Understanding eye movements on mobile devices for better presentation designs of search results

Jaewon Kim[1] (corresponding), Paul Thomas[2,1], Ramesh Sankaranarayana[1], Tom Gedeon[1], and Hwan-Jin Yoon[3]

[1] Research School of Computer Science, The Australian National University
{jaewon.kim, ramesh.sankaranarayana, tom.gedeon}@anu.edu.au

[2] CSIRO, Australia
paul.thomas@csiro.au

[3] Statistical Consulting Unit, The Australian National University
hwan-jin.yoon@anu.edu.au

Abstract

Compared to the early versions of smart phones, recent mobile devices have bigger screens that can present more web search results. Several previous studies have reported differences in user interaction between conventional desktop computer and mobile device-based web searches, so it is imperative to consider the differences of user behaviour for web search engine interface design on mobile devices. However, it is still unknown how the diversification of screen sizes on hand-held devices affects how users search. In this paper, we investigate search performance and behaviour on three different small screen sizes: early smart phones, recent smart phones, and phablets. We found no significant difference with respect to the efficiency of carrying out tasks, however participants exhibited different search behaviours: less eye-movement within top links on the larger screen, fast reading with some hesitation before choosing a link on the medium, and frequent use of scrolling on the small screen. This result suggests that the presentation of web search results for each screen needs to be designed considering the difference in search behaviour. At the end of this paper, we suggest several ideas for presentation design for each screen size.

Introduction

With the rapid increase in the popularity of mobile phones such as smart phones, mobile internet usage has soared by 67% from Sept., 2013 to Aug., 2014, and the use of

Address correspondence to Jaewon Kim, Research School of Computer Science, College of Engineering and Computer Science, The Australian National University, Canberra ACT 0200, Australia. E-mail: jaewon.kim@anu.edu.au

hand-held devices (i.e. mobile phones and tablets) has grown rapidly from 21.9% to 35.3% worldwide, although accessing the web by using desktops remains at 64.6% (Statcounter Global Stats, 2014). From previous studies that compared user behaviour and search performance on a desktop monitor versus a mobile device screen (Jones, Buchanan, & Thimbleby, 2003; Kim, Thomas, Sankaranarayana, Gedeon, & Yoon, 2014), it is believed that the interface design for web searches on mobile devices needs to be different from that on a desktop monitor. Therefore, there have been recent efforts to improve this design by understanding user interaction with small devices (Lagun, Hsieh, Webster, & Navalpakkam, 2014; Guo, Jin, Lagun, Yuan, & Agichtein, 2013).

One interesting trend is the enlargement of screen sizes on mobile devices during the last few years. In the early versions of smart phones which first allowed people to access Internet search engines, the screen size (diagonal) was generally less than 4 inches and displayed only two or three search results for the searchers (e.g. Samsung Galaxy S1 and Apple iPhone 3 or 4). In contrast to these devices, recent smart phones are equipped with a larger screen of 4.5 inches (e.g. Apple iPhone 6) and have a higher resolution. Another recent device is a *phablet* (a portmanteau word combination of the words phone and tablet) that has a screen size of over 5.4 inches (e.g. Samsung Galaxy Note 4 or Apple iPhone 6 Plus). Because of the wider screen sizes, the more recent mobile devices can display four to six search results on the first page without the need to scroll.

This phenomenon (i.e. the enlargement of screen sizes) can be explained by needs in the mobile market. That is, players in the market (e.g. manufacturers, their third parties, and end users) wanted a bigger screen than the early mobile phones, even if they came with some disadvantages such as heavier weight and lower mobility. We cannot say whether or not the needs were caused by a desire for a better web search experience, and users have various purposes for smart phones such as games, entertainment, search, social networking, and education. It is clear, however, that web searching is one major activity for using smart phones (Pew Research Center, 2013; Adwords, 2015), and we need to investigate the effects of the change to determine if there is any difference in user interaction as the main goal of this paper. When improving the search engine interface design for mobile devices, in contrast to conventional monitors, we might need to consider different designs for each mobile device that has a different screen size. This investigation could enhance search engine results pages (SERPs) on small devices for better search experience. In this paper, we explore user web-search performance and behaviour on three different sizes of screens (3.6, 4.7, and 5.5 inches for earlier smart phones, recent smart phones, and phablets, respectively, as shown in Figure 1) using eye-tracking technology. We adopt search speed, search accuracy, and user satisfaction as the user performance metrics, similarly to previous studies (Kim et al., 2014; Lagun et al., 2014) and employ implicit data such as fixation and scanning patterns on SERPs to understand user behaviour.

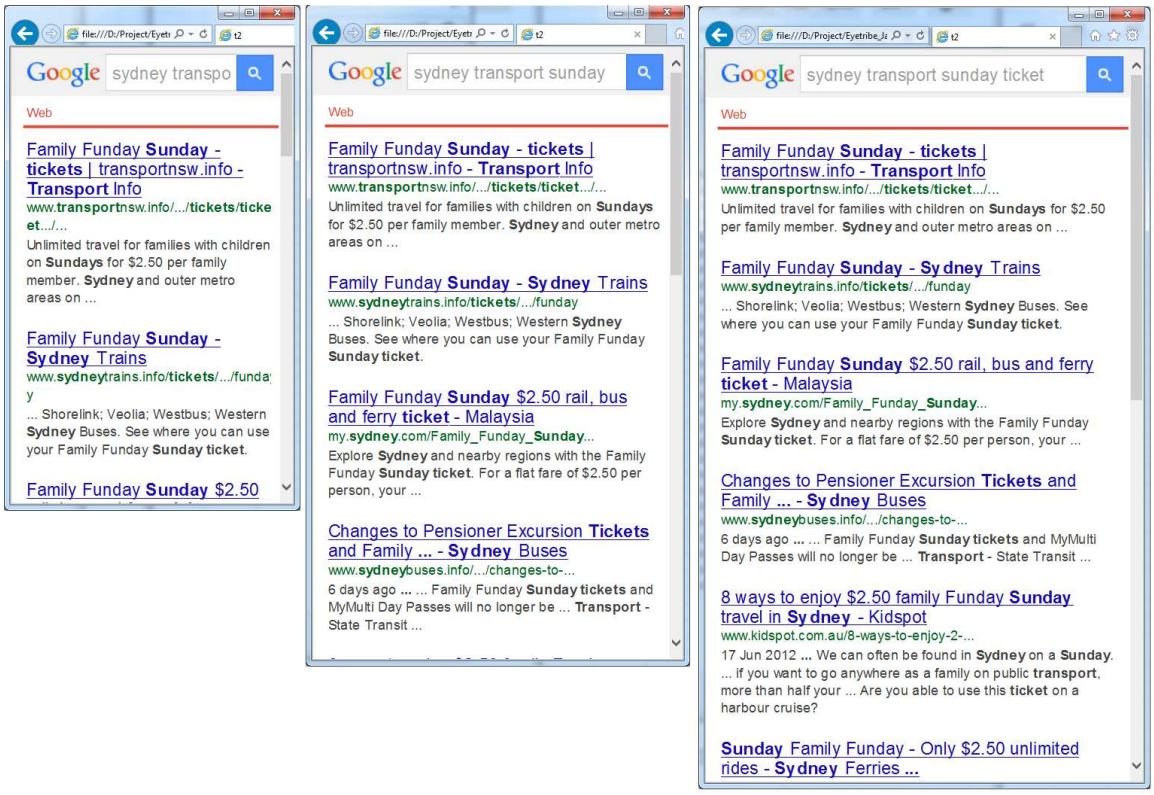

Figure 1. Examples of search engine results pages on three different screens. The screen sizes are 3.6 inches (378×672) for an early smart phone (left), 4.7 inches (495×880) for a recent smart phone (middle), and 5.5 inches (585×1040) for a phablet (right), respectively. *Note:* All aspect ratios are 16:9 and the screen sizes are measured along the diagonal.

