Proportional Search Interface Usability Measures

Mika Käki
Department of Computer Sciences
FIN-33014 University of Tampere, Finland
mika.kaki@cs.uta.fi
+358 3 215 6181

ABSTRACT

Speed, accuracy, and subjective satisfaction are the most common measures for evaluating the usability of search user interfaces. However, these measures do not facilitate comparisons optimally and they leave some important aspects of search user interfaces uncovered. We propose new, proportional measures to supplement the current ones. *Search speed* is a normalized measure for the speed of a search user interface expressed in answers per minute. *Qualified search speed* reveals the trade-off between speed and accuracy while *immediate search accuracy* addresses the need to measure success in typical web search behavior where only the first few results are interesting. The proposed measures are evaluated by applying them to raw data from two studies and comparing them to earlier measures. The evaluations indicate that they have desirable features.

Author Keywords

Search user interface, usability evaluation, usability measure, speed, accuracy.

ACM Classification Keywords

H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces: Evaluation/methodology.
H3.3. Information Storage and Retrieval: Information Search and Retrieval: Search Process.

INTRODUCTION

In order to study the usability of search user interfaces we need proper measures. In the literature, speed, accuracy and subjective satisfaction measures are common and reveal interesting details. They have, however, a few shortcomings that call for additional measures.

First, comparing results even within one experiment—let alone between different experiments—is hard because the measures are not typically normalized in the research reports but multiple raw numbers (like answers found and time used) are reported. Of course, unbiased comparison between studies will always be difficult as the test setup has a big effect on the results, but the problem is compounded by the presentation of multiple task dependent measures. A good measure would be as simple as possible, yet it must not discard relevant information.

Second, the current measures do not reveal the sources of speed differences. In particular, the relation between speed and accuracy may be hard to understand since the current measures for those dimensions are completely separate. For example, it is essential to know if the increase in speed is due to careless behavior or better success.

Third, in the web environment, a typical goal for a search is to find just a few *good enough* answers to a question. This is demonstrated by studies that show that about half of the users only view one or two result pages per query [11]. Current search user interface usability measures do not capture the success of such a behavior very well.

In order to address these problems, we present three new proportional, normalized usability measures. The new measures are designed for the *result evaluation* phase of the search process [10] where real users are involved. *Search speed* is a normalized speed measure expressed in answers per minute. It makes within study comparisons simple and between studies bit more feasible. *Qualified search speed* is a combination of speed and accuracy measures that reveals the tradeoff between speed and accuracy. It shows the source of speed differences in terms of accuracy and is also measured in answers per minute. *Immediate search accuracy* is a measure that captures the success of result evaluation when only the first few hits are interesting. These new measures are evaluated by applying them to data from real experiments and comparing them to conventional measures.

RELATED WORK

In usability evaluations, the measurements are typically based on the three major components of usability: effectiveness, efficiency, and satisfaction [3, 4]. International ISO 9241-11 standard [4] defines *effectiveness* as the “accuracy and completeness with which the users achieve specified goals” and *efficiency* as “resources expended in relation to the accuracy and completeness with which users achieve goals”. According

NordiCHI '04, October 23-27, 2004 Tampere, Finland
Copyright 2004 ACM 1-58113-857-1/04/10... $5.00

to the standard, efficiency measure divides the effectiveness (achieved results) by the resources used (e.g. time, human effort, or cost). In this work, we will leave satisfaction measures out of the discussion and concentrate on objective quantitative measures.

Usability measurements are strongly domain dependent. In the search user interface domain effectiveness is typically measured in terms of accuracy (which is recognized as an example measure in the ISO standard as well). Time (speed of use) is typically used as the critical resource when calculating the efficiency.

In the following we will discuss measuring practices in typical studies evaluating search user interfaces. Note that although almost every study in the information retrieval community deals with searching, they tend to focus on system performance [8] and thus only a few studies are mentioned here.

Speed Measures

The basic approach for measuring the speed is simply to measure the time required for performing a task, but the actual implementation differs from study to study. In early evaluations of the Scatter/Gather system by Pirolli *et al.* [6], times were recorded simply on a task basis. In the results they reported how many minutes it took, on average, to complete a task. In the study by Dumais *et al.* [2], roughly the same method was used, except that the times were divided into categories according to the difficulty of the task. Sebrechts *et al.* [9] used a different categorization method where task execution times were divided into categories according to the subject's computer experience.

Time measurements can also be recorded in a somewhat reversed manner as Pratt and Fagan [7] did. They reported how many results users found in four minutes. This is close to measuring speed (achievement / time), but this normalization to four minutes is arbitrary and does not facilitate comparisons optimally. In a study by Dennis *et al.* [1], the time to bookmark a result page was measured and only one page was bookmarked per task. This setup makes the comparison fairly easy since the reported time tells how much time it takes to find a result with the given user interface. However, this desirable feature was caused by the setup where only one result was chosen, and other types of tasks were not considered.

