Search String Analysis from a Socio-economic Perspective

Theo Mc Donald and Pieter Blignaut

Department of Computer Science, University of the Free State, PO Box 339, Bloemfontein, South Africa, 9300
Theo.sci@mail.uovs.ac.za

Abstract. Search string analysis has implications for developing better designs of Web interfaces and search engines. It is expected that millions of new users from Africa will enter the Internet arena in the not so distant future. Most of these users will be from countries with a low socio-economic standing. In order to determine the effect of socio-economic standing on the search behaviour of grade ten learners, their search characteristics were analysed. This study found that there is a difference in the search behaviour between novice users with a low socio-economic standing and those with a high socio-economic standing. These differences, however, only lasted for the first few sessions, where after all users showed the same search behaviour.

Keywords: Search strings, socio-economic status, Web searching, search characteristics.

1 Introduction

Internet access is becoming available to more and more people on a daily basis. So far people on the wrong side of the digital divide had to sit and watch as those on the other side reap the benefits of the information highway. African Internet usage has trebled to over 12 million since 2000 with 133 users per 10 000 people on the African continent [1] and 139 per 10 000 for the Black population in South Africa [2] having access to the Internet. These figures still compare very unfavourably with the world figure of 1460 per 10 000 people. It reflects even worse if it is compared with the figure of 3 680 per 10 000 for Europe and 6 800 per 10 000 for North America [1].

Maybe that is about to change. At the recent World Summit of the Information Society held in Tunis, Tunisia, the "100 dollars lap-tops for each student "- campaign was launched by the UN Secretary General Kofi Annan. The goal of this campaign is to provide the machines free of charge to children in poor countries who cannot afford computers of their own. It is hoped that the lime-green machines can be placed in the hands of millions of school children around the globe [3].

Given that the above initiatives come into fruition, millions of new users from Africa will enter the Internet arena in the not so distant future. Most of these users will eventually resort to finding information on the Web. What better way is there to do that than utilizing a search engine? Knowing the answers to questions like "Who is

C. Stephanidis (Ed.): Universal Access in HCI, Part I, HCII 2007, LNCS 4554, pp. 234–242, 2007.

searching?", "How are the searches conducted?" and "How successful are the searches?" are all important for several reasons. It has implications for developing better designs of Web interfaces and search engines. Additionally, it has implications in the education of search techniques when the difficulties with the search process are better understood.

A reasonable body of research in respect of searching the Web has accumulated over the last decade. Most of this research focused on the developed world. As far as can been determined, this paper is a start in order to tell the story of the developing world.

In the following sections some background of relevant research in this area will be given and it will be followed by the methods used and the results of the current study. The paper will be concluded with a discussion and conclusion.

2 Previous Research on Searching the Web

The research in the area of how users search the Web can broadly be grouped into large scale studies based on the analysis of log files of prominent search engines and small scale studies that focused more on different groupings of users. The users in the large data sets from the search engines are mostly anonymous and therefore the entries in the log contain only the identification of the users' machine, the time of the entry and the query. These studies determine the number of queries in a session, the number of words in a query, the pages browsed during the query, the information searched for and advanced features used in the query. These studies have the advantage of large data sets from an operational environment, but the user demographics are not available. In the small scale studies data is captured by observation or at the client side. The different groupings of users have to perform different tasks and then their searching patterns are evaluated and compared. In these studies the demographics of the users are known, but the number of users is low. The research in this paper will be based on something in between, having the best of both worlds.

In an overview paper on search trends from 1997 to 2003, Spink and Jansen [4] reported findings from an ongoing series of studies analyzing large-scale data sets obtained from the Excite search engine. They found a steady increase from 1996 to 1999 in the mean number of terms in queries from 1.5 to 2.6 and 2.4 by 2003. An increase in the number of Boolean operators was also identified. Most users searched one query only. The average session included 1.6 queries. More than 70% of the time a user only views the top ten results. In general users view about five Web documents per query.

Another large scale study was the Alta Vista study of Silverstein, et al. [5] who studied approximately one billion search requests over a period of six weeks. They showed that Web users type in short queries, mostly look at the first ten results only and seldom modify the query.