After first surveying previous studies related to our research, we describe the user study and data collection including the experimental procedure. We then explain our measures and the methods adopted for the experiment. Our results and a discussion about the findings are presented along with several limitations of this user study. At the end, we conclude by addressing the implication of these findings with possible presentation designs for each screen size.

Related work

For the purposes of this experiment, we can classify the previous studies into two types. The first type addresses *web search behaviours on SERPs* including eye-tracking analysis, and the second concerns *user interaction on mobile devices*.

Web search behaviours on SERPs

As addressed in the previous section, user web-search performance and behaviour using desktop monitors can be different from that using mobile devices. However, there has not yet been a significant amount of investigation on small screen web search, and hence search behaviour using conventional monitors is worthy of considering as background

knowledge. First, several studies have investigated general behaviour on SERPs. Granka, Joachims, and Gay (2004) used an eye-tracker to investigate the positions of user fixation before the first click on a SERP. They found that users tend to concentrate on links beyond the selected link, and the links they pay most attention to are almost always the first and second links of search results. A similar result that comes from Joachims, Granka, Pan, Hembrooke, and Gay (2005) indicates that subjects prefer to scan SERPs from top-to-bottom. An investigation of content manipulation on SERPs was also conducted in their study. They focused on the effects of the results relevance ranking, which suggested that users tend to scan more links and click lower ranks for a bad ranking. Guan and Cutrell (2007) found a similar result that indicates that the search efficiency decreased when the relevant links were positioned at lower ranks. The above studies suggested the reason to be user bias. Cutrell and Guan (2007) studied the effects of content manipulation (i.e. title, snippets, and URL). They found that richer snippets enhanced aspects of search performance such as speed and accuracy for an informational task but decreased the performance of navigational tasks. Lorigo et al. (2008) analysed the results of Granka et al. (2004) and Lorigo et al. (2006) in more detail, and adopted several of the methodologies from the studies to investigate differences in search behaviour on two famous search engines. Their findings suggested strong similarities between search behaviours, and they also found that search behaviour is highly related to the ranking condition.

Other studies have classified SERP task types and investigated user behaviours with respect to these types. Broder (2002) developed a taxonomy to investigate the purpose of three classes of web search queries: informational, navigational, and transactional. The results suggested that search engines need to consider a user's web search goal for better user satisfaction. Following this, Lorigo et al. (2006) adopted two task types (informational and navigational tasks) to investigate the difference in user behaviour between the task types. This classification of task type has been broadly used after its introduction (e.g. Kim et al., 2014; Dumais, Buscher, & Cutrell, 2010; Guan & Cutrell, 2007). They found that informational tasks needed more time to complete than navigational tasks. In addition, they investigated user behaviours such as complete, linear, and strictly linear patterns as well as skip and regression by using sequences of fixations such as the compressed sequence and minimal scanpaths. They found that half of the participants exhibited the skip (jumping over one link) and regression (jumping back at least one link) patterns in their gaze sequence. This was explained by noting that users did not follow the rank order of the search engine in their examination of search results.

Several studies have classified user search behaviour according to gaze patterns. Klöckner, Wirschum, and Jameson (2004) found that more than half of participants used a 'depth-first' strategy, while the remaining exhibited a 'breadth-first' or 'mixed' strategy (about 10% and 20%–30%, respectively). Aula, Majaranta, and Räihä (2005) defined two kinds of search strategy: 'economic' and 'exhaustive'. They divided the patterns by whether a user scanned less than or more than half of the visible results before making a selection. Dumais et al. (2010) then extended this classification by adding 'economic-ads', that is, users who regularly look at advertisements. According to their findings, both economic groups spent more time on the top three links than the exhaustive users did. In addition, the exhaustive group showed a slower scanning patten of reading links.

User interaction on mobile devices

Several researchers have investigated user interaction on small screens in web search, although these studies did not consider eye-movements. Jones et al. (2003) compared usability among three sizes of interface: mobile phone, handheld computer (e.g. PDA), and conventional desktop interfaces. They found that users take more time to complete tasks and exhibit lower task success rates on smaller screens. They also suggested several improvements for web search design on small screens. Raptis, Tselios, Kjeldskov, and Skov (2013) studied users' perceived usability, task completion times (efficiency), and task completion rates (effectiveness) during information seeking tasks with a particular application on three different screen sizes of mobile devices (3.5, 4.3, and 5.3 inches). Their findings suggested that users on screens larger than 4.3 inches exhibited better search efficiency on the task completion times, but the other results (perceived usability and effectiveness) showed no significant difference among the screens. Guo et al. (2013) compared user interactions on web search documents between a touch-enabled mobile device and a desktop computer with a mouse and keyboard. They investigated touch interactions such as gestures, zooming, swiping, and inactive time on a small screen to improve web search ranking. One of their major findings was that user behavioural signals such as periods of inactivity are significantly correlated to the relevance of web documents.

Several studies adopted eye-trackers to analyse user interaction on small screens. Drewes, De Luca, and Schmidt (2007) investigated gaze interaction for controlling applications on a handheld device using dwell time and gaze gestures and Biedert, Dengel, Buscher, and Vartan (2012) investigated text interaction and reading on a mobile phone screen, although their research was not about web search tasks. Kim et al. (2014) studied the differences of user performance and behaviour for web search tasks based on large and small screens (for a desktop and mobile device, respectively). Although they adopted an emulator with a mouse for the mobile-sized screen, this study was able to compare user interaction according to screen size. They found that there is no significant difference in search speed on SERPs between the screen sizes; however, more hesitant behaviours with complicated scanpaths such as skip, regression, and trackback were exhibited on the smaller screen. Recently, Lagun et al. (2014) studied the effect of relevance in Knowledge Graph (KG) results (e.g. famous person and place) and Instant Answer (IA) results (e.g. the weather today) on a real mobile device by recording eye-movements. Their results indicated that a user's gaze activities tend to increase when KG was irrelevant, and users need less time to complete tasks with less scrolling when IA was the relevant condition. They also found that the second link received more gaze time attention than the first link, unlike the results on desktop screens that showed a top-to-bottom pattern. Although search factors such as KG and IA are interesting, we focus on the organic results (i.e. titles, snippets, URLs), because SERP with the additional result types is still not common, but only efficient for some queries such as famous films or buildings. In addition, the factors (especially KG) occupied most of the space on SERP in a screen for early smart phones (3.6 inches or similar).

The research presented in this paper investigates the differences in web search performance and behaviour on three sizes of small screens while taking into account the findings of previous studies. The terms and definitions of search behaviour that that were adopted

for our study are discussed in the Measurements section.

