What Snippet Size is Needed in Mobile Web Search?

Jaewon Kim[†], Paul Thomas[§†], Ramesh Sankaranarayana[†], Tom Gedeon[†], Hwan-Jin Yoon[¶]

[†]Research School of Computer Science, [¶]Statistical Consulting Unit
The Australian National University
Canberra, Australia
{jaewon.kim, ramesh.sankaranarayana, tom.gedeon, hwan-jin.yoon}@anu.edu.au

[§]Microsoft
Canberra, Australia
pathom@microsoft.com

ABSTRACT

A snippet (content summary for a web page) is one of the main elements in a search result page. Search engines have been improved to reduce users' effort in web search, e.g., providing flexible snippet sizes by considering the purpose of the search and suggesting predicted answers. In most cases, search engines for mobile devices present two or three lines of snippet for each result link. Some studies suggest that long snippets provide a better search experience on desktop screens, but this may not be true for mobile devices because of the smaller screen.

We conducted a user study to investigate what size of snippet is appropriate for mobile devices. Our findings suggest that users with long snippets on mobile devices exhibit longer search times with no better search accuracy for informational tasks. This is caused by the longer reading time, frequent scrolling with bigger viewport movements, and greater time consumption for searching and reading one result. The overall findings suggest that, unlike desktop users, mobile users are best served by snippets of two to three lines.

CCS Concepts

•**Information systems** → **Search interfaces; Web search engines;** •**Human-centered computing** → *Touch screens; Laboratory experiments;*

Keywords

Mobile web search; snippet length; search engine result page

1. INTRODUCTION

Web searching is a major activity on mobile devices [1] and its usage is increasing [24]. Most search engines display *search engine result pages* (SERPs) with several result links, which mainly contain the title, URL, and *snippet* (also known as the summary, caption, or document surrogate). The snippet occupies a larger space in each result, compared to the title and URL.

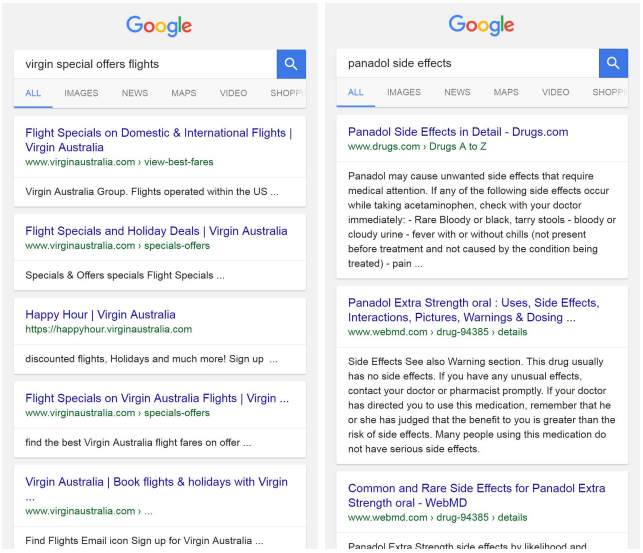

Figure 1: Examples of initial SERPs with short (left) and long (right) snippets.

Current commercial mobile search engines often provide a knowledge graph (KG) to present relevant long information or an instant answer (IA) for popular user queries. Some search engines control the snippet length by predicting the users' goals through analysing the queries. In addition, some mobile SERPs provide a few lines of snippet for each result link that can be expanded to six or more lines by using the 'view more' buttons. The above ideas and technologies are clearly implemented to reduce user effort to reach some particular bit of information or destination web site, and for a better search experience.

Although search engines have been developed with the above-mentioned technologies, most search engines for mobile devices typically provide two or three lines of snippet for a result link; the information in the snippets does not seem to differ from the displays in desktop monitors. Furthermore, when the KG and IA are not relevant, the effect is

CHIIR '17, March 07-11, 2017, Oslo, Norway

DOI: http://dx.doi.org/10.1145/3020165.3020173

worse user satisfaction with increased search times for users on mobile devices [17].

Some studies have investigated the effect of snippet lengths on desktop screens [4, 13, 21]. Some of their results suggest that long snippets are more useful for finding a particular piece of information [4], and that snippets are currently too short to provide sufficient information in many cases [13].

Several studies [11, 14, 16, 22] have suggested that search behaviour can vary by screen size. Therefore, the appropriate snippet length for a mobile device may be different from that of a desktop machine. We conducted an experiment to investigate the effect of snippet length on a mobile device with three different lengths (see Figure 1 for examples of short and long snippets) and two different task-types.

Considering the results from previous studies regarding snippet lengths on desktop screens, we formulated three hypotheses prior to conducting the experiment. First, although there might be some difference due to the smaller screen, we imagined that long snippets would have a positive effect in the search time when looking for a particular piece of information, similar to the results in the previous study [4]:

- H1. For finding a particular piece of information, users will take less search time with long snippets, and need more time with short snippets.

Second, we expected that the users would focus on different elements in the same way they do on a desktop screen [4]:

- H2. Among the title, URL, and snippet, users will consider the snippet as the most important element for finding a particular bit of information, and the URL for reaching a requested web page.

Third, even though we guessed that the search time would be reduced with long snippets, the user satisfaction might differ because of the small screen and familiarity with the snippet length of current search engines:

- H3. Users will express better satisfaction with medium snippets for finding a particular piece of information, and with short snippets for reaching a requested web page.

2. BACKGROUND

We introduce three necessary general lines of background knowledge. The first concerns *reading and scanning SERPs*, the second is related to *search behaviour by snippet length*, and the third addresses *mobile web search*.

2.1 Reading and scanning SERPs

Eye-tracking is a popular method of investigating user interaction in web search, which provides information on the user's gaze, e.g., user attention, saccade, and pupil dilation [12, 23]. Although much work has examined user interaction with web searches using gaze data, we would like to mention some studies related to general search behaviours, which are helpful for reading this paper.

A few studies have investigated broad scanning patterns on SERPs, e.g., the areas that attract searchers' attention and the sequence of the interest. Hotchkiss et al. [9] found that users make a "golden triangle" pattern: the most popular area in first time visits to a SERP, and suggest that considering this pattern is important for SERP design, because a user's interest is dramatically reduced outside of the golden triangle. The result of another study [20] indicates that users exhibit an "F-shaped pattern" while searching, i.e., they scan one vertical stripe followed by two horizontal stripes. This study suggests some guidelines for better web page design, e.g., the most important things should be stated within the first two paragraphs.

Some studies investigated users' scanning patterns involved in their first click decision. Granka et al. [8] focused on how users explore the result links in SERPs above and below the selected link. Their finding suggests that ranks one and two receive most of the user's attention, similar to the results of Hotchkiss [9] and Nielsen [20]. They also found that users tend to scan the links above the selected link. However users often use different patterns near the page fold with some further observation below the selected link.

A study by Joachims et al. [10] presented results similar to the findings of Granka et al. [8] and additionally suggested that users are affected by rank order, after analysing the effects of three different manipulated rank orders (normal, swapped, and reversed). The results of both studies broadly indicate that users read the search results with a top-to-bottom scanning pattern, although Thomas et al. [28] found that some users began their exploration from a different position rather than the top rank on SERPs.

A few studies examined search behaviours according to the user's goals in web search. Broder [3] classified task types into *informational*, *navigational*, and *transactional* web searches with purposes of finding particular information, reaching a specific website, and performing some web-activity, e.g., an online purchase, respectively.

Lorigo et al. [18] conducted an experiment to investigate user behaviours with the informational and navigational tasks. Their findings suggest that users take more time to complete informational tasks. Both informational and navigational tasks have been commonly used (e.g., [8, 10, 14]), and the task types are one main consideration in this study.

In addition, Lorigo et al. defined a *compressed sequence* and a *minimal scanpath* using the fixation sequence. The original scanpath is the sequence of all fixations on a SERP. We can extract the compressed sequence by aggregating consecutive fixations on the same object, and the minimal scanpath can be obtained by removing the previous visits from the compressed sequence. For example, if we assumed that the original scanpath is 2-2-2-1-1-2-3-3-2-4-4, the compressed sequence would be 2-1-2-3-2-4 (length: 6), and the minimal scanpath would be 2-1-3-4 (length: 4). Both the compressed sequence and minimal scanpath have been adopted by several studies to investigate users' search strategies. In this paper, they are valuable measurements for analysing the sequence of movements in user attention.

2.2 Search behaviour by snippet length

Some researchers studied the effects of snippets in SERPs. Paek et al. [21] conducted a user study to compare the usability and user preferences regarding different methods of displaying snippet information, e.g., *normal view*: the full web page was shown by clicking the title; *instant view*: an expanded snippet was additionally displayed by a mouse click; and *dynamic view*: an effect similar to the instant view by mouse hovering over a particular result. They found that

the instant view exhibited faster task completion than the normal view, and about half of the participants preferred the instant view.

A study by Cutrell and Guan [4] focused on the effects of snippet length. They examined search behaviours with three different snippet lengths on a desktop screen (short, medium, and long). Their findings indicated that users tend to spend less search time with the long snippets for informational tasks, whereas the long snippets for navigational tasks required more time. They also found that long snippets for informational tasks led the user to look at fewer links; however the opposite pattern was observed for navigational tasks.

Kaisser et al. [13] conducted two experiments to estimate the preferred snippet length according to answer type (e.g., person, time, and place), and to compare the results of the preferred snippet length to users' preferences in a user study to investigate whether the preferred snippet length could be predicted. Their results suggested that the preferred snippet length depends on the answer type, and users tend to express better satisfaction with the estimated preferred snippet length. Overall their findings indicate that longer snippets may be more useful if the snippets are relevant for the query.

2.3 Mobile web search

Several studies of mobile web search provide reasons for investigating the effect of snippet length on mobile devices. A few studies evaluated the search performance and behaviour between mobile devices and desktop monitors. Jones et al. [11] investigated the search performance of a mobile phone, a personal digital assistant (PDA), and a desktop monitor. Their findings indicated that small screens lead to worse search speed and accuracy. In addition, Kim et al. [14] examined user performance and behaviour on small and large screens for mobile devices and desktop monitors using an eye tracker. They found that users with a small screen were slower to complete tasks due to more reading time, having more difficulty extracting information and displaying less eye movement.

Some studies focused on search behaviour with small screens. Raptis et al. [22] conducted a user study to investigate the effects of three different mobile device screen sizes (3.5, 4.3, and 5.3 inches). They evaluated three variables: perceived usability, task completion times (for efficiency), and task completion rates (for effectiveness). They found that users with the smallest screen needed more time to complete tasks, although no effect was found related to the perceived usability and task completion rate.

Recently, Kim et al. [16] investigated search performance, behaviour, and user satisfaction on three sizes of mobile phone screens (3.6 inches for early smartphones, 4.7 inches for recent smartphones and 5.5 inches for phablets). They saw different behaviours on each screen size, that is, the small size: a higher chance of scrolling with the worst user satisfaction; the medium size: fast information extraction with some hesitation before selecting a link; and the large size: less eye movements on top links.

The above work on user interaction in mobile web search broadly suggest that search behaviour can differ according to screen size, and search engines should consider the difference, which supports the need for this study.

Table 1: Examples of task descriptions and queries.

Informational
• Panadol is a brand of pain reliever. What are some side-effects of Panadol? Is a rash one of them? (panadol side effects) • You are interested in some facts about the Golden Gate bridge in U.S. In what year was the bridge construction completed? (golden gate bridge) • Which two countries played for the 4th match in the Cricket World Cup 2011? (cricket world cup 2011 dates)
Navigational
•You are interested in shoes from Adidas. Find the official Adidas homepage.(adidas shoes [country]) • Find the web page where you can apply for a saving account on the Citibank website. (citibank new account) • You bought a laptop from Sony and something doesn't work as expected. Find the page for Sony technical support.(sony laptop technical support)

3. USER STUDY

This section describes the experimental design and the procedure.

