A Metric-Based Analysis of Web Sites in Serbia: First Findings

ZORICA M. MIHAJLOVIĆ¹, and NIKOS E. MASTORAKIS²
¹College of Information and Communication Technology
Zdravka Čelara 16, 11000, Belgrade
SERBIA
²Department of Computer Science
Military Institution of the University of Education
Hellenic Naval Academy
Hatzikyriakou Avenue, 18539, Piraeus
GREECE

Abstract: - Web engineering can be defined as a disciplined and systematic approach to development, deployment and maintenance of high-quality web applications. It borrows many principles, processes, methods and tools from software engineering and measurements and metrics are among them. The research on web metrics already done provides evidence that even simple HTML tag based metrics can be useful in evaluating quality attributes of web applications and in information retrieval. In this paper we study a number of web sites in Serbia in terms of HTML metrics gathered from the sites’ web pages in order to compare the statistics of metrics with the results coming from a more advanced domain of the Web and presented in a previous work. We want to determine if our web pages are high-quality pages. The study covers web sites from educational and financial topics. Our analysis shows that the results obtained from web sites of banks in Serbia are very close to the results given in the previous work. The foreign banks in Serbia contribute to this mostly. In addition, our analysis finds out that schools of engineering of the University of Belgrade have the best web sites when the statistical features of HTML metrics are compared to those given in the previous work. A special tool for extracting metrics from HTML code automatically, named SMeterH, has been used in our analysis.


1 Introduction
Nowadays, the World Wide Web is present everywhere in our lives. Primarily introduced to enable publishing various contents, in the meantime it becomes a supplier of a lot of quite different services: from e-commerce to distance learning. Developing web applications for such services is a complex endeavor and thus it evolves to a new engineering discipline – web engineering. Web engineering can be defined as a disciplined and systematic approach to development, deployment and maintenance of high-quality web applications [9].

Web engineering (WE) borrows many principles, processes, methods and tools from software engineering (SE) and measurements and metrics are among them [9]. Metrics in WE can be classified in the same way as it is done in SE: product metrics and process metrics [10]. Product metrics are associated with quality attributes of web applications. We use them as a mean to assure web application quality directly. Process metrics are related to WE process improvement. We use them as another mean to produce high-quality web applications, this time indirectly. Besides this, we use process metrics in a more practical way: to estimate effort of web application design [9].

Web product and process metrics are both in the focus of WE research. Product metrics related
to quality are presented in [7], [11] and [1] (among others), and process metrics are studied in [8] (among others). Ivory et al. computed statistical features of HTML metrics measured in web pages judged to be the best by international experts [7]. These statistics can serve as guidelines for future web designs. Vittorini and Di Felice did a similar study on web pages from the specific domain of courseware and web sites whose relevance were testified by the work carried out by both international organizations and research projects [11]. Counter to the authors of [7] and [11] who considered metrics of final web applications, Abrahao et al. studied metrics of models used in web design as early indicators of web application maintainability [1]. On the other hand, Mendes et al. analyzed process metrics that help web designers to estimate design and authoring effort in web projects [8].

Metrics are important even in the case when the other point of view on the Web is considered: the user’s point of view. Users are interested in retrieving a nugget of information they need from large bodies of information. Metrics can facilitate this information retrieval by classifying web pages into different categories [2].

The number of web metrics proposed is large. Some of them are proposed starting from the GQM (Goal-Question-Metric) paradigm [10] and theoretically validated [1] and the others are proposed in a less formal way [7], [11], [8], [2]. The most of them are empirically validated [7], [11], [8], [1], [2].

It is important to notice that the research on web metrics already done provides evidence that even simple HTML tag based metrics can be useful as product metrics [7] and in classification for information retrieval [2]. The other conclusion is that metrics values differ between different topics on the Web (educational, financial, etc.) [7], [11].

In this paper we study a set of web sites in a locale domain of the Web – Serbia, in terms of HTML metrics in order to compare them with the results shown in [7]. Our goal is to find out if web sites in Serbia are similar to those coming from a more advanced domain of the Web or not. In other words, we want to determine if our web pages are of high-quality. To answer this question we consider web sites from two topics: educational and financial topics. The results obtained are encouraging although further improvements are possible. Our first findings give us an empirically validated and relatively objective conclusion on the quality attributes of the web pages selected.

This paper is structured as follows. Section 2 gives more details about the work that has inspired us to this analysis mostly [7]. Section 3 presents the way how we collected data in our study. Section 4 compares our results to those given in [7]. Section 5 shows the additional analysis on our set of web pages in terms of HTML metrics. Finally, Section 6 describes our conclusions.