User study and data collection

In this section, we present the experimental design and procedure and describe the participants, tasks, and equipment. In addition, we discuss the data collection methodology and post processing.

Participants

A total of 20 subjects (10 males and 10 females, aged 24–44 years) from the local University campus voluntarily participated in the experiment. Two participants (females) were excluded from the analysis because of technical issues (e.g. calibration problems). All participants were rated themselves expert or good at finding information using web search engines and did so frequently, and most of them had experience using mobile devices for web searches.

Tasks

Each participant performed a total of nine search tasks (see the descriptions and queries in Table 1) for given initial queries. Three tasks were performed on each screen. As can be seen from the table, we varied the task category, including categories such as weather, science, and sports. Although several previous papers adopted two task types, as introduced in Lorigo et al. (2006), we prepared only informational tasks and excluded the navigational tasks, the goal of which is simply to reach a particular website. The reason for this choice is that when we prepared the navigational type tasks for the experiment, all the relevant links were located in the top link. We obtained the initial SERPs from the Google mobile search engine and then removed the images and unnecessary links so that all tasks showed the same kind of content as shown in Figure 1: titles, snippets, and URLs only. The tasks and initial queries were cached in the system. We also confirmed that all tasks had relevant links that included the right answer(s) within the top three ranks with no manipulation of the rank order in order to ensure an equal balance of task difficulty across all tasks. All tasks were easily solved within 1–2 minutes.

Table 1: Task descriptions and queries.

Category	Task description	Initial task query
Technology	iPhone6 is recently out. In what memory sizes can you get it? (3 kinds)	iPhone 6 specs
Sport	Which two countries will play for the first match in the cricket world cup 2015?	cricket world cup 2015 dates
Science	What is the number of Copper (Cu) in the periodic table?	a periodic table copper
Architecture	You are interested in some facts about the Sydney tower. What is its height?	Sydney tower height
Weather	When does daylight-saving time end in Australia? (any applied states such as NSW, ACT, or VIC)?	2015 daylight savings
Politic	How many seats are there in the Australian parliament for MPs (elected by the Australian people)?	Australian parliamentary seats
Transport	You have heard there is a very cheap transport deal in Sydney on Sunday. What is the name of this, and the price?	Sydney transport Sunday ticket
Law	You want to buy cigarettes for your friend when you come back from overseas. How many cigarettes can you bring?	cigarettes Australia customs
Education	When does ANU's first semester 2015 start?	ANU 2015 dates

Design and procedure

A total of nine tasks on three types of screens were shown to each participant. To control the effects of task and screen presentation order, we adopted a Latin-square method for screen order (each group which has 3 participants faced same screen order), and the task presentation order was randomized using the Williams-Latin square across participants (so each task was first for two participants, second for two participants, and so on). We ensured that font size and type of content on the SERPs were the same for each task in order to focus on the effect of screen size.

We started the experiment with a short conversation to relax the participants and assure them that it was not a test. All subjects were given the same instructions and asked to understand their rights such as withdrawing at any time, before signing a consent form. To become familiar with solving the tasks, they conducted three sample tasks, one for each screen size, and were able to ask questions. We then calibrated their eye gaze with 16-point calibration provided by the eye-tracker, such that the tracking accuracy was within 0.5 degrees of the visual angle. After calibration, the subjects were presented with the first task description and initial query. They were presented with 10 search results on the initial SERP when they pressed the 'start task' button. The tasks were considered completed when they found the desired answer on web documents, and spoke them vocally. After the task, participants were asked to score their satisfaction with the usability of each screen size from 1 to 7. This procedure was scripted, and repeated for nine tasks. At the end of the experiment, the user scored their overall search experience on each screen size and responded

to several questions on a post-experiment questionnaire. A time notice was given 3 minutes after starting each task. The time limit was determined through a preliminary experiment with 5 participants, and we decided that it is sufficient time to reach the answers. After that, the participants were able to decide whether to spend more time finding the answer or to move to the next question, although no one exceeded the time limit. The running time for the experiment was less than 20 minutes to complete everything from the instruction to post-questionnaire. Participants were not paid for their time.

Apparatus

All search tasks were obtained from the Google mobile search engine and shown using Internet Explorer 11. We conducted the experiment on a MS Surface Pro 3 (Microsoft, 2014) for the three screen sizes to configure the same search environment such as font size, ppi (pixel per inch). The device was touch-sensitive, thus scrolling, zooming and "clicking" were possible via touch. The only difference was screen size. Eye gaze was recorded by EyeTribe (The Eye Tribe, 2014) and analysed by custom Visual Basic scripts, in which windows events were gathered by the cached web program. The resolutions (diagonal sizes) of each screen were 378×672 (3.57 inches), 495×880 (4.67 inches), and 585×1040 (5.52 inches) for small-, medium-, and large-size screens, respectively. As addressed earlier, the sizes were inspired by the screen sizes of an early version of a smart phone such as the iPhone 4, recent smart phone such as the iPhone 6, and phablets such as the iPhone 6 Plus or Galaxy Note 4. A scroll bar was placed on the right side of the browser, although we excluded the pixels of the bar from the total visible space. In this setting, the large screen presented about six search results, whereas the medium screen displayed about four results and the small screen displayed two and half results. There were limitations such as the resolutions of screen, the device size, and users' mobility during the experiment. The limitations are discussed in detail in the Conclusions and future work section.

Data collection and post processing

We collected data from two sources. To obtain the users' gaze points, we used an eye-tracker that recorded eye gaze 60 times per second (60Hz). The eye-tracker recorded the x- and y-coordinates of the user's gaze and the system time in a log file. Other data consisted of window events such as the location of a participant's click, how much s/he used the scroll function, and which task s/he was looking at as well as the usability score for each screen with system times. We embedded custom Java script codes into the HTML files of the cached tasks. We then combined the two data sources using the system time (to the millisecond) as the primary key value. The merged data was stored as an Excel file and then extracted via a Visual Basic application (VBA). For the fixation duration, we used a commonly used algorithm; dispersion-threshold identification (I-DT, see Salvucci & Goldberg, 2000 for the definition) with 40 pixels for the dispersion threshold and 100 ms for the duration threshold. By definition, the dispersion threshold was obtained from the distance between the eyes and the screen. As to details of the dispersion threshold value, we assumed that the distance was 60 cm, and the accuracy was 0.5 degrees because the valid operating range of the eye-tracker is 45–75 cm and the system was calibrated for an error of less than 0.5 degrees.

Measurements

In this section, we explain what we measured, how the measures operate, and our reasons for adopting these particular measures. This base knowledge is necessary for the results in the next section.

Search performance

We measured user search performance by search speed and accuracy. The search speed was measured as the time to the first click on a SERP and also by task completion. The search accuracy was calculated according to whether a participant's answer was right or not.

Search speed Two search speeds were measured: *elapsed time to the first click* and *task completion duration*. Because all participants were presented with the same SERP content whereas the web documents presented after a click could vary, we considered the elapsed time to the first click on a SERP as the primary search speed. This was calculated from the time the SERP contents were presented, to the first click. We believe that measuring the elapsed time to the first click is helpful for understanding the efficiency of SERP interface design.

