberton	36	HTML & Web design	No
Edit	Web Design in a Nutshell : A Desktop Quick Reference	Jennifer Niederst, Richard Koman	23.96	HTML & Web design	No

Add New Previous | 1 | Next

Home Registration Shopping Cart Sign In Administration

This dynamic site was generated with CodeCharge

First Page



Category All ▾ Recommended All ▾ Search

Books

Edit	Title	Author	Price	Category	Recommended
Edit	HTML 4 for the World Wide Web Visual Quickstart Guide	Elizabeth Castro	15.99	HTML & Web design	No
Edit	1001 Web Site Construction Tips and Tricks	Richard Schwartz, Kris Jamsa, D. Rummie Connally	39.96	HTML & Web design	No
Edit	Il Ladro di merendine	Andrea Camilleri	10	Giallo	Yes
Edit	Il cane di terracotta	Andrea Camilleri	10	Giallo	Yes
Edit	10 piccoli indiani	Agatha Christie	15	Giallo	No
Edit	Poirot a Styles Court	Agatha Christie	-5	Giallo	Yes
Edit	Sipario Nero	C. Woolrich	10	Giallo	No
Edit	Il caso Saint fiacre	G. Simenon	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

Add New Previous | 2 | Next

Home Registration Shopping Cart Sign In Administration

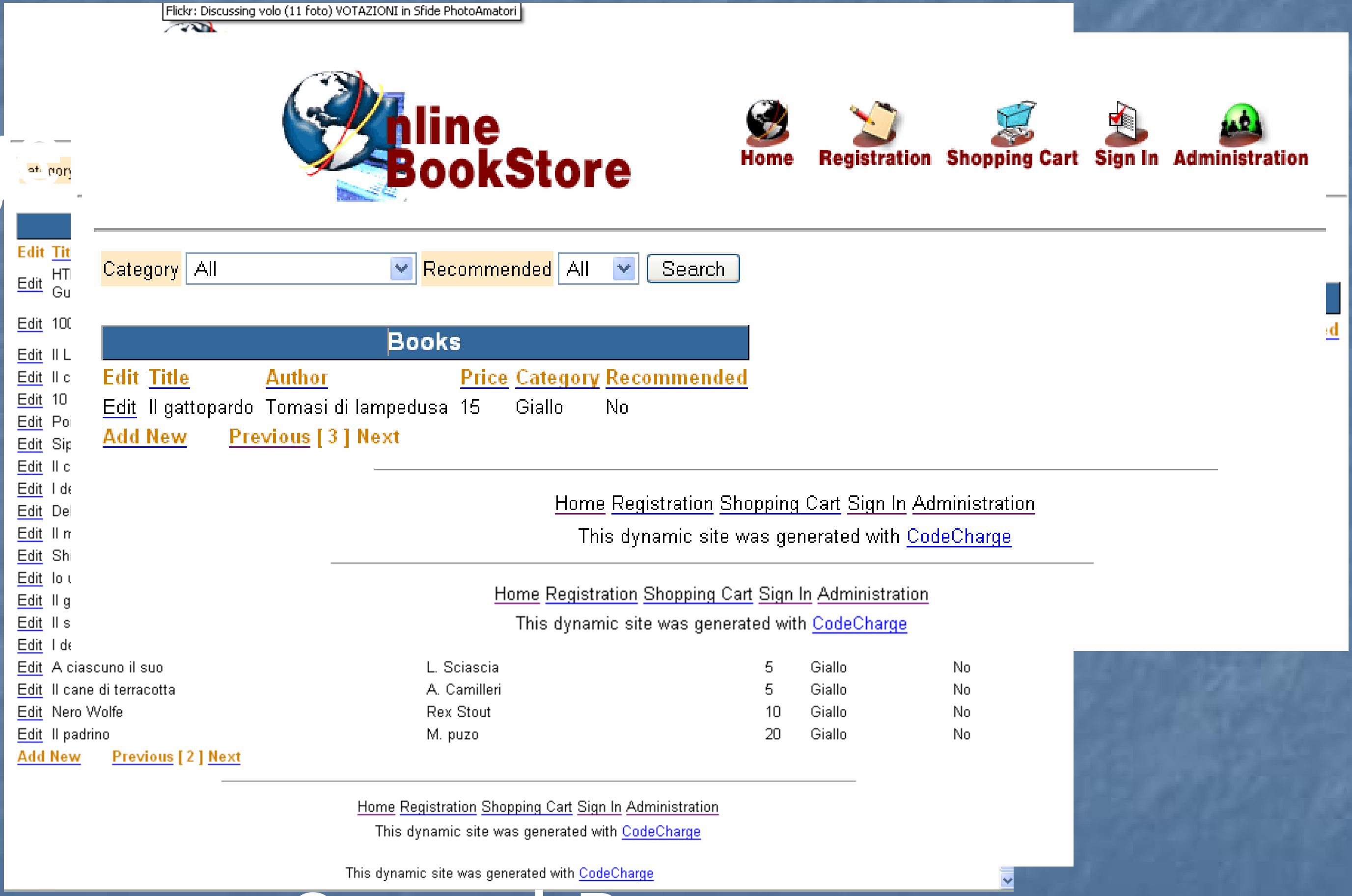
This dynamic site was generated with CodeCharge