2 Starting Work

Ivory et al. considered about two thousand web pages gathered from web sites included in the competition “The Webby Awards 2000” [7]. These web pages were rated by international experts through a rigorous evaluation process in order to select the winners. Eleven simple HTML metrics were extracted from the web pages, including the number of words on a page, the number of hyperlinks, the number of graphics and the number of clusters. A specially automated tool was developed to compute the eleven metrics. Web sites from six topical categories were selected including educational and financial topics. The home page and the web pages from different levels in each web site were used.

The overall score of evaluation was used to define two groups of web pages for the further analysis: “good pages” (top 33% of web pages) and “not good pages” (bottom 67% of web pages). The goal was to determine if there were significant differences between the two groups – both overall and within each category. Further, predictive models for identifying good web pages were made. The results showed that the classification accuracy was higher when categories were taken into account.

The analysis revealed that the number of words was significantly correlated with the most of the other HTML metrics and thus was used to divide web pages into three groups depending on their size: small pages, medium pages and large pages. Then the means and standard deviations of the metrics for small, medium and large pages were computed. A part of these results are presented in Table 1. The significant results were found for the number of graphics for small web
The general conclusion was that metrics were good predictors of web pages rated to be good by experts. In addition, the analysis gave the empirical validation that different metrics were important for web sites from different topics.

Table 1 Means and standard deviations for good and not good web pages for small, medium and large web pages according to [7]

<table>
<thead>
<tr>
<th>Metric</th>
<th>Small Pages</th>
<th>Medium Pages</th>
<th>Large Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (Standard Dev.)</td>
<td>Mean (Standard Dev.)</td>
<td>Mean (Standard Dev.)</td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>11.4 (15.0)</td>
<td>24.88 (26.2)</td>
<td>0.005 25.3 (25.8)</td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>74.6 (16.0)</td>
<td>35.68 (36.2)</td>
<td>0.202 61.1 (51.7)</td>
</tr>
</tbody>
</table>

3 Data Collection

In our study we selected web sites from educational and financial topics in Serbia. Educational web sites were web sites of nineteen schools and faculties of the University of Belgrade [4]: Faculty of Biology, Faculty of Sports and Physical Education, School of Electrical Engineering, Faculty of Political Sciences, Faculty of Pharmacy, Faculty of Security, Faculty of Philosophy, Faculty of Physics, Faculty of Organizational Sciences, Faculty of Geography, Faculty of Civil Engineering, Faculty of Chemistry, Faculty of Mechanical Engineering, Faculty of Mathematics, Faculty of Agriculture, Faculty of Law, Faculty of Forestry, Faculty of Technology and Metallurgy and Teacher’s Training Faculty. This was not a complete set of the members of the University but those that had similar web site designs.

We considered the home page and a number of web pages from the first level of each educational web site. The collection consisted of 200 pages and of an approximate average of 10 web pages per site. The number of web pages is in accordance with the numbers of pages for different categories given in [7] and the number of web documents from the case study presented in [11]. The average is also in accordance with the average of 11 pages per site shown in [7].

Financial web sites were web sites of twenty-nine commercial banks in Serbia according to the list of the National Bank of Serbia [5]. It is a subset of the complete list of about forty banks that is suitable for our analysis. We measured the home page and a number of web pages from different levels of each financial site. This collection consisted of 235 web pages and of an approximate average of 8 web pages per site. This is similar to the numbers computed for the educational web sites.

To compute HTML metrics of web pages we used a specially developed tool named SMeterH. This tool extracts seventeen simple metrics from HTML code automatically. The metrics can be divided into three groups: (i) line-based metrics, (ii) tag-based and element-based metrics, and (iii) special metrics.

The first group of line-based metrics includes the number of characters, the number of all the lines and the number of empty lines of a page. The third group of special metrics includes the number of tables, the number of graphics and the number of hyperlinks.

The second group of tag-based and element-based metrics implemented in SMeterH was created on the basis of the specification of HTML 4.01 [3]. It includes the number of tags of a page, such as <html>, and the number of elements, such as <title></title>. Generally, an element has a start tag and an end tag and thus it forms “a paired element”. In addition, there are some exceptions: an empty element (an element without the end tag) is such an exception. Because of this, the tool distinguishes paired elements, such as <a></a>, from empty elements, such as <hr>. Another exception is related to
paired elements. Some of them must have start
tags and end tags (<a></a>, for instance) and the
others have optional end tags (<p>, for instance).
We named elements from the first category
“mandatory paired elements”. We separated
elements from the second category into two
groups: “fully paired elements” if there is an end
tag in an element, and “partially paired elements”
if there is no such a tag. The tool counts elements
from these groups separately.

