A HUMAN FACTORS EXPERIMENT WITH STUDENTS' GOOGLE SEARCHING

S M Zabed Ahmed ***Tania Hossain***

Abstract

This paper reports on a human factors experiment with students' Google searching. Two different student groups (novice and experienced) volunteered in this study. They carried out five search tasks and their performance was recorded through a computer screen recording application. Data was captured on the time taken, error rates and success score. After completion of search tasks, they all completed a questionnaire on their satisfaction with Google. The performance data showed that overall experienced students performed better. Differences were significant across all performance measures between groups. Students held neither highly positive nor highly negative perceptions about Google. Experienced students were significantly more satisfied with Google than the naive group. The findings could be used to redesign the present Google search engine.

Keywords : Search Engine; Google; Research Methodology

1. Introduction

The growth of the Internet has been phenomenal, particularly since 1993 with the introduction of the World Wide Web (also known as WWW, or simply the web) as a user-friendly graphical interface to Internet resources. The amount of information on the web is growing rapidly as well as the number of new users. The web has also created new challenges for information retrieval. The search engines are evolved to deal with these new challenges. The content, organization, comprehensiveness and search features of search engines vary. Each engine, depending on the size of its index and its searching, provides a service to its users different from each other.

Today, search engines became the de facto method of obtaining information from the web. So many search engines exists; some are large and provide multiple features, while others are smaller and offer limited functionality. Oppenheim et al. (2000) classified search engines as robots, directories, meta search and software tools. Rowley and Farrow (2000) classified them in a very similar way, as keyword, directories, meta search, and subject gateways. Some search engines combine characteristics of more than one of the above categories.

Although the search engines search an enormous volume of information at apparently impressive speed, often they have been the subject of widespread criticism. The reasons include their unfriendly interfaces and search facilities, lack of clear search instructions, lack of transferability, etc. This paper applied human factors criteria to assess students' performance and satisfaction with Google (http://www.google.com).

2. Previous research

A number of studies compared the effectiveness of various search engines. The majority of these studies have been carried out with popular robot-based engines using different search queries and analyzing varying number of results for relevance. For a comprehensive review of such studies, interested readers should consult Chu and Rosenthal (1996) and Oppenheim et al. (2000). Readers can also check latest reviews available at: http://www.searchengineshowdown.com and http://searchenginewatch.com.

A number of comparative evaluation studies involved Google. For example, Ljosland (1999) made a small comparison among three search engines: AltaVista, AlltheWeb and Google. Twelve queries were sent to the engines, and the ten first retrieved documents were evaluated using a three-point relevance scale: relevant (1), partly relevant (0.5), and not relevant (0). Google performed best both in terms of relevant and partly relevant documents. More recently, Mathes (2004) compared four popular search engines that focus exclusively on news content: Altavista News, Daypop, Google News, and Yahoo News. This experiment was designed to evaluate how four popular news search engines compared in both the relevance of results returned and the nature of the sources of those results. The results showed that Altavista and Google had the top overall relevance score, with Altavista just slightly higher. Daypop and Yahoo were third and fourth respectively.

The majority of the comparative studies into the effectiveness of search engines have involved to recall and precision or surrogates of these measures such as relative recall. Lieghton and Srivastava (1999) however, noted that many of these studies reported conflicting conclusions as to which services are better. In addition, many studies had either been based upon small set test, or did not report their methodology. They suggested that an unbiased set of queries must be developed to test search engines objectively. Several studies used other criteria, such as response time, number of web pages covered and coverage, freshness/broken links, search syntax, subject areas/ choice of query, search options, etc. for the evaluation of web search engines.

It has been however, strongly argued that search engines should be evaluated from a user's perspective rather than evaluation based on recall and precision (Dong and Su, 1997; Oppenheim et al. 2000). A human factors evaluation criteria, such as time taken to complete a benchmark tasks, errors, success of the task and subjective satisfaction with the interface, is likely to help designers to get more information about users' needs and goals, thus helping them designing such systems. This study therefore, applied a human factor evaluation criterion to Google in order to assess student's performance and satisfaction with the engine.