Central Page

A Classification Example

Last Page

Single Page



Flickr: Discussing volo (11 foto) VOTAZIONI in Sfide PhotoAmatori

nline BookStore

Category: All — Recommended: All — Search

Books					
Edit	Title	Author	Price	Category	Recommended
Edit	Il gattopardo	Tomasi di lampedusa	15	Giallo	No

Add New Previous [3] Next

Home Registration Shopping Cart Sign In Administration

This dynamic site was generated with CodeCharge

Home Registration Shopping Cart Sign In Administration

This dynamic site was generated with CodeCharge

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

Add New Previous [2] Next

Home Registration Shopping Cart Sign In Administration

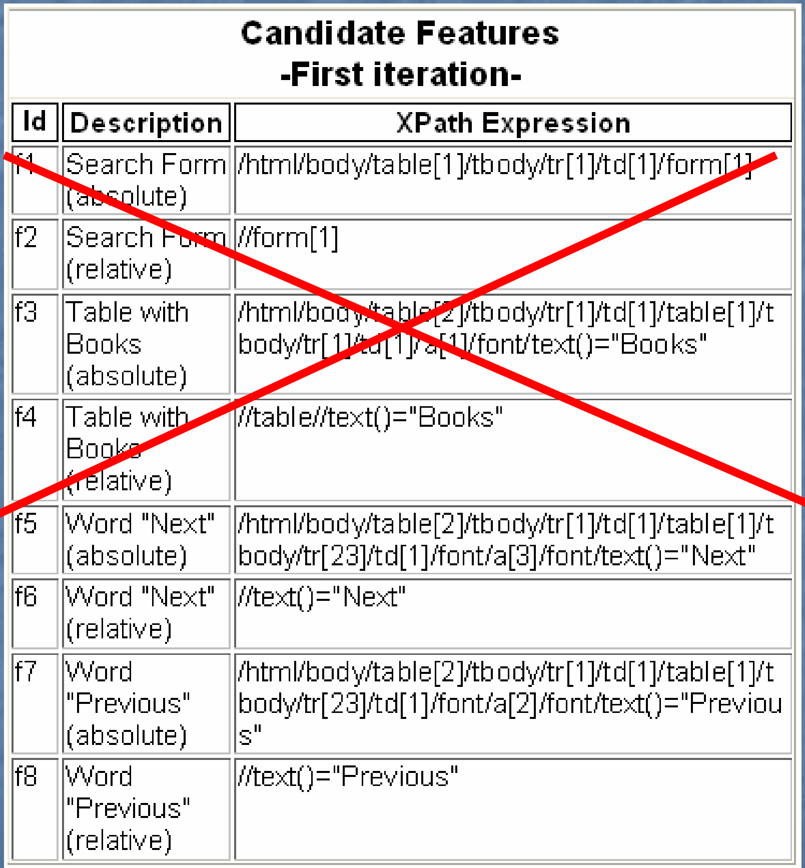
This dynamic site was generated with CodeCharge

This dynamic site was generated with CodeCharge

Central Page

First Page

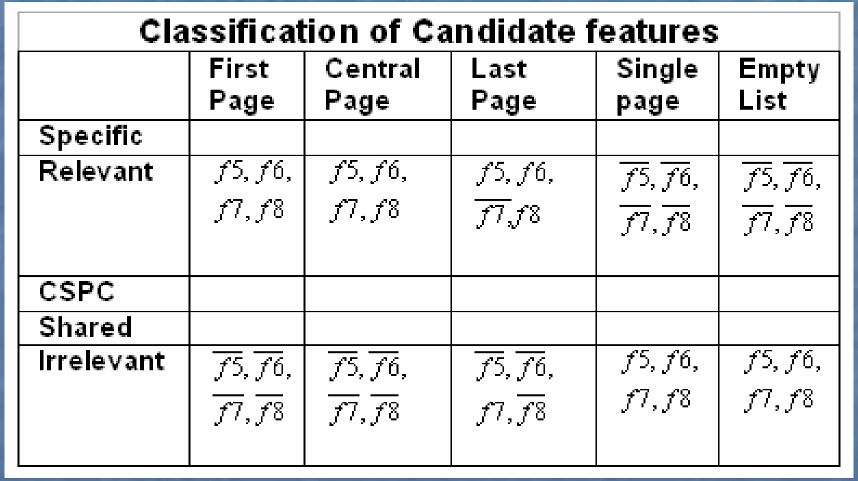
Candidate features



Candidate Features
-First iteration-

Id	Description	XPath Expression
~~f1~~	~~Search Form (absolute)~~	~~/html/body/table[1]/tbody/tr[1]/td[1]/form[1]~~
~~f2~~	~~Search Form (relative)~~	~~//form[1]~~
~~f3~~	~~Table with Books (absolute)~~	~~/html/body/table[2]/tbody/tr[1]/td[1]/table[1]/tbody/tr[1]/td[1]/a[1]/font/text()="Books"~~
~~f4~~	~~Table with Books (relative)~~	~~//table//text()="Books"~~
f5	Word "Next" (absolute)	/html/body/table[2]/tbody/tr[1]/td[1]/table[1]/tbody/tr[23]/td[1]/font/a[3]/font/text()="Next"
f6	Word "Next" (relative)	//text()="Next"
f7	Word "Previous" (absolute)	/html/body/table[2]/tbody/tr[1]/td[1]/table[1]/tbody/tr[23]/td[1]/font/a[2]/font/text()="Previous"
f8	Word "Previous" (relative)	//text()="Previous"

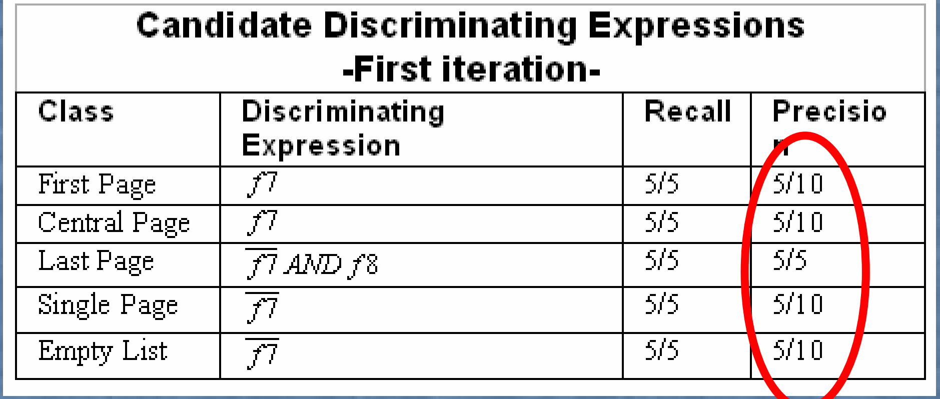
- 8 candidate features have been considered
- f1, f2, f3, f4 have been discarded (always true)
- Negated features have been introduced



Classification of Candidate features

	First Page	Central Page	Last Page	Single page	Empty List
Specific					
Relevant	$f5, f6, f7, f8$	$f5, f6, f7, f8$	$f5, f6, \overline{f7}, f8$	$\overline{f5}, \overline{f6}, \overline{f7}, \overline{f8}$	$\overline{f5}, \overline{f6}, \overline{f7}, \overline{f8}$
CSPC					
Shared					
Irrelevant	$\overline{f5}, \overline{f6}, \overline{f7}, \overline{f8}$	$\overline{f5}, \overline{f6}, \overline{f7}, \overline{f8}$	$\overline{f5}, \overline{f6}, f7, \overline{f8}$	$f5, f6, f7, f8$	$f5, f6, f7, f8$