The intersection between the set of eleven
metrics used in the previous study [7] and the set
of seventeen metrics extracted by SMeterH gives
two metrics: (i) the number of graphics and (ii)
the number of hyperlinks. Thus the previous
study can be repeated on our collection of web
pages in terms of these two metrics only.

4 Comparison with the Starting
Work
To compare the metrics values of web pages in
Serbia with those in [7], we both divided our
educational web pages and financial pages into
three groups concerning their size: small pages,
medium pages and large pages. This separation
was based upon the number of characters of a
page. We decided to use the number of characters
since this metric is the most similar (in our
opinion) to the number of words used in [7] for
the same purpose. To divide pages we used the
PROBE method defined in the Personal Software
Process – PSP [6], although in a modified way.
Originally, the PROBE method is introduced for
estimating software size in the PSP. According
to the method, the sample of programs is divided
into five groups. In our analysis the sample of
web pages was divided into three groups.

Our set of educational web pages was divided
in such a way that those that had the number of
characters between 8298 and 19869 were put into
medium pages, those that had the number less
than 8298 characters were put into small pages
and those that had the number more than 19869
characters were put into large pages. On the other
hand, our set of financial web pages was divided
in such a way that the range between 10744 and
27223 characters was determined for medium
pages.

The means and standard deviations of the
metrics were then computed for small, medium
and large pages for our educational web pages
(Table 2) and financial web pages (Table 3).

The additional goal of our study is to
determine if there are special groups of schools or
banks that are better than others when web page
quality is considered and to find out such groups
if they are present. In order to meet the additional

table 2. means and standard deviations for
educational web pages

<table>
<thead>
<tr>
<th>Metric</th>
<th>Small Pages</th>
<th>Medium Pages</th>
<th>Large Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Schools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>3.68</td>
<td>7.07</td>
<td>19.44</td>
</tr>
<tr>
<td>(5.06)</td>
<td>(6.7)</td>
<td>(22.76)</td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>11.86</td>
<td>24.92</td>
<td>41.76</td>
</tr>
<tr>
<td>(10.71)</td>
<td>(16.97)</td>
<td>(36.05)</td>
<td></td>
</tr>
<tr>
<td><strong>Schools of Sciences</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>2.85</td>
<td>5.55</td>
<td>6.36</td>
</tr>
<tr>
<td>(2.75)</td>
<td>(6.69)</td>
<td>(7.66)</td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>11.73</td>
<td>25.1</td>
<td>35</td>
</tr>
<tr>
<td>(9.59)</td>
<td>(22.64)</td>
<td>(52.06)</td>
<td></td>
</tr>
<tr>
<td><strong>Schools of Engineering</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>9.15</td>
<td>16.4</td>
<td></td>
</tr>
<tr>
<td>(6.94)</td>
<td>(9.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>32.5</td>
<td>57.6</td>
<td></td>
</tr>
<tr>
<td>(15.6)</td>
<td>(15.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Schools of Social Sciences and Hum.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>(14.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Schools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4.39)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>22.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5.27)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IT Oriented Schools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>5.88</td>
<td>15.09</td>
<td></td>
</tr>
<tr>
<td>(6.94)</td>
<td>(11.34)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>19.88</td>
<td>45.09</td>
<td></td>
</tr>
<tr>
<td>(10.53)</td>
<td>(35.53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Other Schools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Graphics</td>
<td>2.29</td>
<td>6.24</td>
<td>17.61</td>
</tr>
<tr>
<td>(2.73)</td>
<td>(5.48)</td>
<td>(24.29)</td>
<td></td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td>6.81</td>
<td>22.85</td>
<td>37.55</td>
</tr>
<tr>
<td>(7.28)</td>
<td>(12.33)</td>
<td>(38)</td>
<td></td>
</tr>
</tbody>
</table>
we assumed that better groups of banks and their web pages could be (i) the group of foreign banks in Serbia and (ii) the group of banks with the best financial reports according to the list of the National Bank of Serbia. Thus we further divided our educational web pages in two ways. First, we divided schools and their belonging web pages into four groups: schools of sciences, schools of social sciences and humanities, schools of engineering sciences and the others. Second, we separated schools and their web pages into two groups: schools oriented towards IT at the most and the others. The results of statistical computations for these groups of schools are also presented in Table 2. Empty fields in Table 2 correspond to small samples of web pages not suitable for statistical computations.