The other indicator of search speed, as used the approaches of previous studies (Cutrell & Guan, 2007; Jones et al., 2003), was the time needed to complete tasks. We calculated this to be a supplementary search speed. Task completion duration is the time from the appearance of the SERP to either the time the participant reached the right answer or the time limit, although no one exceeded this time. This includes the time spent on all web documents and SERPs.

Search accuracy Participants were given only one chance to speak an answer, so that they would be careful in deciding the answer. We assigned the search accuracy to be '1' if a user found the correct answer on the first attempt. Otherwise, if a subject found the wrong answer or reached the time limit, the search accuracy was assigned a score of '0'. We have considered other measurements from previous studies. One study (Cutrell & Guan, 2007) considered whether the user clicked on the 'best' result, and the other study (Kim et al., 2014) measured only whether a participant found the right answer on the first selected web document. However, we could not decide on one best link because all tasks had at least two relevant links that included the right answer. In addition, we focused on whether a user obtained the right answer within the time limit, rather than on whether participants clicked on the link that contained the right answer, and then found the answer on the first selected page.

Search behaviour

We measured data for search behaviour such as fixation duration, scanpaths and scanning direction. In addition to this, we also considered other window events such as click patterns and scrolls. Some types of data have intrinsic meaning, however others need to be consolidated with other data to be meaningful.

Fixation duration Fixation duration has several implications in the understanding of search behaviour. One common belief is that a longer average fixation duration indicates that it is more difficult to obtain information (Just & Carpenter, 1976; Rayner, 1998). Because all tasks have the same components (10 ranks including titles, snippets, and URLs as well as the periphery, as shown in Figure 1), we assigned 10 areas of interest (AOIs) to each SERP to investigate user attention. We measured the mean fixation duration for each SERP as well as for the 10 AOIs of each task.

Click pattern We recorded click points to investigate where a participant finally selected the answer, along with a recording of which links on the SERP they read according to the fixations. This is the main measurement regarding how much participants were biased toward the rank order by the search engine. In addition, this is the most important value needed to determine the trackback value, as explained in the subsection on Trackback.

Scanpath To determine the eye-movement sequence, we considered two kinds of scanpaths, as introduced in Lorigo et al. (2006): *compressed* and *minimal* scanpaths. If we assume that the original scanpath (consisting of the numbered AOIs of fixations on a SERP ordered by time) is 2-2-1-1-2-3-3-4-5-5-4, then the compressed sequence is 2-1-2-3-4-5-4 (length 7), formed by aggregating subsequent fixations. The minimal scanpath is 2-1-3-4-5 (length 5), formed by removing repeat visits from the compressed sequence. The compressed sequence includes the revisits the user has made, and the minimal scanpath does not.

Scanning direction Given the above two definitions, we may analyse five types of scan patterns. Kim et al. (2014) refined the measurement methods from previous studies (Lorigo et al., 2006; Dumais et al., 2010) and described three main methods: *complete* if the user inspected all of the links above the selected link, *linear* if the minimal path is monotonically increasing, and *strictly linear* if the compressed sequence is monotonically increasing without any skips or regressions. As Kim et al. (2014) defined, we measured two additional variables derived from both linear and strictly linear methods, called *linear/ID* (linear or immediate decision) and *strictly linear/ID* (strictly linear or immediate decision) to consider the cases where users looked at only one link and selected it immediately.

Skip and regression We investigated *skip* and *regression* patterns using the compressed sequence and minimal scanpaths across the screens. As defined in a previous study (Kim et al., 2014), we define a skip as a jump of more than one rank (e.g. from rank 3 to 5) and a regression as a jump back of at least one rank (e.g. from 6 to 5). Both behaviours are significant when determining the five scan patterns in the scanning direction and represent how carefully a user scans the SERP.

Scroll As can be seen from Figure 1, users need to use the scroll function to see the lower links on the page. (The number of visible ranks were about two and a half, four, and six on the small, medium, and large screens, respectively). We measured the proportion of scrolling across the screen, which could indicate its usability.

Trackback We used *trackback*, introduced by Kim, Thomas, Sankaranarayana, and Gedeon (2012) to investigate how much additional effort a user expends before making a selection on an SERP. Generally, users tended to exhibit a top-to-bottom scanning pattern and to scan links beyond their choices in previous studies (Granka et al., 2004; Joachims et al., 2005). Participants in our study displayed the same pattern. We measure trackback as the distance between the selected link and farthest link visited.

Questionnaire measures

Questionnaires are a useful supplement to determine a participant's personal information and experience regarding the experiment. We asked participants to fill in a post-experiment questionnaire after completing all tasks. The questionnaire included several questions about age, gender, search convenience on each screen, level of task difficulty, past personal usage of search engines, personal skill with search engines, and personal skill with search on mobile devices.

In addition, to obtain opinions of search experience, which is one important way to investigate SERP interface design (Lagun et al., 2014), we adopted a 7-point Likert scale that ranged from 'Extremely dissatisfied' to 'Extremely satisfied' (valued as 1 and 7, respectively) after completing each task. As a result, we obtained nine sets of satisfaction scores, three results for each screen per participant.

Results and discussion

We first analysed the dataset of 162 tasks (54 tasks for each screen) using the measures introduced in the previous section. We also investigated relationships between search speed and some of the search behaviour in detail. We mainly focused on whether there is a significant screen size effect on search performance and behaviour; we discuss implications of the findings using other results, where needed. At the end of this section, we present a general discussion that reviews the findings for each screen, and address the limitations that we should consider.

We employed analysis of variance (ANOVA) with block structure (participants) for search speed and fixation duration with a log-transformation $log(x + 1)$, so that 0 maps to 0, to maintain the normality assumption, if necessary. We used generalized linear models (GLMs) (McCullagh & Nelder, 1989) with a binomial distribution and the logit link function for binary data, and with a Poisson distribution and logarithm link function for count data. We also used generalized linear mixed models (GLMMs) (Breslow & Clayton, 1993) for binary data and count data with participants as a random effect, σ_p^2. GLMM is an extension to GLM in which the linear predictor has the fixed effects as well as random effects. Thus, if there is no random effect, we used GLM rather than GLMM. In these analyses, we used the GenStat version 17 statistical package (VSN International, 2014).

Power analysis

We carried out post-hoc analyses with the significant level $\alpha = 0.05$ to confirm the power of our design (Chow, Wang, & Shao, 2007). Using the log-transformed time to first click as an example: the standard errors of the difference in means (SE of differences) was observed as 0.0386. In this case, a sample of just seven participants would give the power,

$1 - \beta \geq 0.95$ for all three comparisons (small/medium, medium/large, and small/large screens). With eighteen participants we would maintain the power, $1 - \beta \geq 0.95$ even with SE of differences as high as 0.0645. This gives us a good deal of confidence in the power of the analyses below.

Search performance

As noted in the Measurement section, we analysed two kinds of explicit data for search performance: search speed, search accuracy. There was no significant screen size effect on search speed and search accuracy.