Accuracy Measures

Accuracy measures are based on the notion of relevance which is typically determined by independent judges in relation to a task. In information retrieval studies, accuracy is typically a combination of two measures: *recall* and *precision*. Recall describes the amount of relevant results found in a search in a relation to all the relevant results in the collection. As a perfect query in terms of recall could return all the entries in the collection, it is counterbalanced with the precision measure. Precision describes how clean the result set is by describing the density of relevant results in it. Precision, like recall, is expressed by a percentage number which states the proportion of relevant targets in the result set.

Recall and precision measures are designed for measuring the success of a query. In contrast, when the success of the result evaluation process is studied, the users need to complete the process by selecting the interesting results. Measures are then based on analyzing the true success of the selections. Recall and precision measures are used here too, but the calculation is different. In these cases recall describes the amount of relevant results selected in relation to the amount of them in the result set. Precision, on the other hand, describes the density of relevant results among the selected results.

Veerasamy and Heikes [13] used such measures (called interactive recall and interactive precision) in their study of a graphical display of retrieval studies. They asked participants to judge the relevance of the results in order to get the users' idea of the document relevance. Pirolli *et al.* [6] used only the precision measure in their test of the Scatter/Gather system. The selection of the results was implemented by a save functionality. Dennis *et al.* [1] used an approach where they reported the average relevance of the results found with a given user interface. Relevant results were indicated by bookmarking them. Further variations of the measures where user interaction is taken into account in accuracy evaluation were proposed and used by Veerasamy and Belkin [12].

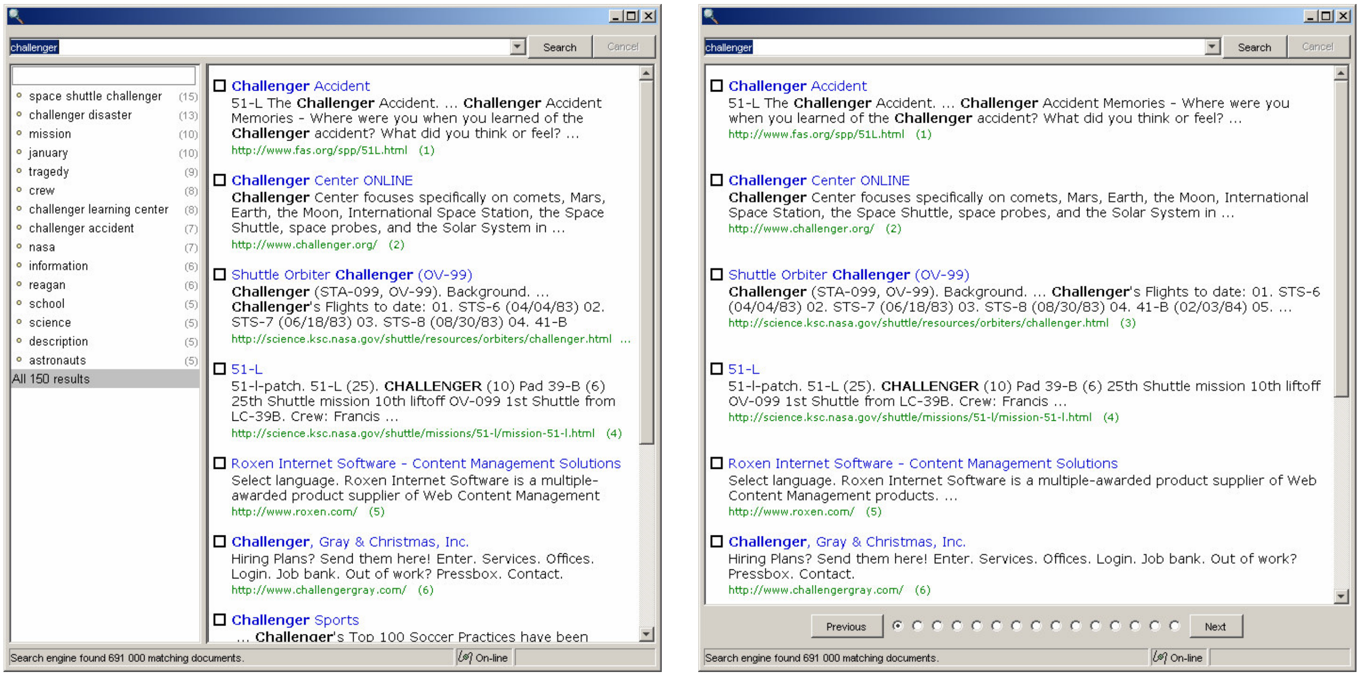
Information Foraging Theory

Stuart Card, Peter Pirolli and colleagues have made extensive research on information foraging theory [5] in Xerox Parc and the results are relevant here as well. In its conventional form information foraging theory states that the rate of gain of valuable information (R) can be calculated using the formula:

$$R = \frac{G}{T_B + T_W} \tag{1}$$

In the formula, G is the amount of gained information, T_B is the total time spent between information patches and T_W is the total time spent within an information patch [5]. An information patch is understood to mean a collection of information such as a document collection, a search result collection or even a single document that can be seen to be a collection of information that requires some actions for digesting the information. In the information foraging process, the forager navigates first between patches and then finds actual meaningful information within a patch. The process is then started over by seeking a new patch.