In accordance with the additional goal, we divided financial web pages in two ways. First, we divided banks and their belonging web pages into four groups: foreign banks and domestic banks. Second, we separated banks and their web pages into two groups: banks with the best financial reports and the others. We computed statistics of metrics for all these groups of banks. From all of the results, only the results of domestic banks are given in Table 3 (numbers signed with *).

<table>
<thead>
<tr>
<th>Metric</th>
<th>Small Pages (Mean / Standard Deviation)</th>
<th>Medium Pages (Mean / Standard Deviation)</th>
<th>Large Pages (Mean / Standard Deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Graphics</td>
<td>11.80 / 12.90* (11.70 / 11.60*)</td>
<td>29.40 / 27.20* (15.10 / 12.80*)</td>
<td>56.20 / 58.40* (38.10 / 39.30*)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.33 / 9.09* (9.22 / 9.26*)</td>
<td>35.40 / 32.90* (25.10 / 19.00*)</td>
</tr>
<tr>
<td>Number of Hyperlinks</td>
<td></td>
<td></td>
<td>62.80 / 66.60* (38.70 / 34.40*)</td>
</tr>
</tbody>
</table>

Table 3 Means and standard deviations for financial web pages (domestic banks*)

To compare the results from our educational web pages to those from [7] we look at Table 2 and Table 1. The significant results given in Table 1 (bolded numbers) are the most important. Table 2 shows that most of the numbers computed for all the schools of the University of Belgrade is not comparable to those in Table 1. The exception is the number of hyperlinks on large pages that is close to the appropriate number for “not good pages” presented in Table 1. When the first way of dividing schools on four groups is considered, the main conclusion is that the number of hyperlinks on large pages extracted from the schools of engineering is close to the appropriate number for “good pages” presented in Table 1: the difference is 5.7%. This is true for the mean values only. The standard deviations are not comparable. When the second way of dividing schools on two groups is considered (IT oriented and others), no special conclusion can be given: the results are similar to those presented for all the schools.

To compare the results from our financial web pages to those from [7] we look at Table 3 and Table 1. Once again, the significant results given in Table 1 are the most important. When the results given for all the web pages are considered, the conclusions are the following:

- the number of graphics on small pages is close to the appropriate number for “good pages” presented in Table 1 (the difference is 3.5%), and
- the number of hyperlinks on large pages is close to the appropriate number for “good pages” presented in Table 1 (the difference is 2.8%).

These are true for the mean values only. When the standard deviations are taken into account, the differences are greater: 19% for the number of graphics on small pages and 23% for the number of hyperlinks on large pages.

When the first way of dividing banks is considered, the conclusion based upon the results shown for domestic banks is that the numbers deviate further from those given in Table 1. Thus we indirectly conclude that foreign banks in Serbia have better web pages than domestic ones. When the second way of dividing banks is considered, the computation done on the group of banks with the best financial reports (not given in
Table 3) shows that the results deviate further from those presented in Table 1. Thus we conclude that this group of banks do not specially contribute to web page quality.

5 Additional Analysis
Collecting HTML metrics from our web pages with SMeterH enabled us to compute the statistics of all the metrics supported by the tool. We used the results of SMeterH to investigate the presence of fully paired HTML elements in web pages even in the cases when partially paired elements are possible. We did this examination because we wanted to determine if the recommendation on discouraging web designers from using partially paired elements given in [3] was followed or not. We conducted the examination on financial web pages and obtained the following results: six banks produced all the web pages with fully paired elements, thirteen did not, and ten banks did it partially. In this third case, there were web pages with fully paired elements and those without them in the same web site. A possible explanation for this case could be the following: different development processes were used to prepare different pages during different times (some developers followed the recommendation and the others did not).

6 Conclusion
The analysis performed on nineteen schools of the University of Belgrade in terms of two HTML metrics of web pages, the number of graphics and the number of hyperlinks, showed that the schools of engineering are the best in the context of web page quality. The analysis performed on twenty nine commercial banks in Serbia in terms of the same metrics showed that their small and large web pages are good and similar to the web pages presented in [7]. Foreign banks in Serbia contribute to this mostly.

The additional analysis performed on banks in Serbia concerning the presence of fully paired HTML elements in web pages showed that the web pages considered can be improved. It is important to notice that the study described in this paper could not be possible without the specially developed tool named SMeterH.

Our future work on the metric-based evaluation of web page quality will go in four directions:

a) we will cover additional topics besides education and finance;
b) we will periodically repeat our analysis on the web sites already measured to find out if there is any improvement or degradation;
c) we will pay attention to the best web sites and their web pages to investigate them in more details,
d) we will include more metrics in the analysis and improve the tool according to new and validated approaches proposed in literature.

References: