COMPUTER
ENGINEERING
PROGRAM

Middle East Technical
University
Northern Cyprus
Campus

eMINE Technical Report Deliverable 4 (D4),

August 2013

Evaluation of Automatic Discovery of Visual Elements of Web Pages

M. Elgin Akpınar
elgin.akpinar@metu.edu.tr
Middle East Technical University,
Ankara, Turkey

Yeliz Yesilada
yyeliz@metu.edu.tr
Middle East Technical University
Northern Cyprus Campus,
Kalkanlı, Güzelyurt, TRNC,
Mersin 10, Turkey

Web pages are typically designed for visual interaction – they include many visual elements to guide the reader. However, when they are accessed in alternative forms such as in audio, these visual elements are not available and therefore they become inaccessible. To address this problem, we have proposed an approach to identify visual elements in a web page and then characterize the semantic role of these web elements. The purpose of this technical report is to discuss the evaluation methodology in detail and present our results, which provide useful information into later applications of web page segmentation, heuristic role detection of web elements and web page transcoding. Our evaluation shows that, our segmentation algorithm has success rate above average and generally, users prefer the detailed segmentation levels to the simplistic segmentation levels. Moreover, our user evaluation on role detection shows that our proposed approach has around 80% receptive accuracy, but the proposed knowledge base could be further improved for better results.

METU
NCC

eMINE

The World Wide Web (web) has moved from the Desktop and now is ubiquitous. It can be accessed by a small device while the user is mobile or it can be accessed in audio if the user cannot see the content, for instance visually disabled users who use screen readers. However, since web pages are mainly designed for visual interaction; it is almost impossible to access them in alternative forms. Our overarching goal is to improve the user experience in such constrained environments by using a novel application of eye tracking technology. In brief, by relating scanpaths to the underlying source code of web pages, we aim to transcode web pages such that they are easier to access in constrained environments.

Acknowledgements

The project is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) with the grant number 109E251. As such the authors would like to thank to (TÜBİTAK) for their continued support. We would also like to thank our participants for their time and effort. Their input is appreciated in improving the proposed method and the knowledge base. We donated 5 TL to UNICEF Turkey for each participant who completed the survey. Finally, we wish to thank Assist. Prof. Dr. Aslı Niyazi and Cansu Telkes for their assistance with the statistics used in this report.

Contents

Middle East Technical University,
Northern Cyprus Campus,
Kalkanlı, Güzelyurt, TRNC,
Mersin 10, TURKEY

Tel: +90 (392) 661 2000
http://www.ncc.metu.edu.tr/

Corresponding author:
M. Elgin Akpınar
elgin.akpinar@metu.edu.tr

1 Introduction

Web pages include many visual elements for visual interaction and guidance for the users. As new web technologies evolve, they provide many utilities, tools and frameworks for designers and developers to create more interactive and, as a result, technically sophisticated web pages. However, this evaluation comes with accessibility issues. When pages are accessed in alternative forms such as in audio with assistive technologies, these visual elements are not available and therefore web pages become inaccessible. Having deep understanding of the semantic structure of a web page is very important for providing better accessibility. Once the roles of the visual elements in a web page layout are identified, they can be used to transcode the page by removing unnecessary elements or reorganise the page structure to improve the accessibility not only for disabled people but also for small screen device users. To achieve this goal, we proposed two methods to first identify meaningful visual elements in a web page[1], and then detect the roles of these elements in the web page[2].

The aim of this technical report is to present the scientific questions addressed in evaluation, the techniques used in evaluation of the data and the results of the evaluation. In order to address our scientific questions, we have conducted a survey based online user evaluation.

The technical report has been organised in the following way: Section 2 describes our survey based user evaluation, its procedure, the material which has been used in this process and the results of the evaluation. Section 3 gives information about the technical evaluation and discuss its results. Finally, Section 4 concludes the technical report.

2 User Evaluation

In order to evaluate our work on web page segmentation and heuristic role detection, we have conducted an online survey which has been available at `http://emine.ncc.metu.edu.tr/eval/survey/`. The main reason for doing this investigation online was to reach more people from different countries, different levels of expertise, age groups, education and professional backgrounds.

2.1 Procedure

We have designed and implemented an online web-based survey application which includes the following four main parts:

Information Sheet: First page in our study included an information sheet about the study. This page mainly included an overview of the study, and some information about the anonymity and the tasks to be completed.

Demographics: When somebody participated in our study, we collected some demographics information about them, for example their gender, experience in web design, education, age range, etc.

Visual Elements Identification: In this part, participants were shown a web page in different levels of segmentation, and they were asked to rate these levels and also rank the levels of segmentation. Based on the best segmentation they have chosen, in the next step they were asked to assign roles to the visual segments in that level.

Complexity	Pages
Low	home.mywebsearch.com/ (0.38),
	www.bing.com/ (0.33),
	www.apple.com/ (0.64),
	www.google.com.tr/ (1.1),
	adf.ly/ (1.43),
	imgur.com/ (1.58)
	wordpress.org/ (1.70),
	babylon.com/ (2.20),
	www.dailymotion.com/tr (2.40),
	www.conduit.com/ (2.70)
Medium	www.microsoft.com/en-us/default.aspx (3.08),
	www.avg.com/tr-tr/homepage (3.11),
	http://www.ebay.com/ (3.12),
	www.youtube.com/ (3.96),
	http://www.adobe.com/ (4.14),
	http://www.zedo.com/ (5.01),
	www.flickr.com/ (5.53),
	http://www.yahoo.com/ (6.43),
	www.alibaba.com/ (6.69),
	en.wikipedia.org/wiki/Main_Page (6.96)
High	www.aol.com/ (7.73),
	www.mediafire.com/ (7.74),
	www.craigslist.org/about/sites/ (8.57),
	ask.com/ (9.42),
	www.bbc.co.uk/ (9.93),
	http://www.uol.com.br/ (10),
	http://www.godaddy.com/ (10),
	http://www.about.com/#!/editors-picks/ (10),
	http://stackoverflow.com/ (10),
	http://www.huffingtonpost.com/ (10)

Table 1: Web pages used in the user study and their complexity scores

Discovering Roles: The participants were provided a list of roles in our knowledge base; however, if the participants could not find the proper role in the list, they were given the chance to specify the exact role of the block.

In overall, the survey application was designed to repeat the last two steps for randomly selected nine pages. The participants were free to leave the survey anytime they wanted; however, we expected them to evaluate at least one page from each complexity group. If a participant left the survey without evaluating at least three pages, we omitted his/her responses, since our study was designed to include pages with different complexity levels and we wanted to make sure that data that we analyse come from participants who evaluated at least one page from each complexity level.

2.2 Materials

The complete survey was designed to include nine randomly chosen web pages from a group of 30 page from different complexity levels, although the participants were not informed about the complexity group of the evaluated pages. In order to choose these 30 pages, we have investigated the complexity of top 100 web from Alexa by using the Visual Complexity Rankings and Accessibility Metrics (VICRAM) framework [3]. For a given

web page, VICRAM assigns a Visual Complexity Score (VCS). For 100 pages, we calculated their VCS and grouped these pages into three: low complexity – pages with VCS lower than three; medium complexity – pages with a VCS between three and seven; and high complexity – pages with a VCS above seven. These 100 pages were then grouped into three based on their VCS, we then randomly selected 10 pages from each complexity group. However, when we were preparing the data for our evaluation, we could not take a screenshot of one of the low complexity pages because it was too dynamic therefore we have 29 pages at the end. This approach gave us a systematic method for choosing pages with different levels of complexity. Table 1 lists these selected pages and their complexity scores.

Although we used top 5 levels of segmentation in our survey, some pages only have 3 or 4 levels since their structures are very simple. For example, `home.mywebsearch.com/` has only 3 levels; `www.bing.com/`, `www.google.com.tr/` and `babylon.com/` have only 4 levels. Although they must be handled separately than the pages with 5 or more levels, they are still significant for our evaluation since there are many pages which have simple structures. Therefore, we represent the results of the pages with 3 and 4 levels separately than the pages with 5 levels.

2.3 Administration

The evaluation conducted as an online survey. The survey application was developed by the author in PHP language with CodeIgniter framework and data stored in a MySQL database.

Our online study was released on the 7th of February 2013. The call for participation was distributed in a number of mailing lists which include CHI announcements, SIGWEB announcements, social media sites such as LinkedIn, Chamber of Computer Engineers in Turkey, and also sent to personal contacts and groups working on web design. Until 25th of July 2013, when the survey was closed for participation, 253 participants completed our study and 34 of them completed at least three pages in overall.

2.4 The Scientific Questions

In our evaluation process, we aimed to gather data to evaluate our proposed segmentation and heuristic role detection algorihthms, based on a set of scientific questions. The main questions addressed in this report are as follows:

1. How successfully does our algorithm divide pages into its segments in a specific level with respect to its previous level?
2. Which level is preferred by the participants among the first 5 levels of segmentation?
3. What is the effect of complexity and user profile on these results?
4. How successfully does our algorithm detects the semantic role of a visual element?

For each question we included the following details:

1. What do we try to find with this scientific question?
2. What is the context of the data which we have related with this scientific question?

Criteria	Value	Count	Percent (%)
Gender	Female	14	41.18
	Male	20	58.82
Age	Under 18	0	0.00
	18-24	8	23.53
	25-34	13	38.24
	35-54	11	32.35
	55+	2	5.88
Web usage	Daily	34	100.00
	Weekly	0	0.00
	Monthly	0	0.00
	Less than once a month	0	0.00
	Never	0	0.00
Current status in web design and development	Worked	25	73.53
	Studied	4	11.76
	Hobby	4	11.76
	Other	1	2.94
Level of expertise	Professional	13	38.24
	Intermediate	16	47.06
	Novice/Beginner	5	14.71
Education	Grade/Primary School	0	0.00
	High/Secondary School	6	17.65
	Associate's Degree	2	5.88
	Bachelor's Degree	6	17.65
	Master's Degree	11	32.35
	Doctorate	9	26.47
	Other	0	0.00

Table 2: Demographic summary of participants

3. Which techniques used to evaluate the results of this scientific question?

4. What are the results and how do we interpret them?

2.5 Participants

The survey results includes many participants with different profiles. Demographic distribution of the participants is given in Table 2. Of our 34 participants, 14 were female and 20 were male. Eight participants were aged between 18-24, 13 of them were 25-34, 11 of them were 25-54 and two of the participants were aged over 55. All of the participants use internet daily. 25 participants have worked in web design and development, four of them studied this subject, three of them are interested in web design as a hobby and one individual selected his current status as other. 13 participants describe their level of expertise in web design and development as professional, 16 individuals as intermediate and five of them as novice/beginner, which provides a balanced number of participants between professional and intermediate level of expertise. Six participants completed high/secondary school, two completed associate's degree, six completed bachelor's degree, 11 completed master's degree and nine completed doctorate.

2.6 Results

There are three complexity groups and each page is in one of these groups. The main assumption is that each participant completes the survey with nine pages in three complexity levels, but most of them left the survey without completing all the pages. However, our random page selection algorithm makes sure that all participants would evaluate pages from each complexity level in three consequitive pages and we only evaluate the data retrieved from the participants who evaluated at least three pages. In other words, we only take into consideration the participants who evaluated at least one page from each complexity levels.

In overall, 327 pages have been rated including 102 pages with low complexity, 113 pages with medium complexity and 111 pages with high complexity. 202 of pages were considered as valid, since their assigners satisfied the minimum requirement of labelling at least one page from all complexity groups, including 68 low, 66 medium and 68 high complexity pages. Morever, 1,946 role assignments have been made and 1,458 were considered in our evaluation as valid assignments, due to the number of pages evaluated by their assigners. 232 roles assigned to pages with low complexity, 580 roles were assigned to pages with medium complexity and 646 roles were assigned to pages with high complexity.

	3 Levels			4 Levels			5 Levels		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Level 1	8	2.125	0.8345	23	3.217	1.5063	171	2.322	1.2398
Level 2	8	3.375	1.0607	23	3.217	1.0426	171	2.795	1.1320
Level 3	8	3.875	1.2464	23	3.609	1.3052	171	3.298	0.9695
Level 4	-	-	-	23	3.609	1.0762	171	3.509	1.0311
Level 5	-	-	-	-	-	-	171	3.637	1.1770

Table 3: Overall rating results

2.6.1 Results of Segmentation Algorithm Evaluation

The Scientific Question I: How successfully does our algorithm divide pages into its segments in a specific level with respect to its previous level?