If we discard the separation of two different types of activities (between and within patches) for simplicity, equation 1 states the information gain rate in terms of time unit. This matches with common practices in the field and is the basis for our proposed measurements as well.



space shuttle challenger	(15)
challenger disaster	(13)
mission	(10)
january	(10)
tragedy	(9)
crew	(8)
challenger learning center	(8)
challenger accident	(7)
nasa	(7)
information	(6)
reagan	(6)
school	(5)
science	(5)
description	(5)
astronauts	(5)
All 150 results	

☐ **Challenger** Accident
51-L The **Challenger** Accident. ... **Challenger** Accident Memories - Where were you when you learned of the **Challenger** accident? What did you think or feel? ...
http://www.fas.org/spp/51L.html (1)

☐ **Challenger** Center ONLINE
Challenger Center focuses specifically on comets, Mars, Earth, the Moon, International Space Station, the Space Shuttle, space probes, and the Solar System in ...
http://www.challenger.org/ (2)

☐ Shuttle Orbiter **Challenger** (OV-99)
Challenger (STA-099, OV-99). Background. ... **Challenger**'s Flights to date: 01. STS-6 (04/04/83) 02. STS-7 (06/18/83) 03. STS-8 (08/30/83) 04. 41-B
http://science.ksc.nasa.gov/shuttle/resources/orbiters/challenger.html ...

☐ 51-L
51-l-patch. 51-L (25). **CHALLENGER** (10) Pad 39-B (6) 25th Shuttle mission 10th liftoff OV-099 1st Shuttle from LC-39B. Crew: Francis ...
http://science.ksc.nasa.gov/shuttle/missions/51-l/mission-51-l.html (4)

☐ Roxen Internet Software - Content Management Solutions
Select language. Roxen Internet Software is a multiple-awarded product supplier of Web Content Management
http://www.roxen.com/ (5)

☐ **Challenger**, Gray & Christmas, Inc.
Hiring Plans? Send them here! Enter. Services. Offices. Login. Job bank. Out of work? Pressbox. Contact.
http://www.challengergray.com/ (6)

☐ **Challenger** Sports
... **Challenger**'s Top 100 Soccer Practices have been

Search engine found 691 000 matching documents.

☐ **Challenger** Accident
51-L The **Challenger** Accident. ... **Challenger** Accident Memories - Where were you when you learned of the **Challenger** accident? What did you think or feel? ...
http://www.fas.org/spp/51L.html (1)

☐ **Challenger** Center ONLINE
Challenger Center focuses specifically on comets, Mars, Earth, the Moon, International Space Station, the Space Shuttle, space probes, and the Solar System in ...
http://www.challenger.org/ (2)

☐ Shuttle Orbiter **Challenger** (OV-99)
Challenger (STA-099, OV-99). Background. ... **Challenger**'s Flights to date: 01. STS-6 (04/04/83) 02. STS-7 (06/18/83) 03. STS-8 (08/30/83) 04. 41-B (02/03/84) 05. ...
http://science.ksc.nasa.gov/shuttle/resources/orbiters/challenger.html (3)

☐ 51-L
51-l-patch. 51-L (25). **CHALLENGER** (10) Pad 39-B (6) 25th Shuttle mission 10th liftoff OV-099 1st Shuttle from LC-39B. Crew: Francis ...
http://science.ksc.nasa.gov/shuttle/missions/51-l/mission-51-l.html (4)

☐ Roxen Internet Software - Content Management Solutions
Select language. Roxen Internet Software is a multiple-awarded product supplier of Web Content Management products. ...
http://www.roxen.com/ (5)

☐ **Challenger**, Gray & Christmas, Inc.
Hiring Plans? Send them here! Enter. Services. Offices. Login. Job bank. Out of work? Pressbox. Contact.
http://www.challengergray.com/ (6)

Search engine found 691 000 matching documents.

Figure 1. Compared user interfaces in our experiment. Category user interface on the left, reference user interface on the right.

The gap that is left in the information foraging theory in a relation to making concrete measurements, is the definition of information gain. The gap is well justified as the definition would unnecessarily reduce the scope of the theory. On the other hand, when we deal with concrete problems, we can be more specific and thus obtain preciseness. This is our approach here: we apply the basic relationships stated in the information foraging theory and provide meaningful ways of measuring the gain. All this is done in the context of evaluating search user interfaces in the search result evaluation phase. We will get back to this topic in the discussions of the new measures to see their relationship to the information foraging theory in more detail.

EXPERIMENT

We will evaluate the proposed measures using data from an experiment of ours. This experiment was conducted to evaluate a new search user interface idea by comparing it to the *de facto* standard solution.

Our proposed user interface used automatically calculated categories for facilitating the result access (Figure 1, left). As the categories we used the most common words and phrases found within the result titles and text summaries (snippets). Stop word list and a simple stemmer were used for improving the quality of the categories (e.g. discarding very common words such as ‘and’ or ‘is’). As the category word (or phrase) selection was based solely on the word frequencies, the categories were neither exclusive nor exhaustive. There was a special built-in category for accessing all the results as one long list. The hypothesis behind the category user interface was such that it would allow users to identify and locate interesting results easier and faster than the conventional solution.

The calculated categories were presented to the user as a list beside the actual result list. When a category was selected from the list, the result listing was filtered to display only those result items that contained the selected word or phrase. There were a total of 150 results that the user could access and from which the categories were computed.

Participants

There were 20 volunteer participants (8 male, 12 female) in the experiment. Their average age was 35 years varying from 19 to 57 years and they were recruited from the local university. Almost all of the participants can be regarded as experienced computer users, but none of them was an information technology professional.

Apparatus

There were two user interfaces to access the search results:

1. *The category interface* (category UI, Figure 1, left) presented the users with a list of 15 automatically generated categories on the left side of the user interface. When the user selected a category, the corresponding results were shown on the right side of the user interface much like in popular e-mail clients.
2. *The reference interface* (reference UI, Figure 1, right) was a Google web search engine imitation showing results in separate pages, ten results per page. The order of the results was defined by the search engine (Google). In the bottom of the window, there were controls to browse the pages in order (*Previous* and

Next buttons) or in random order (a radio button for each page). There were 15 pages so that the participants could access a total of 150 results.

Design and Procedure

The experiment had *search user interface* as the only independent variable with two values: category UI and reference UI. The values of the independent variable were varied within the subjects and thus the analysis was done using repeated measures tools. As dependent variables we measured: 1) time to accomplish a task in seconds, 2) number of results selected for a task, 3) relevance of selected result in a three step scale (relevant, related, not relevant), and 4) subjective attitudes towards the systems.

The experiments were carried out in a usability laboratory. One experiment lasted approximately 45 minutes and contained 18 (9+9) information seeking tasks in two blocks: one carried out with the category interface and the other using the reference interface. The order of the blocks and the tasks were counterbalanced between the participants. For each task, there was a ready-made query and users did not (re)formulate the queries themselves. This kind of restriction in the setup was necessary to properly focus on measuring the success in the result evaluation phase of the search.

The actual task of the participant was to “collect as many relevant results for the information seeking task as possible as fast as you can”. The participants collected results by using check boxes that were available beside each result item (see Figure 1).

In the test situation there were two windows in the computer desktop. The *task window* displayed information seeking tasks for the participants who were instructed to first read the task description, then push the ‘Start’ button in the task window and promptly proceed to accomplish the task in the *search window*. Upon task completion (participant’s own decision or time-out), the participants were instructed to push the ‘Done’ button in the task window. The time between ‘Start’ and ‘Done’ button presses was measured as the total time for the task. This timing scheme was explained to the participants. Time for each task was limited to one minute.

Accuracy measures are based on ratings done by the experimenter (one person). The rating judgments were made based solely on the task description and the very same result title and summary texts that the participants saw in the experiment. Actual result pages were not used because it would have added an extra variable into the design (result summary vs. page relation), which we did not wish. All the tasks had at least two defining concepts like in “Find *pictures* of *planet Mars*”. For relevant results, all of the concepts was required to be present in some form (different wording was of course allowed). Related results were those where only the most dominant concept was present (e.g. planet Mars). Rest of the results was considered to be not relevant.

RESULTS

For comparing the proposed measures we present here the results of our experiment using the conventional measures: time, number of results, and precision. The time measure did not reveal very interesting results, because the test setup limited the total time for one task to one minute. Thus the mean times for conditions were close to each other: 56.6 seconds ($sd = 5.5$) for the category UI and 58.3 seconds ($sd = 3.5$) for the reference UI. The difference is not statistically significant as repeated measures analysis of variance (ANOVA) gives F(1,19) = 3.65, *ns*.

In contrast, number of results revealed a difference. When using the category UI the participants were able to find on average 5.1 ($sd = 2.1$) results per task whereas using the reference UI yielded on average 3.9 ($sd = 1.2$) selections. The difference is significant since ANOVA gives F(1,19) = 9.24, $p < .01$.

Precision measure gave also a statistically significant difference. When using the category UI on average 65% ($sd = 13$) of the participants’ selections were relevant in a relation to the task. The corresponding number for the reference UI was 49% ($sd = 15$). ANOVA gave F(1,19) = 14.49, $p < .01$.

The results are compatible with other studies done with similar categorizing search user interfaces. For example, Pratt and Fagan [7] have also reported similar results in favor of categorizing user interface. When categories work, they enhance the result evaluation process by reducing the number of items that need to be evaluated. Users find interesting looking categories and evaluate only the results within those categories. Concentration of relevant documents in the interesting categories is higher than in the whole result set.

SEARCH SPEED

Definition

In order to make the comparison of speed measures easier, we suggest a proportional measure. When the search time and number of results are combined into one measure, just like in measuring physical speed by kilometers or miles per hour, we get a search user interface **search speed** measure expressed in *answers per minute (APM)*. It is calculated by dividing the number of answers found by the time it took to find them:

$$search\ speed = \frac{answers\ found}{minutes\ searched} \qquad (2)$$

In relation to the ISO-9241-11 standard this is an efficiency measure whereas the plain number of answers is an (simple) effectiveness measure. In terms of information foraging theory, we replace the G term in equation 1 with number of results found and the time is normalized to minutes. This concretizes the rate (R) in equation 1 to be answers per minute. The structure of equations 1 and 2 is essentially the same.

Whenever two (or more) measures are reduced into one, there is a risk of loosing relevant information. This is the case here as well. The proposed measure does not make the distinction between a situation where one answer is found in 10 seconds and a situation where four answers are found in 40 seconds. In both cases the speed is 6 answers per minute and the details of the situation are lost in the measurement. However, we feel that speed measure is nevertheless correct also in this case. The situation can be compared to driving 50 km/h for 10 or 40 minutes. The traveled distance is different, but the speed is the same. This means that proposed speed measure does not apply in every situation and attention must be paid in measurement selection.

The problem of reducing two measures into one has also been thoroughly discussed by Shumin Zhai [14] in the context of input devices. He points out that reduction of two Fitts' law variables (*a* and *b*) in calculating throughput of an input device leads to a measure that is dependent of the task. The same problem does not apply here as our situation is not related to Fitts' law. However, our measure is dependent on the task, but it is not dependent of the used time or the number of results collected like previous measures.

Evaluation

In order to evaluate the suggested measure it was applied to the results of Scatter/Gather evaluation by Pirolli *et al.* [6]. In their experiment the task was to find relevant documents for a given topic. The table below summarizes the results (SS = similarity search, SG = scatter/gather):

Measurement	SS	SG
Original		
Time used in minutes	10.10	30.64
Number of answers	16.44	12.26
Search speed		
Answers per minute	1.62	0.40

The first two rows show the actual numbers reported in the paper while the third row shows the same results in answers per minute. It is arguably easier to understand the relationship between the two user interfaces from the normalized search speed measure. It communicates that the SS condition was roughly four times faster than the SG condition. The relation is hard to see from the original results. In addition, measurements can be easily related to one's own experiences with similar user interfaces because of the normalization.

In the second table below, the search speed measure is applied to the data from our own experiment. Here the difference between raw numbers and normalized measure is not as large as in the previous example because the time used for the tasks is roughly the same in both cases due to the test setup. Nevertheless, the suggested measure makes the comparison easier. Note also that the fairly large difference with the speeds in the experiment by Pirolli *et al.* is presumably due to experiment set-up (tasks, conditions, equipment, etc.).

Measurement	Category UI	Reference UI
Raw numbers		
Time used in minutes	0.94	0.97
Number of answers	5.1	3.9
Search speed		
Answers per minute	5.4	4.0

When an analysis of variance is calculated on the answers per minute measure, we see a bit stronger result compared to the conventional measures where just the number of results revealed significant difference. Here ANOVA gives F(1,19) = 11.3, p < .01. Slight increase in the F statistic is due to the combination of two measures that both have a difference in the same direction. In summary, search speed measures the same phenomena as the previously used measures (it is calculated from the same numbers) and it can make distinctions between the measured objects.

QUALIFIED SEARCH SPEED

Definition

Previously used recall and precision measures do not directly tell where possible speed differences come from or what the relation between speed and accuracy is. The suggested **qualified search speed** measure refines the search speed measure with categories of relevance to address this shortcoming. To keep the measure understandable and robust, we use only two or three categories of relevance. Like the previous measure, the qualified search speed is also measured in answers per minute, with a distinction that the speed is calculated separately for each relevance category according to the equation 3. There *RCi* stands for relevance category *i* (typical categories are e.g. relevant and irrelevant).

$$qualified\ search\ speed_{RCi} = \frac{answers\ found_{RCi}}{minutes\ searched} \qquad (3)$$

Note that the sum over all relevance categories equals to the normal search speed.

When qualified search speed is described in information foraging terminology, we can see that the gain is now defined more precisely than with search speed. While search speed takes into account only the number of results, qualified search speed adds the quality of the results into the equation. In essence, this gives us a more accurate estimate of the gain of information, and thus a more accurate rate of information gain. Note that this shows also in the rate magnitude: rate is now stated in (e.g.) number of relevant results per minute.

Evaluation

When the qualified search speed measure is applied to the data of our experiment and compared to the simple measure of precision, a few observations can be made. First, the proposed measure preserves the statistically significant

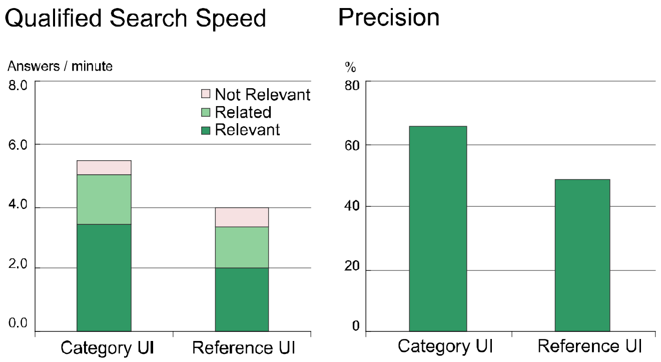


Figure 2. Qualified search speed measure compared to precision measure of data gathered in our own study.

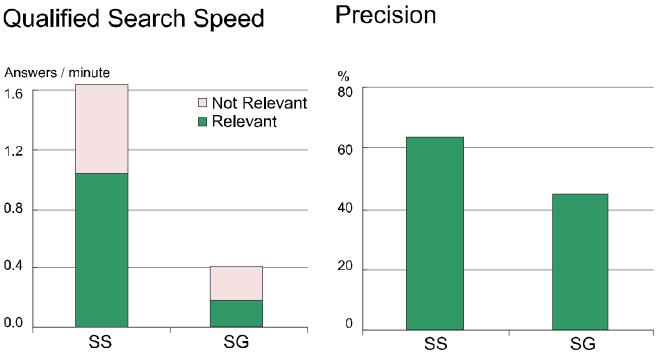


Figure 3. Qualified search speed measure compared to precision measure in the Scatter/Gather study [4].

difference that was observed with the conventional precision measure. ANOVA for the speed of acquiring relevant results gives $F(1,19) = 32.4$, $p < .01$.

Second, both measures (Figure 2) convey roughly the same information about the precision of the user interfaces including: 1) with the category UI more than half of the selected results were relevant whereas with the reference UI about half of the results were relevant, and 2) using the category UI participants were more successful in terms of precision. However, with the suggested qualified search speed measure, the amplitude of difference in precision is not obvious and thus the new measure cannot replace the old one.

Third, in addition to what can be seen in the precision chart, the qualified search speed chart (Figure 2) reveals some interesting data. It shows that the improvement in speed is due to the fact that participants have been able to select more *relevant* results while the proportion of not relevant results decreased a bit. The same information could surely be acquired by combining conventional speed and precision measures, but when the information is visible in one figure it is arguably easier to find such a relationship. Note also that although the new measure is mainly concerned about the accuracy of use, it informs the reader simultaneously about the speed of use as well.

Figure 3 makes a comparison between the new measure and the original precision measure using the data collected in the Scatter/Gather experiment [6]. Here it is worthwhile to note that even though precision measures are close to those in the previous example, the qualified search speed measure reveals large differences between the conditions. Qualified search speed seems to reveal the tradeoff between accuracy and speed convincingly in this case. We can also notice that both conditions here are much slower than those in Figure 2 as the qualified search speed is normalized just like the simpler search speed.

It is notable that qualified search speed does not measure the same phenomena as precision and thus they are not replaceable. We can image a situation where high qualified speed is associated with low precision and vice versa. In reality this could happen when users try to be very precise in one condition and very fast in another. On the other hand, we saw that qualified evaluation speed can make clear distinctions between user interfaces, which is a compulsory quality for a useful measure.

IMMEDIATE ACCURACY

Definition

The last suggested measure captures the success of typical web search behavior. In such a task, the user wants to find a piece of information that would be good enough for an information need and overall speed and accuracy are not as important as quick success. The measure is called **immediate accuracy** and it is expressed as a success rate. The success rate states the proportion of cases where at least one relevant result is found by the n^{th} selection. For applying the measure, the order of each result selection must be stored and the relevance of them must be judged against the task. The selections for each task and participant are then gone through in the order they were made and the frequency of first relevant result finding is calculated for each selection (first, second, and so on). When this figure is divided by the total number of observations (number of participants * number of tasks) we get the percentage of first relevant result found per each selection. Equation 4 shows the calculation more formally, there n stands for n^{th} selection.

$$immediate\ accuracy_n = \frac{number\ of\ first\ relevant\ results_n}{total\ number\ of\ observations} \quad (4)$$

When the figures calculated with equation 4 are plotted into a cumulative line chart (Figure 4) we can see when at least one relevant result is found on average. For example, (in Figure 4) after the second selection in 79 % of the cases at least one relevant result is found when using the category user interface. Notice also that the lines do not reach the 100 % value. This means that in some of the cases the users were not able to find any relevant results.

When looking back to information foraging theory, this measure takes us to a different approach compared to the previous ones. This measure abandons time as the limiting

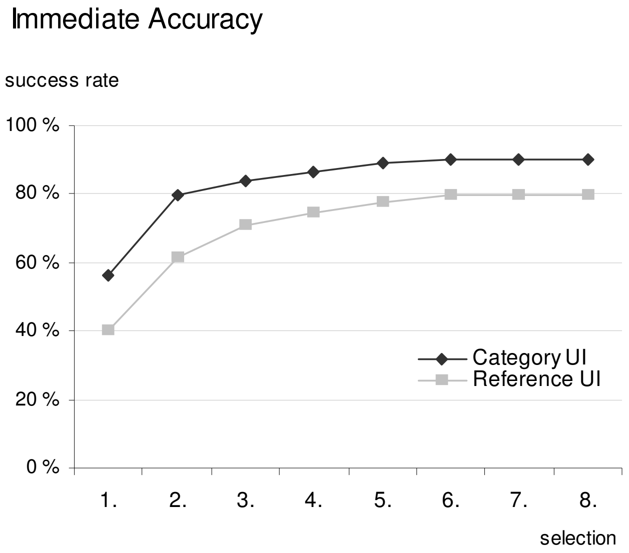


Figure 4. Immediate accuracy of category UI and reference UI. The measure shows the proportion of the cases where a relevant result have been found at *n*th selection.

resource against which the gain is compared and replaces it by selection ordinal (remember that ISO standard leaves the choice of resource up to the domain). As this new resource is discrete in nature, the expression of the measure as a single figure (rate) becomes hard and thus, for example, a cumulative chart is preferred for easily interpretable figures. From another perspective of information foraging theory, we can say that immediate accuracy is a measure for estimating the beginning of the within patch gain slope. Note, that it is only an estimation of the *beginning* of the slope as all subsequent relevant selections are discarded in this measure. In this view, we define an information patch to be a search result set.

Evaluation

The evaluation is based only on our own data because the measure requires information that is typically not reported in the publications. Figure 4 shows that the user orientates faster while using the category UI as the first selection produces already a relevant result in 56 % of the cases. In contrast, the reference UI produces a relevant result in 40 % of the first selections. By the second selection, the difference is bit greater since in 79 % of the cases the users have found at least one relevant result with the category UI, while the corresponding number for the reference UI is 62 %.

In the analysis of cumulative data, the most interesting points are those where the difference between compared measurements changes. Change points are most important because cumulative figures will preserve the difference if not further changes happen. In our case the difference is made at the first selection and remains virtually the same afterwards. This difference is statistically significant as ANOVA gives F(1,19) = 12.5, p < .01 and it is preserved throughout the selections (F(1,19) ≥ 10.4, p < .01 for all subsequent selections).

Findings of Spink *et al.* [11] stated that users only select one or two results per query. Immediate accuracy allows us to see the success of the studied user interface in such a case. We can focus on a given selection and quickly see the success rate at that point. Note that this kind of information is not available using the conventional accuracy measures and straightforward speed measures.

Immediate Success Speed

Another fairly simple and obvious way for measuring immediate success would be to record the time to the first relevant result. We did try this measure as well, but found a problem.

In our experiment, the average time to find the first relevant result was practically the same in both cases (20 and 21 seconds for category and reference UI respectively) and there was no statistically significant difference. This could, of course, be the true situation, but the amount of relevant results suggested the opposite.

The problem comes from the fact that the first relevant result is not always found. With the category UI users were not able to find a single relevant result for a task in 10% of the cases whereas the same number for reference UI was 21%. We felt that this is a big difference and that it should be visible in the measurement as well. However, we were not able to come up with a reasonable solution for normalizing the time measurement in this respect and thus the measurement is not promoted as such.

In addition, the results of Spink *et al.* [11] suggest that the time to first relevant result is not very important for the search process. Since searchers tend to open only one or two results, the time does not seem to be the limiting factor, but the number of result selections is. This supports also the choice of immediate accuracy over the time to the first relevant result.

DISCUSSION

Our goal was to provide search user interface designers, researchers, and evaluators with additional measures that would complement the current ones. The first problem with them is that result comparison is hard, even within one experiment. Proportional measures makes within study comparisons easy and in addition they let readers relate their previous experience better to the presented results. We proposed normalized *search speed* measure that is expressed in answers per minute. As the measure combines two figures (number of answers and time searched) into one proportional number, it makes the comparisons within an experiment easy and between experiments bit more feasible.

The second shortcoming of the current measures is the fact that it is difficult to see the tradeoff between speed and accuracy. To address this problem, we proposed the *qualified search speed* measure that divides the search speed measure into relevance categories. The measure allows readers to see what the source of speed in terms of accuracy is. In the evaluation we showed that conventional measures may only tell the half of the story. For instance, in the case of the Scatter/Gather experiment the precision measure showed only moderate difference between the systems whereas qualified speed revealed a vast difference in the gain of relevant results. Combining speed and accuracy measures is particularly effective in such a case as it eliminates the need to mentally combine the two measures (speed and accuracy).

The third weakness of the current measures is their inability to capture users' success in typical web search behavior where the first good enough result is looked for. We proposed the *immediate accuracy* measure to solve this flaw. Immediate accuracy shows the proportion of the cases where the users are able to find at least one relevant result per n^{th} result selection. It allows readers to see how well and how fast the users can orient themselves to the task with the given user interface. As the measurements are made based on finding the first relevant result, the reader can compare how well different solutions support users' goal of finding the first relevant answer (and presumably few others as well) to the search task.

The proposed measures are not intended to replace the old measures, but rather to complement them. They lessen the mental burden posed to the reader as important information of different type (e.g. speed, accuracy) is combined into one proportional measure. In summary, the proposed measures capture important characteristics of search user interface usability and communicate them effectively.

The issue of making comparisons between experiments is not completely solved by these new measures. We feel that the problem is not in the properties of the new measures but in the nature of the phenomena to be measured. In the context of search user interfaces the test settings have a huge effect on the results that cannot be solved simply with new measures. One solution for the problem could be test setup standardization. In the TREC interactive track such an effort have been taken, but it seems that the wide variety of research questions connected to searching cannot be addressed with a single standard test setup.

ACKNOWLEDGMENTS

This work was supported by the Graduate School in User Centered Information Technology (UCIT). I would like to thank Scott MacKenzie, Kari-Jouko Räihä, Poika Isokoski, and Natalie Jhaveri for invaluable comments and encouragement.

REFERENCES

1. Dennis, S., Bruza, P., McArthur, R. Web Searching: A Process-Oriented Experimental Study of Three Interactive Search Paradigms. *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 2, 2002, 120-133.
2. Dumais, S., Cutrell, E., Chen, H. Optimizing Search by Showing Results in Context. *Proceedings of ACM CHI'01 (Seattle, USA)*, ACM Press, 2001, 277-284.
3. Frøkjær, E., Hertzum, M., Hornbæk, K. Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? *Proceedings of ACM CHI'2000 (The Hague, Netherlands)*, ACM Press, 2000, 345-352.
4. ISO 9241-11: *Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability,* International Organization for Standardization, March 1998.
5. Pirolli, P. and Card, S. Information Foraging. *Psychological Review,* 1999, Vol. 106, No. 4, 643-675.
6. Pirolli, P., Schank, P., Hearst, M., Diehl, C. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. *Proceedings of ACM CHI'96 (Vancouver, Canada)*, ACM Press, 1996, 213-220.
7. Pratt, W., Fagan, L. The Usefulness of Dynamically Categorizing Search Results. *Journal of the American Medical Informatics Association,* Vol. 7, No. 6, Nov/Dec 2000, 605-617.
8. Saracevic, T. Evaluation of Evaluation in Information Retrieval. *Proceedings of ACM SIGIR'95 (Seattle, USA)*, ACM Press, 1995, 138-146.
9. Sebrechts, M., Vasilakis, J., Miller, M., Cugini, J., Laskowski, S. Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. *Proceedings of ACM SIGIR'99 (Berkeley, USA)*, ACM Press, 1999.
10. Shneiderman, B., Byrd, D., Croft, B. Clarifying Search: A User-Interface Framework for Text Searches. *D-Lib Magazine*, January 1997.
11. Spink, A., Wolfram, D., Jansen, M., and Saracevic, T.: Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology,* 2001, Vol. 52, No. 6, 226-234.
12. Veerasamy, A., Belkin, N. Evaluation of a Tool for Visualization of Information Retrieval Results. *Proceedings of ACM SIGIR'96 (Zurich, Switzerland)*, ACM Press, 1996, 85-92.
13. Veerasamy, A., Heikes, R. Effectiveness of a Graphical Display of Retrieval Results. *Proceedings of ACM SIGIR'97 (Philadelphia, USA)*, ACM Press, 1997, 236-244.
14. Zhai, S. *On the validity of Throughput as a Characteristic of Computer Input*. IBM Research Report, RJ 10253, IBM Research Division. August 2002.