Discriminating Expressions – First iteration

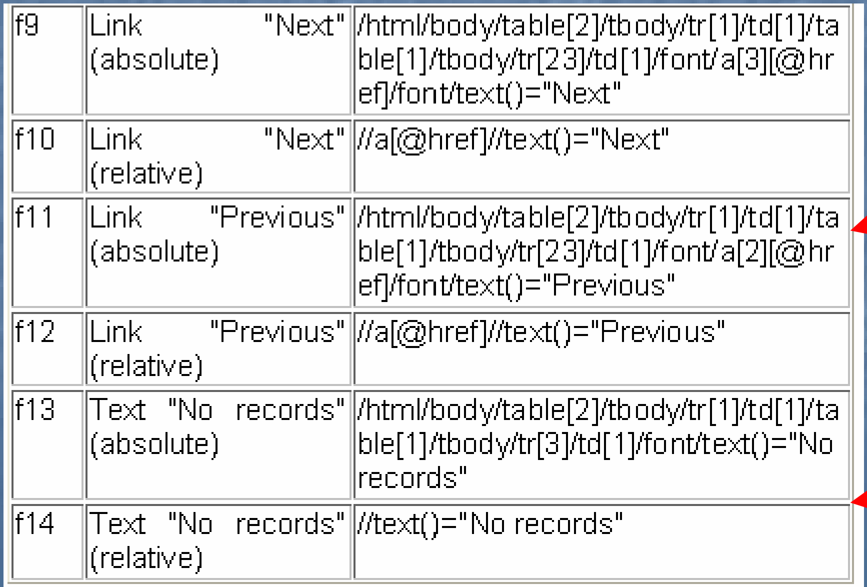


Candidate Discriminating Expressions
-First iteration-

Class	Discriminating Expression	Recall	Precision
First Page	$f7$	5/5	5/10
Central Page	$f7$	5/5	5/10
Last Page	$\overline{f7}$ AND $f8$	5/5	5/5
Single Page	$\overline{f7}$	5/5	5/10
Empty List	$\overline{f7}$	5/5	5/10

Training Set Validation

There is not enough precision!



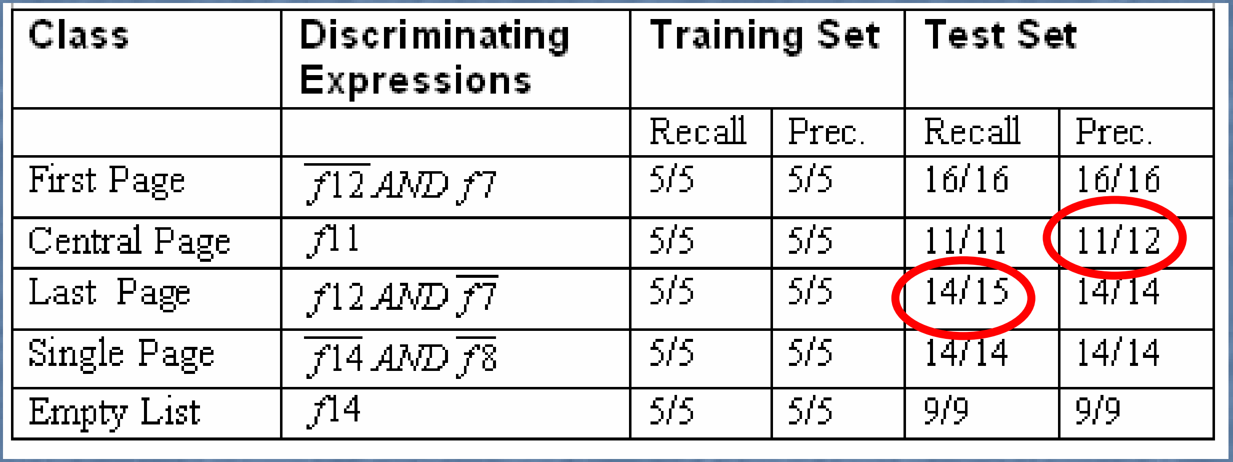
f9	Link "Next" (absolute)	/html/body/table[2]/tbody/tr[1]/td[1]/table[1]/tbody/tr[23]/td[1]/font/a[3][@href]/font/text()="Next"
f10	Link "Next" (relative)	//a[@href]//text()="Next"
f11	Link "Previous" (absolute)	/html/body/table[2]/tbody/tr[1]/td[1]/table[1]/tbody/tr[23]/td[1]/font/a[2][@href]/font/text()="Previous"
f12	Link "Previous" (relative)	//a[@href]//text()="Previous"
f13	Text "No records" (absolute)	/html/body/table[2]/tbody/tr[1]/td[1]/table[1]/tbody/tr[3]/td[1]/font/text()="No records"
f14	Text "No records" (relative)	//text()="No records"