3. Research Methodology

Twenty-four postgraduate and undergraduate students of the Department of Information Science and Library Management, Dhaka University, Bangladesh took part in this study. They were divided into two groups: novice (Sn) and experienced (Se). The Sn group (n=12) comprised of undergraduate students and they all reported not having used the Internet before this experiment. The Se group (n=12) on the other hand, consisted of both undergraduate and postgraduates students. They all had varying levels of Internet use experience.

Students came once at a time for the experiment. At the beginning, they were given a brief description of the purpose of the study and the experimental procedures of the session that would be followed. Since novice users had not used Internet earlier, they were given 10 minutes to explore Google searching. The objective was to familiarize them with the interface so that they felt comfortable in performing the actual search tasks. For experienced students, this was not needed, as they were already familiar with the web. Students were also told that if any task took more than 10 minutes to complete, they would be stopped. If they felt that they would be unable to complete search task and wanted to move on, this would be allowed. They were then given the search tasks (see below) and told to try to work on their own.

WinCam 2000 was used to record screen activities of each student's entire search session. It recorded how each student searched Google. After completion of all tasks, students were asked to complete a questionnaire on their satisfaction with the Google. This questionnaire was designed taking items from QUIS (Chin et al., 1988).

3.1 Search tasks

Students were given the following five search tasks. These tasks were obtained from a survey conducted with web users at Dhaka University, Bangladesh.

Task 1: What is the total collection of Dhaka University Library?

Task 2: Find two job sites in Bangladesh

Task 3: Find the website of the World Education Services (WES)

Task 4: Where is the headquarters of the British Council?

Task 5: Find information about SAARC Agricultural Information Centre

Most of these tasks are general in nature. Few searches require using phrases and Boolean operators in search terms.

3.2 Variables studied

The data collected from this study was analysed, according to the following performance and satisfaction measurement criteria.

3.2.1 Performance variables:

- Time taken: The total time taken to complete each search task. These times were extracted from the computer screen recordings.
- Error rates: Total number of errors made was tabulated from screen recordings.
- Success score: Successful completion of each search task, as well as requested termination, and termination as a result of the twenty-minutes time limit was counted from screen recordings.

3.2.2 Subjective satisfaction

The Questionnaire for User Interface Satisfaction (QUIS) data was used to determine users' subjective satisfaction with Google on a 7 point scale.

3.3 Hypotheses

The null hypotheses developed for the study were:

H1: There is no difference between novice (Sn) and experienced (Se) students in total time taken to complete tasks.

H2: There is no difference between Sn and Se students in total number of errors made.

H3: There is no difference between Sn and Se students in total success score of search tasks.

H4: There is no difference between Sn and Se students in subjective satisfaction with Google.

4. Results of the study

4.1 Performance variables:

The time taken to complete each search task was rounded to the nearest minute. The task completion time included both task completion time, instances of requested termination and termination as a result of the ten-minute time limit. Table 1 shows the average time taken to complete each search task by both Sn and Se students.

	Time taken (mins.)		*Error rates*		*Success score*	
	Sn	*Se*	*Sn*	*Se*	*Sn*	*Se*
Task 1	4.45 (1.30)	3.47 (1.05)	2.00 (0.85)	1.42 (1.44)	0.42 (0.51)	0.67 (0.49)
Task 2	4.30 (1.29)	2.39 (1.89)	2.08 (1.00)	0.75 (0.97)	0.17 (0.39)	0.75 (0.45)
Task 3	3.20 (1.62)	2.38 (1.60)	1.42 (1.73)	0.92 (1.00)	0.67 (0.49)	0.92 (0.29)
Task 4	4.53 (0.75)	3.55 (1.25)	1.83 (0.72)	1.25 (0.97)	0.00 (0.00)	0.58 (0.51)
Task 5	2.78 (2.11)	1.75 (1.44)	1.08 (0.67)	0.50 (0.90)	0.50 (0.52)	0.92 (0.29)
Overall	19.27(4.53)	13.54 (5.29)	8.41 (2.64)	4.83 (2.91)	1.75 (0.75)	3.83 (1.47)

Table 1: Overall performance results

The number of errors made by two search groups was counted separately. Table 1 shows the average number of errors made by both novice and experienced groups. The novice group made more errors in completing tasks than the experienced group. Novices are particularly poor in using Boolean operators and phrase searching. Most of them entered the search terms as they appear in the search tasks.

The success score of a search task was scored as 1 if the search task was successful or O if it was unsuccessful. No partial Credit was given. So, the maximum average success score for a task was 1, if all searchers in the group were successful. Table 1 shows the average success score by each group. Overall experienced students performed better than the novice users. It can be seen that experienced students took less time, made few errors and were more successful in completing the search tasks. The independent sample t-test results for time taken, errors and success score is shown in Table 2.

Performance variables	t-value	df	Sig. (2-tailed)
Time taken	-2.847	22	.009
Error rate	-3.152	22	.005
Success score	4.376	22	.000

Table 2: The independent t-test results for time taken, error rates, and success score

The result showed that there were significant differences in time taken, errors and success score between novice and experienced students. Thus, the null hypotheses H1, H2 and H3 are rejected.

4.2. Subjective satisfaction:

Means and standard deviations (in parentheses) of data collected through QUIS are shown in Table 3. Analysis of the QUIS data revealed that students held neither highly positive nor highly negative perceptions of Google.

Question	***Sn***	***Se***	***Question***	***Sn***	***Se***
Overall reactions			*Terminology and Feedback*		
Terrible vs. wonderful	5.83 (1.64)	6.50 (0.79)	Simple and natural dialogue	4.91 (2.02)	6.33 (0.89)
Unimpressive vs. impressive	5.63 (1.68)	6.33 (0.98)	Terms used in the system	5.66 (1.61)	5.00 (0.73)
Difficult vs. Easy	5.50 (1.62)	6.00 (1.70)	Position of message	5.41 (1.50)	5.81 (0.88)
Inefficient vs. efficient	5.00 (1.62)	6.25 (0.87)	Prompts for input	5.75 (1.54)	5.91 (1.08)
Useless vs. useful	6.08 (1.08)	6.58 (0.67)	Inform about work progress	5.41 (1.37)	6.25 (0.75)
Unfriendly vs. friendly	6.16 (0.93)	6.91 (0.75)	Error messages	5.28 (1.27)	5.63 (1.43)
Frustrating vs. satisfying	5.66 (1.66)	5.91 (1.78)	*Learning*		
Ineffective vs. powerful	6.00 (1.12)	6.08 (1.31)	System learning	5.90 (1.22)	6.16 (1.40)
Dull vs. stimulating	6.16 (1.02)	6.16 (0.93)	Exploring by trial and error	5.30 (1.15)	5.09 (1.98)
Rigid vs. flexible	5.33 (1.96)	6.00 (1.70)	Remembering commands	5.90 (1.10)	6.25 (0.87)
Screen			Performing tasks is simple	5.27 (1.10)	6.17 (0.71)
Reading Characters	5.75 (1.35)	6.33 (0.88)	Help messages on the screen	6.36 (0.80)	5.90 (1.29)
Onscreen information	5.58 (0.79)	6.16 (1.40)	Help access	5.72 (1.42)	6.10 (0.88)
Information arrangement	5.63 (1.68)	6.33 (0.89)	*System capabilities*		
Easy to find information	5.81 (1.16)	5.72 (1.55)	System speed	5.18 (1.67)	4.33 (2.05)
Screen sequencing	5.33 (1.55)	5.91 (1.16)	System reliability	5.27 (1.34)	4.50 (1.98)
Screen back track	6.16 (1.19)	5.83 (1.89)	Correcting mistakes	5.10 (1.38)	6.08 (1.31)
Back to main screen	6.16 (1.19)	6.58 (1.16)	Designed for all levels of users	5.81(0.98)	5.33 (1.88)

Table 3: Overall QUIS results

The Mann-Whitney U-test was carried out to see the differences between novice and experienced students regarding subjective satisfaction with Google. The result of the test (Table 4) showed that subjective satisfaction by the novice and experienced users with Google differed significantly. Thus, the null hypothesis H4 is rejected.

	Group	Mean Rank	Sum of Ranks	Mann-Whitney U	Wilcoxon WZ	Asymp. Sig (2-tailed)
Subjective Satisfaction	*Sn*	25.55	843	335.5	843	0.001
	Se	41.45	1368			

Table 4: Mann-Whitney results for subjective satisfaction

6. Discussions and conclusions

This paper discusses the results of a human factors experiment with students' Google searching. Twenty-four undergraduate and postgraduate students were assigned to perform five information search tasks using Google. WinCam 2000, a screen recording application, was used to record students' Google performance. Students' subjective satisfaction with the Google was also collected through QUIS. Although overall search performance was poor, the results showed that students were able to complete simple search tasks. However, there were some tasks that even experienced students found difficult to complete and they could not retrieve relevant information using Google. The major difficulty for both students groups had been in using Boolean operators. Novices were particularly poor in using Boolean searches. Most of them entered search terms as they appeared in the tasks. It is evident that Google failed to assist the students in their information search process, as even the experienced students were unable to retrieve relevant information. This suggests that Google searching is difficult. The need for better design that contributes better performance by both novice and experienced students remains.

The results showed significant performance differences between novice and experienced students across all performance measures. Experienced students took less time, made fewer errors and were more successful in completing the search tasks than naive students. The QUIS also showed that experienced students were significantly more satisfied with Google than novice students. It is quite expected that whose who are more proficient with the system are more likely to be satisfied with it. It would be interesting to see if novices' satisfaction changes over an extended period of time.

With the rapid proliferation of the Internet, millions of users now have access to the web though various search engines. These users typically have no search experience or training. Today, search engines became an integral part of obtaining information from the Internet. Google is certainly a leader in the Internet world that serves millions of users. However, it needs improvements to serve its clientele better.

References

1. Chin, J. P., Diehl, V. A. and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In: CHI '88: Proceedings of the Conference on Human Factors in Computing Systems, 15-19 May, Washington, DC. New York: ACM, 213-218
2. Chu, H., Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. Proceedings of the ASIS Annual Meeting. Available at: http://www.asis.org/annual-96/ElectronicProceedings/chu.html
3. Dong, X. Y. and Su, L. T. (1997). Search engines on the World Wide Web and information retrieval from the Internet: a review and evaluation. Online and CD-ROM Review, 21(2), 67-82
4. Lieghton, H. V. and Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). Journal of the American Society for Information Science, 50(10), 870-881
5. Ljosland, M. (1999). Evaluation of web search engines and the search for better ranking algorithms. SIGIR'99 Workshop on Evaluation of Web Retrieval. Available at: http://www.aitel.hist.no/~mildrid/dring/paper/SIGIR.html
6. Oppenheim, C., Morris, A., McKnight, C., and Lowley, S. (2000). The evaluation of WWW search engines. Journal of Documentation, 56(2), 190-211
7. Rowley, J. E. and Farrow, J. (2000). Organizing Knowledge: An Introduction to Managing Access to Information. Burlington, VT: Gower
8. Mathes, A. (2004). A Comparison of WWW News Search Engines. Available at: http://www.adammathes.com/academic/search-engines/news/

[All web references were checked on 28 August, 2006]