What we try to find: The participants of the survey are given top 5 levels of segmentation and asked to rate the success of the segmentation in a particular level. The participants had 5 choices: 1. Extremely Poor, 2. Below Average, 3. Average, 4. Above Average 5. Excellent. In this question, we aim to find how successfully our segmentation algorithm divides a web page into its visual blocks.

Context of the data: The data used in this question includes the rating results of each participant on 5 levels of segmentation.

Technique of evaluation: We simply calculate the mean and standart deviation values for the samples of rating results for each level.

Results: The results of the perceived success of our segmentation algorithm based on participant responses are given in Table 3.

Discussion: According to the results in Table 3, the success rate of our algorithm increases while the level goes deeper in the page structure. For the pages with only three levels, success rate of Level 1 is below average and success rates of Level 2 and Level 3 are above average. For the pages with only four levels, success rate of all levels are above average and it increases while the level number increases. Although mean of Level 1 and Level 2; Level 3 and Level 4 are equal, the standard deviation of Level 2 and Level 4 are lower; therefore, less variance occurs for Level 2 and Level 4. Finally, for the pages with five or more levels, the success rate of Level 1 and Level 2 are below average, while success rate of Level 3, Level 4 and Level 5 are above average. Similarly, the success rate increases while the number of levels increases. These results indicate that, segmentation in lower levels is rated as below average, while segmentation in middle or higher levels is rated as above average. The success rate, in general, increases as parallel to the level number.

	Low Complexity			Medium Complexity			High Complexity		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
Level 1	37	2.405	1.2793	66	2.349	1.2215	68	2.250	1.2504
Level 2	37	2.945	1.1772	66	2.833	1.1580	68	2.677	1.0851
Level 3	37	3.513	1.0440	66	3.303	0.9762	68	3.177	0.9133
Level 4	37	3.405	0.9267	66	3.606	1.1077	68	3.471	1.0144
Level 5	37	3.621	1.0633	66	3.576	1.2286	68	3.706	1.1977

Table 4: Rating results for complexity groups with 5 levels

The Scientific Question II: How successfully does our algorithm divide pages into its segments for a specific complexity group?

What we try to find: In this question, we try to find the perceived success of our segmentation algorithm for low, medium and high complexity pages separately.

Context of the data: The responses of participants includes page complexity information which are in three groups: low, medium and high complexity.

Technique of evaluation: Each participant was expected to evaluate 3 pages in each complexity group. Similar to the overall rating analysis, we calculated the mean and standart deviation results separately for each complexity group.

Results: The results of the perceived success of our segmentation algorithm with respect to the complexity of the pages, are given in Table 4.

Discussion: According to the results in Table 4, success rates increases while the level number increases, except for Level 4 of low complexity group and Level 5 of medium complexity group. Chart 5 represents the rating results for low, medium and high complexities with overall results. As can be seen visually in the chart, segmentation in lower levels is rated as below average while segmentation in middle and higher levels is rated as above average. It seems possible that this parallel increase is due to a correlation with level preferrence of the participants, which will be investigated with the Scientific Question XI.

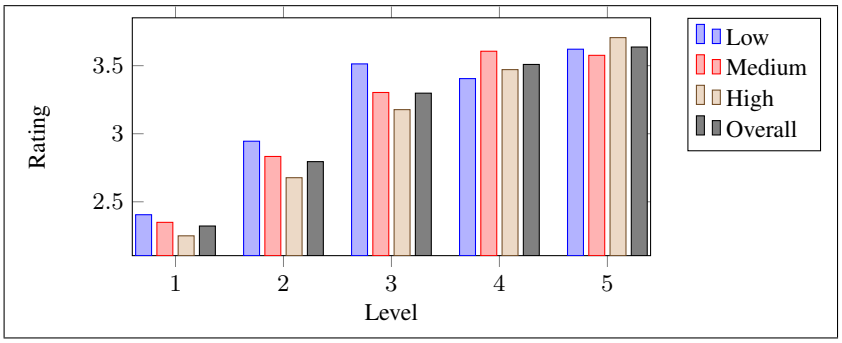

Chart 5: Bar chart for rating results

	Paired Differences							
				95% Confidence Interval of the Difference				
Pairs	Mean	Std. Dev.	Std. Err. Mean.	Lower	Upper	t	df	Sig. (2-tailed)
Low - Med. (L. 1)	0.412	1.083	0.186	0.034	0.790	2.218	33	0.034
Low - Med. (L. 2)	0.289	0.949	0.163	-0.042	0.620	1.776	33	0.085
Low - Med. (L. 3)	0.162	1.022	0.175	-0.195	0.518	0.923	33	0.363
Low - Med. (L. 4)	-0.167	1.135	0.204	-0.583	0.250	-0.817	30	0.420
Low - Med. (L. 5)	0.139	1.469	0.300	-0.482	0.759	0.463	23	0.648
Low - High (L. 1)	0.314	0.921	0.158	-0.008	0.635	1.986	33	0.055
Low - High (L. 2)	0.338	0.872	0.150	0.034	0.642	2.262	33	0.030
Low - High (L. 3)	0.255	1.075	0.184	-0.120	0.630	1.383	33	0.176
Low - High (L. 4)	0.022	1.045	0.188	-0.362	0.405	0.115	30	0.910
Low - High (L. 5)	0.160	1.217	0.248	-0.354	0.674	0.643	23	0.527
Med. - High (L. 1)	-0.098	0.789	0.135	-0.373	0.177	-0.724	33	0.474
Med. - High (L. 2)	0.049	0.795	0.136	-0.228	0.326	0.359	33	0.722
Med. - High (L. 3)	0.093	0.829	0.142	-0.196	0.382	0.655	33	0.517
Med. - High (L. 4)	0.172	1.043	0.179	-0.192	0.536	0.959	33	0.345
Med. - High (L. 5)	-0.113	1.129	0.194	-0.507	0.281	-0.582	33	0.564

Table 6: Paired T Test results on rating data

The Scientific Question III: What is the effect of complexity on rating responses?

What we try to find: In this question, we try to find whether participants have rated the levels with respect to the complexity of the page, or they rated them independent of the complexity.

Context of the data: For each level rating response, we know which page was rated by which participant. By using this information, we can group the responses with respect to the page complexity and participant. At the end of this pre-process, we have a list of participants and their responses for each level in complexity group.

Technique of evaluation: Each participant has been evaluated different number of pages. In order to reduce these numbers to equal numbers of pages for each participant, first of all we calculated the overall response of a participant for a specific level in each complexity groups. Then, by using SPSS, we applied Paired T Test over samples of same level but different complexity groups (low - medium, low - high and medium - high). Our significance level (α) is 0.05 our null hypothesis is that, there is a statistically significant differecence between the results of two complexity groups.

Results: The results of Paired T Test for each pair of levels, are shown in Table 6.

Discussion: According to the significance values in Table 6, most of which are higher than $\alpha = 0.05$ (expect for level 1 in low - medium complexities and level 2 in low - high complexities), we shall reject the null hypothesis; there is no statistically significant difference between the results of two complexity groups. In general, therefore, it seems that the complexity of a page is not a significant criteria for rating the segmen-

Participant Profile		Level 1			Level 2			Level 3		
		Mean	N	Std. Dev.	Mean	N	Std. Dev.	Mean	N	Std. Dev.
Age	18-24	2.33	3	0.58	3.33	3	1.15	4.00	3	1.73
	35-54	2.00	5	1.00	3.40	5	1.14	3.80	5	1.10
Education	High/Secondary School	2.00	2	0.00	3.00	2	1.41	3.50	2	2.12
	Associate's Deg.	1.00	1	0.00	3.00	1	0.00	3.00	1	0.00
	Bachelor's Deg.	3.00	2	0.00	4.00	2	0.00	5.00	2	0.00
	Master's Deg.	2.50	2	0.71	4.00	2	1.41	4.00	2	1.41
	Doctorate	1.00	1	0.00	2.00	1	0.00	3.00	1	0.00
Gender	Female	2.00	5	1.00	3.20	5	0.84	3.80	5	1.10
	Male	2.33	3	0.58	3.67	3	1.53	4.00	3	1.73
Level of Expertise	Professional	2.67	3	0.58	4.00	3	1.00	4.33	3	1.15
	Intermediate	2.00	3	1.00	3.33	3	1.15	4.33	3	1.15
	Novice/Beginner	1.50	2	0.71	2.50	2	0.71	2.50	2	0.71
Current Status	Worked	2.17	6	0.98	3.50	6	1.05	4.00	6	1.10
	Studied	2.00	1	0.00	4.00	1	0.00	5.00	1	0.00
	Other	2.00	1	0.00	2.00	1	0.00	2.00	1	0.00

Table 7: Rating results for pages with 3 levels with respect to user profiles

tation results. Another finding was that, for any complexity group, there is a clear trend of increasing in rating results with respect to the increasing in level number.

The Scientific Question IV: How does the success of our segmentation algorithm changes when we group the participants into their expertise, current status in web design, age, gender, highest education completed and internet usage?

What we try to find: At the beginning of the survey, we collect some demographic information of the participants. This information includes age, gender, education level, expertise in web design and development, current status involved in web design and usage of the internet. Using this demographic data, we can group the participants in different levels, such as their professional background. In this question, we try to find how different groups of participants perceived the success of our algorithm.

Context of the data: Each participant has a demographic profile in our database. When rating result table and participant profile table are joined on user ID column, it is possible to relate participant responses with the profile information of the participants. We can group the participants with respect to their age, gender, education, level of expertise and current status in web design and development.

Technique of evaluation: First of all, we group the participants based on a selected criteria, such as age or expertise. Then we apply the same technique in calculation of overall success on the reduced set of data. At the end, we find the mean value and standart deviation for the selected group of participants.

Results: The results of the perceived success of our segmentation algorithm, grouped by different participant profiles, are presented in Table 7, Table 8 and Table 9. The results are discussed with the results of the Scientific Question V.

Participant Profile		Level 1			Level 2			Level 3			Level 4		
		Mean	N	Std. Dev.	Mean	N	Std. Dev.	Mean	N	Std. Dev.	Mean	N	Std. Dev.
Age	18-24	2.75	4	1.71	2.50	4	1.00	3.00	4	1.41	3.75	4	0.50
	25-34	2.75	8	1.58	3.00	8	0.76	3.75	8	1.28	3.38	8	0.92
	35-54	3.78	9	1.48	3.78	9	1.20	4.00	9	1.32	4.00	9	1.32
	55+	3.50	2	0.71	3.00	2	0.00	2.50	2	0.71	2.50	2	0.71
Education	High/Secondary School	3.33	3	1.53	2.67	3	1.15	3.00	3	1.73	3.67	3	0.58
	Associate's Degree	3.00	1	0.00	4.00	1	0.00	4.00	1	0.00	5.00	1	0.00
	Bachelor's Degree	2.50	6	1.52	3.17	6	1.17	3.67	6	0.82	4.33	6	0.82
	Master's Degree	2.67	6	1.63	3.00	6	0.89	3.83	6	1.47	3.17	6	0.75
	Doctorate	4.00	7	1.15	3.57	7	1.13	3.57	7	1.62	3.43	7	1.51
Gender	Female	2.64	11	1.29	2.82	11	0.75	3.09	11	1.04	3.64	11	0.92
	Male	3.75	12	1.54	3.58	12	1.16	4.08	12	1.38	3.58	12	1.24
Level of Expertise	Professional	3.00	8	1.69	3.25	8	1.28	3.75	8	1.58	3.50	8	1.41
	Intermediate	3.42	12	1.38	3.25	12	0.97	3.50	12	1.17	3.75	12	0.97
	Novice/Beginner	3.00	3	2.00	3.00	3	1.00	3.67	3	1.53	3.33	3	0.58
Current Status	Worked	2.86	14	1.51	3.21	14	1.19	3.64	14	1.34	3.79	14	1.25
	Studied	4.33	3	1.15	3.33	3	1.15	3.67	3	1.53	3.67	3	0.58
	Hobby	3.50	6	1.52	3.17	6	0.75	3.50	6	1.38	3.17	6	0.75