Turning to small scale studies, Amed et al. [6] reported on the performance and satisfaction of novice and expert users with the Web of Science interface. The users had to perform seven search tasks and their performance was recorded through transaction logging and screen recording. Overall, the experienced users performed

better than the novice group. They also found the number of search terms used for each task to be between 1.90 and 4.00. Hölscher and Strube [7] also used expert and novice users in their study. They compared the search queries of 12 expert participants doing web-based information seeking tasks to a large sample of average users from a German search-engine. They found a difference in the average query length of the experts (3.64 words) and the average users (1.66 words). They also found that experts made use of advanced search options much more frequently than average users.

Because the Web is a global phenomenon, it is important to determine if there are regional differences in searching behaviour. To the best of our knowledge only a few such studies are reported in the literature. Spink et al. [8] studied the search characteristics of users from the United States and Europe. They analyzed large data sets from the Excite search engine (US) and the FAST search engine (German). Their results show differences between the two populations in terms of searching behaviour and topics searched. Kralisch and Mandl [9] found that the cultural background of users have a great impact on their search strategies. They used data gathered from logs of a large international E-Health website.

Mc Donald and Blignaut [10] tested black and white university students on how successful they searched the website of the university. No differences were found between the two groups. Blignaut and Mc Donald [2] reported on the performance of black and white grade ten school children doing different tasks on the Internet. No difference between black and white experienced users was found, but significant differences between black and white inexperienced users were found. Whereas the research in that paper focused on the success of the search, this paper will focus on the characteristics of how the school children conducted the search. This paper endeavours to add to the research on search string analysis by reporting on the differences in search behaviour of different socio-economic groups. For the purposes of this study high socio-economic standing will mean those that can afford computer equipment and Internet access and low socio-economic standing will mean those that currently cannot afford computers and Internet access.

3 Methodology

All the grade ten learners from six different schools were brought to computer laboratories of the University of the Free State in order to be tested on various aspects of Internet usage. Three of the schools were from the traditionally advantaged communities with the availability of computers and Internet access (the so-called “haves”). The other three schools were from the traditionally disadvantaged schools with a lack of ICT access and located in a lower socio-economic area (the “have-nots”). The numbers from the different schools are shown in Table 1. The learners were brought to the university in groups of 50-60 and in this way 527 learners were tested. A session is defined as a task attempted by the user by supplying at least one search string. A query is a search string entered by a user in order to complete a task. A term is the words used for each search string.

Table 1. Dataset of school children tested

School	Number	Sessions	Queries	Terms
1 – have not	132	563	821	2884
2 – have not	80	201	263	1111
3 – have not	68	264	364	1402
4 – have	97	567	895	3251
5 – have	105	484	854	3467
6 – have	45	271	462	1783
TOTAL	**527**	**2350**	**3659**	**13898**

On arrival at the university, the purpose of the test was explained to the learners. Then a short introduction and demonstration of the search engine (Google) was given. This ensured that all the learners had at least a basic understanding of how to search the Web. All the machines in the laboratory were prepared to have the Google home page on the display. Additionally, an in-house developed software tool was loaded to capture all keystrokes. This application filled the bottom 10% of the screen and the Google search engine the rest. The application required the users to answer some questions on their demographic details (gender, age, computer experience, Internet experience, language and language proficiency). Thereafter, the tasks that had to be done by means of the Google search engine were displayed. After completing a task, the user had to type in the answer. The Google home page was reloaded after each task. Because each task was separately dealt with, each task can be seen as a normal search session.

The learners had to complete seven tasks. The tasks were selected in such a way that it would be of interest to the test subjects. Two of the tasks were sports related, two were political related, two were entertainment related and one was arts related. The tasks had variable levels of complexity. Three of the tasks could be considered as easy, two were of medium difficulty and two could be considered as difficult. Three of the tasks were more relevant for the high socio-economic group, three were more relevant for the low socio-economic group and one was independent of socio-economic status (SES).

