Validation of a Web Application by Using a Limited Number of Web Pages

Doru Anastasiu Popescu
University of Pitesti, Faculty of Mathematics and Computer Science, Romania
dopopan@yahoo.com

Maria Catrinel Dănăuţă
School of Electronics And Computer Science, University of Southampton, United Kingdom
cd4g09@ecs.soton.ac.uk

Abstract

In this paper, we are trying to introduce a method of selection of some web pages from a web application, which will be verified by using different validating mechanisms. The number of selected web pages cannot be higher than a previously established constant. The method of selection of these web pages must assure the highest possible quality of the verification of the entire application. The error detection of these web pages will automatically lead to the error detection in other pages. This fact will be realised by using an equivalence relation among the web pages of the web application.

Keywords: Relation, Algorithm, Tags, Verification, GARWA

1. Introduction

In order to verify the web applications, different static or dynamic validators can be used. Many of these validators can be freely used on the Internet, with an online connection (for example [9], [10], [11]). Most of these validators associate an execution to the verification on only one web page. From this point of view, in this paper we will introduce an equivalence relation between two web pages of a web application (which will be called in the rest of the paper WA), relation used by the authors in [1] as well, in section 1. Using this equivalence relation, we can obtain a partition of the set web pages from WA and construct an undirected balanced graph associated to this partition (section 3), called GARWA. By using this graph, we will introduce in section 4 a new algorithm of determining a fixed number of web pages from WA that verify WA the best.

2. Defining an equivalence relation among the web pages of a web application

Let $P = \{p_1, p_2, ..., p_n\}$ the web pages of the web application, called WA, and Tg a set of unimportant tags. For example, Tg can contain the tags
, </BR>, <P>, which are often used in texts. Tg can be the empty set as well.

For any web page p_i, le T_i be the sequence of tags from p_i which are not in Tg (by sequence we understand that the tags are in the order they appear in p_i).

Definition

For two web pages p_i and p_j are in the relation ~ (written $p_i \sim p_j$) if and only if $T_i = T_j$.

Example

Next, we will consider three web pages p_1, p_2 and p_3 that correspond to the files **Pag1.html**, **Pag2.html** and **Pag3.html** respectively, with the following source code:

Pag1.html

```
<HTML>
<HEAD>
<TITLE>Web page 1</TITLE>
</HEAD>
<BODY>
<p> Picture 1
```

```
<IMG SRC="picture.jpg">
</BODY>
</HTML>
```

Pag2.html

```
<HTML>
<HEAD>
<TITLE>Web page 2</TITLE>
</HEAD>
<BODY>
<B> <p>Picture 2 </p> </B>
<IMG SRC="picture.jpg">
</BODY>
</HTML>
</BODY>
</HTML>
```

Pag3.html

```
<HTML>
<HEAD>
<TITLE>Web page 3</TITLE>
</HEAD>
<BODY>
<FONT COLOR=red>Picture 3 </FONT>
<FONT SIZE=4 COLOR=red>Picture 3 </FONT>
</BODY>
</HTML>
</BODY>
</HTML>
```

Let us consider Tg={ , , <B>, </B>}. Then:

T_1=(<HTML>, <HEAD>, <TITLE>, </TITLE>, </HEAD>, <BODY>, <IMG SRC="picture.jpg">, </BODY>, </HTML>)

T_2=(<HTML>, <HEAD>, <TITLE>, </TITLE>, </HEAD>, <BODY>, <IMG SRC="picture.jpeg">, </BODY>, </HTML>, </BODY>, </HTML>)

T_3=(<HTML>, <HEAD>, <TITLE>, </TITLE>, </HEAD>, <BODY>, <FONT COLOR=red>, </FONT>, <FONT SIZE=4 COLOR=red>, </FONT>, </BODY>, </HTML>).

Notice that $T_1=T_2$ and, consequently, $p_1 \sim p_2$, $T_1 \neq T_3$, so p_1 is not in the relation ~ with p_3.

Note

The relation ~, defined above, is an equivalence relation. The verification of this property is immediate.

Using the previous note, we can determine a partition of the set of web pages P. We will call C_1, C_2, ..., C_m the equivalence classes associated to ~ in P. This way, $P=C_1 \cup C_2 \cup ... \cup C_m$ and $C_i \neq \emptyset$, for any $i \in \{1, 2, ..., m\}$.