Table 8: Rating results for pages with 4 levels with respect to user profiles

Participant Profile		Level 1			Level 2			Level 3			Level 4			Level 5		
		Mean	N	Std. Dev.	Mean	N	Std. Dev.	Mean	N	Std. Dev.	Mean	N	Std. Dev.	Mean	N	Std. Dev.
Age	18-24	2.14	35	1.54	2.46	35	1.24	3.14	35	1.09	3.54	35	1.20	3.86	35	1.24
	25-34	2.25	75	1.25	2.84	75	1.24	3.35	75	0.99	3.36	75	1.01	3.59	75	1.15
	35-54	2.53	51	1.06	2.88	51	0.91	3.29	51	0.90	3.78	51	0.88	3.76	51	1.16
	55+	2.40	10	0.70	3.20	10	0.63	3.50	10	0.71	3.10	10	1.10	2.60	10	0.70
Education	High/Secondary Sch.	2.84	19	1.71	3.00	19	1.29	3.47	19	1.07	3.95	19	1.08	3.84	19	1.34
	Associate's Degree	2.60	10	1.35	2.50	10	0.97	3.00	10	0.67	3.60	10	0.70	3.10	10	0.99
	Bachelor's Degree	1.67	36	0.89	2.33	36	0.96	3.19	36	0.92	3.72	36	1.03	4.28	36	0.97
	Master's Degree	2.37	57	1.17	2.89	57	1.22	3.16	57	1.10	3.12	57	1.12	3.16	57	1.19
	Doctorate	2.49	49	1.16	3.00	49	1.04	3.53	49	0.82	3.61	49	0.84	3.76	49	1.01
Gender	Female	2.11	74	1.20	2.64	74	1.01	3.19	74	0.87	3.59	74	0.96	3.72	74	1.15
	Male	2.48	97	1.25	2.92	97	1.20	3.38	97	1.04	3.44	97	1.08	3.58	97	1.20
Level of Expertise	Professional	2.20	71	1.21	2.68	71	1.22	3.11	71	1.14	3.14	71	1.19	3.31	71	1.30
	Intermediate	2.47	77	1.23	2.86	77	0.97	3.45	77	0.79	3.82	77	0.85	3.86	77	1.10
	Novice/Beginner	2.22	23	1.35	2.96	23	1.36	3.35	23	0.88	3.61	23	0.66	3.91	23	0.73
Current Status	Worked	2.08	124	1.10	2.63	124	1.06	3.25	124	1.02	3.50	124	1.12	3.66	124	1.26
	Studied	3.62	21	1.16	3.33	21	1.20	3.48	21	0.87	3.76	21	0.83	3.57	21	1.03
	Hobby	2.54	24	1.28	3.25	24	1.22	3.46	24	0.78	3.33	24	0.70	3.50	24	0.88
	Other	1.00	2	0.00	2.00	2	1.41	2.50	2	0.71	3.50	2	0.71	4.50	2	0.71

Table 9: Rating results for pages with 5 and more levels with respect to user profiles

	Level 1	Level 2	Level 3	Level 4	Level 5
Age	X^2(12, N=202)= 22.846, **p=0.029**	X^2(12, N=202)= 26.924, **p=0.008**	X^2(12, N=202)= 8.868, p=0.714	X^2(15, N=202)= 29.379, **p=0.014**	X^2(15, N=202)= 31.187, **p=0.008**
Education	X^2(16, N=202)= 34.291, **p=0.005**	X^2(16, N=202)= 15.224, p=0.508	X^2(16, N=202)= 23.734, p=0.095	X^2(20, N=202)= 34.527, **p=0.023**	X^2(20, N=202)= 41.885, **p=0.003**
Gender	X^2(4, N=202)= 7.618, p=0.107	X^2(4, N=202)= 10.343, **p=0.035**	X^2(4, N=202)= 10.677, **p=0.03**	X^2(5, N=202)= 4.186, p=0.523	X^2(5, N=202)= 2.774, p=0.735
Level	X^2(8, N=202)= 9.507, p=0.301	X^2(8, N=202)= 13.492, p=0.096	X^2(8, N=202)= 23.693, **p=0.003**	X^2(10, N=202)= 31.698, **p=0**	X^2(10, N=202)= 22.441, **p=0.013**
Status	X^2(12, N=202)= 38.231, **p=0**	X^2(12, N=202)= 17.747, p=0.124	X^2(12, N=202)= 10.165, p=0.601	X^2(15, N=202)= 27.267, **p=0.027**	X^2(15, N=202)= 16.647, p=0.34

Table 10: Pearson's Chi Square Test results for participant groups

The Scientific Question V: Is there a correlation between participant groups and rating responses?

What we try to find: In this question, we try to find whether some trends exist for some participant groups on segmentation rating responses.

Context of the data: Data consists of the participant profile for all criteria and their rating responses for each level of each page.

Technique of evaluation: For each criteria, such as age or education, we have applied Pearson's Chi Square Test to find whether a correlation exists between participant groups and their responses. Our significance level (α) is 0.05. Our null hypothesis is that there is no relationship between participant profiles and their responses. Our alternative hypothesis is that there is a relationship between participant profiles and their responses.

Results: The results of Pearson's Chi Square Tests for each criteria, are given in Table 10. The results which satisfy p-value $\leq \alpha$ are highlighted and they indicate that, null hypothesis can be rejected in their cases.

Discussion: As can be seen from the Table 10, age is a significant criteria in Level 1, Level 2, Level 4 and Level 5; education in Level 1, Level 4 and Level 5; gender in Level 2 and Level 3; level of expertise in Level 3, Level 4 and Level 5; and finally, current status in web design and development in Level 1 and Level 4. In this part, we only presented and discussed the results of the pages with 5 levels and the pages with 3 or 4 levels have similar results with these pages.

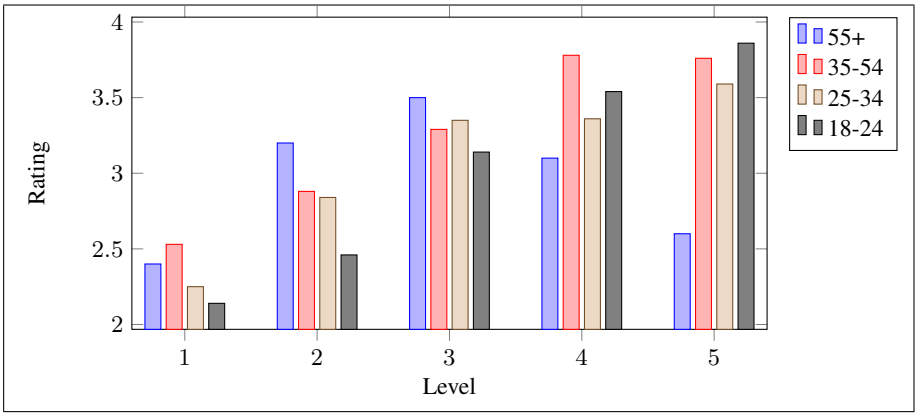

Chart 11: Bar chart for rating results grouped by age

Chart 11 shows the rating results categorised based on age criteria. From the chart we can see that, the participants aged between 18 and 24 gave the highest rates for Level 5, while the participants aged between 35 and 54 gave the highest rates for Level 4. For participants aged between 18 - 24 and 25-34, we can observe an increase on rating results when we go from Level 1 to Level 5. For participants aged between 35 and 54, the rating results increase until Level 4, but they decrease in Level 5. For participants over 55, the rating results increase until Level 3 and they decrease until Level 5 to a close rating with Level 1. It is apparent from this chart and Pearson's Chi-Square Test that, there is a significant difference between the rating results of different participant groups. The results of this analysis indicate that older participants rated our segmentation algorithm as more successful in middle levels, while younger participants rated our segmentation algorithm as more successful in higher levels.

The rating results which are categorised based on gender criteria are presented in Chart 12. From the data in the chart, it is apparent that, female participants gave higher rating results than male participants for higher levels, while male participants gave higher rating results than female participants for lower levels. The rating results for both female and male participants increase from Level 1 to Level 5.

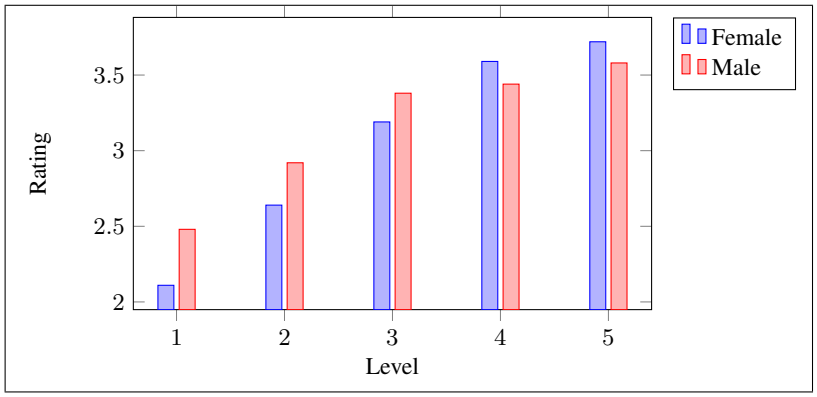

Chart 12: Bar chart for rating results grouped by gender

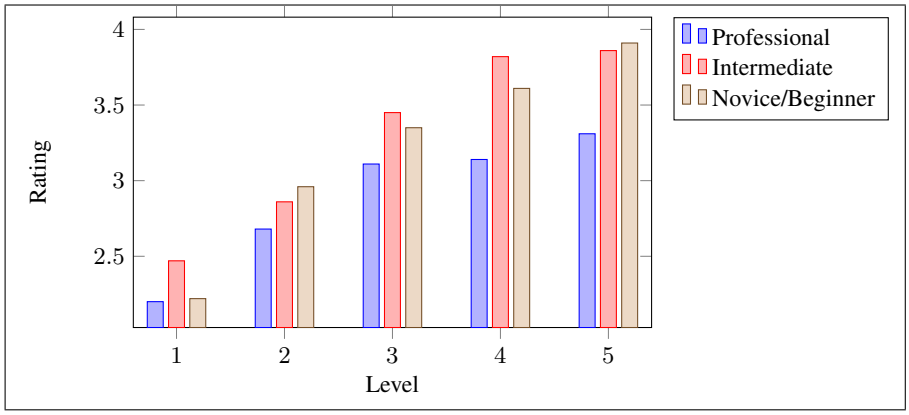

Chart 13: Bar chart for rating results grouped by level of expertise

The rating results categorised based on level of expertise criteria are provided in Chart 13. It is apparent from the chart that, for each participant group, the rating results increase from Level 1 to Level 5. Among three participant groups, the lowest rating results belong to professional participants. A possible explanation for this might be that, professional participants have been involved in web design and development, and have knowledge and experience more than other participant groups. Therefore, they may have responded in more selective and critical approach. As can be seen in the chart and Pearson's Chi-Square Test results suggest, rating results change based on level of expertise and professional participants rate our segmentation algorithm as less successful, when we compare the results to other expertise groups.

Chart 14 shows the rating results which are categorised based on current status of the participants involved in web design and development. It is important to note that, among 34 participants, only one of them selected his/her status as other. Therefore,