The learners' demographic details, the start and end time of each task, the search strings used, all the links followed as well as the answers to each question were written to a database. From these the search characteristics of the users were calculated (see Table 2).

For each task (session) the number of search strings (queries per session) used to perform the task, the number of search strings with advanced options and the duration of the session were determined. In addition, for each search string, the number of words (terms per query) in the search string, the number of unique links followed (pages viewed) for the search string and the number of result pages viewed were determined. These measurements were then used as dependent variables and statistically analyzed to determine if there were significant differences between these variables and the categorical factors of socio-economic standing (SES) (have/have not) and Internet Experience (IE) (novice/intermediate/experienced). Users with 5 or less previous exposures to the Internet were regarded as novice users. Users who had 6-20 previous Internet exposures were regarded as intermediate users while users who

used the Internet more than 20 times prior to this study were regarded as experienced. In all cases a factorial ANOVA was done. The analysis was firstly done for all users that at least attempted the task by entering a search string. Thereafter, the analysis was repeated using only the learners who had completed the task correctly.

Table 2. Search characteristics per school

Statistic	1	2	3	4	5	6	ALL
Mean terms per query	03.5	04.2	03.9	03.6	04.1	03.9	03.8
% with one term per query	10.3	13.3	12.4	03.9	02.9	01.3	06.3
% with two terms per query	37.2	30.4	34.3	27.8	17.3	24.7	27.9
% with three or more terms	52.5	56.3	53.3	68.3	79.7	74.0	65.8
Mean number of queries per session	01.5	01.3	01.4	01.6	01.8	01.7	01.6
% with one query per session	69.5	77.1	70.1	59.4	57.2	59.8	64.1
% with two queries per session	22.0	17.9	22.7	29.8	22.5	22.1	23.7
% with more than two queries	08.5	05.0	07.2	10.8	20.2	18.1	12.1
Mean number of pages viewed per query	02.9	02.3	02.4	02.0	01.7	02.1	02.2
% viewed one summary page	97.9	98.5	98.9	94.9	98.1	97.0	97.2
% viewed two summary pages	02.0	00.4	00.8	03.9	01.6	02.6	02.2
% viewed more than two	00.1	01.1	00.3	01.2	00.2	00.4	00.5
% advanced queries	03.4	01.1	01.9	01.7	02.7	01.6	02.8

4 Results

Only 102 of the 3659 (2.79%) search strings contained advanced search parameters. Because of the low number, it was not further statistically analyzed. The first part of the statistical analysis included all users who at least attempted a task. The results for the different variables are shown in Table 3.

Table 3. Statistical results of the different variable using all learners

Variable	**N**	**Effect**	**F**	**p**
Number of search strings per task	2350	SES	7.44	0.01*
		IE	2.02	0.13
		SES*IE	4.44	0.01*
Number of words per search string	3659	SES	0.00	0.98
		IE	1.04	0.36
		SES*IE	1.82	0.16
Number of pages viewed per search string	3659	SES	26.2	0.00*
		IE	2.62	0.07
		SES*IE	5.25	0.01*
Number of result pages per search string	3659	SES	5.00	0.03*
		IE	0.86	0.43
		SES*IE	0.64	0.53

For the number of search strings used per task, both the socio-economic standing and the interaction effect were significant (α=0.05). The means for the have-nots and the haves were 1.45 and 1.65 respectively. Tukey's HSD indicated that the significant interaction was because the novice have-nots differed significantly (α=0.05) from the novice, intermediate and experienced users in the haves-group.

The number of words per search string was not significant for any of the effects but the number of pages viewed per search string was significant (α=0.05) for the socio-economic standing and the interaction effect. In this case, the means for the have-nots and haves were 1.51 and 1.11 respectively. Tukey's HSD showed the interaction effect was also because of a significant difference between novice have-nots and all the have-groups. There was also a significant difference between the intermediate have-nots and the intermediate users in the have-group.

For the number of result pages viewed per search string, the socio-economic standing was significant (α=0.05) with the means for the have-nots and the haves being 1.03 and 1.05 respectively.