If $p \in C_i$, then $C_i=\{q \in P \mid q \sim p\}$, for any $i \in \{1, 2, ..., m\}$.

3. Constructing an undirected balanced graph associated to a WA

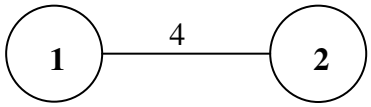
Next, we will use the notations from section 2. From every set C_i, $i \in \{1, 2, ..., m\}$ we will select an element, that we choose to call q_i (because of the method of defining these sets, it does not matter which element we choose). Obviously, there will exist any $j \in \{1, 2, ..., n\}$, such that $q_i=p_j$.

The graph that we will construct will be called GARWA (Graph Associated Relation in the Web Application) and will have as nodes the web pages q_1, q_2, ..., q_m. GARWA will have an edge between any two nodes. The cost of an edge between the web pages q_i and q_j will be maximum$\{|T_i|,|T_j|\}$-h, where h is the length of the longest common sequence of tags from T_i and T_j, and $|T_i|$, $|T_j|$ are the number of elements from the sets T_i and T_j, respectively.

For the example in section 1, we have m=2, $C_1=\{p_1, p_2\}$, $C_2=\{p_3\}$. The GARWA of this WA has two nodes corresponding to the web pages $q_1=p_1$ and $q_2=p_3$. The cost of the only edge in

GARWA is equal to 4=12-8, because the biggest length of a common sequence from T_1 and T_3 is 8 (such a sequence is written in bold in T_1 and T_3) and maximum{|T_1|, |T_3|}=maximum{9,12}=12.
T_1=(**<HTML>**, **<HEAD>**, **<TITLE>**, **</TITLE>**, **</HEAD>**, **<BODY>**, <IMG SRC="picture.jpg">, **</BODY>**, **</HTML>**)
T_3=(**<HTML>**, **<HEAD>**, **<TITLE>**, **</TITLE>**, **</HEAD>**, **<BODY>**, <FONT COLOR=red>, </FONT>, <FONT SIZE=4 COLOR=red>, </FONT>, **</BODY>**, **</HTML>**).

Figure 1 shows the GARWA for this example.



GARWA

Figure 1

4. Algorithm of determining the web pages that will be used to verify the WA

Using the notations from sections 2 and 3, we will consider a natural number, different from 0, k≤m. We are intending of determining k from the web pages in the WA that will lead to a better verifying of WA. The quality of the verification of a WA, by using exactly k web pages, can be thought of from two points of view:

I) The maximisation of the number of web pages that will implicitly be verified (by verifying a web page q_i from C_i, we practically implicitly verify all the web pages in C_i)

II) The maximisation of the number of verified tags in all the web pages of the WA.

The first criterion has been introduced in [1] by the authors. We will next deal with criterion II).

In order to solve this problem, we propose a Greedy algorithm.

Note

Because of the method of defining the cost of an edge in the GARWA, we note the fact that if we want to maximise the number of verified tags by verifying two web pages, we need to choose from the GARWA two nodes i and j with the property that the cost of [i,j] is maximum.

By using the previous note, our problem is reduced to determining a set M with k nodes from GARWA, with the property that the sum of the edges' cost with both ends in M is maximum.

Example

For the next GARWA (see Figure 2):

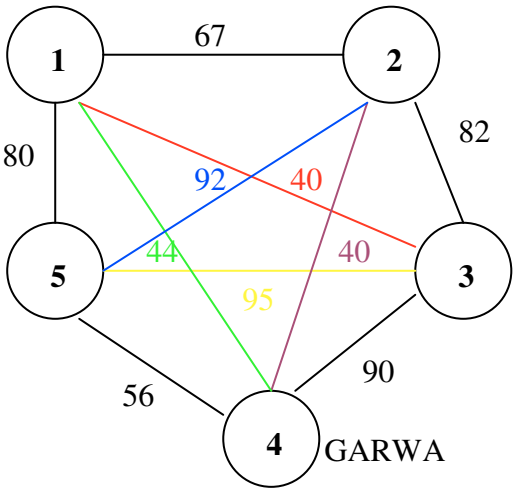


Figure 2

If we consider k=3, then M={2, 3, 5}, because the sum of the edges' cost [2,3], [2,5] and [3,5] is maximum, as in 92+95+82=269.