Search speed We analysed two search speeds: elapsed time to first click on a SERP as the main factor, and task completion duration as a supplement. We applied ANOVA with using a log-transformation. As shown in Table 2, there is no significant screen size effect on the two kinds of search speed, although the mean of time to first click for medium size of screen was much longer (10.47 s) followed by small (9.12 s) and large (7.7 s) screens. It has been believed that task completion duration decreases as screen size is reduced (Jones et al., 2003; Raptis et al., 2013; Kim et al., 2014). These results, however, indicate that participants exhibited similar search speeds on SERPs as well as on completing tasks.

Table 2: Search performance and behaviour.

		Mean values			Statistics			
		L	M	S	*p*-value	L	M	S
Search performance								
Search speed	Time to first click [s]	7.70	10.47	9.12	0.195			
	Task completion duration [s]	20.89	24.79	23.08	0.655			
Search accuracy	Correct answer rate [%]	94.44	98.15	94.44	0.589			
Search behaviour								
Fixation duration on SERP	Per task [s]	3.97	5.60	5.53	0.087			
	Per link [s]	2.16	1.89	2.49	*	ab	a	b
Clicks	Ranks	1.39	1.52	1.46	0.697			
Scanpath	Minimal scanpath	2.06	2.76	2.26	**	a	b	a
	Compressed sequence	3.33	5.50	4.35	**	a	b	ab
	Compressed minus minimal	1.28	2.74	2.09	*	a	b	ab
Scanning direction	Complete rate [%]	96.30	94.44	100	**	a	a	b
	Linear rate [%]	46.30	31.48	57.41	*	ab	a	b
	Strictly linear rate [%]	11.11	1.85	9.26	0.087			
	Linear/ID rate [%]	81.48	55.56	79.63	**	a	b	a
	Strictly linear/ID rate [%]	46.30	25.93	31.48	*	a	b	a
Skip and regression	Skip [%]	14.81	22.22	7.41	0.087			
	Regression [%]	53.70	74.07	68.52	*	a	b	ab
Scroll	Scrolled rate [%]	3.70	20.37	35.19	***	a	b	b
Trackback	Count	1.07	1.91	1.28	***	a	b	a
Search satisfaction	7-point Likert scale	5.24	4.91	4.20	***	a	a	b

* Significant at 0.05 level.
** Significant at 0.01 level.
*** Significant at 0.001 level.
Note: SERP denotes search engine results page, and L, M, and S denotes large, medium, and small, respectively
Note: Labels a and b indicate the type of result, 'a' type is significantly different from 'b', but not different to 'ab'.

Search accuracy For the accuracy data, we used a GLM with a binomial distribution. Similar to a result from Kim et al. (2014), the result indicates that there is no significant effect on search accuracy among different mobile-sized screens. As can be seen from the mean values in Table 2, the search accuracy consists of extremely high correct-answer rates across all screens (more than 94%). Because all tasks include correct answers in top three ranks, this seems to be caused by a strong bias toward the rankings provided by search engines. For further analysis, it is possible that we need to conduct an additional experiment by manipulating the ranking order, as done in several previous studies (Guan & Cutrell, 2007; Joachims et al., 2005). However, we may glimpse the reason by analysing the scroll effect and fixations on each AOI, noting how participants concentrated on the top links without using the scroll function. This is discussed in the next subsections.

Search behaviour

We can determine several differences in search behaviour across screen sizes, although there was no significant effect in search performance. We discuss the possible implications and investigate some of the differences in detail by also looking at search speed results.

Fixation duration We applied ANOVA with using a log-transformation to mean fixation duration on SERPs. Although Table 2 shows that the effect is near the significant level, there is no significant difference with respect to screen size. This means that subjects expended similar effort to obtain enough information to make a selection on the small screens (Just & Carpenter, 1976; Rayner, 1998; Dumais et al., 2010), unlike the difference between a monitor and small screen in the study by Kim et al. (2014). For further analysis, we investigate how much effort the participants made to read one link by connecting this result with other search behaviour. This is discussed in the next subsection on scanpath.

Although the mean fixation duration per task showed no significant difference, we can obtain interesting findings by investigating the fixation time on AOIs by applying ANOVA. There is a significant difference on AOI fixation duration ($F_{(2,340)} = 4.23$, $p < 0.05$). Figure 2 shows how long a participant's eyes stayed on each AOI on each screen to obtain information. Similarly to the results on a desktop monitor (Granka et al., 2004; Cutrell & Guan, 2007), the fixation durations decreased for the lower-ranked results. In this graph in particular, we can see that participants rarely spent attention on links under rank 4 (about 98% of total fixation duration). As mentioned in the previous subsection on search accuracy, this may be one reason for the very high rates of search accuracy achieved by clicking on the top three links with strong bias toward the rank order. We see some special characteristic for the small screen. Because the small screen showed only two or two and a half visible links on the SERP, the fixation duration on AOI 1 was much higher than the other AOIs on the screen, and there is a significant difference on the fixation duration of link 1 (SED = 0.2005). The above findings may be connected to scrolling, therefore this is discussed in detail in the subsection below on scroll.

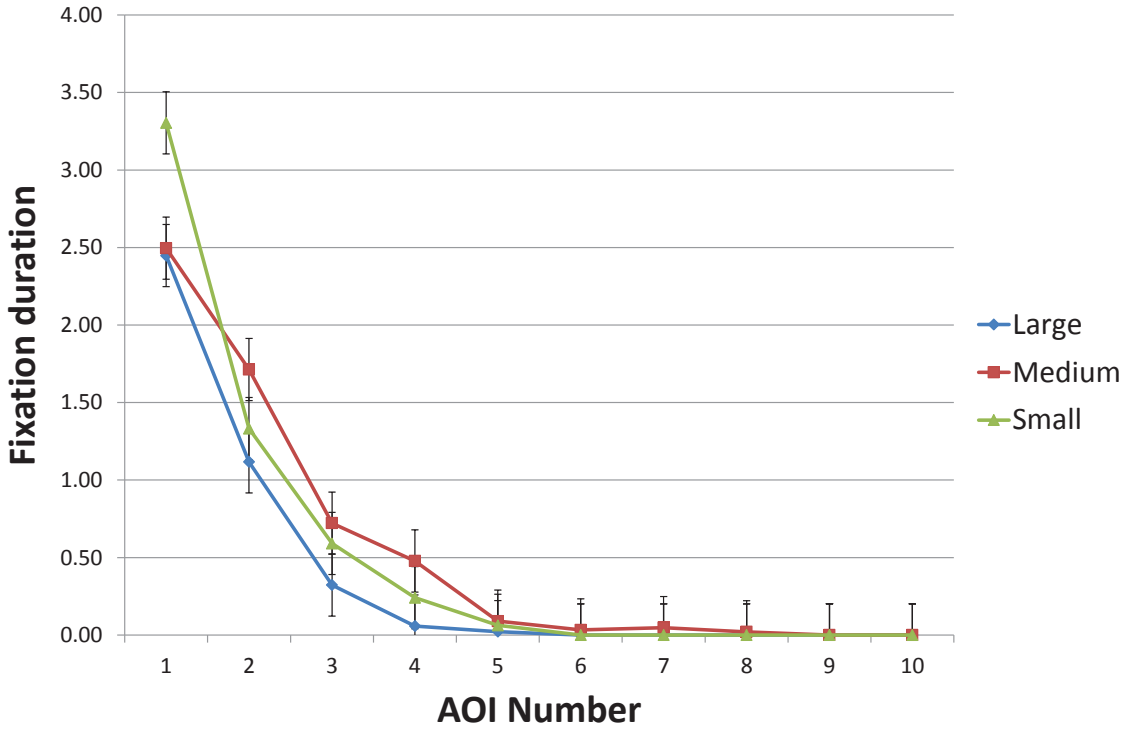

Figure 2. Mean fixation duration on each AOI (second/user/task), SED = 0.2005.