In the second part of the statistical analysis, only the users who had completed the various tasks correctly, were included. This was done in order to level the playing field between the have and have-not groups. The results for the number of search strings used per task are displayed in Table 4.

A significant difference (α=0.05) was found for task 1 only, specifically with regard to the effect of socio-economic standing. The means for this task were 1.40 and 1.73 for the have-nots and haves respectively.

Table 4. Number of search strings per task for each task (p values)

Task	N	SES	IE	SES * IE
1	219	0.02*	0.21	0.86
2	152	0.66	0.48	0.50
3	43	0.37	0.91	0.08
4	177	0.23	0.21	0.86
5	75	0.37	0.90	0.47
6	94	0.59	0.53	0.56
7	171	0.92	0.56	0.86

The statistical results for the number of words per search string are shown in Table 5. For task 1 socio-economic standing was significant (α=0.05) (means: have-nots 2.42 and haves 3.22). The interaction between socio-economic standing and Internet experience was also significant. Tukey's HSD post-hoc analysis test showed that the significant effect can be attributed to a significant difference between have-not novices and have-not experienced users (0.02) as well as between the have-not novices and all experience levels in the haves-group. There was also a significant difference between have-not intermediate users and novice users in the haves-group (0.02) as well as between have-not intermediate users and experienced users in the haves-group (0.03).

For task 2, Internet experience was significantly different between the groups. The means were 3.79, 3.32 and 4.52 for novices, intermediate and experienced users respectively.

Table 5. Number of words per search string for each task (p values)

Task	N	SES	IE	SES * IE
1	342	0.00*	0.07	0.04*
2	257	0.74	0.01*	0.11
3	62	0.52	0.71	0.54
4	231	0.84	0.07	0.43
5	107	0.57	0.44	0.90
6	172	0.49	0.19	0.23
7	242	0.08	0.10	0.33

The results for the number of unique pages viewed per search string are shown in Table 6. In the case of task 2 there was a significant difference for socio-economic standing. The means were 1.45 for the have-nots and 0.82 for users in the have-groups. For task 5 there is an interaction effect which can be attributed to a small number of users in some of the cells (n=1).

Table 6. Number of unique pages viewed per search string for each task (p values)

Task	N	SES	IE	SES * IE
1	342	0.13	0.68	0.70
2	257	0.00*	0.15	0.41
3	62	0.56	0.20	0.15
4	231	0.09	0.28	0.36
5	107	0.46	0.76	0.03*
6	172	0.14	0.76	0.20
7	242	0.33	0.78	0.70

The statistical results for the number of result pages viewed can be seen in Table 7. No significant differences were found for any of the tasks. This is because the large majority of the users viewed one result page only.

Table 7. Number of result pages viewed per search string for each task (p values)

Task	N	SES	IE	SES * IE
1	342	0.32	0.17	0.18
2	257	0.15	0.97	0.97
3	62	0.48	0.58	0.58
4	231	0.27	0.58	0.87
5	107	0.70	0.52	0.32
6	172	0.21	0.58	0.58
7	242	//	//	//

// No variance

5 Discussion

When comparing the search characteristics (Table 2) of learners with the results from other studies, some interesting observations can be made. The average number of search strings per session of 1.56 is basically the same as the 1.6 reported by Spink and Jansen [4] for Web search engine users. It is, however, lower than the 2.3 and 2.9 reported for US and European users as reported by Spink, et. al. [8]. In the 2004 paper they reported that two in three users submitted only one query and six in seven did not go beyond two queries. This is confirmed by this study with respective figures of 64% and 88%. Silverstein, et al [5] reported 63.7% sessions with just one request.

According to Spink and Jansen [4] the mean query length for Excite users increased from 1.5 in 1996 to 2.6 in 1999 and 2.4 in 2003. For US users and European users the figures were 2.6 and 2.3 respectively [8]. The mean number of terms per query for this study was 3.80. This figure is more in line with the figure of 3.64 found by Hölscher and Strube [7] in a small scale study for expert users. In their study the average for a large number of users from the Fireball search engine log file was 1.66. It would appear that when users are given an explicit task to search for on the Web, they use more words per query as reported by log files. There is a tendency, especially in the case of novice users, to repeat the question as the search string.