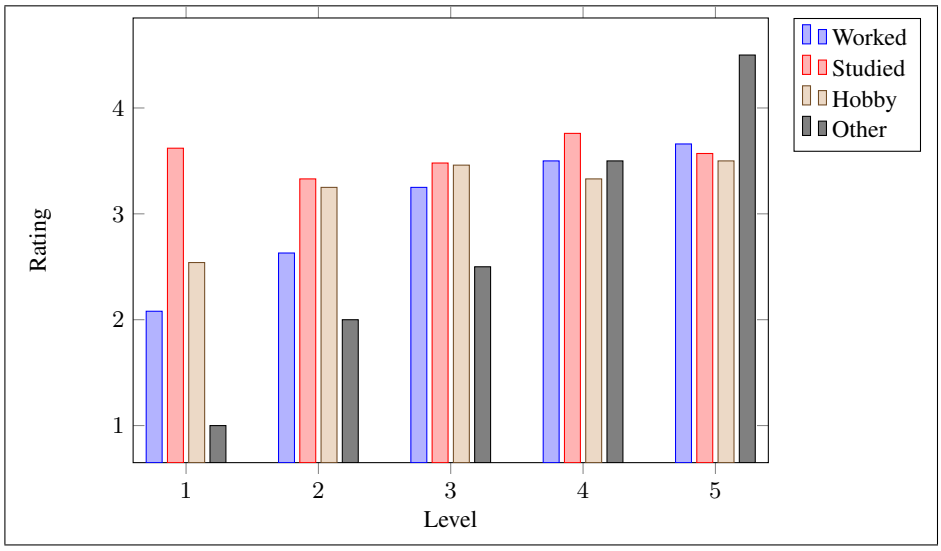

Chart 14: Bar chart for rating results grouped by current status

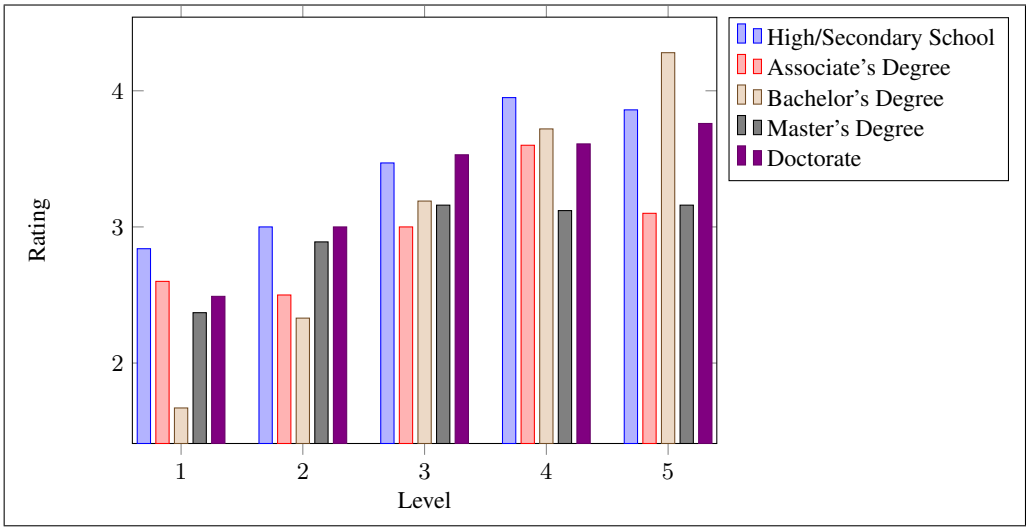

Chart 15: Bar chart for rating results grouped by education

the difference between the bar labeled as 'Other' and the other bars in the chart can be omitted. From this data we can see that, the survey conducted with the participants who have worked in web design and development, resulted in the lowest values of segmentation rating. A possible explanation for this is similar with the level of expertise criteria. The participants who worked in web design and development may have responded more selective and critical than others. It is somewhat surprising that, the rating results of the participants who have studied in this field are in the same range (between average and above average) for each level, unlike other results in which at least one level appears in a lower or higher range. It is difficult to explain this result, but it might be due to the fact that, this participant group examined segmentation process as a whole, rather than dividing the task to the levels separately.

Finally, Chart 15 illustrates the rating results categorised based on education criteria. The participants who completed high/secondary school gave the highest rating results for Level 4 among other participant groups and their rating results increase until Level 4, but decrease in Level 5. The rating results of associate's degree group decrease from Level 1 to Level 2, increase in Level 3 and Level 4, but decrease again in Level 5. Bachelor's degree group gave the highest rating results for Level 5 among other participant groups and their rating results increase from Level 1 to until Level 5. The rating results of master's degree group increase from Level 1 to Level 3, and Level 3, Level 4 and Level 5 have close values. Finally, the rating results of doctorate group increase from Level 1 to Level 5; however, Level 3, Level 4 and Level 5 have close values. The participants who completed bachelor's degree gave the highest rating results for Level 5 and lowest rating results for Level 1. Associate's degree and master's degree groups rated all levels as less successful than other groups. Although the rating results stay at the same ratio for other groups, it changes for bachelor's degree group. While the rating results for Level 1 and Level 2 are lower than other groups, they become higher at Level 3, Level 4 and Level 5.

Level	Pages with 3 levels			Pages with 4 levels			
	Best Level		Worst Level	Best Level			Worst Level
1	1 (12%)	0 (0%)	7 (87,5%)	6 (26.1%)	2 (8.7%)	2 (8.7%)	13 (56.5%)
2	3 (37,5%)	4 (50,0%)	1 (12,5%)	0 (0%)	6 (26.1%)	16 (69.6%)	1 (4.3%)
3	4 (50,0%)	4 (50,0%)	0 (0%)	5 (21.7%)	12 (52.2%)	4 (17.4%)	2 (8.7%)
4	-	-	-	12 (52.2%)	3 (13.0%)	1 (4.3%)	7 (30.4%)

Table 16: Ranking results for pages with 3 and 4 levels

	Best Level				**Worst Level**
Level 1	12 (7,0%)	14 (8,2%)	13 (7,6%)	18 (10,5%)	114 (66,7%)
Level 2	23 (13,5%)	15 (8,8%)	16 (9,4%)	108 (63,2%)	9 (5,3%)
Level 3	22 (12,9%)	23 (13,5%)	112 (65,5%)	10 (5,8%)	4 (2,3%)
Level 4	31 (18,1%)	92 (53,8%)	19 (11,1%)	22 (12,9%)	7 (4,1%)
Level 5	83 (48,5%)	27 (15,8%)	11 (6,4%)	13 (7,6%)	37 (21,6%)

Table 17: Ranking results for pages with 5 and more levels

The Scientific Question VI: Which level is preferred by the participants among the first 5 levels of segmentation?

What we try to find: The participants are provided top 5 levels of segmentation and asked to rank the levels based on their best and worst level selection. We try to find which level is selected as the best level by the majority of the participants. The result of this part may be used in transoding application, when we need to decide on which segmentation level is most suitable for segmentation and transcoding.

Context of the data: Similar to the previous question, the data in this question includes the ranking results of each participant for 5 levels. The ranking values ranges from 1 to 5, where 1 indicates the best level and 5 indicates the worst level.

Technique of evaluation: In this part, we have calculated the frequencies and percentages of ranking results for all responses.

Results: The results of participant preference on level selection are given in Table 16 and Table 17.

Discussion: It is apparent from Table 16 and Table 17 that, participants selected highest available level as the best level, and lowest level as worst level. It is therefore likely that participants preferred higher levels as their best level of segmentation than lower levels. This result may be related to the fact that, higher levels have more detailed segmentation and simpler segments, while lower levels are consisted of larger segments. This finding has important implications for developing a transcoding method based on web page segmentation, since it suggests the most preferred level of segmentation.

The Scientific Question VII: Which level is preferred by the participants among the first 5 levels of segmentation for a specific complexity group?

What we try to find: The aim of this question is to calculate participant ranking results for low, medium and high complexity pages separately.

Compl.	Level	Best Level				Worst Level
Low	1	2 (5.4%)	4 (10.8%)	2 (5.4%)	6 (16.2%)	23 (62.2%)
	2	6 (16.2%)	4 (10.8%)	5 (13.5%)	21 (56.8%)	1 (2.7%)
	3	7 (18.9%)	7 (18.9%)	19 (51.4%)	3 (8.1%)	1 (2.7%)
	4	5 (13.5%)	20 (54.1%)	6 (16.2%)	4 (10.8%)	2 (5.4%)
	5	17 (45.9%)	2 (5.4%)	5 (13.5%)	3 (8.1%)	10 (27.0%)
Medium	1	4 (6.1%)	5 (7.6%)	8 (12.1%)	7 (10.6%)	42 (63.6%)
	2	10 (15.2%)	5 (7.6%)	6 (9.1%)	42 (63.6%)	3 (4.5%)
	3	8 (12.1%)	8 (12.1%)	45 (68.2%)	3 (4.5%)	2 (3.0%)
	4	18 (27.3%)	31 (47.0%)	4 (6.1%)	10 (15.2%)	3 (4.5%)
	5	26 (39.4%)	17 (25.8%)	3 (4.5%)	4 (6.1%)	16 (24.2%)
High	1	6 (8.8%)	5 (7.4%)	3 (4.4%)	5 (7.4%)	49 (72.1%)
	2	7 (10.3%)	6 (8.8%)	5 (7.4%)	45 (66.2%)	5 (7.4%)
	3	7 (10.3%)	8 (11.8%)	48 (70.6%)	4 (5.9%)	1 (1.5%)
	4	8 (11.8%)	41 (60.3%)	9 (13.2%)	8 (11.8%)	2 (2.9%)
	5	40 (58.8%)	8 (11.8%)	3 (4.4%)	6 (8.8%)	11 (16.2%)

Table 18: Ranking results for complexity groups

Context of the data: The data includes ranking results for each page and page complexity information which are in three groups: low, medium and high complexity.

Technique of evaluation: Each participant was expected to evaluate 3 pages in each complexity group. In this part, we have calculated the frequencies and percentages of ranking results for all responses separately for each complexity group.

Results: The results of participant preference on level selection according to the page complexities, are given in Tab. 18.

Discussion: From the Table 18 we can see that, for each complexity group, participants selected highest level as their best level, and lowest level as their worst level. Contrary to our expectations, this analysis did not find a significant difference between best level preferences for different complexity groups. However, further work is required to establish this.

The Scientific Question VIII: What is the effect of complexity on level ranking?

What we try to find: In this question, we aim to investigate the effect of complexity in level ranking results of the participants. In other words, we are trying to find whether participants ranked differently with respect to the complexity of the page or they tend to rank the levels independent of the complexity of the page.

Context of the data: Each ranking record in our database is related with a page ID and each page has a complexity group. Therefore, it is possible to group the ranking results based on the complexity groups. The data obtained from overall ranking evaluation consists of five sets for 1st, 2nd, 3rd, 4th and 5th best level of segmentation for each levels, which represents the number of participants who ranked a level in a specific place.

Technique of evaluation: In order to find the effect of complexity on best level selection, we compared to sets of results from different complexities. Therefore, we used T-Test

	Paired Differences							
				95% Confidence Interval of the Difference				
	Mean	Std. Dev.	Std. Err. Mean	Lower	Upper	t	df	Sig. (2-tailed)
High - Med. (L.1)	0.081	1.359	0.173	-0.264	0.426	0.467	61	0.642
High - Med. (L.2)	0.194	1.114	0.141	-0.089	0.476	1.368	61	0.176
High - Med. (L.3)	0.000	0.905	0.115	-0.230	0.230	0.000	61	1.000
High - Med. (L.4)	0.129	1.048	0.133	-0.137	0.395	0.970	61	0.336
High - Med. (L.5)	-0.403	1.273	0.162	-0.727	-0.080	-2.493	61	0.015
High - Low. (L.1)	0.000	1.495	0.253	-0.514	0.514	0.000	34	1.000
High - Low. (L.2)	0.411	1.523	0.204	0.003	0.819	2.018	55	0.048
High - Low. (L.3)	0.097	0.953	0.121	-0.145	0.339	0.799	61	0.427
High - Low. (L.4)	0.097	1.445	0.184	-0.270	0.464	0.527	61	0.600
High - Low. (L.5)	-0.210	1.747	0.222	-0.653	0.234	-0.945	61	0.348
Med. - Low (L.1)	-0.143	1.630	0.275	-0.703	0.417	-0.519	34	0.607
Med. - Low (L.2)	0.214	1.436	0.192	-0.170	0.599	1.117	55	0.269
Med. - Low (L.3)	0.095	1.027	0.129	-0.163	0.354	0.736	62	0.465
Med. - Low (L.4)	-0.032	1.459	0.184	-0.399	0.336	-0.173	62	0.863
Med. - Low (L.5)	0.190	1.674	0.211	-0.231	0.612	0.903	62	0.370

Table 19: Pairwise T Test results on ranking data

in calculation. The returned results are about 0.05 and 0.03; however, we realised that, the participants who evaluated both sets are the same and two sets are related with each other. Finally, we decided to use pairwise T-Test for this relation. Our significance level (α) is 0.05. Our null hypothesis is that, there is a relation between complexity group and ranking results.

Results: The results of Paired T Test for each pair of levels, are presented in Tab. 19. As table shows, at the α = 0.05 level of significance, we can conclude that there is no statistically significant difference between two complexities and the differences between complexity means are likely due to chance and not likely due to the IV manipulation. Therefore, we shall reject the null hypothesis.

Discussion: As the results in Table 19 indicate, there is no statistically significant difference between two complexities. This finding was unexpected and suggests that complexity is not a major factor in best level preference of the participants.

The Scientific Question IX: How do expertise, current status in web design, age, gender, highest education completed and usage of the internet affect the preference on the best segmentation level?

What we try to find: As in the previous question, we aim to find the decision of different participant groups on best level of segmentation.

Context of the data: Similar to the previous question, we have the ranking results of participants which can be grouped based on given participant profiles.

Technique of evaluation: The technique which will be used in this part is similar to the previous technique, expect that, we have grouped the data according to the participant profiles.

Results: The results of the participant preference on level selection, grouped by different participant profiles, are given in Table 20, Table 21, Table 22, Table 23 and Table 24. These results are discussed with the results of the Scientific Question X.