Click pattern Using a GLMM with a Poisson distribution for the analysis, we found that a participant's click patterns were mostly distributed on the top three links with no significant differences. The chance of clicking the top three links for the first choice was, overall, about 95%. Although the top links included the relevant answer, this proportion is very high. This phenomenon may be explained by the fact that users are strongly biased by the rank order from the Google search engine, as found in previous works (Joachims et al., 2005; Kim et al., 2014).

Scanpath We can obtain the compressed sequence by aggregating subsequent fixations from the raw scanpath, and the minimal scanpath is given by removing the repeat visits from the compressed sequence. We adopted a GLMM with a Poisson distribution to analyse the three kinds of scanpath metrics. First, as seen in Table 2, we determined that there are significant differences for the minimal scanpath length ($\sigma_p^2 = 0.055$, $\chi^2 = 12.21$, df $= 2$, $p < 0.01$) and compressed sequence length ($\sigma_p^2 = 0.134$, $\chi^2 = 11.39$, df $= 2$, $p < 0.01$). The difference in minimal scanpath length is found between the medium screen and other screens. The compressed sequence length results are little different from the minimal scanpath length results. The difference occurs only between the medium and large screens. With the above behaviours, we also investigated how often participants revisited a link, where their eye remained to read. We analysed the difference length of the values between both behaviours using a GLMM with a Poisson distribution. Similarly to the results of the

compressed sequence length, there is a significant difference ($\sigma_p^2 = 0.288$, $\chi^2 = 9.23$, df = 2, $p < 0.05$) between the large and medium screens. As a consequence, the results of the scanpath behaviours indicate that subjects tended to read more links on the medium screen with a higher revisit count.

As mentioned in the subsection on fixation duration, we investigated how much effort participants spent to obtain information from one link (fixation duration per link) by using the minimal scanpath length and mean fixation duration. The fixation duration per link is significantly different between the medium and small screens ($F_{(2,142)} = 3.62$, $p < 0.05$). This indicates that participants chose to spend less time to read a link on the medium screen than on the small screen.

We also investigated the relationship between the SERP search speed (time to first click) and the fixation duration per link. Figure 3 shows that there is clearly a positive relationship between both variables on the three screens. That is, all three lines show a pattern where the fixation duration per link increases as the elapsed time to first click increases. This also indicates that the fixation duration per link on the medium screen is lower than that on the small and large screens. Given the above results, we can conclude that participants on the medium screen viewed more links and frequently revisited them; however, they needed to spend less time extracting information per link.

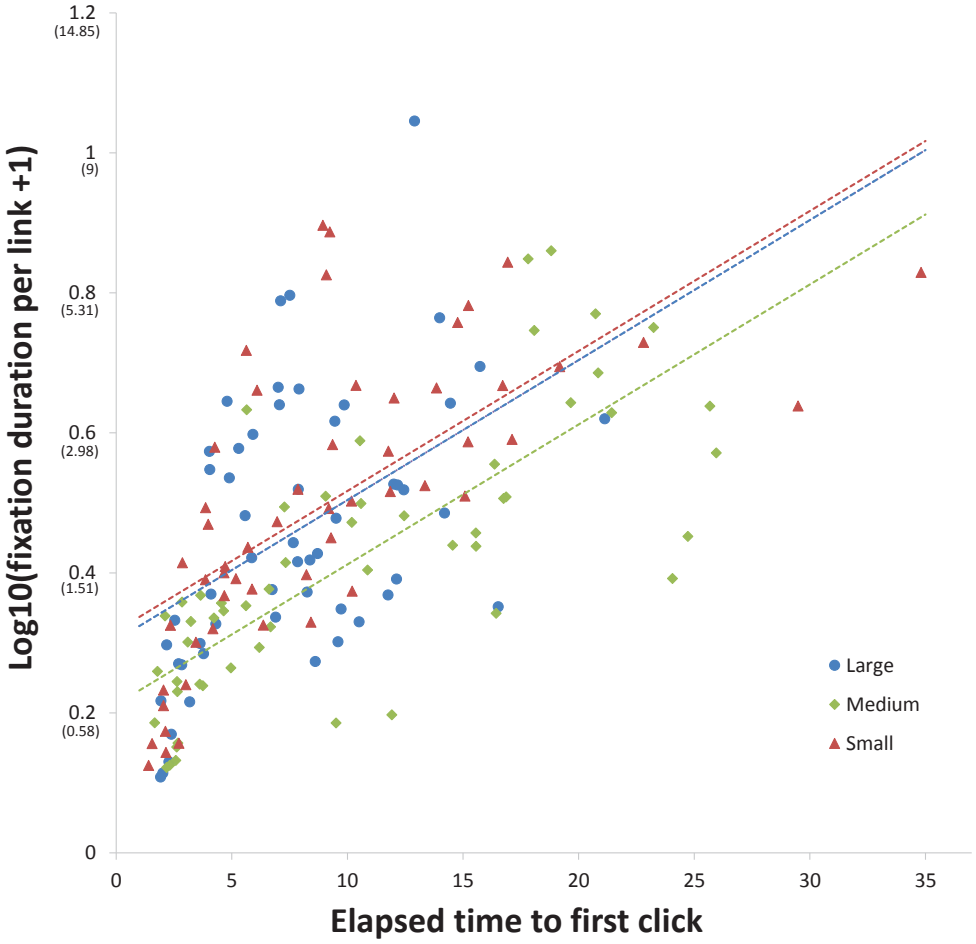

Figure 3. Relationship between search speed (elapsed time to first click) and fixation duration per link [sec]: The numbers on the y-axis are log-transformed (with back-transformed values). Intercept (SE: standard error) of the large screen is 0.304 (0.025), $p < 0.001$, intercept (SE) of the medium screen is 0.212 (0.028), $p < 0.001$, intercept (SE) of the small screen is 0.317 (0.026), $p < 0.001$, and the common slope (SE) is 0.020 (0.002), $p < 0.001$.

Scanning direction With respect to behaviours in the scanning direction, we investigated differences in how users scan links on the three sizes of screen: complete (if fixations are made on all of the links up to a selection), linear (if the minimal path is monotonically increasing), and strictly linear (if the compressed sequence is monotonically increasing with no skips and no regressions). We adopted a GLMM with a binomial distribution for the scanning direction analysis. Table 2 indicates that there is a significantly different effect on the complete pattern ($\sigma_p^2 = 4.585$, $\chi^2 = 11.86$, df $= 2$, $p < 0.01$). This difference is caused by the result on the small screen. That is, although the proportions of complete patterns on the other screens are also very high (over 94%), this indicates that a user on a small screen needed to look at all links more carefully before choosing the correct link.