3.1 Participants

We recruited 30 participants with varying backgrounds, e.g., chemistry, biology, computer science, education, history, and law, from inside and outside a university campus. We excluded six of them due to low eye gaze tracker calibration accuracy, leaving us with 24 participants (13 male) aged 22–42 (mean: 29.4, standard deviation (SD): 5.7).

Using 7-point Likert-scale questions (1: completely unfamiliar/bad, 7: completely familiar/good), participants were asked how familiar they were with search engines and how good they were at using mobile devices. The participants marked high scores for both the first (mean: 6.25, SD: 0.6) and second (mean: 5.79, SD: 1.0) questions. That is, most participants considered themselves familiar with search engines and good at using mobile devices.

3.2 Tasks

Each participant completed twelve tasks (six each of informational and navigational tasks; see Table 1 for examples of the tasks) which were derived from those of Dumais et al. [6]. The task categories varied, e.g., chemistry, sports, travel, history, and law, and the tasks were simple. Every task had at least two relevant links within the top three ranks, and included a correct answer/destination web page.

We extracted the SERP rank orders with titles and URLs from one of the most popular mobile search engines (Google). However, we generated the different snippet lengths using Nutch for crawling and Solr for snippets with the Lucene library [27] using the highlight function, because we were unable to find a search engine that provided long snippets (six or more lines) when we designed the experiment[1]. Each task had 10 search result links with titles, URLs and snippets.

We prepared three different snippet lengths: short with one line, medium with two or three lines, and long with six

[1]Yandex recently introduced longer snippets with 'read more' buttons if the users want to see more than the typical snippet lines. This is not available via an API at the time of writing.

to seven lines as shown in the short and long snippet examples in Figure 1. With the snippet length manipulation, the initial SERPs displayed 5.0 (4.6–5.4), 3.7 (3.4–3.9), and 2.4 (2.3–2.7) result links above the page fold with short, medium and long snippets, respectively. The long snippet includes the medium snippet, and the short snippet is part of the medium snippet.

3.3 Design and procedure

In this experiment, we adopted a within-subject design to investigate the effect of the two main treatments: task type (2) × snippet length (3). Each participant completed 12 tasks, including six tasks for each task type, and two of the six tasks included the same type of snippet length (i.e., a set of 'SSMMLL' for informational and navigational tasks. S, M, and L denote short, medium, and long, respectively). To minimize the carry-over effect, we randomized the task order within each task type. In addition, the orders for task type and snippet length were counter-balanced, and every task was evenly shown with the three different snippet sizes across the participants.

After they agreed and signed the consent form, we showed them three sample tasks with each snippet length to familiarize them with solving the tasks. We then calibrated their gaze recording using a 9-point procedure, and the task lists were shown on the screen.

When the participants clicked the first task on the list, the description and initial query for the task were displayed. After this, the participants could proceed to the first SERP. Once they announced the desired information or reached the requested web page, we considered the task to be completed. After each task, the participants were asked about their satisfaction with the snippet length using a 7-point Likert scale (1: completely dissatisfied, 7: completely satisfied). The cycle was continued to the last (12th) task.

At the end of the experiment, participants were asked to fill out a post-experiment questionnaire, which included basic information, such as age, background, familiarity with search engines and using mobile devices, their preferred snippet length, and their thoughts about the most important SERP element (title, URL, or snippet) for each task type. Our participants spent about 25–30 minutes in the laboratory room to complete everything from the welcome introduction to filling out the questionnaire.

3.4 Apparatus

We adopted an iPhone 6 plus (5.5 inches with a 1080×1920 pixel resolution) for the experiment, which has a popular screen size [19]. The mobile phone was connected to the main system as a secondary monitor using the Twomon software [5] and search results were displayed through Internet Explorer. To collect the gaze data, we used Facelab 5 [26], and we analyzed the data using Eyeworks [7].

4. RESULTS

We obtained data from 288 tasks (144 tasks for each task type, and 48 tasks for each snippet length within each task type). We focused on the effects both by task types (*informational tasks* (ITs) and *navigational tasks* (NTs)) and snippet length (short, medium and long).