			Level 1	Level 2	Level 3
Age	18-24	Best Level	1 (33.3%)	1 (33.3%)	1 (33.3%)
		2nd Level	0 (0%)	1 (33.3%)	**2 (66.7%)**
		Worst Level	**2 (66.7%)**	1 (33.3%)	0 (0%)
	35-54	Best Level	0 (0%)	2 (40%)	**3 (60%)**
		2nd Level	0 (0%)	**3 (60%)**	2 (40%)
		Worst Level	**5 (100%)**	0 (0%)	0 (0%)
Gender	Female	Best Level	0 (0%)	1 (20%)	**4 (80%)**
		2nd Level	0 (0%)	**4 (80%)**	1 (20%)
		Worst Level	**5 (100%)**	0 (0%)	0 (0%)
	Male	Best Level	1 (33.3%)	**2 (66.7%)**	0 (0%)
		2nd Level	0 (0%)	0 (0%)	**3 (100%)**
		Worst Level	**2 (66.7%)**	1 (33.3%)	0 (0%)
Education	High/Secondary Sch.	Best Level	0 (0%)	1 (50%)	1 (50%)
		2nd Level	0 (0%)	1 (50%)	1 (50%)
		Worst Level	**2 (100%)**	0 (0%)	0 (0%)
	Bachelor's Degree	Best Level	0 (0%)	**1 (100%)**	0 (0%)
		2nd Level	0 (0%)	0 (0%)	**1 (100%)**
		Worst Level	**1 (100%)**	0 (0%)	0 (0%)
	Master's Degree	Best Level	1 (50%)	0 (0%)	1 (50%)
		2nd Level	0 (0%)	1 (50%)	1 (50%)
		Worst Level	1 (50%)	1 (50%)	0 (0%)
	Associate's Degree	Best Level	0 (0%)	1 (50%)	1 (50%)
		2nd Level	0 (0%)	1 (50%)	1 (50%)
		Worst Level	**2 (100%)**	0 (0%)	0 (0%)
	Doctorate	Best Level	0 (0%)	0 (0%)	**1 (100%)**
		2nd Level	0 (0%)	**1 (100%)**	0 (0%)
		Worst Level	**1 (100%)**	0 (0%)	0 (0%)
Level of Expertise	Professional	Best Level	1 (33.3%)	1 (33.3%)	1 (33.3%)
		2nd Level	0 (0%)	1 (33.3%)	**2 (66.7%)**
		Worst Level	**2 (66.7%)**	1 (33.3%)	0 (0%)
	Intermediate	Best Level	0 (0%)	0 (0%)	**3 (100%)**
		2nd Level	0 (0%)	**3 (100%)**	0 (0%)
		Worst Level	**3 (100%)**	0 (0%)	0 (0%)
	Novice/Beginner	Best Level	0 (0%)	**2 (100%)**	0 (0%)
		2nd Level	0 (0%)	0 (0%)	**2 (100%)**
		Worst Level	**2 (100%)**	0 (0%)	0 (0%)
Current Status	Worked	Best Level	1 (16.7%)	2 (33.3%)	3 (50%)
		2nd Level	0 (0%)	**3 (50%)**	3 (50%)
		Worst Level	**5 (83.3%)**	1 (16.7%)	0 (0%)
	Studied	Best Level	0 (0%)	0 (0%)	**1 (100%)**
		2nd Level	0 (0%)	**1 (100%)**	0 (0%)
		Worst Level	**1 (100%)**	0 (0%)	0 (0%)
	Other	Best Level	0 (0%)	**1 (100%)**	0 (0%)
		2nd Level	0 (0%)	0 (0%)	**1 (100%)**
		Worst Level	**1 (100%)**	0 (0%)	0 (0%)

Table 20: Ranking results for pages with 3 levels, with respect to user profiles

			Level 1	Level 2	Level 3	Level 4
Age	”18-24”	Best Level	0 (0.0%)	0 (0.0%)	1 (25.0%)	**3 (75.0%)**
		2nd Level	**2 (50.0%)**	0 (0.0%)	1 (25.0%)	1 (25.0%)
		3rd Level	1 (25.0%)	**3 (75.0%)**	0 (0.0%)	0 (0.0%)
		Worst Level	1 (25.0%)	1 (25.0%)	**2 (50.0%)**	0 (0.0%)
	”25-34”	Best Level	2 (25.0%)	0 (0.0%)	2 (25.0%)	**4 (50.0%)**
		2nd Level	0 (0.0%)	1 (12.5%)	**6 (75.0%)**	1 (12.5%)
		3rd Level	0 (0.0%)	**7 (87.5%)**	0 (0.0%)	1 (12.5%)
		Worst Level	**6 (75.0%)**	0 (0.0%)	0 (0.0%)	2 (25.0%)
	”35-54”	Best Level	3 (33.3%)	0 (0.0%)	1 (11.1%)	5 (55.6%)
		2nd Level	0 (0.0%)	3 (33.3%)	**5 (55.6%)**	1 (11.1%)
		3rd Level	0 (0.0%)	**6 (66.7%)**	3 (33.3%)	0 (0.0%)
		Worst Level	**6 (66.7%)**	0 (0.0%)	0 (0.0%)	3 (33.3%)
	”55+”	Best Level	1 (50.0%)	0 (0.0%)	1 (50.0%)	0 (0.0%)
		2nd Level	0 (0.0%)	**2 (100.0%)**	0 (0.0%)	0 (0.0%)
		3rd Level	1 (50.0%)	0 (0.0%)	1 (50.0%)	0 (0.0%)
		Worst Level	0 (0.0%)	0 (0.0%)	0 (0.0%)	**2 (100.0%)**
Gender	Female	Best Level	1 (9.1%)	0 (0.0%)	3 (27.3%)	**7 (63.6%)**
		2nd Level	1 (9.1%)	2 (18.2%)	**6 (54.5%)**	2 (18.2%)
		3rd Level	2 (18.2%)	**8 (72.7%)**	1 (9.1%)	0 (0.0%)
		Worst Level	**7 (63.6%)**	1 (9.1%)	1 (9.1%)	2 (18.2%)
	Male	Best Level	5 (41.7%)	0 (0.0%)	2 (16.7%)	5 (41.7%)
		2nd Level	1 (8.3%)	4 (33.3%)	**6 (50.0%)**	1 (8.3%)
		3rd Level	0 (0.0%)	**8 (66.7%)**	3 (25.0%)	1 (8.3%)
		Worst Level	**6 (50.0%)**	0 (0.0%)	1 (8.3%)	5 (41.7%)
Education	High/Secondary School	Best Level	0 (0.0%)	0 (0.0%)	1 (33.3%)	**2 (66.7%)**
		2nd Level	**2 (66.7%)**	0 (0.0%)	0 (0.0%)	1 (33.3%)
		3rd Level	1 (33.3%)	**2 (66.7%)**	0 (0.0%)	0 (0.0%)
		Worst Level	0 (0.0%)	1 (33.3%)	**2 (66.7%)**	0 (0.0%)
	Bachelor’s Degree	Best Level	**1 (100.0%)**	0 (0.0%)	0 (0.0%)	0 (0.0%)
		2nd Level	0 (0.0%)	**1 (100.0%)**	0 (0.0%)	0 (0.0%)
		3rd Level	0 (0.0%)	0 (0.0%)	**1 (100.0%)**	0 (0.0%)
		Worst Level	0 (0.0%)	0 (0.0%)	0 (0.0%)	**1 (100.0%)**
	Master’s Degree	Best Level	0 (0.0%)	0 (0.0%)	0 (0.0%)	**6 (100.0%)**
		2nd Level	0 (0.0%)	0 (0.0%)	**6 (100.0%)**	0 (0.0%)
		3rd Level	0 (0.0%)	**6 (100.0%)**	0 (0.0%)	0 (0.0%)
		Worst Level	**6 (100.0%)**	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Associate’s Degree	Best Level	1 (16.7%)	0 (0.0%)	3 (50.0%)	2 (33.3%)
		2nd Level	0 (0.0%)	1 (16.7%)	3 (50.0%)	2 (33.3%)
		3rd Level	0 (0.0%)	**5 (83.3%)**	0 (0.0%)	1 (16.7%)
		Worst Level	**5 (83.3%)**	0 (0.0%)	0 (0.0%)	1 (16.7%)
	Doctorate	Best Level	**4 (57.1%)**	0 (0.0%)	1 (14.3%)	2 (28.6%)
		2nd Level	0 (0.0%)	**4 (57.1%)**	3 (42.9%)	0 (0.0%)
		3rd Level	1 (14.3%)	3 (42.9%)	3 (42.9%)	0 (0.0%)
		Worst Level	2 (28.6%)	0 (0.0%)	0 (0.0%)	**5 (71.4%)**

Table 21: Ranking results for pages with 4 levels, with respect to user profiles (Part 1: Age, gender and education)

			Level 1	**Level 2**	**Level 3**	**Level 4**
Level of Expertise	Professional	Best Level	2 (25.0%)	0 (0.0%)	2 (25.0%)	**4 (50.0%)**
		2nd Level	0 (0.0%)	3 (37.5%)	**4 (50.0%)**	1 (12.5%)
		3rd Level	0 (0.0%)	**5 (62.5%)**	2 (25.0%)	1 (12.5%)
		Worst Level	**6 (75.0%)**	0 (0.0%)	0 (0.0%)	2 (25.0%)
	Intermediate	Best Level	3 (25.0%)	0 (0.0%)	2 (16.7%)	**7 (58.3%)**
		2nd Level	2 (16.7%)	3 (25.0%)	**6 (50.0%)**	1 (8.3%)
		3rd Level	2 (16.7%)	**8 (66.7%)**	2 (16.7%)	0 (0.0%)
		Worst Level	**5 (41.7%)**	1 (8.3%)	2 (16.7%)	4 (33.3%)
	Novice/Beginner	Best Level	1 (33.3%)	0 (0.0%)	1 (33.3%)	1 (33.3%)
		2nd Level	0 (0.0%)	0 (0.0%)	**2 (66.7%)**	1 (33.3%)
		3rd Level	0 (0.0%)	**3 (100.0%)**	0 (0.0%)	0 (0.0%)
		Worst Level	**2 (66.7%)**	0 (0.0%)	0 (0.0%)	1 (33.3%)
Current Status	Worked	Best Level	2 (14.3%)	0 (0.0%)	2 (14.3%)	**10 (71.4%)**
		2nd Level	1 (7.1%)	3 (21.4%)	**9 (64.3%)**	1 (7.1%)
		3rd Level	0 (0.0%)	**11 (78.6%)**	2 (14.3%)	1 (7.1%)
		Worst Level	**11 (78.6%)**	0 (0.0%)	1 (7.1%)	2 (14.3%)
	Studied	Best Level	1 (33.3%)	0 (0.0%)	1 (33.3%)	1 (33.3%)
		2nd Level	1 (33.3%)	1 (33.3%)	0 (0.0%)	1 (33.3%)
		3rd Level	1 (33.3%)	1 (33.3%)	1 (33.3%)	0 (0.0%)
		Worst Level	0 (0.0%)	1 (33.3%)	1 (33.3%)	1 (33.3%)
	Hobby	Best Level	**3 (50.0%)**	0 (0.0%)	2 (33.3%)	1 (16.7%)
		2nd Level	0 (0.0%)	2 (33.3%)	**3 (50.0%)**	1 (16.7%)
		3rd Level	1 (16.7%)	**4 (66.7%)**	1 (16.7%)	0 (0.0%)
		Worst Level	2 (33.3%)	0 (0.0%)	0 (0.0%)	**4 (66.7%)**

Table 22: Ranking results for pages with 4 levels, with respect to user profiles (Part 2: Level of experise and current status in web design and development)