This can be connected to skip behaviour in the subsection on skip and regression. We next looked into the linear and strictly linear patterns. There is a significantly different effect on the linear pattern ($\sigma_p^2 = 0.629$, $\chi^2 = 8.29$, df = 2, $p < 0.05$), but there is no effect on the strictly linear pattern. This difference can be observed between the medium and small screens. The difference is that users exhibited a stronger top-to-bottom pattern on the small screen, even if there were skip or regression behaviours. However, because there were a large proportion of immediate decisions across all three screens (meaning that a subject looked at only one link and selected it) we added this pattern into the linear and strictly linear patterns for further analysis. In the cases of linear/ID and strictly linear/ID patterns, we found the same significant differences (linear/ID: $\sigma_p^2 = 1.059$, $\chi^2 = 13.62$, df = 2, $p < 0.01$; strictly linear/ID: $\sigma_p^2 = 1.732$, $\chi^2 = 7.61$, df = 2, $p < 0.05$). Both behaviours show that participants on the medium screen were less likely to follow a top-to-bottom pattern whether or not there was a skip and regression while searching. The difference between both linear and strictly linear and both linear/ID and strictly linear/ID seems to come from the immediate decision on the large screen. The participants on the large screen exhibited an ID pattern about 35% of the time, whereas they had this pattern about 23% and 22% of the time on the medium and small screens, respectively. Thus, the bigger proportion on the large screen produces significant effects between the large and medium screens on the linear/ID and strictly linear/ID, although there is not such a different effect on linear and strictly linear.

Skip and regression Using a GLM with a binomial distribution, there is no significant effect on skip (a jump of more than one rank) rates across the screens, whereas a clear difference was found on the regression (a jump back of at least one rank) rate ($\sigma_p^2 = 1.732$, $\chi^2 = 7.61$, df = 2, $p < 0.05$) by a GLMM with a binomial distribution. For the case of skip rate, the effect is near the significant level ($p = 0.087$), and the skip rate mean values are lower on the small screen and higher on the medium screen. On the other hand, we determined that participants exhibited a higher regression rate on the medium screen than on the large one. Skip behaviour is connected to the complete pattern, as we earlier concluded. That is, lower mean skip rates on the small screen probably impacted the higher complete rate, although it was only near the significant level. In addition, we may say that the higher skip and regression rates on the medium screen contributed to the higher differences on the link revisits (i.e. compressed sequence length minus minimal scanpath length) and on the minimal scanpath length between the medium and large screens.

Scroll Using a GLMM with a binomial distribution, a significant difference on scroll rates can be observed ($\sigma_p^2 = 1.969$, $\chi^2 = 15.90$, df = 2, $p < 0.001$). The subjects exhibited a low scrolling rate on the large screen (3.7%), whereas it happened more often on the medium and small screens (20.37% and 35.19%, respectively). By connecting this to user bias as addressed in the search accuracy, these low proportions of scroll rates across the screens seem to support the above explanation: participants were strongly biased toward rank order. In addition, when we merge this result with AOI fixations, the expectation in the paragraph fixation duration is verified: subjects did not need to scroll to read the top 1–3 ranks on the large screen, whereas they had to use this function to look below the page fold of the small screen.

We also investigated how scrolling (scrolled or non-scrolled cases) affected the search speed for different size of screens. We used a linear mixed model (LMM) with participants as random effects, which is same as the GLMM with normal distribution and identity link function. As shown in Table 3, there are significant effects of scrolling on search speed (elapsed time to first click) for the medium and small screens ($\sigma_p^2 = 0.061$, $\chi^2 = 12.67$, df $= 1$, $p < 0.001$, and $\sigma_p^2 = 0.019$, $\chi^2 = 27.91$, df $= 1$, $p < 0.001$, respectively), whereas no significant effect was found on the large screen. This result indicates that users on smaller than 4.7 inch screens clearly spend more time using a scroll function, and decreasing the scroll effects on both screens may improve search speed on SERPs.

Table 3: Effects of scrolling on search speed (elapsed time to first click) (s)

	Mean values		
Screen size	Non-scrolled	Scrolled	p-value
Large	7.73	6.95	0.530
Medium	8.66	17.54	*
Small	6.01	14.84	***

* Significant at 0.05 level.
*** Significant at 0.001 level.

Trackback A significantly different effect on trackback was observed using a GLMM with a Poisson distribution ($\sigma_p^2 = 0.191$, $\chi^2 = 16.24$, df $= 2$, $p < 0.001$). Participants on the medium screen recorded a higher trackback value, which is related to scanpath. As we determined from previous behaviours, a user looked at more links with a higher revisit rate on the medium screen, which is caused by the higher regression rate, skip rate, and compressed minus minimal value. The trackback is a summarised value of these behaviours, and this result indicates that the subjects spent more effort before making a decision on the medium screen.

Questionnaire measures

At the end of the experiment, participants were asked several questions. All subjects responded that they were computer science post-graduate students, although they had different research areas. Fourteen out of eighteen answered that the tasks were easy to solve, whereas the rest thought they were not easy, but also not difficult. Thirteen replied that carrying out the tasks on the large screen was the most convenient, although two voted for the medium screen and three thought that there was no difference across the screens. However, none of the participants thought that the small screen was convenient. In addition, all participants use a search engine at least once a day and believed that they were good at using search engines. Lastly, 16 out of 18 replied they were good at controlling mobile devices, although two of them had no experience using mobile phones for web search.

Applying GLMM with a Poisson distribution, there is a significant effect of screen size on user satisfaction ($\sigma_p^2 = 0.051$, $\chi^2 = 28.07$, df $= 2$, $p < 0.001$). By comparing mean values

among the screens with the standard error of the difference (SED), a difference is shown for the small screen. This means that participants felt it was less convenient to perform tasks on the small screen than it was on the large and medium screens.

General Discussion

In this section, we summarize the implications for the screen sizes by reviewing the results of each behaviour. First, several factors of search performance and behaviour were not significantly different across the three kinds of screen. Participants spent similar times on SERPs, on web documents after clicking, and even on obtaining information per task. Furthermore, they indicated a strong bias toward the ranking order provided by the search engine on all screens by selecting top links and mostly spending time reading the top AOIs.

However, several significant differences were observed for other factors. First, compared to the interactions on the medium screen, participants on the large screen read fewer links with fewer revisits despite the fact that there was no difference in the time spent reading each link. They also exhibited a higher immediate decision rate by looking at only one link, as well as a top-to-bottom pattern with lower regression, less scrolling, and lower track values. In general, we can explain this as caused by the fact that the screen of a phablet allows users to look at only a few top links with little eye-movement away from the top links.

Second, we found several differences in behaviour on the medium screen. Subjects needed less time to obtain information from one link; however, they visited more links, with frequent revisits, with higher trackback values, and with higher regression rates than on the large screen. Therefore, they could draw a relatively lower top-to-bottom pattern. It could be said that users on the recent smart phone, which has a 4.5 inches screen or similar, tend to exhibit hesitation when making decisions, although they scanned many visible links and read them more quickly.

Finally, participants on the small screen exhibited only a few significantly different behaviours compared to those on other screens, although the screen size obtained the worst score for search satisfaction. When we look at the mean values in search behaviour, most values are between those of the large and medium screens without significant differences. However, to comment on the behaviours on the small screen in general, we can say that users with screens the size of early smart phones tend to hesitate a little when choosing a link on a SERP with long link reading times, and high scroll function use caused by the less-visible links, although they spent a lot of time reading the first link.