To investigate the hypotheses, we measured search time for H1, user attention for H2, and user satisfaction for H2 and H3. We explored search behaviour, e.g., scroll rates, with viewport movements and scanpath, as introduced in Section 2. In further investigation, we analysed how search behaviour was related to search performance and user attention.

We adopted several analysis techniques. First, we employed analysis of variance (ANOVA) for continuous data, e.g., search time and fixation duration, with a log-transformation $log(x+1)$ to maintain the normality assumption. Second, we adopted generalized linear mixed models (GLMMs) [2] with a binomial distribution and a logit link function for binary data, and a Poisson distribution and logarithm link function for countable data. Third, for the score data from the 7-point Likert scale, we used a linear mixed model (LMM) [30]. Finally, linear regression [25] was adopted to analyse the relationships between two dependent variables (e.g., scanpath and search time).

We acknowledge that there may be individual differences in our participants' familiarity with web search engines and mobile devices. To consider the individual difference, we used a block structure (subject) for ANOVA, and adopted a GLMM and LMM instead of a generalized linear model (GLM) and a linear model (LM) because observed random effects between subjects (σ_s^2) were greater than standard errors (SEs) in all variables.

All analyses were conducted using the GenStat statistical package [29].

4.1 Search performance and user attention

To test H1, we first adopted *task completion duration* (the total time required to complete a task) as search time; this is the same approach as in Cutrell and Guan's study [4]. Only task type significantly affected task completion duration ($F_{(1,259)} = 172.22$, $p < 0.001$). As can be seen in Table 2, the time spent for ITs was almost twice that for NTs (36–41 s for ITs vs 18–21 s for NTs) and participants exhibited no difference on time spent by snippet length. However, because each task included several relevant links on a SERP, the time spent on the linked web pages varied considerably according to the design for either a full site or a mobile-friendly version. The effect of different resolutions among web pages is beyond the search engine's control. Therefore we considered the *time to first click* as the main search time in this study instead of task completion duration, and compared search time to search behaviour. Note that the time to first click is very similar to the task completion duration minus the time spent on web documents, because participants needed only one click to reach the answer for most of the tasks (overall 94%).

We also found significant effects due to task type, snippet length and their interaction on time to first click ($F_{(1,259)} = 32.58$, $p < 0.001$, $F_{(2,259)} = 5.63$, $p < 0.01$, and $F_{(2,259)} = 3.36$, $p < 0.05$, respectively). This indicates that participants with ITs needed more time to decide (mean 18.9 s vs 13.6 s, for averages of ITs and NTs, respectively), and long snippets led participants to stay longer on the SERPs (mean 14.6 s, 15 s, and 19.3 s, for short, medium, and long, respectively). The effect of the interaction of the two treatments seems to be caused by long snippets for ITs (about 8 s higher, as shown in Table 2). As can be seen in Figure 2, participants with ITs exhibited the highest time consumption with long snippets, although the time spent for NTs does not follow this pattern.

We calculated search accuracy as another search perfor-

Table 2: Search performance, behaviour, and satisfaction for each task type, broken down by snippet length.

		ITs			NTs			*p*-value		
		S	M	L	S	M	L	Task type	Snippet length	Interaction
Search performance	Task completion duration [s]	36.37	41.20	40.98	18.89	20.43	21.10	***	0.24	0.80
	Time to first click [s]	16.24	16.35	24.03	12.90	13.69	14.47	***	**	*
	Search accuracy	0.85	0.85	0.94	1.00	0.98	1.00	*	0.22	0.97
Fixation duration	Total [s]	7.71	7.98	11.69	6.71	6.87	6.98	***	0.06	*
	On titles [s]	4.20	3.48	3.60	3.65	3.60	2.79	0.15	*	0.24
	On URLs [s]	1.55	1.44	1.41	1.68	1.30	1.41	0.42	0.29	0.92
	On snippets [s]	1.96	3.06	6.68	1.38	1.97	2.79	***	***	**
	per link [s]	1.85	2.52	3.56	1.92	2.51	2.66	*	***	*
Search behaviour	Scroll rate	0.29	0.21	0.46	0.17	0.13	0.25	**	**	0.89
	Viewport movement (pixel)	540	675	1148	640	422	1021	0.20	**	0.72
	Compressed sequence length	9.56	7.50	8.27	7.13	6.42	5.75	***	0.12	0.61
	Minimal scanpath length	4.17	3.23	3.13	3.46	2.79	2.50	***	***	0.81
	Revisit	5.40	4.27	5.15	3.67	3.63	3.25	**	0.66	0.60
User satisfaction	7-point Likert scale	4.46	5.35	4.67	5.25	5.21	4.23	0.60	***	***