			Level 1	Level 2	Level 3	Level 4	Level 5
Age	18-24	Best Level	2 (5.7)	4 (11.4)	1 (2.9)	3 (8.6)	**25 (71.4)**
		2nd Level	4 (11.4)	2 (5.7)	3 (8.6)	**24 (68.6)**	2 (5.7)
		3rd Level	1 (2.9)	2 (5.7)	**27 (77.1)**	4 (11.4)	1 (2.9)
		4th Level	1 (2.9)	**26 (74.3)**	4 (11.4)	2 (5.7)	2 (5.7)
		Worst Level	**27 (77.1)**	1 (2.9)	0 (0)	2 (5.7)	5 (14.3)
	25-34	Best Level	4 (5.3)	15 (20.0)	12 (16.0)	13 (17.3)	**31 (41.3)**
		2nd Level	8 (10.7)	9 (12.0)	10 (13.3)	**36 (48.0)**	12 (16.0)
		3rd Level	7 (9.3)	1 (1.3)	**51 (68.0)**	10 (13.3)	6 (8.0)
		4th Level	7 (9.3)	**45 (60.0)**	2 (2.7)	15 (20.0)	6 (8.0)
		Worst Level	**49 (65.3)**	5 (6.7)	0 (0)	1 (1.3)	20 (26.7)
	35-54	Best Level	6 (11.8)	2 (3.9)	5 (9.8)	13 (25.5)	**25 (49.0)**
		2nd Level	1 (2.0)	4 (7.8)	6 (11.8)	**27 (52.9)**	13 (25.5)
		3rd Level	5 (9.8)	9 (17.6)	**32 (62.7)**	3 (5.9)	2 (3.9)
		4th Level	5 (9.8)	**33 (64.7)**	4 (7.8)	5 (9.8)	4 (7.8)
		Worst Level	**34 (66.7)**	3 (5.9)	4 (7.8)	3 (5.9)	7 (13.7)
	55+	Best Level	0 (0)	2 (20.0)	**4 (40.0)**	2 (20.0)	2 (20.0)
		2nd Level	1 (10.0)	0 (0)	**4 (40.0)**	**5 (50.0)**	0 (0)
		3rd Level	**5 (50.0)**	**4 (40.0)**	2 (20.0)	2 (20.0)	2 (20.0)
		4th Level	0 (0)	**4 (40.0)**	0 (0)	0 (0)	1 (10.0)
		Worst Level	4 (40.0)	0 (0)	0 (0)	1 (10.0)	**5 (50.0)**
Gender	Female	Best Lev.	3 (4.1)	9 (12.2)	9 (12.2)	15 (20.3)	**38 (51.4)**
		2nd Lev.	7 (9.5)	6 (8.1)	6 (8.1)	**42 (56.8)**	13 (17.6)
		3rd Lev.	3 (4.1)	5 (6.8)	**55 (74.3)**	7 (9.5)	4 (5.4)
		4th Lev.	6 (8.1)	**53 (71.6)**	4 (5.4)	8 (10.8)	3 (4.1)
		Worst Lev.	**55 (74.3)**	1 (1.4)	0 (0)	2 (2.7)	16 (21.6)
	Male	Best Lev.	9 (9.3)	14 (14.4)	13 (13.4)	16 (16.5)	**45 (46.4)**
		2nd Lev.	7 (7.2)	9 (9.3)	17 (17.5)	**50 (51.5)**	14 (14.4)
		3rd Lev.	10 (10.3)	11 (11.3)	**57 (58.8)**	12 (12.4)	7 (7.2)
		4th Lev.	12 (12.4)	**55 (56.7)**	6 (6.2)	14 (14.4)	10 (10.3)
		Worst Lev.	**59 (60.8)**	8 (8.2)	4 (4.1)	5 (5.2)	21 (21.6)
Level of Expertise	Professional	Best Level	4 (5.6)	9 (12.7)	13 (18.3)	10 (14.1)	**35 (49.3)**
		2nd Level	7 (9.9)	9 (12.7)	6 (8.5)	**38 (53.5)**	11 (15.5)
		3rd Level	7 (9.9)	5 (7.0)	**46 (64.8)**	8 (11.3)	5 (7.0)
		4th Level	8 (11.3)	**46 (64.8)**	3 (4.2)	12 (16.9)	2 (2.8)
		Worst Level	**45 (63.4)**	2 (2.8)	3 (4.2)	3 (4.2)	18 (25.4)
	Intermediate	Best Level	6 (7.8)	8 (10.4)	8 (10.4)	17 (22.1)	**38 (49.4)**
		2nd Level	5 (6.5)	5 (6.5)	11 (14.3)	**43 (55.8)**	13 (16.9)
		3rd Level	4 (5.2)	10 (13.0)	**51 (66.2)**	7 (9.1)	5 (6.5)
		4th Level	9 (11.7)	**49 (63.6)**	6 (7.8)	6 (7.8)	7 (9.1)
		Worst Level	**53 (68.8)**	5 (6.5)	1 (1.3)	4 (5.2)	14 (18.2)
	Novice/ Beginner	Best Level	2 (8.7)	6 (26.1)	1 (4.3)	4 (17.4)	**10 (43.5)**
		2nd Level	2 (8.7)	1 (4.3)	6 (26.1)	**11 (47.8)**	3 (13.0)
		3rd Level	2 (8.7)	1 (4.3)	**15 (65.2)**	4 (17.4)	1 (4.3)
		4th Level	1 (4.3)	**13 (56.5)**	1 (4.3)	4 (17.4)	4 (17.4)
		Worst Level	**16 (69.6)**	2 (8.7)	0 (0)	0 (0)	5 (21.7)

Table 23: Ranking results for pages with 5 and more levels, with respect to user profiles (Part 1: Age, gender and level of experise)

			Level 1	Level 2	Level 3	Level 4	Level 5
Education	High/ Secondary School	Best Level	2 (10.5)	4 (21.1)	1 (5.3)	3 (15.8)	**9 (47.4)**
		2nd Level	4 (21.1)	2 (10.5)	3 (15.8)	**8 (42.1)**	2 (10.5)
		3rd Level	1 (5.3)	2 (10.5)	**11 (57.9)**	4 (21.1)	1 (5.3)
		4th Level	1 (5.3)	**10 (52.6)**	4 (21.1)	2 (10.5)	2 (10.5)
		Worst Level	**11 (57.9)**	1 (5.3)	0 (0)	2 (10.5)	5 (26.3)
	Bachelor's Degree	Best Level	0 (0)	0 (0)	1 (2.8)	2 (5.6)	**33 (91.7)**
		2nd Level	0 (0)	0 (0)	0 (0)	**34 (94.4)**	2 (5.6)
		3rd Level	0 (0)	1 (2.8)	**35 (97.2)**	0 (0)	0 (0)
		4th Level	0 (0)	**35 (97.2)**	0 (0)	0 (0)	1 (2.8)
		Worst Level	**36 (100.0)**	0 (0)	0 (0)	0 (0)	0 (0)
	Master's Degree	Best Level	5 (8.8)	12 (21.1)	10 (17.5)	11 (19.3)	**19 (33.3)**
		2nd Level	9 (15.8)	8 (14.0)	8 (14.0)	**22 (38.6)**	10 (17.5)
		3rd Level	6 (10.5)	4 (7.0)	**34 (59.6)**	9 (15.8)	4 (7.0)
		4th Level	5 (8.8)	**30 (52.6)**	4 (7.0)	14 (24.6)	4 (7.0)
		Worst Level	**32 (56.1)**	3 (5.3)	1 (1.8)	1 (1.8)	**20 (35.1)**
	Associate's Degree	Best Level	2 (20.0)	0 (0)	1 (10.0)	**5 (50.0)**	2 (20.0)
		2nd Level	0 (0)	2 (20.0)	4 (40.0)	2 (20.0)	2 (20.0)
		3rd Level	1 (10.0)	**4 (40.0)**	**5 (50.0)**	0 (0)	0 (0)
		4th Level	1 (10.0)	**4 (40.0)**	0 (0)	2 (20.0)	**3 (30.0)**
		Worst Level	**6 (60.0)**	0 (0)	0 (0)	1 (10.0)	**3 (30.0)**
	Doctorate	Best Level	3 (6.1)	7 (14.3)	9 (18.4)	10 (20.4)	**20 (40.8)**
		2nd Level	1 (2.0)	3 (6.1)	8 (16.3)	**26 (53.1)**	11 (22.4)
		3rd Level	5 (10.2)	5 (10.2)	**27 (55.1)**	6 (12.2)	6 (12.2)
		4th Level	11 (22.4)	**29 (59.2)**	2 (4.1)	4 (8.2)	3 (6.1)
		Worst Level	**29 (59.2)**	5 (10.2)	3 (6.1)	3 (6.1)	9 (18.4)
Current Status	Worked	Best Level	5 (4.0)	10 (8.1)	16 (12.9)	22 (17.7)	**71 (57.3)**
		2nd Level	8 (6.5)	9 (7.3)	11 (8.9)	**76 (61.3)**	20 (16.1)
		3rd Level	8 (6.5)	10 (8.1)	**88 (71.0)**	10 (8.1)	8 (6.5)
		4th Level	12 (9.7)	**89 (71.8)**	5 (4.0)	13 (10.5)	5 (4.0)
		Worst Level	**91 (73.4)**	6 (4.8)	4 (3.2)	3 (2.4)	20 (16.1)
	Studied	Best Level	4 (19.0)	**6 (28.6)**	3 (14.3)	5 (23.8)	3 (14.3)
		2nd Level	6 (28.6)	5 (23.8)	3 (14.3)	3 (14.3)	4 (19.0)
		3rd Level	3 (14.3)	3 (14.3)	**11 (52.4)**	3 (14.3)	1 (4.8)
		4th Level	1 (4.8)	6 (28.6)	4 (19.0)	**7 (33.3)**	3 (14.3)
		Worst Level	**7 (33.3)**	1 (4.8)	0 (0)	3 (14.3)	**10 (47.6)**
	Hobby	Best Level	3 (12.5)	7 (29.2)	3 (12.5)	4 (16.7)	**7 (29.2)**
		2nd Level	2 (8.3)	1 (4.2)	9 (37.5)	**11 (45.8)**	3 (12.5)
		3rd Level	0 (0)	3 (12.5)	**11 (45.8)**	6 (25.0)	2 (8.3)
		4th Level	5 (20.8)	**11 (45.8)**	1 (4.2)	2 (8.3)	5 (20.8)
		Worst Level	**14 (58.3)**	2 (8.3)	0 (0)	1 (4.2)	**7 (29.2)**
	Other	Best Level	0 (0)	0 (0)	0 (0)	0 (0)	**2 (100.0)**
		2nd Level	0 (0)	0 (0)	0 (0)	**2 (100.0)**	0 (0)
		3rd Level	0 (0)	0 (0)	**2 (100.0)**	0 (0)	0 (0)
		4th Level	0 (0)	**2 (100.0)**	0 (0)	0 (0)	0 (0)
		Worst Level	**2 (100.0)**	0 (0)	0 (0)	0 (0)	0 (0)

Table 24: Ranking results for pages with 5 and more levels, with respect to user profiles (Part 2: Education and current status in web design and development)

	Level 1	Level 2	Level 3	Level 4	Level 5
Age	X^2(12, N=171)= 27.16, p=0.007	X^2(12, N=171)= 30.792, p=0.002	X^2(12, N=171)= 32.882, p=0.001	X^2(12, N=171)= 16.446, p=0.172	X^2(12, N=171)= 24.678, p=0.016
Education	X^2(16, N=171)= 42.207, p=0	X^2(16, N=171)= 42.045, p=0	X^2(16, N=171)= 38.734, p=0.001	X^2(16, N=171)= 50.59, p=0	X^2(16, N=171)= 50.114, p=0
Gender	X^2(4, N=171)= 5.923, p=0.205	X^2(4, N=171)= 6.441, p=0.169	X^2(4, N=171)= 7.465, p=0.113	X^2(4, N=171)= 1.907, p=0.753	X^2(4, N=171)= 2.848, p=0.583
Level	X^2(8, N=171)= 3.138, p=0.925	X^2(8, N=171)= 9.346, p=0.314	X^2(8, N=171)= 10.282, p=0.246	X^2(8, N=171)= 6.602, p=0.58	X^2(8, N=171)= 6.649, p=0.575
Status	X^2(12, N=171)= 30.239, p=0.003	X^2(12, N=171)= 26.214, p=0.01	X^2(12, N=171)= 24.43, p=0.018	X^2(12, N=171)= 29.14, p=0.004	X^2(12, N=171)= 29.294, p=0.004

Table 25: Pearson's Chi Square Test results for participant groups

The Scientific Question X: Is there a correlation between participant groups and ranking responses?

What we try to find: In this question, we try to find whether some trends exist for some participant groups on segmentation ranking responses.

Context of the data: Data consists of the participant profile for all criteria and their ranking responses for each level of each page.

Technique of evaluation: For each criteria, such as age or education, we have applied Pearson's Chi Square Test to find whether a correlation exists between participant groups and their responses. Our significance level (α) is 0.05. Our null hypothesis is that there is no relationship between participant profiles and their responses. Our alternative hypothesis is that there is a relationship between participant profiles and their responses.

Results: The results of Pearson's Chi Square Tests for each criteria, are given in Tab. 25. The results which satisfy p-value $\leq \alpha$ are highlighted and they indicate that, null hypothesis can be rejected in their cases.

Discussion: According to the results in Table 25, age, education and current status in web design and development seems to be significant criteria in level ranking.

