Measuring and Ensuring Similarity of User Interfaces: the Impact of Web Layout

Sebastian Heil[1], Maxim Bakaev[2], Martin Gaedke[1]

[1]Technische Universität Chemnitz, 09107 Chemnitz, Germany
{sebastian.heil,martin.gaedke}@informatik.tu-chemnitz.de
[2]Novosibirsk State Technical University, 630073 Novosibirsk, Russia
bakaev@corp.nstu.ru

Abstract. Given the rapid update cycles in modern web information systems and the abundance of legacy software being migrated to the web, controlling similarity between user interfaces (UI) is an actual problem of interaction engineering. The similarity (consistency) aspect is also increasingly considered in computer-aided design, where it is included in the optimized goal function, to minimize re-learning effort for users. In this paper, we explore the impact of the proposed layout distance measure, which is calculated for different levels of hierarchy in web UIs, which we identify as: Region – Block – Group – Element. To support our approach, we conducted an experimental pilot study in the context of an ongoing medical information system (IS) web migration project. The regression analysis suggests that layout distance (particularly, its *orientation* dimension) does have effect on web UI similarity as perceived by users. The results can be used by web engineers, in particular to smoothen the transition between versions of a UI for users and IS operators.

Keywords: Similarity Measure, User Interface, Web Migration, HCI.

1 Introduction

Web information systems (WIS) increasingly supersede existing information systems (IS), even in domains where traditional desktop-based IS were common, such as medical software or in the financial sector [1]. In particular, cloud-based SaaS offerings gain more popularity. As deployment is simplified by the centralized architecture, development cycle lengths are significantly reduced. Updates in the production versions of SaaS WIS can be applied up to several times a day. Given these rapid update cycles and the abundance of legacy software being migrated to the web, controlling similarity between user interfaces (UI) is an actual problem of interaction engineering. Being able to determine the similarity between different versions of a user interface allows to control the amount of change during update cycles and to ensure smoother transitions and reduced learning efforts when migrating user interfaces to the web.

Migrating a legacy desktop IS into a web IS implies a fundamental paradigm shift which involves changes in all parts of the application: presentation (user interface), application logic and persistence (database) [2]. However, existing approaches like

adfa, p. 1, 2011.

[1], [3] mainly focus on code and data migration and often disregard problems and costs associated with changes to the user interface and user interaction. The transformation or re-development of a user interface as a combination of HTML, CSS and Javascript source code does not only change the internal structure, but also impacts the visual appearance and user interaction.

As a first step towards supporting developers to measure and control user interface similarity, in this paper we explore the impact of layout, which is arguably the top consideration when migrating a UI to the web. In section 2, some related works are examined and our approach for measuring distances between interfaces based on counting and visual area measurements is outlined. In section 3, we describe the pilot experimental study that we have undertaken to support our approach, while the analysis of the results is presented in section 4.

2 Related Work

Direct applications of UI similarity seem to be scarce in HCI, where it is overshadowed by the concept of consistency, whose importance is widely recognized, but which conceptually remains rather vague [4]. Consistency is, in a way, similarity within a single interface – between different screens, elements, minor and major conventions, etc. The reason we chose to speak of not consistency but similarity is that the latter is already extensively used in the AI field, in particular for case-based reasoning (CBR) that is gaining increased popularity in intelligent and recommender systems. The main idea is that a new problem is identified and the search for similar but already resolved problems is undertaken, in the assumption that similar problems have similar solutions.

In HCI, practical application of CBR methods remains somehow limited, in particular because interaction problems' formal identification remains unresolved, although in one of our research works we proposed to employ concepts of dedicated web design support ontology to describe the interaction context [5]. An important milestone in the field was the development of the SUPPLE system that is capable of auto-generating user interfaces with a model-driven approach. The authors introduced an *interface dissimilarity* metric that was included in the optimized goal function, so that new interfaces produced by the system resemble the old ones, familiar to the users [6]. The metric was determined as linear combination of factors {0/1} reflecting whether or not the two considered interface widgets are similar according to a certain criterion. The authors also put forward a list of widget features: language, orientation of data presentation, primary manipulation method, widget geometry, etc.; but their groping and layout don't seem to get further adequate consideration. We'd like to note that with the use of the totally same widgets designers would be able to create radically different interfaces, although it may be less of a problem in case of interfaces auto-generated by an intelligent system from the same interface model.

Thus it so far remains unclear whether robust quantitative identification of web interfaces is feasible, but we believe such an undertaking will be of considerable potential use for AI methods advance in HCI.

