S.O.S.: Does Your Search Engine Results Page (SERP) Need Help?

Maximilian Speicher*
Technische Universität
Chemnitz
09111 Chemnitz, Germany
speim@hrz.tu-chemnitz.de

Andreas Both
R&D, Unister GmbH
04109 Leipzig, Germany
andreas.both@unister.de

Martin Gaedke
Technische Universität
Chemnitz
09111 Chemnitz, Germany
gaedke@cs.tu-chemnitz.de

ABSTRACT

Over the past 20 years, search engines have become *the* entry point of the WWW. Due to evolving needs for different and new kinds of information, the interfaces of search engine results pages (SERPs) change over time. Thus, their usability must be continuously evaluated to ensure user satisfaction and competitive edge. As no complete solution exists, we present *S.O.S.:* the *SERP Optimization Suite*. Our approach comprises (a) *WaPPU*, which is a near real-time tool for evaluating web interfaces based on usability scores, and (b) a *catalog of best practices* that maps bad scores to potential causes and corresponding adjustments for optimization. During a case study in which we assessed and optimized a real-world SERP, S.O.S. has proven its feasibility and effectiveness by significantly improving the SERP's usability.

Author Keywords

best practices; evaluation; metrics; optimization; scores; search engine results pages; usability

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g., HCI): User Interfaces—*Evaluation/methodology, Style guides, User-centered design*

INTRODUCTION

Ever since the launch of *Archie*[1]—which is considered as the first of its kind—in 1990, search engines have become *the* entry point of the WWW. Numerous users do not even make use of the address bar of their browser anymore, but just type where they want to go into the search box of the default home page. For example, the *Start Page* of *Mozilla Firefox* is powered by Google[2] and simply shows their well-known search box. Also, the address bars of all modern browsers interpret non-URL input as search queries, which are forwarded to the default search engine defined in the browser settings. In this way, it is nowadays possible to browse the Web completely without URLs, thanks to *search engines*.

While the early search engines were relatively simple applications and provided limited functionality—i.e., they performed a full-text search on an index of HTML documents—, requirements towards search engines have evolved constantly. Nowadays, there is a demand for different types of results, such as news, images and videos. Search engines also more and more tend to answer queries directly on their *search engine results pages* (SERPs) and provide such answers in dedicated info boxes (cf. *Google Knowledge Graph*[3]). Finally, search engines as well have to somehow incorporate advertisements into their interfaces to ensure they can make profit and offer their services for free. All of this brings up numerous requirements concerning the delivery and arrangement of all different kinds of content. As a result, SERPs evolve continuously since they are the search engines' prime interfaces for ultimately presenting the desired information to the user.

To date, there is no form of a SERP interface that can be generally considered to be optimal. Although Google clearly stands at the top (number 1 in the *Alexa* top 500[4]), particularly trending search engines such as *qwant.com* and *duckduckgo.com* show different approaches towards SERP interfaces. To give just one example, besides other obvious differences, all three search engines feature completely different look & feels concerning the aforementioned info boxes (Figure 1). Thus, it is crucial for search engine owners to continuously evaluate the usability of their SERPs regarding the evolving needs of users and the introduction of new kinds of information.

In today's industry, SERPs are often evaluated using A/B testing set-ups that try to increase quantitative metrics like clicks on advertisements. These are popular due to their better efficiency compared to traditional usability evaluation methods, such as heuristic inspections [20] or lab studies. Yet, A/B tests usually lack effectiveness in determining usability [19]. For evaluating SERPs in an efficient *and* effective manner, we propose *S.O.S.:* the ***S**ERP **O**ptimization **S**uite.* S.O.S. comprises two interconnected components: (1) *WaPPU*, which is a tool for performing *usability-based* A/B tests, and (2) a

*The contents of this paper were developed while Maximilian Speicher stayed at Unister GmbH as an industrial PhD student.

[1] `http://people.lis.illinois.edu/~chip/projects/timeline/1990archie.htm` (Sep 1, 2014).

[2] `http://www.google.com/firefox` (Sep 1, 2014).

[3] `http://www.google.com/insidesearch/features/search/knowledge.html` (Sep 1, 2014).

[4] `http://www.alexa.com/topsites` (Sep 1, 2014).

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3145-6/15/04...$15.00
http://dx.doi.org/10.1145/2702123.2702568

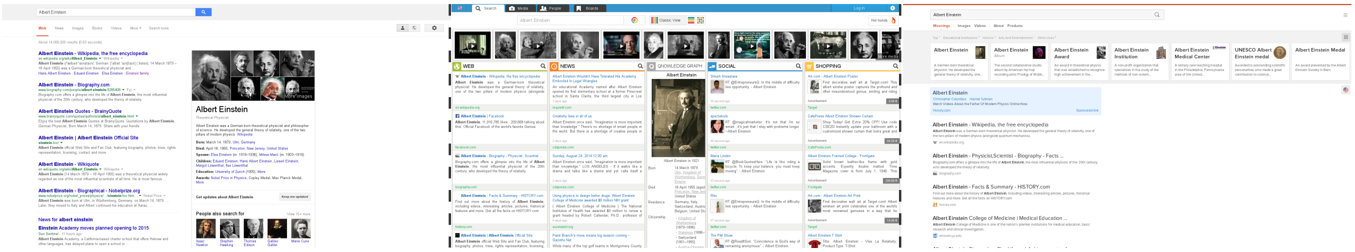


Figure 1. Comparison of the SERP interfaces of Google (left), Qwant (middle) and DuckDuckGo (right). They particularly differ concerning the look & feel and placement of the info box about "Albert Einstein".

catalog of best practices w.r.t. SERP usability. Our approach is able to determine usability scores of a SERP interface and for suboptimal scores automatically point out potential causes and corresponding countermeasures.

Speicher et al. have developed WaPPU ("*Wa*s that *P*age *P*leasant to *U*se") as a general tool for A/B testing based on quantitative usability scores [23]. Although existing work (e.g., [3, 8, 9]) already builds on user interaction analysis for better efficiency, *WaPPU* is the first approach to perform A/B tests based on seven usability factors as the target metrics (e.g., *readability* = -0.5 for Interface A). These metrics are based on a questionnaire as well as real users' interactions with the interface. Speicher et al. have proven the feasibility and effectiveness of WaPPU in a user test with two real SERP interfaces [23]. Yet, a major shortcoming of A/B testing approaches is the fact that improvements to an interface need to be performed *a priori*. That is, the old *and* the new version of the interface must be available before the usability-based A/B test starts. To date, there is no state-of-the-art tool that is able to propose optimizations to an interface based on usability scores or user interactions—like "Your rating for *readability* is negative, please adjust your font size"—, which would be highly desirable.

Therefore, we propose an extension to WaPPU to form S.O.S. Since the mapping between a usability score / vector of user interactions y and causes for bad usability x is *not* a bijective function $f(x) = y$, it is technically not possible to infer the exact necessary optimizations for an assessed interface. To give just one example, a bad score of readability might be caused by the wrong font size and line height while the font itself is good. Yet, it could also be caused by the wrong font while font size and line height are good. Thus, the proposed extension is a catalog of best practices that—in the context of SERPs—, maps a bad rating y of a usability factor to a *set* of potential causes C and corresponding countermeasures C': $f(y) = \{C, C'\}$. This catalog has been determined in a three-step process that followed the *find–fix–verify* pattern described by Bernstein et al. [5] to ensure it is well-founded. First, we have reviewed well-known SERP interfaces and extracted common best practices as well as potential problems from these. Second, we asked ten dedicated experts to review and revise this basic catalog. Third, each revision has been approved or rejected by three additional, independent experts. We have applied rules from the final catalog to a real-world SERP that had been evaluated using WaPPU. This case study suggests the effectiveness of S.O.S., as the redesigned SERP shows significant approvements for certain usability factors.

In the following, we discuss related work and motivate this paper by introducing WaPPU and shortcomings of A/B testing in more detail. In the subsequent section, we describe the development of the catalog of best practices. The workflow of S.O.S. is presented after that. Then, we explain the corresponding case study before discussing our work and giving concluding remarks.

BACKGROUND AND RELATED WORK

A/B testing [19] is a common method to evaluate changes to an interface. That is, the original version of the interface is compared against a modified version (e.g., advertisements vs. no advertisements) w.r.t. a pre-defined target metric (e.g., clicks on a specific link). The two versions that are compared are also referred to as the "interfaces-under-test" in the remainder of this paper.

Usability is a relative concept that can only be evaluated within a specified context. As Brooke [7] puts it: "Usability is not a quality that exists in any real or absolute sense." In this paper, the notion of "usability" is based on ISO 9241-11 [1], which we further specify as follows: We consider the usability *in use* [2] of a single webpage (which we refer to as the "web interface"), as perceived by real users. Moreover, we consider *do-goals* (achieving tasks) rather than *be-goals* (feeling special, being satisfied etc.) [13]. In accordance with [15], we remove "satisfaction" from the original definition of [1], as it has to be considered separate from usability in use.

Today, there are numerous approaches to **interaction-based usability evaluation**, such as [3, 8, 9] or the commercial tool *m-pathy*[5]. All of these try to infer usability / user experience issues from user's behavior on a website. This is partly done by visualizing interactions ([3], m-pathy) and partly by comparing users' logs to pre-defined optimal logs for a given task ([8, 9]). However, the above approaches still require manual inspection of the results by developers or experts. In particular, they do not provide the evaluator with a *quantitative* measurement of the performance of the investigated interface.

Contrary to the above, methods for **metric-based usability evaluation** intend to assess interfaces based on scores for usability. Nebeling et al. [16] developed a set of metrics specifi-

[5] `http://www.m-pathy.com/` (Sep 11, 2014).

cally aimed at large-screen contexts. Their metrics—which are determined based on a webpage's rendered layout—include small-text ratio and media-content ratio, to name only two. Speicher [21] has transferred these metrics to small-screen contexts and also provides an additional metric for a webpage's "overall quality". The *System Usability Scale* (SUS) [7] is a lightweight instrument for assessing arbitrary interfaces that provides a score between 0 and 100. It poses ten questions to the user that have to be answered on a 5-point Likert scale. *AttrakDiff*[6] is a similar tool that, however, assesses the user experience of a web application on a two-dimensional scale (pragmatic and hedonic quality), and thus builds on a different instrument. It is also specifically aimed at carrying out A/B tests. Although the above solutions provide different kinds of scores for evaluating interfaces, they are based on either static analysis of a rendered layout or users' answers to a questionnaire. None of them specifically aims at deriving such scores from user interactions.

The goal of the WaPPU tool contained in S.O.S. is to combine the two general approaches described above—i.e., evaluating interfaces based on user behavior and providing corresponding usability scores.

Concerning **usability optimization** of interfaces, *decision-theoretic optimization* [10] is a general approach. With this method, users' preferences are elicited using either *example critiquing*, i.e., giving the user the option to change parts of an interface, or *active elicitation*, i.e., explicitly asking a user for their preferences. In this way, it is possible to learn a cost function that can be optimized through linear optimization. Examples for systems implementing this technique are ARNAULD [10] and SUPPLE [11]. Similar to the above, *Crowd-Adapt* [17] gives users the chance to express their preferences by adjusting the layout of a webpage (rearranging elements, changing the font size etc.). In this way, interface optimizations are crowdsourced for different viewing contexts. *W3Touch* [18] also builds on the crowdsourcing idea, however, in the specific context of touch devices. The tool engages implicit feedback by, e.g., scaling text according to users' average zooming factor. While all of the above are promising tools for improving usability through interface optimization, with the exception of W3Touch they follow fundamentally different approaches compared to S.O.S. Also, none specifically aims at the case of optimizing SERPs.

When it comes to **SERP optimization**, [6] describe the *layout definition problem.* That is, given results from different domains and different result types, they try to compute an optimal layout. However, this is an *a priori* approach resulting in a complete *design strategy* for the SERP. Therefore, it is more likely involved in the early stages of the design process, e.g., before an A/B test is carried out.

Concerning **best practices**, there is a vast number of general checklists for web interfaces, such as *Userium*[7]. Also, *Userfocus* present a rather general, 20-item set of best practices for search usability[8]. However—to the best of our knowledge—there is no well-founded catalog of best practices for optimizing SERP usability. In particular, there is none that relates causes for bad usability and corresponding countermeasures to an actual usability evaluation.

The tool that is probably closest to our S.O.S. approach has been described in [12]. The proposed approach detects *bad usability smells* based on *a priori* assumptions and then provides the developer with a report of existing problems. Contrary to our work, Grigera et al. focus on rather high-level usability problems—like long navigation paths—in the context of complete web applications. Yet, their tool might be a useful complement to S.O.S., particularly if there occur intersections between bad usability smells and suboptimal usability scores reported by WaPPU.

MOTIVATION: WAPPU

WaPPU is an ongoing research project that has been introduced by Speicher et al. [23]. The prototype[9] has been developed as a general tool to (a) enable A/B tests that are based on *usability* as the prime target metric, and (b) enable the inference of the corresponding usability scores from user interactions. Its aim is to be more effective in determining usability than conversion-based[10] A/B tests while retaining most of their efficiency. In fact, A/B tests are popular in industry due to their advantages over traditional methods of usability evaluation, particularly concerning efficiency [19]. This is underpinned by a variety of available (commercial) A/B testing tools, such as *Google Analytics*[11] and *Visual Website Optimizer*[12], to name only two.

WaPPU evaluates seven usability factors: (1) Informativeness, (2) Understandability, (3) Confusion, (4) Distraction, (5) Readability, (6) Information Density, and (7) Accessibility.

These factors (or items) are taken from a usability instrument that has been specifically tailored to the needs of the tool [22]. That is, it is cost-effective and designed for inferring ratings of its items directly from user interactions [22]. In the context of WaPPU, the usability factors are what we refer to as usability *metrics* in the remainder of this paper. They also form an *overall metric for usability* by summing up ratings of all seven items.

User Interaction Tracking

WaPPU tracks a total of 27 features, among which are, e.g., *cursor speed, length of the cursor trail, number of hovers* and *page dwell time*. The developer has to include a client-side jQuery plug-in and define components (e.g., navigation, sidebar etc.) for which the features shall be tracked. All defined components have to be present in both interfaces-under-test

[6] `http://attrakdiff.de/index-en.html` (Sep 11, 2014).

[7] `https://userium.com/` (Sep 11, 2014).

[8] `http://www.userfocus.co.uk/resources/searchchecklist.html` (Sep 11, 2014).

[9] `https://github.com/maxspeicher/wappu-service` (Sep 2, 2014).

[10] For example, the number of clicked advertisements or the number of completed checkout processes.

[11] `http://www.google.com/analytics/` (Sep 2, 2014).

[12] `https://vwo.com/` (Sep 2, 2014).

of the A/B test, i.e., the high-level structure of both interfaces has to be the same. The collected interactions are then sent to the server-side WaPPU service in suitable time intervals.

Collecting Usability Judgments

WaPPU provides the option to automatically display a questionnaire when a user leaves a web interface. This questionnaire corresponds to the seven usability factors above and can be included in any of the interfaces-under-test included in an A/B test. It is necessary to display the questionnaire in at least one involved interface, as otherwise WaPPU is not able to predict usability scores from user interactions. In the current implementation, the user answers the seven questions based on a 3-point scale: ☹, 😐 or ☺ (bad–neutral–good). These seven *usability judgments* are then sent to the server side together with the tracked user interactions. In near real-time, the WaPPU service automatically learns one model for each usability factor based on existing machine learning techniques. Moreover, separate models are learned for different user contexts since factors such as the screen size have an impact on the tracked interaction features. The current prototype considers *screen size* and the presence of *ad blockers* as dimensions of user context and uses an incremental version of the Naïve Bayes classifier [14] for learning the models. This means that in total, seven models are provided per A/B test and detected user context. Once these models are of good enough[13] quality, the questionnaire can be removed and the usability of both involved interface versions is inferred from user interactions alone.

The WaPPU Dashboard

WaPPU summarizes the metrics and scores of an A/B test in a dedicated dashboard (Figure 2). For the seven individual factors, a score between -1 and $+1$ is given, corresponding to the bad–neutral–good scale (☹, 😐, ☺) described above. The score of the overall usability metric is then determined by summing up all factors, i.e., it has a value between -7 and $+7$ that is normalized to a value between 0% and 100%. All scores are visualized together with their standard deviations. The overall score is provided in analogy to SUS [7]. [4] found that the university grade analog (100–90 points correspond to an “A”, 89–80 points to a “B” etc.) is a good rule-of-thumb for interpreting SUS scores. Moreover, they state that “products which are at least passable have [...] scores above 70” [4]. Thus, we set **70%** as the lower bound for a *good* usability score in WaPPU. This corresponds to individual factor scores of **0.4**, i.e., if all usability factors have a score of 0.4, we get an overall score of 70%.

If an interface-under-test features the questionnaire, the eight usability scores displayed in the dashboard are derived directly from users' answers. Otherwise, WaPPU predicts the scores based on the interactions collected on the respective interface and the available models. For example, if interface “A” displays the questionnaire, WaPPU learns the seven models M_i based on the corresponding answers and interactions $\vec{I}$ tracked on “A”: $M_i \leftarrow \mathit{learn}(\text{answer}_\text{i}^\text{A}, \vec{I}^A)\ \forall i \in \text{usability factors}$. The usability scores of interface “B” are then inferred from these models and the interactions collected on “B”: $\text{score}_i^B = M_i(\vec{I}^B)\ \forall i \in \text{usability factors}$.

Additionally, the WaPPU dashboard features a traffic light indicating statistically significant differences between the interfaces-under-test. That is, WaPPU applies a *Mann–Whitney U test* to the (predicted) overall usability scores produced by all involved users. In case the usability scores of the two interfaces-under-test are statistically equal—i.e., no definite statement about which one is better can be made—the traffic light is yellow, otherwise it is red (“A” better) or green (“B” better).

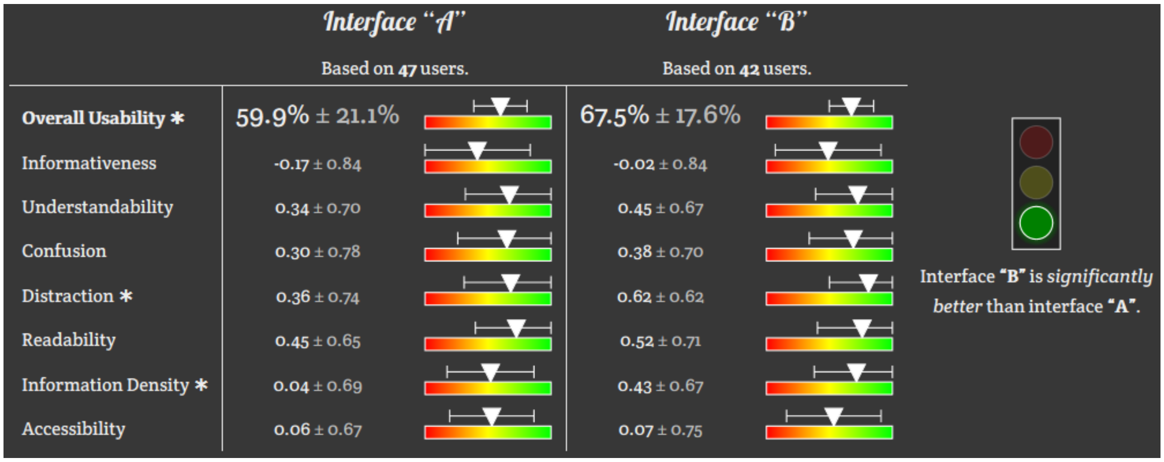
Limitations

While WaPPU has demonstrated its feasibility and effectiveness in a user study [23], the tool has a major shortcoming. Because it follows the principle of A/B testing, both versions of the interface to be tested have to be present before the parallel A/B test starts. Particularly, improvements to an interface need to be performed *a priori*, i.e., before the developer is provided with quantitative evidence. This might lead to lots of test being carried out that cannot detect a significant difference between the two versions of the interface. Therefore, it would be highly desirable if WaPPU was able to propose improvements to an interface based on its measured usability.

A straightforward idea would be to use the interactions tracked by WaPPU for inferring the causes of bad usability scores. Once the specific causes would be known, corresponding improvements could be proposed. Yet, a bad score of a usability metric can have numerous different reasons, i.e., the mapping between the score y and the cause x is not a bijective function $f(x) = y$. As a representative example, let us consider bad *informativeness* on a SERP—i.e., the particular information the user is looking for is not present or only present in parts. Bad informativeness can have two main reasons: (1) the desired information is in fact not present on the SERP, and (2) the desired information is hidden in such a way that the user cannot find it (which might be due to a bad layout). In both cases, the user would give a rating of -1 (☹) for the item *informativeness*. Yet, they only consider the fact that the desired information is not present on the SERP from their point of view. Particularly, the user does not care about the specific cause for this lack of information. In [23], it is reported that on SERPs, bad informativeness is indicated by, e.g., a lower relative amount of hovers over the list of search results. However, in analogy to the given rating, the user's behavior would not change depending on the specific *cause* for a missing piece of information. This is because the user has no knowledge about this cause, which means that his interactions can only be influenced by the resulting higher-level fact that their desired information is not present.

Thus, we propose to extend WaPPU with a catalog of best practices for optimizing the usability of SERPs. For each of the seven usability metrics, this catalog maps bad scores y to a *set* of potential causes C and corresponding countermeasures C': $f(y) = \{C, C'\}$. Together, WaPPU and the catalog form *S.O.S.*—the SERP Optimization Suite. We will also describe a new workflow which ensures that developers can react to bad usability scores based on the catalog.

[13] The definition of “good enough” is up to the developer in this case.



	Interface "A" Based on **47** users.	*Interface "B"* Based on **42** users.
Overall Usability ✱	59.9% ± 21.1%	67.5% ± 17.6%
Informativeness	-0.17 ± 0.84	-0.02 ± 0.84
Understandability	0.34 ± 0.70	0.45 ± 0.67
Confusion	0.30 ± 0.78	0.38 ± 0.70
Distraction ✱	0.36 ± 0.74	0.62 ± 0.62
Readability	0.45 ± 0.65	0.52 ± 0.71
Information Density ✱	0.04 ± 0.69	0.43 ± 0.67
Accessibility	0.06 ± 0.67	0.07 ± 0.75

Interface **"B"** is *significantly better* than interface **"A"**.

Figure 2. Screenshot of the WaPPU dashboard showing the evaluation of the A/B test carried out during the case study. It shows the scores (mean and standard deviation) of each usability factor and the overall usability metric for the two involved SERPs (Interface "A" = original SERP, Interface "B" = redesigned SERP, * significant difference).

BEST PRACTICES FOR SERP USABILITY

The catalog for optimizing the usability of SERPs has been determined in a three-step process. This process corresponds to the *find–fix–verify* pattern described by Bernstein et al. in [5]. In the first (the 'find') phase, we have reviewed SERPs of eight popular search engines and extracted common best practices and shortcomings. From these, we have compiled a first version of our catalog. In the second (the 'fix') phase, we have asked ten dedicated experts to review the catalog and propose adjustments as required from their point of view. Finally, in the third (the 'verify') phase, each adjustment from the second phase has been reviewed by another three independent experts, who could either approve or reject it. The final catalog has then been created by only making those adjustments that were approved by at least two experts from the third phase. By making use of a state-of-the-art pattern [5] and having our catalog thoroughly reviewed by a total of 20 experts—which according to [20] find more than 75% of present errors—, we believe the final outcome to be well-founded and a valid instrument for usability optimization on SERPs.

All resources, including the SERP reviews, the first version of the catalog, the anonymized revised catalogs, the anonymized reviews of the adjustments and the *complete final catalog* can be accessed at `http://vsr.informatik.tu-chemnitz.de/demo/SOS`. The three phases applied for creating the catalog will be described in detail in the following.

'Find' Phase

In the first phase, we have reviewed the top five search engines as ranked by Alexa[14]: (1) *google.com*, (2) *yahoo.com*, (3) *baidu.com*, (4) *yandex.ru*, and (5) *bing.com*. Moreover, we have considered three trending search engines that have been intensively and repeatedly discussed among friends and colleagues—i.e., (A) *qwant.com*, (B) *duckduckgo.com*, and (C) *ecosia.org*. These follow new approaches (all-in-one search, anonymous search and green search) and as of September 3, 2014, have considerable *Klout*[15] scores (A: 53, B: 81, C: 63) and numbers of *Twitter*[16] followers (A: 1,544, B: 32.4K, C: 5,095).

	Web/Interface Design		**Web Development/Engineering**	
Phase	work	skills	work	skills
'Fix'	4	3	3.5	3
'Verify'	3	3	5	4
	Usability / User Experience		**Search Engines**	
	work	skills	work	skills
'Fix'	4	2	2	2
'Verify'	3.5	2	2.5	3

Table 1. Experts ratings of how much their work/studies are concerned with each of the given fields (1 = not at all, 5 = very much) and how they rate their own skills (1 = no knowledge, 4 = expert). Numbers given are median values.

From these eight SERP interfaces, we have extracted common best practices (approaches that could be found in the majority of the investigated SERPs); and shortcomings (approaches followed only by a minority of the SERPs that we considered to be problematic from the usability perspective). Best practices included, e.g., search suggestions / related search queries, semantic results and generous use of white space. Shortcomings included not clearly identifiable advertisements, infinite scrolling and overloaded results, among others. From this, we compiled a first (basic) version of our catalog of best practices. This basic version contained a total of 40 potential causes for bad scores as well as 61 corresponding countermeasures for the seven usability metrics.

'Fix' Phase

In the second phase, the basic catalog was given to ten dedicated experts in the fields of web/interface design, web development/engineering, user experience / usability and search

[14] `http://www.alexa.com/topsites` (Aug 17, 2014)

[15] `https://klout.com/home` (Sep 3, 2014).

[16] `https://twitter.com/` (Sep 3, 2014).

engines. Each one was asked to review the catalog and make adjustments as necessary from their point of view—i.e., by adding new causes/countermeasures or by altering or removing existing ones. They were also provided with examples of SERPs for the query "Albert Einstein" (Google, Bing, DuckDuckGo).

The experts performed a total of 110 changes (additions, removals, alterations) to the basic catalog, which is an average of 11 changes per expert. In [20], Nielsen and Molich state that "evaluations from several evaluators [...] do rather well, even if they only consist of three to five people." In analogy to this, we assume that in the 'fix' phase with 10 experts ($\gg$3–5), *at least* 71% [20] of the errors/problems in the basic catalog were found. Particularly, as the setting of revising a catalog of best practices given interface examples is not fundamentally different from heuristic evaluation of interfaces.

The experts' ratings of their skills in each of the relevant fields is given in Table 1. Four of the experts worked in academia, four worked in industry and seven were (graduate) students (multiple answers possible). Five considered themselves to be a practitioner, one to be a researcher, three said "half/half" and one said "don't know". Nine of the experts were male (one female) at an average age of 27.1 (σ=1.85); five owned a Bachelor's and the remaining five a Master's degree.

'Verify' Phase

In the last phase defined by the *find–fix–verify* pattern [5], the catalogs from phase two were anonymized and randomly assigned to another ten independent experts (three per expert) who were disjoint from the experts above. In this way, every adjustment proposed for the basic catalog was reviewed three times. The experts were instructed to compare the revised catalogs to the basic version and approve or reject each adjustment individually. Thus, valid adjustments were found by majority decision (at least two approvals). In total, 44 adjustments (40%) were approved with a 3–0 vote, 37 (33.64%) were approved with a 2–1 vote, 9 (8.18%) were rejected with a 3–0 vote and 20 (18.18%) were rejected with a 2–1 vote. In case two contradictory adjustments to the same cause/countermeasure had been approved (a proposed change vs. a proposed removal), we counted the overall votes (by six experts), which happened four times. Moreover, in case of a draw (3–3), the item was—*in dubio pro reo*—not removed from the catalog, which was the case thrice.

The experts' ratings of their skills in each of the relevant fields is given in Table 1. Six of the experts worked in academia and five in industry (multiple answers possible). Three considered themselves to be a practitioner, one to be a researcher and six said "half/half". Nine of the experts were male (one female) at an average age of 29.5 (σ=3.47); eight owned a Master's and one a PhD degree (one stated "other degree").

Final Catalog

The final catalog is the semantic union[17] of (a) all approved adjustments from the revised catalogs, and (b) the remaining items from the basic catalog. Since we considered the *find–fix–verify* pattern [5] as well as Nielsen and Molich's findings [20], we believe that our catalog is well-founded and a valid extension to the WaPPU tool during search engine design and development. In Figure 5, we provide an *exemplary* excerpt of the final catalog that consists of the causes and countermeasures related to the case study described below. For each usability factor, the potential *causes* for a bad rating are numbered with *Roman* numerals and *countermeasures* are numbered with *Arabic* numerals. Items that were added or adjusted by the experts and approved in the 'verify' phase have their cumulative votes given in square brackets *[pro–contra]*.

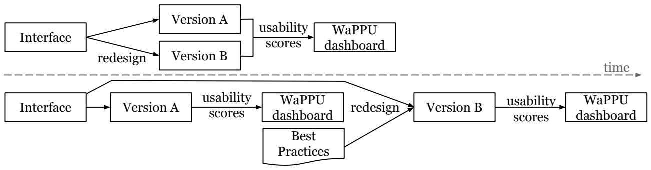
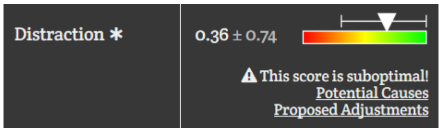


Figure 3. Usability-based A/B testing (above) vs. S.O.S.-supported sequential A/B testing (below).


Distraction ✱

0.36 ± 0.74

⚠ This score is suboptimal!

Potential Causes

Proposed Adjustments

Figure 4. We have equipped the dashboard of the WaPPU prototype to display warnings next to suboptimal scores. Clicking the links leads to the corresponding parts of the catalog of best practices.

S.O.S.: THE SERP OPTIMIZATION SUITE

The catalog (Figure 5) is intended to be an extension to the WaPPU service *in the context of search engine design and development*. Together, they form *S.O.S.*—the SERP Optimization Suite. Yet, when making use of S.O.S., there is the need for a different workflow compared to regular A/B testing. With the usual A/B testing workflow, both versions of the tested interface are available before the test starts. That is, the redesign intended to improve usability happens *a priori* and the two versions of the interface are evaluated in parallel (Figure 3). This is one of the main problems of standard A/B testing, as, e.g., Nielsen [19] points out. Contrary, when using S.O.S., we want to have empirical evidence of the original interface's usability first. Only *after* that, we identify potential causes for bad usability scores using the catalog and apply corresponding countermeasures, thus forming a redesigned interface. Then, also the new version of the interface is evaluated using WaPPU to investigate the changes made.

The WaPPU tool is technically not restricted to running A/B tests in parallel. Rather, the actual temporal distribution of users to the interfaces-under-test is completely up to the developer. That is, it is easily possible to direct 100% of the users to only one version of the tested interface. The A/B testing workflow to be used with S.O.S. is therefore to (1) first direct all users to interface A, (2) then consult the catalog of best practices and perform a redesign, and (3) finally direct all users to interface B to validate the redesign (Figure 3). To support this new workflow, we have already integrated according means into the WaPPU dashboard (Figure 4).

[17] If two or more causes/countermeasures express the same meaning in a different way, they are included as one item in the final catalog.

A. **Informativeness**
- I. bad index quality (desired result[s] not present on page)
- II. bad ranking quality (desired result[s] not present or ranked too low)
- III. desired result not clearly identifiable:
  - a. inappropriate title and/or abstract
  - b. too many other results
  - c. too much content other than results
- 2. ⑨ provide search suggestions / related search terms
- 6. clarify layout:
  - b. ①④⑧ clearly separate results from other content such as ads (or remove the latter) *[2–1]*
  - c. ①②③ reduce amount of content other than results
  - e. ④ clearly separate title from abstract *[3–0]*

B. **Understandability**

C. **Confusion**
- II. too many irrelevant results *[3–0]*
- III. too much content other than results
- IV. too many advertisements which are poorly marked as such; no clear separation between advertisements / sponsored results and real results *[6–0]*
- 2. ⑧ clearly highlight results and mark other content as such
- 5. ①②③ reduce amount of content other than results

D. **Distraction**
- I. too much content other than results
- III. too many images
- IV. non-results more salient than results *[2–1]*
- V. overloaded results (e.g., displaying secondary information, social media buttons etc.)
- 1. ①②③ reduce amount of content other than results
- 3. ③ reduce amount of images
- 4. ③④⑤⑥⑦⑧ ensure that results are more salient than other content *[3–0]*
- 5. clarify presentation of results:
  - a. ② reduce to: title, URL, abstract

E. **Readability**
- I. wrong font size or character spacing (too small / too large) *[9–0]*
- II. wrong line height (too small / too large) *[6–0]*
- V. text not properly grouped (e.g., via white space or separation lines) *[3–0]*
- VII. inconsistent alignment of results and/or other elements of the page
- 1. ⑤ adjust font size or character spacing (or offer according option to reader) *[11–1]*
- 2. ⑤ adjust line height *[9–0]*
- 5. add white space:
  - a. ④⑦ between title, URL and abstract of result
  - b. ④ between results
  - c. ④⑦ between results and other content
- 7. ⑥⑦ align results and other elements of the page consistently

F. **Information Density**
- III. too much content other than results; content that is not related to results *[3–0]*
- IV. too little white space
- V. missing visual hierarchy with salient results *[2–1]*
- VI. overloaded results (e.g., displaying secondary information, social media buttons etc.)
- 3. ①②③ reduce amount of content other than results
- 4. add white space:
  - a. ④⑦ between title, URL and abstract of result
  - b. ④ between results
  - c. ④⑦ between results and other content
- 5. ⑥⑦⑧ introduce contrast and visual hierarchy to separate results from content other than results *[3–0]*
- 10. clarify presentation of results:
  - a. ② reduce to: title, URL, abstract
- 11. ③ remove unnecessary icons/abbreviations or add explaining tooltips *[3–0]*

G. **Accessibility**
- I. too much scrolling effort for user:
  - a. too much content other than results, especially above results
  - d. bad ranking quality (desired result[s] not present or ranked too low)
- II. desired result(s) not immediately identifiable:
  - b. missing visual hierarchy *[2–1]*
  - c. missing contrast between results and other content *[3–0]*
- 1. reduce scrolling effort:
  - a. ① reduce amount of content other than results, especially above results
- 2. better highlight results / improve result presentation:
  - b. ⑧ introduce contrast and visual hierarchy to separate results from content other than results *[2–1]*
  - c. ③④⑤⑥⑦⑧ ensure that results stand out against other content

Figure 5. Excerpt of the final version of the catalog, showing the best practices applied during the case study (the circled numbers refer to Figure 7).

CASE STUDY

We made use of S.O.S. in a case study during which we evaluated a real-world SERP with the help of WaPPU.[18] The study was carried out in terms of a remote asynchronous user study with 81 users who triggered 198 search requests.

The catalog of best practices was taken into account for redesigning the original SERP, which reached only suboptimal usability scores in the evaluation. Our results show that, after identifying issues and taking appropriate countermeasures based on the catalog, the new version of the SERP performs significantly better concerning *distraction, information density* and also the *overall usability metric*. Furthermore, all remaining usability factors reached better scores after the redesign.

Evaluation: Original SERP

The case study was carried out with the SERP (for web search results) of a real-world search engine currently developed by the R&D department of the cooperating company. In the following, we report results for the largest and most representative group of users involved, i.e., novel users (they had never used the search engine before) with HD screens (N=89). During the A/B test, the original version of the SERP showed a suboptimal performance concerning its *overall usability*, with a score of 59.9% (σ=21.1%) based on 47 users. While this cannot be considered as "bad" in general, it is clearly below the threshold for good usability—which is 70% (cf. Section "The WaPPU Dashboard" above)—and thus not a satisfactory outcome from the company's point of view. Particularly, the evaluation revealed problems with respect to *informativeness* (m=-0.17, σ=0.84)[19], *information density*

[18] The case study has also been described by Speicher et al. [23], however, with a focus on evaluating the feasibility of WaPPU.

[19] The range of the score of an individual factor is $[-1, 1]$.

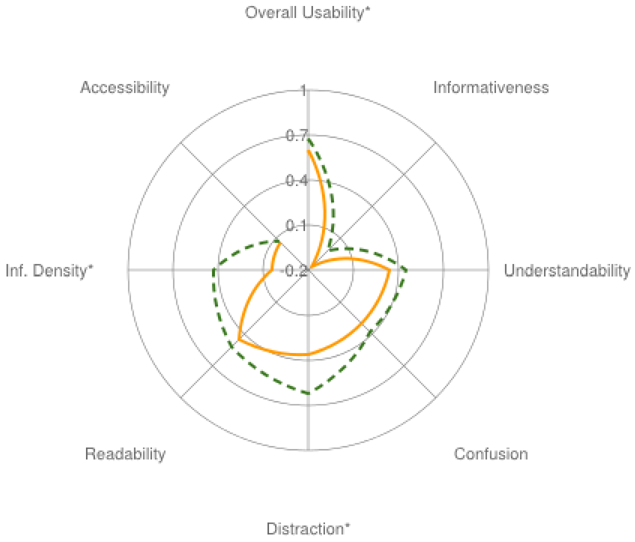


Figure 6. Results of the case study. The original SERP is given in orange, the redesigned SERP is given in dashed green (* significant difference).

(m=0.04, σ=0.69) and *accessibility* (m=0.06, σ=0.67). Contrary, the best-scoring factor was *readability* with a score of 0.45 (σ=0.65). In total, the scores of six out of seven factors were suboptimal, as they lay below the threshold of 0.4. The complete results for the original SERP are given in Figure 2 (left) and Figure 6.

Redesign

Based on the catalog of best practices, the original version of the SERP was evaluated by three experts in front-end design (one graphic designer, one interaction designer and one PhD student with focus on human–computer interaction). According to the workflow of S.O.S. and the usability scores provided by WaPPU, they identified a number of usability issues (the detailed items can be looked up in Figure 5):

- bad result quality (A.I./II., G.I.d.)
- relevant information not clearly identifiable (A.III., C.II., D.IV., G.II.)
- lack of white space (C.IV., E.V., F.IV./V.)
- too much other content above the list of results (A.III.c., C.III., D.I., F.III., G.I.a.)
- bad typography (E.I./II.)
- no clear structure (E.VII.)
- overloaded results (D.III./V., F.VI.).

To be able to apply a broader range of best practices in the case study, we also considered the factor *readability* for optimization, although its score was above the threshold of 0.4. Based on the identified causes for bad usability, the experts applied the corresponding adjustments given by the catalog (the circled numbers refer to Figure 7):

- ① reducing the amount of advertisements (A.6.b./c., C.5., D.1., F.3., G.1.a.)
- ②③ reducing results to title, URL and abstract; particularly by removing social media buttons (A.6.c., C.5., D.1./3./4./5.a., F.3./10.a./11., G.2.c.)
- ④⑤ adjusting results regarding white space, font size and line height (A.6.b./e., D.4., E.1./2., E.5.a./b./c., F.4.a./b./c., G.2.c.)
- ⑥ better aligning results (D.4., E.7., F.5., G.2.c.)
- ⑦ clearer separation between images and text (D.4., E.5.a./c., E.7., F.4.a./c., F.5., G.2.b.)
- ⑧ visualizing and separating results more clearly (A.6.b., C.2., D.4., F.5., G.2.b./c.)
- ⑨ optimizing the display of related search terms (A.2.)

A comparison between the old and the new SERP as well as details about the applied adjustments are given in Figure 7.

Evaluation: New SERP

Like the original version of the SERP, the redesign resulting from the above changes was evaluated using WaPPU. With respect to *overall usability,* the new version reached a score of 67.5% (σ=17.6%) based on 42 users. This was still below the threshold of 70%, but makes for a significant improvement ($p<0.05$, W=782)[20]. Furthermore, the item *distraction* raised by 0.26 points to a score of 0.62 (σ=0.62, $p<0.05$, W=798.5). A third significant improvement was achieved by the item *information density*, which gained 0.39 points to reach a new score of 0.43 (σ=0.67, $p<0.01$, W=692). Besides, the scores of all remaining usability items improved as well, however, not with statistical significance. Overall, four out of seven factors lay above the threshold of 0.4 after the redesign, which is an increase by three. The complete results for the redesigned SERP are given in Figure 2 (right). Moreover, Figure 6 illustrates that the redesign clearly dominates the original SERP across all usability factors.

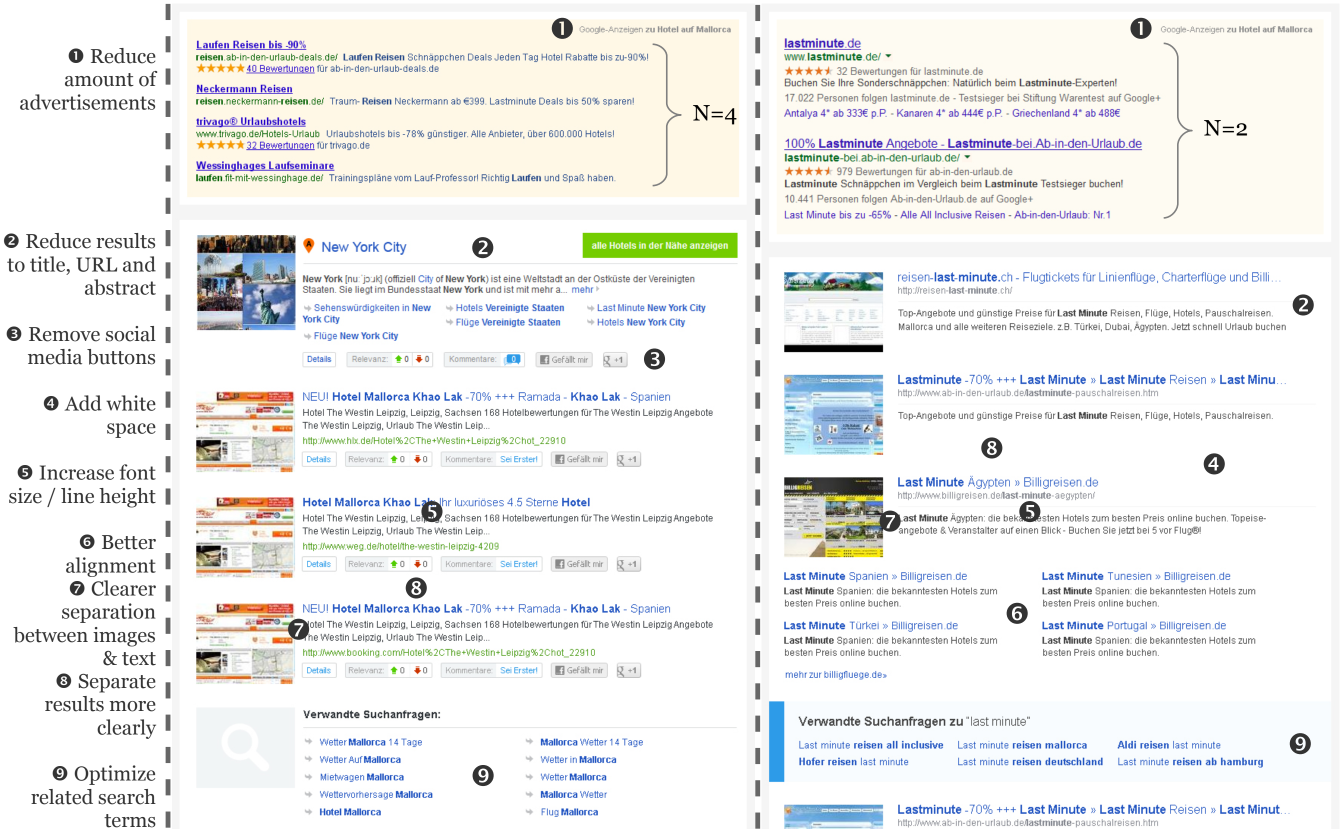
Although only a fraction of the catalog of best practices was applied during the case study, the above results are a reasonable proof of effectiveness. The experts asked to redesign the SERP applied a set of rather subtle adjustments (Figure 7) that were particularly aimed at reducing *distraction* while improving *readability* and *information density*. The fact that two of these as well as the *overall usability* metric showed significant improvements is a clear indication of the feasibility and effectiveness of S.O.S. and the contained catalog of best practices.

DISCUSSION

S.O.S. has shown its good potential for evaluating and improving the usability of SERPs. While a first case study underpins the feasibility of the approach, it still has certain shortcomings.

First of all, our presented catalog is restricted to best practices for *SERPs*. Contrary, WaPPU serves as a *general* tool for evaluating the usability of any kind of web interface. Thus, it would be desirable and possible to also provide a general catalog that is applicable to a broader range of interfaces (cf. *userium.com*), or a collection of more specific catalogs for different types of interfaces. Besides rules that are specifically aimed at SERPs (e.g., A.I./II., C.2.), our catalog already contains a subset of general rules (e.g., E.I./II., F.11.). However, compiling a general catalog or a set of more specific

[20] All tests of significance were carried out in terms of *Mann–Whitney U tests* (α=0.05).


❶ Reduce amount of advertisements
❷ Reduce results to title, URL and abstract
❸ Remove social media buttons
❹ Add white space
❺ Increase font size / line height
❻ Better alignment
❼ Clearer separation between images & text
❽ Separate results more clearly
❾ Optimize related search terms
N=4
N=2
❶
❷
❸
❹
❺
❻
❼
❽
❾

Google-Anzeigen **zu Hotel auf Mallorca**

Laufen Reisen bis -90%
reisen.ab-in-den-urlaub-deals.de/ **Laufen Reisen** Schnäppchen Deals Jeden Tag Hotel Rabatte bis zu-90%!
★★★★★ 40 Bewertungen für ab-in-den-urlaub-deals.de

Neckermann Reisen
reisen.neckermann-**reisen**.de/ Traum- **Reisen** Neckermann ab €399. Lastminute Deals bis 50% sparen!

trivago® Urlaubshotels
www.trivago.de/Hotels-Urlaub Urlaubshotels bis -78% günstiger. Alle Anbieter, über 600.000 Hotels!
★★★★★ 32 Bewertungen für trivago.de

Wessinghages Laufseminare
laufen.fit-mit-wessinghage.de/ Trainingspläne vom Lauf-Professor! Richtig **Laufen** und Spaß haben.

New York City

alle Hotels in der Nähe anzeigen

New York [nuːˈjɔːɹk] (offiziell City of **New York**) ist eine Weltstadt an der Ostküste der Vereinigten Staaten. Sie liegt im Bundesstaat **New York** und ist mit mehr a... mehr

Sehenswürdigkeiten in **New York City** · Hotels **Vereinigte Staaten** · Last Minute **New York City**
Flüge **New York City** · Flüge **Vereinigte Staaten** · Hotels **New York City**

Details · Relevanz: 0 0 · Kommentare: 0 · Gefällt mir · +1

NEU! **Hotel Mallorca Khao Lak** -70% +++ Ramada - **Khao Lak** - Spanien
Hotel The Westin Leipzig, Leipzig, Sachsen 168 Hotelbewertungen für The Westin Leipzig Angebote The Westin Leipzig, Urlaub The Westin Leip...
http://www.hlx.de/Hotel%2CThe+Westin+Leipzig%2Chot_22910
Details · Relevanz: 0 0 · Kommentare: Sei Erster! · Gefällt mir · +1

Hotel Mallorca Khao Lak Ihr luxuriöses 4.5 Sterne **Hotel**
Hotel The Westin Leipzig, Leipzig, Sachsen 168 Hotelbewertungen für The Westin Leipzig Angebote The Westin Leipzig, Urlaub The Westin Leip...
http://www.weg.de/hotel/the-westin-leipzig-4209
Details · Relevanz: 0 0 · Kommentare: Sei Erster! · Gefällt mir · +1

NEU! **Hotel Mallorca Khao Lak** -70% +++ Ramada - **Khao Lak** - Spanien
Hotel The Westin Leipzig, Leipzig, Sachsen 168 Hotelbewertungen für The Westin Leipzig Angebote The Westin Leipzig, Urlaub The Westin Leip...
http://www.booking.com/Hotel%2CThe+Westin+Leipzig%2Chot_22910
Details · Relevanz: 0 0 · Kommentare: Sei Erster! · Gefällt mir · +1

Verwandte Suchanfragen:

Wetter **Mallorca** 14 Tage · **Mallorca** Wetter 14 Tage
Wetter Auf **Mallorca** · Wetter in **Mallorca**
Mietwagen **Mallorca** · Wetter **Mallorca**
Wettervorhersage **Mallorca** · **Mallorca** Wetter
Hotel Mallorca · Flug **Mallorca**

Google-Anzeigen **zu Hotel auf Mallorca**

lastminute.de
www.**lastminute**.de/
★★★★★ 32 Bewertungen für lastminute.de
Buchen Sie Ihre Sonderschnäppchen: Natürlich beim **Lastminute**-Experten!
17.022 Personen folgen lastminute.de - Testsieger bei Stiftung Warentest auf Google+
Antalya 4* ab 333€ p.P. - Kanaren 4* ab 444€ p.P. - Griechenland 4* ab 488€

100% **Lastminute** Angebote - **Lastminute**-bei.Ab-in-den-Urlaub.de
lastminute-bei.ab-in-den-urlaub.de/
★★★★★ 979 Bewertungen für ab-in-den-urlaub.de
Lastminute Schnäppchen im Vergleich beim **Lastminute** Testsieger buchen!
10.441 Personen folgen Ab-in-den-Urlaub.de auf Google+
Last Minute bis zu -65% - Alle All Inclusive Reisen - Ab-in-den-Urlaub: Nr.1

reisen-**last-minute**.ch - Flugtickets für Linienflüge, Charterflüge und Billi...
http://reisen-**last-minute**.ch/
Top-Angebote und günstige Preise für **Last Minute** Reisen, Flüge, Hotels, Pauschalreisen. Mallorca und alle weiteren Reiseziele. z.B. Türkei, Dubai, Ägypten. Jetzt schnell Urlaub buchen

Lastminute -70% +++ **Last Minute** » **Last Minute** Reisen » **Last Minu**...
http://www.ab-in-den-urlaub.de/**lastminute**-pauschalreisen.htm
Top-Angebote und günstige Preise für **Last Minute** Reisen, Flüge, Hotels, Pauschalreisen.

Last Minute Ägypten » Billigreisen.de
http://www.billigreisen.de/**last-minute**-aegypten/
Last Minute Ägypten: die bekanntesten Hotels zum besten Preis online buchen. Topeiseangebote & Veranstalter auf einen Blick - Buchen Sie jetzt bei 5 vor Flug®!

Last Minute Spanien » Billigreisen.de
Last Minute Spanien: die bekanntesten Hotels zum besten Preis online buchen.

Last Minute Tunesien » Billigreisen.de
Last Minute Spanien: die bekanntesten Hotels zum besten Preis online buchen.

Last Minute Türkei » Billigreisen.de
Last Minute Spanien: die bekanntesten Hotels zum besten Preis online buchen.

Last Minute Portugal » Billigreisen.de
Last Minute Spanien: die bekanntesten Hotels zum besten Preis online buchen.

mehr zur billigfluege.de»

Verwandte Suchanfragen zu "last minute"

Last minute **reisen all inclusive** · Last minute **reisen mallorca** · **Aldi reisen** last minute
Hofer reisen last minute · Last minute **reisen deutschland** · Last minute **reisen ab hamburg**

Lastminute -70% +++ **Last Minute** » **Last Minute** Reisen » **Last Minut**...
http://www.ab-in-den-urlaub.de/**lastminute**-pauschalreisen.htm

Figure 7. A visual comparison of the original (left) and redesigned SERP (right) evaluated during the case study. Please note that although in this specific image advertisements take more space on the redesigned page, the actual number of advertisements was reduced from four to two. The above are not screenshots of the actual interfaces in the browser, but the official models provided by the cooperating company's graphic designers.

ones is not a trivial process, as their validity must be ensured in analogy to the approach presented in this paper. This is planned as our prime direction of future work.

Second, we are only able to provide *best practices* for optimizing SERP usability. This means that identifying concrete usability problems and adequate adjustments still involves manual work by designers that is already done today without the help of S.O.S. Yet, Nielsen and Molich [20] found that even "good evaluators may sometimes overlook easy problems". Such "easy" problems (like a wrong font size) are often easily overseen and become obvious only when they are pointed out (by a colleague or, in our case, a catalog of best practices). Thus, we believe that S.O.S. is a valuable tool even for experienced designers. Another important point is that our suite facilitates communication with non-designers, as it relates concrete directives for optimization to usability metrics and scores. In this way, necessary adjustments to an interface can be better communicated towards superiors and other company officials.

Since during the case study we could only identify a fraction of causes for bad usability and corresponding countermeasures, not the complete catalog of best practices has been tested. Therefore, a more thorough case study involving each item of the catalog should happen as part of future work. However, it has to be noted that our catalog also includes a number of high-level instructions for adjusting the backend of a search engine—like improving the index or ranking quality—, which are not trivial to perform. Still, the results of the case study highlight the potential of our approach.

Finally, our catalog points out that advertising often stands in contrast to good usability. However, as one expert involved in verifying the catalog stated, "removing advertisements is unrealistic, as they are the only source of revenue [for search engines]". Therefore, although no advertisements would be optimal from the usability perspective, an adequate balance has to be found. This is necessary to ensure that search engines are able to provide free services to their users.

CONCLUSIONS

We have presented S.O.S.—the SERP Optimization Suite, which consists of the WaPPU tool [23] and a catalog of best practices for identifying and eliminating usability problems on SERPs. While WaPPU provides scores for seven different usability factors, for each of these our catalog contains potential causes for a suboptimal score (i.e., below 0.4) and proposes corresponding countermeasures. This relates necessary adjustments for optimization to usability metrics and scores and thus facilitates communication with non-designers and company officials. However, S.O.S.'s catalog of best practices can also be used as a standalone tool for designing user-friendly SERPs from scratch. During a case study, S.O.S. was used to evaluate and redesign the SERP of a real-world search engine. For the original SERP, WaPPU detected suboptimal scores for six out of seven usability factors as well as an overall score of 59.9%. After applying a selection of countermeasures from the catalog, the redesigned SERP yielded good usability scores (>0.4) for four out of seven factors. Also the overall score improved significantly to 67.5%.

Regarding future work, we particularly intend to apply our approach to web interfaces beyond SERPs. This includes the development of a general catalog as well as corresponding user studies. Also, S.O.S. could be integrated with crowdsourcing platforms (e.g., *Amazon MTurk API*[21]). In this way, tasks for identifying usability problems and selecting appropriate adjustments based on the catalog could be automatically posted to MTurk by WaPPU once the score of a factor drops below 0.4. This would be a considerable step into the direction of fully automatic usability evaluation *and* optimization.

ACKNOWLEDGMENTS

We thank all experts who contributed to our catalog of best practices. Special thanks go to R. Frankenstein, B. Freiberg, V. Nguyen, T. Stangneth and K. Zatinschikow, who helped with the case study. This work has been supported by the ESF and the Free State of Saxony.

Gefördert aus Mitteln der Europäischen Union

Europa fördert Sachsen. ESF Europäischer Sozialfonds

REFERENCES

1. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, 1998.

2. ISO/IEC 25010:2011: Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models, 2011.

3. Atterer, R., Wnuk, M., and Schmidt, A. Knowing the User's Every Move – User Activity Tracking for Website Usability Evaluation and Implicit Interaction. In *Proc. WWW* (2006).

4. Bangor, A., Kortum, P. T., and Miller, J. T. An Empirical Evaluation of the System Usability Scale. *IJHCI 24*, 6 (2008).

5. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: A Word Processor with a Crowd Inside. In *Proc. UIST* (2010).

6. Bozzon, A., Brambilla, M., and Comai, S. A Characterization of the Layout Definition Problem for Web Search Results. In *OTM Workshops* (2010).

7. Brooke, J. SUS: A "quick and dirty" usability scale. In *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, Eds. Taylor and Francis, 1996.

8. Carta, T., Paternò, F., and Santana, V. F. Web Usability Probe: A Tool for Supporting Remote Usability Evaluation of Web Sites. In *Proc. INTERACT* (2011).

9. de Vasconcelos, L. G., and Baldochi Jr., L. A. Towards an Automatic Evaluation of Web Applications. In *Proc. SAC* (2012).

10. Gajos, K., and Weld, D. S. Preference Elicitation for Interface Optimization. In *Proc. UIST* (2005).

11. Gajos, K., Wobbrock, J. O., and Weld, D. S. Improving the Performance of Motor-Impaired Users with Automatically-Generated, Ability-Based Interfaces. In *Proc. CHI* (2008).

12. Grigera, J., Garrido, A., and Rivero, J. M. A Tool for Detecting Bad Usability Smells in an Automatic Way. In *Proc. ICWE (Demos)* (2014).

13. Hassenzahl, M. User Experience (UX): Towards an experiential perspective on product quality. In *Proc. IHM* (2008).

14. John, G. H., and Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In *Proc. UAI* (1995).

15. Lew, P., Olsina, L., and Zhang, L. Quality, Quality in Use, Actual Usability and User Experience as Key Drivers for Web Application Evaluation. In *Proc. ICWE* (2010).

16. Nebeling, M., Matulic, F., and Norrie, M. C. Metrics for the Evaluation of News Site Content Layout in Large-Screen Contexts. In *Proc. CHI* (2011).

17. Nebeling, M., Speicher, M., and Norrie, M. C. CrowdAdapt: Enabling Crowdsourced Web Page Adaptation for Individual Viewing Conditions and Preferences. In *Proc. EICS* (2013).

18. Nebeling, M., Speicher, M., and Norrie, M. C. W3Touch: Metrics-based Web Page Adaptation for Touch. In *Proc. CHI* (2013).

19. Nielsen, J. Putting A/B Testing in Its Place, 2005. `http://www.nngroup.com/articles/putting-ab-testing-in-its-place/`, retrieved September 1, 2014.

20. Nielsen, J., and Molich, R. Heuristic Evaluation of User Interfaces. In *Proc. CHI* (1990).

21. Speicher, M. W3Touch: Crowdsourced Evaluation and Adaptation of Web Interfaces for Touch. Master's thesis, ETH Zurich, Switzerland, 2012.

22. Speicher, M., Both, A., and Gaedke, M. Towards Metric-based Usability Evaluation of Online Web Interfaces. In *Mensch & Computer Workshopband* (2013).

23. Speicher, M., Both, A., and Gaedke, M. Ensuring Web Interface Quality through Usability-based Split Testing. In *Proc. ICWE* (2014).

[21] `https://www.mturk.com/mturk/welcome` (Sep 10, 2014).