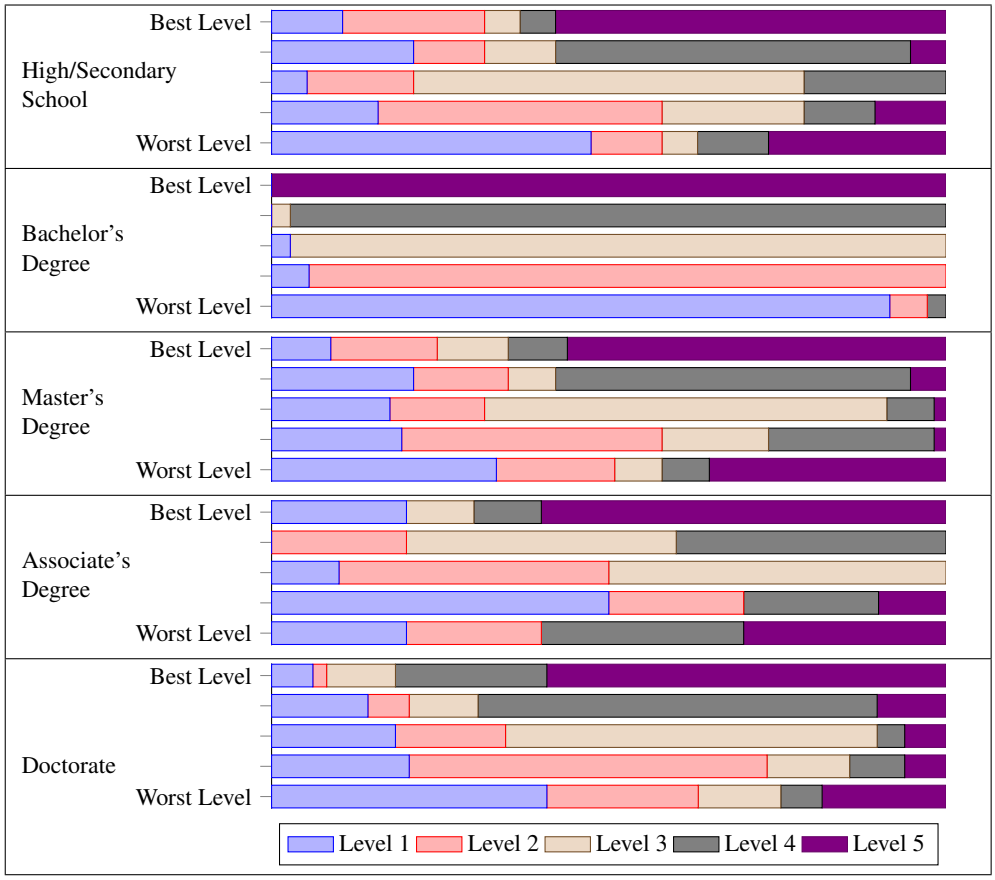

Chart 26: Stacked bar chart for ranking results grouped by education

Chart 26 shows the ranking results for pages with 5 and higher levels which are categorised based on education criteria. According to the results, participants who completed high/secondary school, bachelor's degree, master's degree or doctorate selected Level 5 as best level and Level 1 as worst level. On the other hand, the participants who completed associate's degree selected Level 4 as their best level and Level 1 as their worst level. Unlike the pages with 5 levels, participants who completed bachelor's degree or doctorate have selected Level 1 as their best level and Level 4 as their worst level. Therefore, age could be a major factor along with page complexity, and this finding is supported with Pearson's Chi-Square Test results.

Chart 27 represents the ranking results for pages with 5 and higher levels which are categorised based on age criteria. Participants aged between 18 - 24, 25 - 34 and 35- 54, have selected Level 5 as their best level and Level 1 as their worst level. On the other hand, participants aged over 55 have selected Level 3 as their best level and Level 5 as their worst level. For pages with only 3 levels, participants aged between 18 and 24 have selected Level 2 as the best level and Level 1 as the worst level, while participants aged between 35 and 54 followed the general trend of highest level as the best level and lowest level as the worst level. For pages with only 4 levels, participants aged between 18 and 24 have selected Level 4 as the best level and Level 3 as the worst level. There are only two participants who are

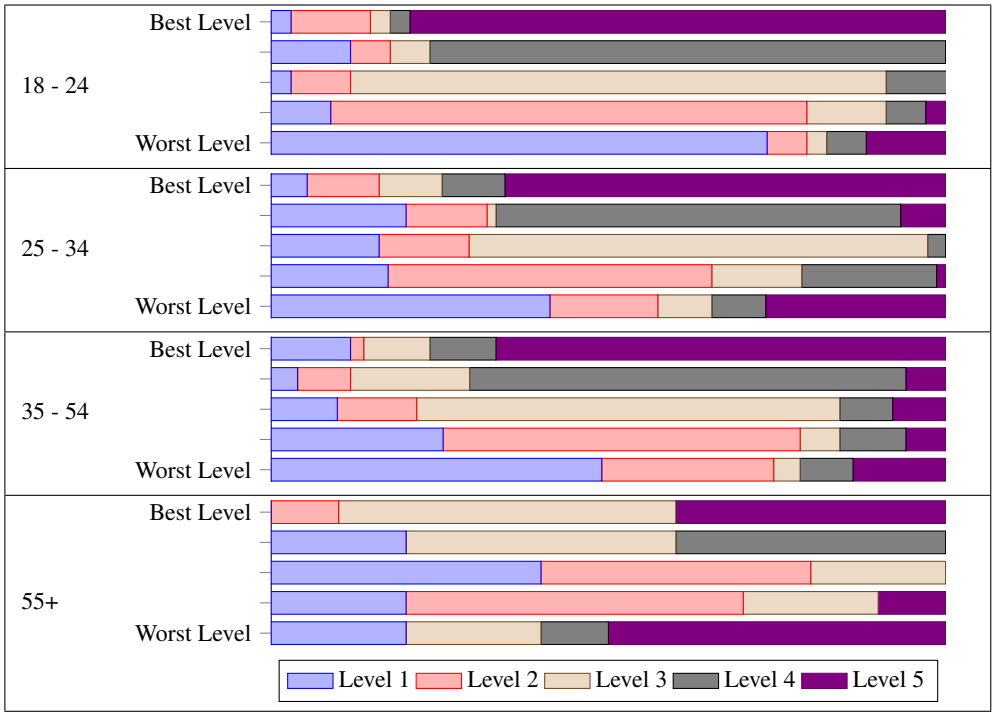

Chart 27: Stacked bar chart for ranking results grouped by age

over 55 and have ranked a page with only 4 levels. Although they agree on Level 5 as the worst level, they disagree on the best level since one of them selected Level 1 while the other selected Level 3. According to the Pearson's Chi-Square Test results, there was a significant positive correlation between age and best level preference of the participants. These findings suggest that the participant aged over 55 have different level preference than the others.

Chart 28 illustrates the ranking results for pages with 5 and higher levels which are categorised based on gender criteria. Both female and male participants selected Level 5 as their best level and Level 1 as their worst level. According to Pearson's Chi-Square Test

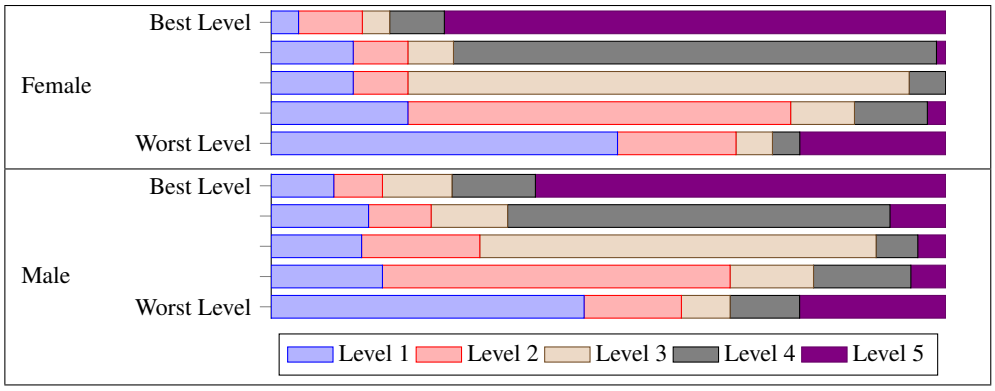

Chart 28: Stacked bar chart for ranking results grouped by gender

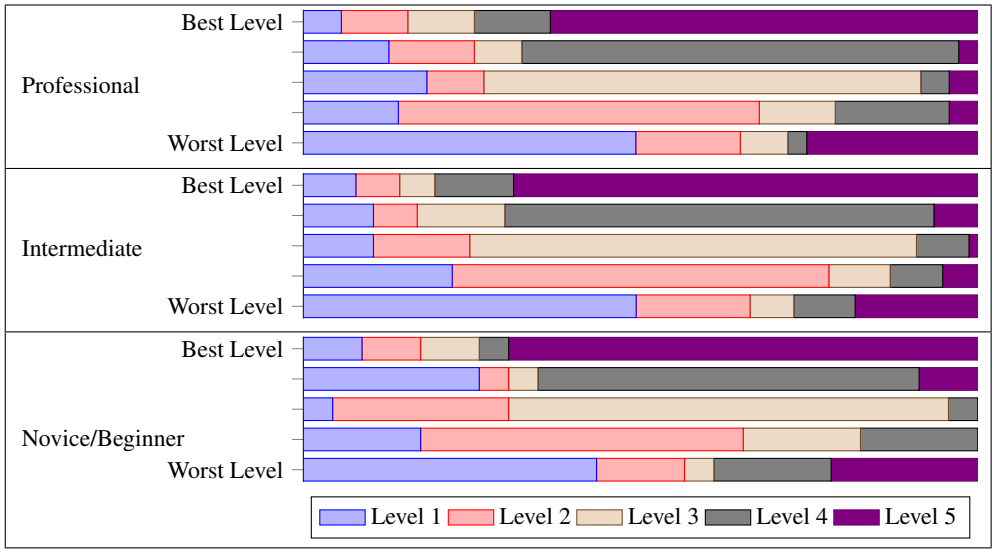

Chart 29: Stacked bar chart for ranking results grouped by level of expertise

results, no significant differences were found between best level preferences of female and male participants.

Chart 29 represents the ranking results for pages with 5 and higher levels which are categorised based on level of expertise criteria. Professional, intermediate and novice/beginner participants have selected Level 5 as the best level and Level 1 as the worst level. Similar to the gender criteria, there is no difference in the preference of three participant groups, therefore, level of expertise is not a significant criteria in best level preference.

Finally, the ranking results for pages with 5 and higher levels which are categorised based on current status involved in web design and development criteria, are shown in Chart 30. According to these results, all participant groups except for the participants who have studied in this field, have selected Level 5 as their best level and Level 1 as their worst level. Unlike the other groups, the participants who have studied web design and development have selected Level 2 as the best level and Level 5 as the worst level. As Pearson's Chi-Square Test results indicate, there was a significant positive correlation between current status and best level preference. It is difficult to explain this result, but it might be related to a simplistic approach of the participants who studied web design.

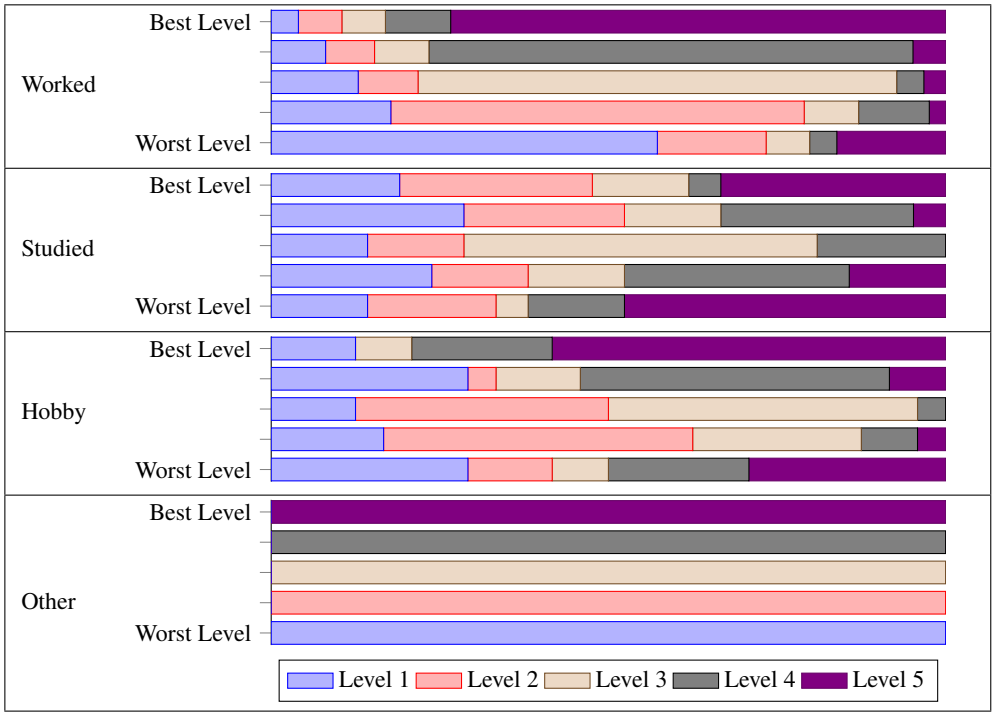

Chart 30: Stacked bar chart for ranking results grouped by current status

		Best Level	2nd Level	3rd Level	4th Level	Worst Level	Total
Excellent	Count	70	23	14	7	5	119
	Expected Count	23,8	23,8	23,8	23,8	23,8	119,0
	% within Ranking	40,9%	13,5%	8,2%	4,1%	2,9%	13,9%
Above Average	Count	67	78	48	24	11	228
	Expected Count	45,6	45,6	45,6	45,6	45,6	228,0
	% within Ranking	39,2%	45,6%	28,1%	14,0%	6,4%	26,7%
Average	Count	29	49	66	57	35	236
	Expected Count	47,2	47,2	47,2	47,2	47,2	236,0
	% within Ranking	17,0%	28,7%	38,6%	33,3%	20,5%	27,6%
Below Average	Count	3	19	38	61	53	174
	Expected Count	34,8	34,8	34,8	34,8	34,8	174,0
	% within Ranking	1,8%	11,1%	22,2%	35,7%	31,0%	20,4%
Extremely Poor	Count	2	2	5	22	67	98
	Expected Count	19,6	19,6	19,6	19,6	19,6	98,0
	% within Ranking	1,2%	1,2%	2,9%	12,9%	39,2%	11,5%
Total	Count	171	171	171	171	171	855
	Expected Count	171,0	171,0	171,0	171,0	171,0	855,0
	% within Ranking	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Table 31: Rating * ranking crosstabulation

The Scientific Question XI: Is there a relation between the rating results and the ranking results?