*Significant at 0.05 level. ** Significant at 0.01 level. *** Significant at 0.001 level.
Note: ITs and NTs denote informational and navigational tasks, and S, M, and L denote short, medium and long snippets, respectively.

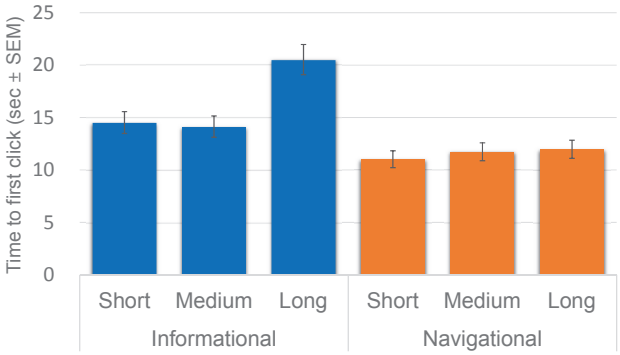

Figure 2: Search time (time to first click). Note that numbers on each y-axis are the values after back-transformation.

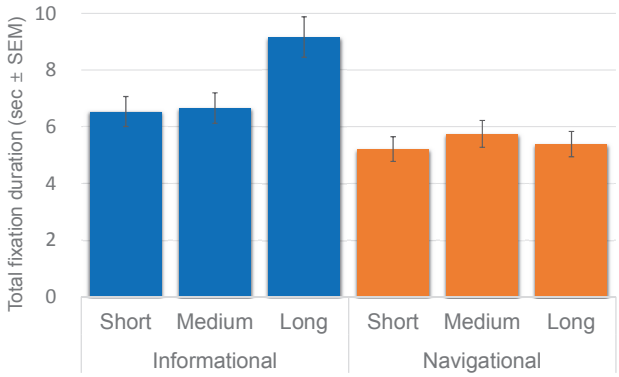

Figure 3: Total fixation duration. Note that numbers on each y-axis are the values after back-transformation.

mance measurement. Each SERP in our tasks included several relevant links with correct answers, unlike the tasks in Cutrell and Guan's study [4], which only contained one best answer. In this paper, therefore, we defined accuracy as the proportion of reaching the correct answer on the first attempt.

A significant difference can be observed in search accuracy between the task types ($\sigma_s^2 = 0.464$, $\chi^2 = 5.10$, df = 1, $p < 0.05$). This indicates that navigational tasks are easier to complete. However, even though we considered the chance at the first attempt, the rates are very high for both task types (the lowest is 85%). In addition, the mean rate for long snippets for ITs is higher than with other snippets (about 9%). However, this result is not statistically significant and does not seem to explain the reason for the longest time spent with long snippets for ITs. Although long snippets might lead to better correct answer rates if the tasks are more difficult than ours, our results suggest that participants with long snippets for ITs did not exhibit better search accuracy, despite spending more time.

To test H2 and further investigate the influences on search time, we measured fixation duration to examine how much effort participants expended to extract the information [12, 23]. Considering the screen size and the distance between participants and the screen, fixations were recorded if a gaze lasted at least 100 ms within a 70 pixel diameter region using algorithms in the Eyeworks software, and we assigned areas of interest (AOIs) to titles, URLs and snippets in each result link.