A new iteration of feature candidature is needed

New features have been added:

1. We need to distinguish when the words “Next” and “Previous” are links or simple labels
2. We need to distinguish between empty page and single page

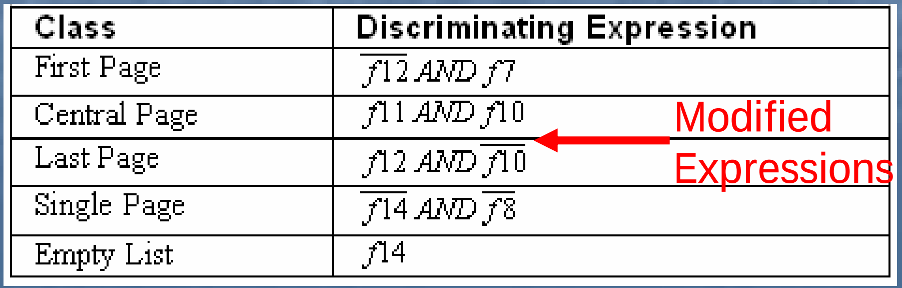
Discriminating Expressions – Second Iteration



Class	Discriminating Expressions	Training Set		Test Set	
		Recall	Prec.	Recall	Prec.
First Page	$\overline{f12}$ *AND* $f7$	5/5	5/5	16/16	16/16
Central Page	$f11$	5/5	5/5	11/11	11/12
Last Page	$f12$ *AND* $\overline{f7}$	5/5	5/5	14/15	14/14
Single Page	$\overline{f14}$ *AND* $\overline{f8}$	5/5	5/5	14/14	14/14
Empty List	$f14$	5/5	5/5	9/9	9/9

Training Set Validation is OK but there are some problems in a page belonging to the Test Set (assigned to Central page instead of Last Page)

The incorrect-assigned page has been added to the Training Set and the Candidate Discriminating Expressions have been evaluated



Class	Discriminating Expression
First Page	$\overline{f12}$ *AND* $f7$
Central Page	$f11$ *AND* $f10$
Last Page	$f12$ *AND* $\overline{f10}$
Single Page	$\overline{f14}$ *AND* $\overline{f8}$
Empty List	$f14$

Modified Expressions

Recall and Precision are, now, 100% both on the Training Set and on the Test Set

Discussion

- Process Effort
  - In the case study, about 100 minutes were needed to obtain the classification model
  - The activities requiring human intervention was the most critical
    - Training and Test Set Collection step
      - The reliability of the classification model depends on the *representativeness* of the Training and Test Sets
      - Web pages resulting from the execution of a Test Suite may be good candidates to fill in the Training and the Test Set
    - Feature Candidature step
      - Heuristic techniques allowing a semi-automatic generation of features are currently under experimentation

Further application scenarios

- Support to testing automation
  - The possible reached pages are arranged in a set of equivalence classes, corresponding to different test results
  - The proposed discriminating expressions make it possible the automatic evaluation of the testing results
- Support to dynamic analysis
  - The recognition of the reached interaction scenarios makes it possible the automatic collection of logs of the visited use cases and scenarios, providing also useful information for user profiling

Conclusions and Future Works

- A process for the definition of a classification model allowing the automatic identification of classes of Built Client pages has been defined and validated
- The classification model allows the run-time identification of the classification of a Built Client page, with short response time and high precision
- Future Works:
  - We are working for the process improvement, in order to reduce the effort needed to human made tasks
  - Further experimentation will be carried out in order to assess the scalability of the approach

Time is over ... Are there any questions?