3 Method

As we mentioned above, we first decided to consider the layout similarity of user interfaces, for which end we propose approaches to determining the *distances* between several interfaces in terms of their visual layout. By analogy with AI-based CBR algorithms, we could try to infer weights for the several potentially meaningful dimensions of the distance, which we identify and quantify as the following:

1. **Orientation:** the share of interface items that have different visual orientation – horizontal, vertical, or other.
2. **Order:** the share of interface items that have different order relative to both their neighbors (start and end are virtual neighbors). This dimension is especially relevant for migration, since it implies that the interfaces in comparison (legacy and web ones) have the same or comparable items.
3. **Density:** the share of interface items that have different visual density of sub-items. It shouldn't matter whether the density is increased or decreased – we shall consider the effect on similarity in the same way.

The interface items that we mention here are somehow close to widgets introduced in interface dissimilarity metric in [6]. They can belong to either level in the modern web interfaces organizational hierarchy, which we see as Region – Block – Group – Element. The relations between the levels may be described with more complex models, but we will so far consider all the levels to be equal in importance and assume that changes in any of them have comparable effects on perceived similarity.

In order to support our approach and provide a first evaluation, we conducted an experimental pilot study, which took place in the context of an ongoing medical IS (patient management system) web migration project.

4 Experiment Description

The experiment scenario is adapted from a research collaboration project with an industrial partner – the migration of medical software system to the web. The migration of existing user interfaces used by doctors and nurses introduces changes both in layout and interaction, and in the health sector there is generally not much time to conduct extensive user training, so similarity becomes especially prominent.

4.1 Experimental Design and the Hypothesis

The experiment had within-subjects design, with main independent variables being the layout distances and dependent variables being the similarity of old and new user interfaces, as perceived by users. We also added additional dependent variables that we outline in more detail below, to more fully capture the users' experience with interfaces. Our main hypothesis, related to the approach we propose for expressing the layout distance measure, is thus the following:

H_0: there's no effect of distance measures on perceived interface similarity.

To evaluate the validity of our approach and the experimental design, as well as the diversity of the subjects' evaluations, we'll also explore the differences in the calculated dimensions of the distance, as well as correlations between the evaluation scales.

4.2 The User Interfaces

We chose three legacy user interfaces representing different levels of complexity:

- User interface screen A is a simple graphical shift schedule (complexity: 1)
- User interface screen B is a calendar for appointment scheduling (complexity: 2)
- User interface screen C is an extensive patient data form (complexity: 3)

For each of them we created web versions, implemented in HTML, CSS and Javascript using Bootstrap and jQuery. In terms of layout, they are copies of their original desktop counterparts with no intentional changes apart from those changes introduced by the migration to the web. We assigned identifiers A0, B0 and C0 to these web user interface versions of A, B and C respectively. Then, for each of the three web interfaces, maintaining their original functionalities, we created three variations by varying one of the three main aspects per variation and assigned identifiers 1, 2 and 3: Orientation (A1, B1, C1), Order (A2, B2, C2), Density (A3, B3, C3).

For the *orientation* variations, we changed the layout from horizontal to vertical and vice versa by repositioning groups of UI elements. To vary *order*, we changed the positions of elements like buttons or text fields (along with their labels) within regions. We did not mix them between different regions like by re-ordering days in the calendar or by moving patient data inputs into the medical billing region, as this would result in an unrealistic user interface that intendedly confuses users. To achieve *density* variations, we replaced color fills in the shift schedule (A) by letters and adjusted position and spacing of elements for the other two interfaces (B, C). Obviously, these changes cannot be regarded as completely independent – e.g. changing the orientation for instance may also results in a change of density. The interfaces can be found at https://vsr.informatik.tu-chemnitz.de/demos/LayoutSimilarity.

4.3 Calculation of Distances

The values for the main independent variables in our experiment, the distances, were determined using the method we proposed above. So, we first determined the total number of items on each level of hierarchy for the web interfaces (results are shown in Table 1). Then, we asked an expert to determine the number of items in the web interfaces that altered relatively to respected desktop versions in regard to orientation, order or density, so that we could determine the distances. Table 2 shows the numbers of items that changed in regions (R), blocks (B), groups (G), and elements (E), as well as the calculated distances for each of the dimension. It should be noted that for *order*, changes in a single region are impossible, so this hierarchy level was not included in the calculation. For *density*, changes within an element are not possible, so the elements level was not included in the calculation.

Table 1. The total numbers of items per hierarchy levels in web interfaces

Interfaces	***Regions***	***Blocks***	***Groups***	***Elements***	***Notes***
A0-A3	1	2	4	7	We consider the calendar inner table as one element.
B0-B3	1	3	7	22	Interiors of small and large calendars are one element each.
C0-C3	1	3	7	48	Each input field and its label are one element.