Limitations Despite our efforts to create a similar environment for the various small screens, we would like to point out that there are several limitations to this study. First, participants were sitting on a chair and freely mobile during the experiment, however they were requested not to move their head too much so that it remained within the boundary of the recording zone. Second, although the tablet was a touch-sensitive screen, it was not the same as smart phones, and users could not move around while holding the device. Third, because participants were recruited from a particular group, these results may not represent the search performance and behaviour of the general public. Lastly, differences in the number of pixels between smart phones and the tablet in this experiment could lead to

different results. We recognize that display resolutions in recent mobile devices are higher than the tablet has.

Conclusions and future work

In this paper, we observed search performance and behaviour across three different sizes of screen. Our findings indicate that several search behaviours and the usability are different across the screens, although participants exhibited a similar search speed and accuracy. Glancing through the results reported in this paper, one might conclude that there is no need to change anything in SERPs because there is no significant difference for search speed and accuracy. However, if we could create better design for SERPs for each type of screen by understanding search behaviour, which could provide users with a better search experience.

We know that adding information into snippets clearly improves the search performance for informational tasks on desktop monitors (Cutrell & Guan, 2007). However, we need to consider the limitations of screen size on mobile devices. We suggest several possible ideas for designing the interface for web searches on the screens for each mobile device, bearing the limitations of our study in our mind. These ideas are also directions for our future work. First, participants in our experiment did not read as many links as there were visible on the screen, less scrolling on the large screen. Therefore, for the screen on a phablet that is 5.5 inches or similar, one promising idea is to *display a knowledge graph* that shows rich information regarding the keywords that may be helpful for search performance, as shown in a previous study (Lagun et al., 2014). Because the screen has enough space (for about six links) to display the KG with several of top links on the initial SERP, one possible design would be to locate the KG at the top of the screen instead of the top three links, moving them below the KG to reduce the effect of decreasing search speed that occurs when the KG is not relevant, as shown in Lagun et al. (2014).

Second, on the medium-sized screen, the subjects exhibited a faster reading speed with hesitant eye-movement to make a decision among the top links. Consequently, we expect that *enriching the content of top links by showing longer snippets* (Cutrell & Guan, 2007) could reduce the hesitation on the recent smart phones (4.7 inches). That is, we could display only three links with rich snippets instead of showing four links. This could reduce the hesitation behaviour by providing additional information to help the user choose a relevant link.

Third, our users reported a low usability for the small screen and had a slow reading performance for each link. In fact, widening the screen size would be the best solution to improve the search satisfaction with reducing use of a scroll, although this is clearly impractical. Therefore, we may suggest three ideas for earlier-type smart phones (3.5 inches), which could contribute to an increase in satisfaction as well as search speed: *making the best use of peripheral vision* and *reducing font size* to display more contents, and *embedding page up and down button on the interface or horizontal page changes instead of a vertical scroll function*, as suggested by Jones, Marsden, Norliza, and Boone (1999). Additionally, displaying only one link that has rich information in a snippet along with one of the above functions on the smallest screen is a promising design to reduce the time consumed by scrolling.

In addition, for general mobile devices, we recommend the idea of *providing a small mark indicating that an item on a SERP links to a mobile-optimised page instead of a full-size page for desktops* as mentioned by Kim et al. (2014), and as initially suggested by Jones et al. (2003). Recently, the Google mobile search engine provides the indicator by marking with "Mobile-friendly" in front of each snippet on SERPs, if the links connect to mobile-optimised pages. This may reduce the user time cost on SERP as well as on web documents after selecting a link, and improve search for all sizes of small screens.

References

Adwords. (2015). *Building for the next moment.* Retrieved from http://adwords.blogspot.com.au/2015/05/building-for-next-moment.html.

Aula, A., Majaranta, P., & Räihä, K. (2005). Eye-tracking reveals the personal styles for search result evaluation. In *Human-computer interaction-INTERACT 2005* (pp. 1058–1061). Springer.

Biedert, R., Dengel, A., Buscher, G., & Vartan, A. (2012). Reading and estimating gaze on smart phones. In *Proceedings of the symposium on eye tracking research and applications* (pp. 385–388).

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3–10.

Chow, S.-C., Wang, H., & Shao, J. (2007). *Sample size calculations in clinical research.* CRC press.

Cutrell, E., & Guan, Z. (2007). What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 407–416).

Drewes, H., De Luca, A., & Schmidt, A. (2007). Eye-gaze interaction for mobile phones. In *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on computer human interaction in mobile technology* (pp. 364–371).

Dumais, S., Buscher, G., & Cutrell, E. (2010). Individual differences in gaze patterns for web search. In *Proceedings of the third symposium on information interaction in context* (pp. 185–194).

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 478–479).

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 417–420).

Guo, Q., Jin, H., Lagun, D., Yuan, S., & Agichtein, E. (2013). Mining touch interaction data on mobile devices to predict web search result relevance. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 153–162).

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 154–161).

Jones, M., Buchanan, G., & Thimbleby, H. (2003). Improving web search on small screen devices. *Interacting with Computers*, *15*(4), 479–495.

Jones, M., Marsden, G., Norliza, M.-N., & Boone, K. (1999). Improving web interaction in small screen displays.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480.

Kim, J., Thomas, P., Sankaranarayana, R., & Gedeon, T. (2012). Comparing scanning behaviour in web search on small and large screens. In *Proceedings of the Australasian document computing symposium* (pp. 25–30). ACM Press.

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H.-J. (2014). Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology*.

Klöckner, K., Wirschum, N., & Jameson, A. (2004). Depth-and breadth-first processing of search result lists. In *CHI'04 extended abstracts on human factors in computing systems* (pp. 1539–1539).

Lagun, D., Hsieh, C.-H., Webster, D., & Navalpakkam, V. (2014). Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval* (pp. 113–122).

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., ... Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, *59*(7), 1041–1052.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using google. *Information Processing & Management*, *42*(4), 1123–1131.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear model* (2nd ed.). Chapman & Hall/CRC.

Microsoft. (2014). *Microsoft surface pro 3.* Retrieved from http://www.microsoft.com/surface/en-au/products/surface-pro-3.

Pew Research Center. (2013). *Cell phone activities 2013.* Retrieved from http://http://www.pewinternet.org/2013/09/19/cell-phone-activities-2013/.

Raptis, D., Tselios, N., Kjeldskov, J., & Skov, M. B. (2013). Does size matter?: investigating the impact of mobile phone screen size on users' perceived usability, effectiveness and efficiency. In *Proceedings of the 15th international conference on human-computer interaction with mobile devices and services* (pp. 127–136).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78).

Statcounter Global Stats. (2014). *Mobile internet usage soars by 67%.* Retrieved from http://gs.statcounter.com/press/mobile-internet-usage-soars-by-67-perc.

The Eye Tribe. (2014). *Theeyetribe.* Retrieved from https://theeyetribe.com/.

VSN International. (2014). *Genstat for windows 17h edition.* VSN International, Hemel Hempstead, UK. Web page: www.vsni.co.uk.