Ph.D. Dissertation

Reverse Engineering
Web Applications

Porfirio Tramontana

University of Naples “Federico II”
Web Applications: open problems

- In the past years, a great request for Web Applications takes place, due to the World Wide Web diffusion making available many services all over the world.

- Web Applications have been developed with immature design methodologies and technologies.

- Nowadays, there is a number of legacy Web Applications needing for maintenance and re-engineering.
Ph. D. Thesis Goals

• Reverse Engineering and comprehension are fundamental tasks needed to efficiently support maintenance, testing and quality assessment of Web Applications

Doctoral Thesis Goals

• To propose models, methods and tools supporting Reverse Engineering and Comprehension of Web Applications
Peculiarities of script-based Web Applications

- Page based
- Client-Server Architecture
- Interpreted languages
- Client pages may be generated “on the fly”
- Client pages are executed in a browser (and the designer doesn’t know what kind of browser will be used)
- HTML interpreters are fault tolerant

... and so on ...
A process for the Reverse Engineering of Web Applications

**Extraction**
- WA Source Code
- WA Execution

**Static Analysis**
- Identification of cloned components
- Identification of Interaction Design Patterns

**Dynamic Analysis**
- Assignment of Concepts
- Functional Clustering
- Business Level UML Diagram Abstractions

**Abstraction**
- Cloned components
- Interaction Design Patterns
- Concepts describing Reverse Engineering artifacts
- Groups of pages realizing Web Application use cases
- Structural and Business Level UML diagrams
- Maintainability assessment


Ph.D. Dissertation Forum – ICSM 2005
Analysis of Web Applications

1) Static analysis of the source code
   A multi-language parser analysing the source code of Server pages, Client pages and Script modules has been realized.
   During the analysis of server pages, facts related to the client pages that are built by server pages are also recorded.
   Static analysis results are stored in an intermediate form and are used to fill a relational database.

2) Dynamic Analysis
   Analysis of Built Client pages in order to add to the database some facts that have been observed by executing the application.

   The reference model adopted is an extension of the one proposed by Conallen for the forward engineering of Web Applications.
Model of Web Applications

Ph.D. Dissertation Forum – ICSM 2005
WARE (Web Application Reverse Engineering) tool

WARE Architecture

WARE Architecture Diagram abstracted by WARE

Functional Clustering of Web Pages

• **Goal:**
  To cluster together subsets of components realizing Web Application functionalities

• **Proposed Technique:**
  Hierarchical clustering algorithm, grouping Web Application pages in subsets, maximizing the cohesion and minimizing the coupling between them


Ph.D. Dissertation Forum – ICSM 2005
Concept Assignment

**Goal:**
- To identify the more relevant concepts in client pages with the purpose to suggest a semantic description of client pages and of functional clusters of pages

**Proposed Technique:**
- Heuristic Algorithms based on Information Retrieval
  - Candidate concepts are searched in textual content of client pages
  - Single common words and short word sequences are candidate to be concepts

Interaction Design Patterns Identification

■ Goal:
  ■ To identify repetitive structures in Web Client pages
    ■ These structures can be related to known Programming Patterns

■ Proposed Technique:
  ■ Statistical methodology based on features extracted in the source code of client pages.
    ■ Presence, quantity and dimension of forms, tables, input fields, frames, common keywords and so on.


Ph.D. Dissertation Forum – ICSM 2005
Identification of cloned components

**Goals:**
- Re-Engineering of cloned components via code transformations
- Classification of Built Client Pages
- Identification of reusable Programming Patterns

**Proposed Techniques:**
- Extraction of features in the structure of Client pages and in the source code of server pages
- Computation of distance measures between pages (Euclidean distance, Levenshtein edit distance)


Ph.D. Dissertation Forum – ICSM 2005
Abstraction of Business Level Models

- **Goals:**
  - To abstract object oriented business level models of Web Applications
- **Proposed Techniques:**
  - Classes and attributes are identified by analysing the data that are exchanged between user, Web pages and databases.
  - Class methods are identified by analysing the functions implemented by cluster of pages.
  - Relationships between classes are identified analysing data structures and data flow among pages.


Ph.D. Dissertation Forum – ICSM 2005
Maintainability Model

- **Goals:**
  - To propose models and methods for the assessment of the maintainability of Web Applications

- **Proposed Models and Techniques:**
  - Adapting to Web Applications the Oman model (thought for traditional applications)
  - Selection of a set of product metrics and proposal of a maintainability index that can be calculated with negligible effort and time

Current and future works

- Techniques for the dynamic analysis of Web Applications
- Accessibility assessment of Client pages
- Migration from Web Applications to Web Services
- Testing of Web Applications
  - Mutation Testing techniques
- Maintainability assessment
  - Definition of ageing measures for Web Applications


Time is over ... Are there any questions?