Table 2. Numbers of changed items and distances (Dist.) per similarity dimensions

	Orientation		***Order***		***Density***	
Interface	*Changed*	*Dist.*	*Changed*	*Dist.*	*Changed*	*Dist.*
A0	1B	0.125	none	0	1G	0.083
A1	1R, 2B, 2G	0.625	1B	0.167	1R, 1G	0.417
A2	none	0	1B	0.167	1B, 1G	0.250
A3	1B	0.125	none	0	2G*	0.167
B0	none	0	none	0	1B, 1G	0.159
B1	1R, 1B	0.333	none	0	1R, 2B, 1G	0.603
B2	none	0	3B	0.333	1B, 1G	0.159
B3	1B	0.083	none	0	1R, 2B, 1G	0.603
C0	none	0	none	0	1R**	0.333
C1	1R, 2B	0.417	none	0	1R	0.333
C2	none	0	4G, 24E (50%)	0.357	1R	0.333
C3	none	0	none	0	1R, 2B	0.556

[*] Although A3 was supposed to be dedicated version with changes in density (1R would be expected), visually the density didn't change with the removal of color fills, according to the evaluating expert.

[**] For the interface screen C, all web versions visually had different density compared to the desktop one, so 1R was assigned for the each version.

4.4 Subjects and Procedure

In our pilot experimental study we employed 7 subjects (of which 2 were female), all of them students majoring in Informatics or staff of Chemnitz Technical University, Germany. Their ages ranged from 21 to 50, average being 28.4 and SD=9.83. All but one participant were of German nationality, and all of the subjects were proficient in English. Before the experiment informed consent to take part in the study was obtained. The subjects didn't have previous experience with the medical WIS, but have good background in software development, however, not related to HCI. As such, they can be rather considered experts than representatives of the system's target users.

We would show the participants how to achieve the specially designed tasks in the legacy UI, and they were then asked to achieve them in one of the web UIs. Our experiment environment would randomly select one of the four web interfaces and display it to participant. When the task list was completed, the participants answered several questions, assessing *difficulty*, *Like* and *similarity* impressions. Then, the entire process was repeated, showing a new task list on another legacy UI and then having the participant replicate it in the four web versions, for the remaining interfaces, overall three times. In order to avoid participants being biased from recognizing A0, B0 and C0 as "basic" versions, we re-numbered all versions in what was visible to the participants. The experimental sessions with each participant were scheduled and conducted for the duration of about one week. Each session lasted about one hour and was performed on the same desktop PC and screen, for the sake of consistency in the interfaces representation.

5 Results

5.1 Descriptive Statistics

In Table 3 we show the values of the distance factors together with the evaluations provided by the participants, per the three scales.

Table 3. Values for the factors and the subjects' evaluations

	Distances			***Evaluations***		
Interface	***Orientation***	***Order***	***Density***	***Difficulty***	***Like***	***Similarity***
A0	0.125	0.000	0.083	1.714	3.429	4.000
A1	0.625	0.167	0.417	2.000	2.571	2.857
A2	**0.000**	**0.167**	**0.250**	2.000	3.571	**4.286**
A3	0.125	0.000	0.167	1.571	2.571	3.714
B0	**0.000**	**0.000**	**0.159**	1.857	3.286	**4.143**
B1	0.333	0.000	0.603	2.143	2.000	3.000
B2	0.000	0.333	0.159	1.714	3.714	3.286
B3	0.083	0.000	0.603	1.857	3.143	3.571
C0	0.000	0.000	0.333	2.143	3.143	3.714
C1	0.417	0.000	0.333	2.857	2.000	2.714
C2	0.000	0.357	0.333	2.143	3.000	4.000
C3	**0.000**	**0.000**	**0.556**	1.286	3.571	**4.143**
Avg.	0.142	0.085	0.333	1.940	3.000	3.619
(SD)	(0.207)	(0.137)	(0.181)	(1.057)	(1.299)	(1.029)

The *difficulty* evaluation had the lowest absolute value (1.940), which is understandable since the employed interfaces were relatively simple, especially given the subjects' proficiency in computers. The greatest standard deviation of the *Like* evaluation (1.299) was also to be expected, as these answers have the highest degree of subjectivity. We detected highly significant positive correlation (Pearson's ρ=0.734, p=0.007) between *Like* and *similarity* evaluations, which may imply that people prefer familiar interfaces, although the experimental environment may have affected this judgement, hinting that similar equals good. The significant negative correlation (ρ=-0.632, p=0.027) between *difficulty* and *Like* should have been expected, as it's well known that in interaction perceived difficulty invokes negative feelings. The negative correlation between *diffi-*

culty and *similarity* was significant at α=0.06 (ρ=-0.563, p=0.057), which supports the assumption that familiar interfaces have lower perceived difficulty.