What we try to find: In this question, we investigate a correlation between ranking results and rating results of the survey which proves the relation between two sets of results.

Context of the data: The data in this section consists of both rating and ranking results which are related over a particular participant. The data sets may consist of either single results for each page, or overall perspective over each complexity groups.

Technique of evaluation: Sample size was large for a non-parametric measure, so a parametric test, Pearson's Chi Square Test was run on rating and ranking data. Data management and analysis was performed using IBM SPSS Statistics 20. Our null hypothesis is that there is no relationship between rating and ranking results. Our alternative hypothesis is that there is a relationship between rating and ranking results.

Results: The results of the test which was performed on SPSS are given in Table 31 and Table 32. As can be seen in Table 32, the p-value is 0; therefore, we can reject the null hypothesis and conclude that there is a relationship between rating and ranking results at 5% significance level.

Discussion: Pearson's Chi-Square Test results indicate that there is a relationship between rating and ranking results. This means that participants rated the pages based on their best level preference. However, these two tasks are different and not necessarily related. Since the survey conducted online, we only introduced the task in our survey by using a reasonable explainotary text and did not have a chance to validate whether a participant understood the task. Therefore, it is possible that participants

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	433.070	16	0
Likelihood Ratio	420.163	16	0
N of Valid Cases	855		

Table 32: Chi-Square Tests

may rank the levels in rating section and select the assignments with respect to their best selection of levels rather than the segmentation success of the page. Since the same segmentation rules apply on each visual block in the page, we expect that, the success of segmentation has similar values in all levels. However, rating results are not similar for the same pages and are parallel to the ranking results as Pearson's Chi Square Test proves.

Complexity Group	System-Expert Evaluation	Receptive Evaluation
Low Comp.	79.82	73.68
Medium Comp.	88.28	79.77
High Comp.	88.47	85.53
Overall	86.83	80.82

Table 33: Accuracy results of role detection algorithm

2.6.2 Results of Role Detection Algorithm Evaluation

The Scientific Question XII: What is the success of our heuristic role detection algorithm?

What we try to find: We try to calculate the success rate of our proposed approach by comparing the system generated roles of web elements and the responses of participants who are provided a list of roles defined in our knowledge base.

Context of the data: The data used in this question includes the roles assignment for specific visual blocks and their corresponding roles which are generated by our system.

Technique of evaluation: First of all, we did strict string comparison between system generated roles and participant assignments. However, the overall results were very unsatisfactory. After investigating the reasons for these low results, which are also discussed in 'Discussion' section, we applied a receptive evaluation on the data, which consists of manual comparison.

Results: Table 33 presents the results obtained from the preliminary analysis of the data obtained from the participants who have worked in or studied web design and development as professional. The success rate of each complexity group and overall result was calculated proportional to the number of visual elements which were evaluated in each page. System - Expert evaluation consists of the comparison of the system results and expert responses with respect to the concept described in the ontology. As can be seen from this table, in overall the system has an accuracy of 86.83% accuracy. We then conducted a strict string comparison between the roles assigned by the system and also the participants. This gave us an average of 28.86% (low complexity pages - 26.32%, medium complexity - 28.99% and high complexity - 29.92%) accuracy. This is mainly because participants use slightly different versions of the role text used to label visual elements. Therefore, we have manually analysed the role assignments given by our participants and compared them with our system assignments. These results shown as "receptive evaluation" in Table 33. In overall, the accuracy rate is 80.82%.

Discussion: Our receptive evaluation shows that our proposed system has more than 80% accuracy rate, however when we do strict string comparison this accuracy becomes around 30%. This is mainly because there are differences in people's perception of roles and their approach in labelling the visual elements. This can also be seen from the majority rule application to the data collected from participants. When the majority rule applied to roles assigned, we see that only %32.58 of the role assignments have the majority of participants' agreements. This could be because of different reasons. It could be because people perceive pages differently or it could be because

people do not typically think about the roles explicitly but they just use the visual elements implicitly, and they were asked to articulate the roles they could not do it well. This could also be because people were not asked to complete tasks on web pages, and they were just given screenshots of web pages. These role assignments could have been done differently if people were asked to complete specific tasks. In our evaluation study, we have also explicitly asked participants to check the underlying source code of web pages and also associated CSS files. Despite our recommendation, participants may not analyse the DOM structure or CSS styles of the pages, which means participants formed their overall assessment only on the visual representation of the blocks. Another reason could be, many roles in our knowledge base could have been perceived very similar, and we did not give explanation of the roles; therefore, participants may not comprehend the difference.

One unanticipated finding from this evaluation was that most participants perceived that many blocks have more than one role in the page layout. We have noticed this from the roles assigned by our participants, they tend to assign more than one role. However, in our study we asked participants to choose only one role from the roles defined in our eMine ontology. Therefore, further studies could be conducted to investigate this further.

Some blocks are combinations of different sub-blocks which have different roles. When we look at the roles assigned by our participants, we have noticed that participants selected only one of them, omitting the remaining meaningful blocks. We have addressed this issue, by manually comparing the roles assigned by our participants and the system.

This study was also very useful for improving our knowledge base. When we look at the roles given our participants, we see that some of these roles do not exist in our knowledge base, therefore the results of this study will be used to improve our knowledge base. Some results also show that some of our concepts are still too generic in our knowledge base. The roles given by our participants showed that it would be much better if they are further defined in a deeper granularity.

	Complexity	High	Medium	Low
Dom Structure Cons.	Memory Usage	20,659.53 KB	16,293.37 KB	9,370.56 KB
	Time	5,069 ms	2,120 ms	730 ms
Visual Block Extr.	Memory Usage	21,610.17 KB	14,351.68 KB	9,535.14 KB
	Time	30 ms	19 ms	4 ms
Content Structure Cons.	Memory Usage	21,519.97 KB	13,711.88 KB	9,699.04 KB
	Time	93 ms	53 ms	21 ms
Total Node Count		3,811	1,888	624
Max. Node Depth		22	21	16
Visited Node Count		2,609	1,328	395
Invalid Node Count		98	84	44
Invisible Node Count		69	35	26
Block Count		569	237	65
Block Depth		13	11	7

Table 34: Performance results for segmentation implementation

3 Technical Evaluation

With a technical evaluation, we have mainly investigated the technical feasibility of the proposed approaches and implementation in the ACTF platform. We checked the performance characterised in terms of total memory usage, time elapsed for role detection of complete pages and total number of blocks calculated. The technical evaluation has been performed on a machine which has following features: Intel®Core™2 Duo CPU T9600 @ 2.80 GHz processor, 2.071.34 MB memory, NVIDIA GeForce GT 220M videocard and Windows 7 32 Bit operating system. Even though, we understand the results presented in this section are specific to our configuration of test machine, they still provide significant information about the overall performance and memory usage of our system.

Technical Evaluation of Segmentation Implementation

The segmentation algorithm executes in 3 iterative steps: DOM structure construction, visual block extraction and content structure construction. In DOM structure construction, web page is parsed and a tree of nodes is constructed. In visual block extraction, nodes are analysed and decided whether to create a visual block for corresponding node. Finally, in content structure construction, visula blocks are organised in a hierarchical structure. Table 34 presents the performance results of the system which consist of cumulative memory usage and response time for each step, total node count, maximum node depth, visited node count, invalid node count, invisible node count, produced block count and block depth for the whole page for each level.

Complexity Group	Total Memory	Total Time	Average Memory per Block	Average Time per Block	Block Count
Low	8,369 KB	6,576 ms	244.29 KB	102.29 ms	65
Medium	7,013 KB	23,799 ms	36.44 KB	102.12 ms	237
High	9,165 KB	54,837 ms	34.28 KB	101.95 ms	569
Overall	8,176 KB	29,157 ms	100.20 KB	102.11 ms	298

Table 35: Performance results for role detection implementation

Technical Evaluation of Role Detection Implementation

Table 35 presents the performance results of the system which consist of memory usage and response time. Total memory usage and total time elapsed in role detection process of a whole page and a single block are given with the average block count for each complexity level.

Discussion

According to the results in Table 34, DOM structure construction step uses higher amount of memory than other steps and executes in longer time. This main reason for this is that, all nodes in a web page are traversed and converted to a suitable format for use in later steps. Therefore, the total memory usage and time are based on casting operation. Reducing this operation may result in better performance results. Since node count changes with respect to the complexity of the pages, low level pages which have less nodes, give better results. Therefore, the node count affects the performance of the segmentation, so that, while node count increases, memory usage and total time ellapsed also increase.

Average time elapsed for role detection of a single block has close values in each complexity level and total time elapsed for role detection of a whole page is proportional to the number of blocks in the page. Total memory consumed for the whole page has also close values for each complexity group, showing that, larger amount of the memory consumed for shared static resources, such as working memory of Jess. Therefore, average memory consumed decreases while the number of blocks increases.

4 Conclusion

This technical report presented the procedure and results of both technical performance evaluation and user evaluation of our proposed approaches. User evaluation investigated the success of our proposed segmentation algorithm and heuristic role detection algoritm and the level preferences of the users. Technical evaluation aimed to assess technical feasibility of the proposed approach and implementation in the ACTF platform.

The analysis of the rating results have shown that, the success of our segmentation algorithm is below average for lower levels of segmets in a web page, while it is above average for middle and higher levels of segmentation in the page. While web page complexity is not a major criteria in rating results, participant characteristics such as age, gender, education, level of expertise and current status involved in web design and development significantly affects the rating results. One of the more significant findings to emerge from this study is

that, users prefers more sophisticated segmentation with small segments to simplistic segmentation with large segments. Web page complexity is still not a major criteria, while age, educational background and current status involved in web design and development significantly affect best level preference. The best level preference that we have identified therefore assists in our transcoding studies, by providing deep understanding on user preferences of segmentation level. Finally, correlation analysis on rating and ranking results revealed that there is a significant correlation between these results.

The user evaluation for heuristic role detection algorithm shows that our proposed system has around 80% receptive accuracy, but the proposed knowledge base could be further improved for better accuracy. The results in performance evaluation suggest that, response time is related to the complexity of the page, while memory consumption is independent of the complexity if shared resources are used.

In conclusion, the analysis presented in this paper contributes an effective understanding for perceived and technical success of our proposed approaches. The study also has gone some way towards enhancing our understanding of level preference of participants, which we believe to be useful in many applications including web page transcoding.

References

[1] M. Elgin Akpinar and Yeliz Yesilada. Vision based page segmentation: Extended and improved algorithm. Technical report, Middle East Technical University Northern Cyprus Campus, 2012.

[2] M. Elgin Akpinar and Yeliz Yesilada. Heuristic role detection of visual elements of web pages. In Florian Daniel, Peter Dolog, and Qing Li, editors, *ICWE*, volume 7977 of *Lecture Notes in Computer Science*, pages 123–131. Springer, 2013.

[3] E. Michailidou. *ViCRAM: Visual Complexity Rankings and Accessibility Metrics*. PhD thesis, 2010.