Algorithm

Step 1. We determine one member from each equivalence class C_1, C_2, …, C_m, called q_1, q_2, …, q_m.
Step 2. We construct the undirected graph GAEWA, more precisely we determine a matrix $c=(c_{ij})_{i,j=1,2,\ldots,m}$, where c_{ij} is the cost of the edge [i,j]; if i=j, $c_{ij}=0$.
Step 3. If k=1, then M$\leftarrow${i}, where i is chosen such that $c_{i1}+c_{i2}+\ldots+c_{im}$=maximum.
Step 4. If k>1, then:

- We determine i si j with c_{ij} = maximum and M$\leftarrow${i,j}.
- For r=1,k-2 do
  We determine a node x$\notin$ M, x$\in${1,2,...,m}, with the property that the sum of the edges' costs with an end in x and the other one in the set M is maximum.
  M$\leftarrow$M$\cup${x}
  End for

Step 5. By using existing validators, such as [9], [10], [11] or the applications introduced in [7], [8], we verify the web pages in the set M.

Notes

1. The operation of determining the equivalence classes involves verifying the relation ~ between any two web pages of the WA and, consequently, has a complexity of $O(n^2 \cdot a)$, where a is the maximum number of characters from a file which contains a web page from the WA.
2. Constructing the matrix c involves using an algorithm which determines the maximum length for a common sequence associated to two sequences of tags (by using Dynamic Programming Method in [2], there is introduced such an algorithm), the complexity being of the order $O(m^2 \cdot b^2)$, where b is the maximum number of tags from a web page in the WA.

5. Conclusions and future work

Simplifying the method of verification and testing of the web applications and maximising the quality of the verification is the principal objective for the time being. We consider that the method we describe in the previous sections is an important step in this matter and, by combining the present techniques of verification and testing with the mode of reducing the objects that need to be tested we can obtain very good results in obtaining web application of very good quality that function correctly. We believe that this idea of selecting certain components from a web application can be developed by using other methods of comparison among the components (that can use notion as those introduces in [3], [4], [7]).

5. References

[1] Catrinel Maria Dănăuţă, Doru Anastasiu Popescu, (2009), Method of reduction of the web pages to be verified when validating a web site, *Buletin Ştiinţific, Universitatea din Piteşti, Seria Matematică şi Informatică*, Nr. 15, pg 19-24.
[2] Cormen T.H., Leiserson C.E., Rivest R.L., Stein C. (1990), Introduction to ALGORITHMS, second edition, MIT Press
[3] Doru Anastasiu Popescu, (2009), Testing web application navigation based on component complexity, *Buletin Ştiinţific, Universitatea din Piteşti, Seria Matematică şi Informatică*, Nr. 15, pg 107-118.
[4] Mao Cheng-ying, Lu Yan-sheng (2006), A Method for Measuring the Structure Complexity of Web Application, *Wuhan University Journal of Natural Sciences*, vol. 11, No. 1
[5] Samir Khuller, Anna Moss, Joseph (Seffi) Naor (1999), The budgeted maximum coverage problem, *Wuhan University Journal of Natural SciencesInformation Processing Letters*, 70, pg.39-45

[6] G. Sreedhar, A.A. Chari, V.V. Ramana (2010), Measuring Qualitz of Web Site Navigation, *Journal of Theoretical and Applied Information Technology*, Vol. 14, Nr. 2.
[7] ZHAO Cheng-li, YI Dong-yun (2004), A Method of Eliminating Noise in Web Pages by Style Tree Model and Its Applications, *Wuhan University Journal of Natural Sciences*, vol. 9, No. 5
[8] Zhongsheng Qian, Huaikou Miao, Hongwei Zeng (2008), A Practical Web Testing Model for Web Application Testing, *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, IEEE
[9] Alpine HTML Doctor: http://www.alpineinternet.com/
[10] Validone HTML/XHTML/… http://www.validome.org/
[11] W3C Markup Validation Service http://validator.w3.org/