5.2 Regression Analysis

Regression analysis for the three factors and *difficulty* evaluation did not find any significant effects (p=0.549, R^2=0.221). In regressions for *Like* and *similarity*, only the *orientation* factor was significant, so the other factors were removed from the models. The resulting model for *Like* (1) was highly significant: p=0.005, R^2=0.561. The model for *similarity* (2) had even greater significance and better fit: p=0.001, R^2=0.670.

$$Like = 3.3 - 2.14 * D_{ORIENT} \quad (1)$$

$$Similarity = 3.92 - 2.14 * D_{ORIENT} \quad (2)$$

5.3 The Layout Distance Measures

The obvious way to calculate the final distance measure for layout is to take the average of *orientation*, *order* and *density* distances, which would result in what we'll call the D_{basic} layout distance. However, we can also take normalized coefficients for all factors in regression for similarity (even though order and density distance factors were not found to be significant) as weights in calculating the overall distance measure, in which case the formula will be the following:

$$D_{LAYOUT} = 0.726 * D_{ORIENT} + 0.162 * D_{ORDER} + 0.112 * D_{DENSITY} \quad (3)$$

The correlation between the two measures was highly significant (ρ=0.862, p<0.001). However, correlation between D_{layout} and *similarity* (-0.831, p=0.001) was found to be higher than for D_{basic} and *similarity* (-0.716, p=0.009). The only another significant correlation at α=0.05 was the negative one between D_{layout} and *Like* evaluation (ρ=-0.727, p=0.007). We further compared the two distance measures by attempting regressions for *similarity* evaluation. The model for D_{basic} was significant (p=0.009, R^2=0.513), but the regression for D_{layout} (4) showed even higher significance and considerably better fit (p=0.001, R^2=0.691).

$$Similarity = 4.07 - 2.89 * D_{LAYOUT} \quad (4)$$

6 Conclusions and Future Work

In our research work we sought to explain why we consider interface similarity to be an important and potentially useful metric. Our assumption was that familiar interfaces, other things being equal, are more usable to users, and this should be considered in interface design and re-design activities. In our current work we focused on interface layout and proposed an approach for quantitative expression of distances between two interfaces in this regard. The considered dimensions included *orientation*, *order* and *density*, and we also sought to determine their relative importance for users and thus contribution to the overall layout distance measure. To support our approach, we designed and conducted a pilot study with 7 subjects and 12 interface screens constructed from 3 legacy interfaces.

The results of the analysis suggest that our hypothesis H_0 could be rejected, and the proposed distance measures do have effect on web interface similarity as perceived by users. The regression model (2) was highly significant (p=0.001) and had reasonably fair R^2=0.670. The only significant layout dimension was *orientation* distance, which predictably had a negative coefficient in the equation. Based on the model, we calculated normalized weights for the three dimensions and determined the overall layout distance metric, D_{layout} (3). Compared to the simple average (D_{basic}), this metric had higher correlation with similarity (ρ=-0.831) and produced considerably better regression (4): p=0.001, R^2=0.691.

Among the limitations of our current study we'd like to note, first of all, the small sample of users and low diversity of interfaces. Although the experiment participants performed specially developed realistic tasks, they were not quite representative of the target user group, had no previous experience with the employed WIS, and there was little interaction with the interfaces. All in all, we are far from asserting that the results of our pilot study can be used directly, but the proposed approach may be still sound. Our plans for future work include further exploration of interface complexity factor and coverage of other aspects of interface similarity.

Acknowledgements. This research was funded by RFBR, according to the research project No. 16-37-60060 mol_a_dk, and supported by the eHealth Research Laboratory funded by medatixx GmbH & Co. KG.

References

1. L. Aversano, G. Canfora, A. Cimitile, and A. De Lucia, "Migrating legacy systems to the Web: an experience report," in Proceedings Fifth European Conference on Software Maintenance and Reengineering, 2001, pp. 148–157.
2. G. Canfora, A. Cimitile, A. De Lucia, and G. a. Di Lucca, "Decomposing legacy programs: a first step towards migrating to client–server platforms," J. Syst. Softw., vol. 54, no. 2, pp. 99–110, Oct. 2000.
3. M. Colosimo, A. de Lucia, R. Francese, and G. Scanniello, "Assessing Legacy System Migration Technologies through Controlled Experiments," in 2007 IEEE International Conference on Software Maintenance, 2007, pp. 365–374.
4. Molich, Rolf, Robin Jeffries, and Joseph S. Dumas. "Making usability recommendations useful and usable." Journal of Usability Studies 2, no. 4 (2007): 162-179.
5. M. Bakaev and T. Avdeenko, "Indexing and comparison of multi-dimensional entities in a recommender system based on ontological approach," Computación y Sistemas, vol. 17 (1), pp. 5-13, 2013.
6. K. Gajos, A. Wu, and D.S. Weld. "Cross-device consistency in automatically generated user interfaces," Proc. 2nd Workshop on Multi-User and Ubiquitous UIs, 2005, pp. 7-8.