The mean number of pages viewed per query for this study was 2.21. In the study of Spink et. al. [8] the corresponding figure for European users was 2.2 and 1.7 for US users. In a different study they reported an average of 2.35 [4]. It seems the figure reported here is quite in line with those of other studies. For the number of result pages viewed (10 results per page), it is a different story. In this study 97.2% of the users viewed only the first result page. This is substantially higher than the 70% reported by Spink and Jansen [4]. In conclusion, when discussing the search characteristic, it appears that the users in this study tend to use more words in the search string and to view only the first result page. Otherwise the characteristics are basically the same as for other studies.

Turning now to the statistical analysis of the socio-economic standing of the pupils, the following observations can be made. When including all the users, the socio-economic effect was significant for all the variables, except for the number of words used in the search string. The have-not users made use of fewer queries, but followed more links to view pages. Interestingly, there was no significant effect for Internet experience for any of the variables. There was, however, an interaction effect between socio-economic standing and Internet experience. The post-hoc tests indicated that this difference was basically between novice and intermediate users from the low socio-economic standing and the rest of the pupils. The difference can be attributed to novice have-nots utilising a much lower number of search strings per session and viewing a much higher number of pages than the other groupings. It seems as if the novice have-nots, rather than changing the search string, attempt to find the answer to the task by following more links. It can also indicate that their formulation of the search string is such that the results returned by the search string do not clearly indicate the most appropriate link to follow.

Considering only the users who had the task correct, the same pattern emerged, but only for the first two tasks. In addition, there was also a significant difference in

socio-economic standing for the number of words used in the search string for task 1 only. In this case the novice have-nots made use of fewer words than the other groupings. Again this can indicate that they have problems to formulate a good search string.

6 Conclusion

Many new users from Africa with a low socio-economic standing will enter the Internet arena in the near future. It is important to know if they will search the Web differently than their Western counterparts. This paper analysed the search characteristics of grade ten high school learners with different socio-economic standings. The results seem to suggest that there is a difference in the search behaviour between novice users with a low socio-economic standing and those with a high socio-economic standing. These differences, however, only lasted for the first few sessions, where after all users showed the same search behaviour. In terms of the many new African users who may enter the internet search arena in the immediate future, it may mean that they would soon be able to find their feet using the current search engines.

References

1. Internet World Stats: Available at (Accessed on August 17, 2005) http://www.internetworldstats.com/af/za.htm
2. Blignaut, P.J., Mc Donald, T.: The Effect of the Digital Divide on Web Searching Behaviour. (Submitted for publication, 2006)
3. UN news releases: $100 computer for children unveiled by UN. (Accessed on 17/10/06) (2005) Available at http://news.mongabay.com/2005/1117-laptop.html
4. Spink, A., Jansen, B.J.: A study of Web search trends. Webology, 1(2), (2004) Available at http://www.webology.ir/2004/v1n2/a4.html
5. Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a large Web search engine query log. ACM SIGIR Forum 33(3), 6–12 (1999)
6. Ahmed, S.M.Z., McKnight, C., Oppenheim, C.: A study of users' performance and satisfaction with the Web of Science IR interface. Journal of Information Science 30(5), 459–468 (2004)
7. Hölscher, C., Strube, G.: Web search behavior of Internet experts and newbies. The International Journal of Computer and Telecommunications Networking 33, 337–346 (2000)
8. Spink, A., Ozmutlu, S., Ozmutlu, H.C., Jansen, B.J.: U.S. versus European web searching trends. ACM SIGIR Forum, 36(2) (2002)
9. Kralisch, A., Mandl, T.: Intercultural aspects of design and interaction with retrieval systems. In: Proceedings of the 11th International Conference on Human-Computer Interaction, Las Vegas, USA (2005)
10. McDonald, T., Blignaut, P.J.: The effect of cultural differences on the efficiency of searches on a university website. In: Proceedings of the 11th International Conference on Human-Computer Interaction, Las Vegas, USA (2005)