We adopted ANOVA to compare the user's attention on titles, URLs, and snippets, and the total duration of the three elements. The task type and the interaction between task type and snippet length had significant effects on total fixation duration ($F_{(1,259)} = 20.86$, $p < 0.001$ and $F_{(2,259)} = 3.27$, $p < 0.05$, respectively). As shown in Table 2, ITs with long snippets received about 4 s more attention than the other snippets. With the interaction of task type and snippet length, the pattern shown in Figure 3, for each task type, and broken down by snippet length, is very similar to the pattern for search time (see Figure 2). Using linear regression, we ran further investigations for the relation between total fixation duration and search time and, we found that the relation was clearly positive (common slope: 0.73, SE: 0.29). This suggests that the increase in search time is mainly explained by an increase in time spent reading.

Before moving our focus to user attention on each element, we thought it would be helpful to confirm the difference of the proportion of total fixation duration of each element, to

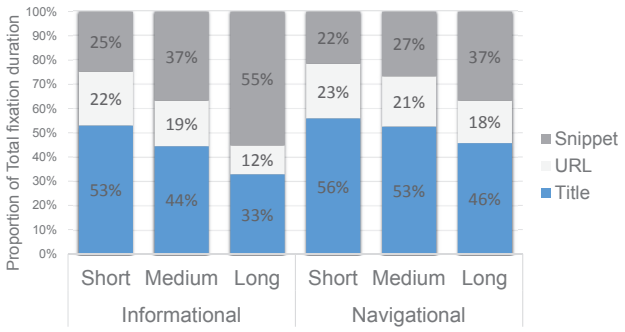

Figure 4: Proportion of total fixation duration for each element of SERPs.

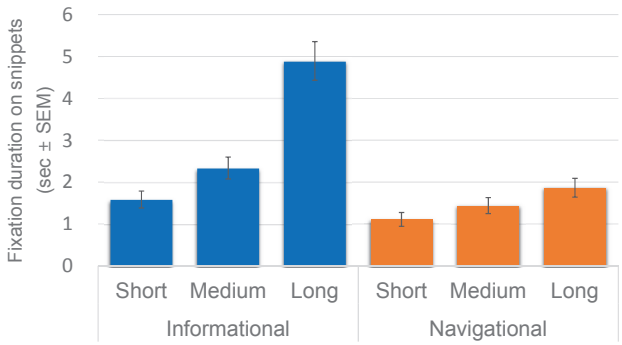

Figure 5: Fixation duration on snippets. Note that the numbers on each y-axis are the values after back-transformation.

better understand overall user attention and also make it possible to test H2.

Figure 4 illustrates how participants distributed their interests for each SERP element according to task type and snippet length. As a general pattern, participants exhibited similar attention for both task types: as the snippet size increased, the title and URL received smaller proportions of attention, whereas the participants tended to look more at snippets. Although participants with NTs also exhibited this pattern, it was clearer for ITs: the proportions of reading the title and URL for ITs were reduced by almost half (from 53.0% and 22.2% to 32.8% and 11.9% for the title and URL, respectively).

Considering the result of total fixation duration (the longer reading time with long snippets for ITs), one possible inference is that participants with long snippets applied additional effort to snippets for ITs, rather than that the proportions of reading the title and URL were absorbed into the snippet reading time. We investigated the inference by analysing user attention on each element.

Fixation duration on the title exhibited a significant difference according to snippet length ($F_{(2,259)} = 3.23$, $p < 0.05$). Although participants paid less attention to the title with long snippets across both task types, as expected from the difference in the proportion of user attention on the title, the maximum difference within each task type was less than 1 s as shown in Table 2. In addition, participants did not exhibit any difference in attention on the URL due to both variables. We now expect that the inference is true: the longest total fixation duration with long snippets for ITs is due to the different fixation duration on snippets, because the difference in proportions of reading both the title and URL had little (about 1 s on the title) or no effect.

Fixation duration on the snippet was affected by task type, snippet length, and their interaction ($F_{(1,259)} = 41.08$, $p < 0.001$, $F_{(2,259)} = 26.54$, $p < 0.001$, and $F_{(2,259)} = 6.06$, $p < 0.01$, respectively). As we inferred, Figure 5 shows that medium snippets for ITs led participants to pay more attention to the snippet than the short snippets did (1.96 s vs 3.06 s in Table 2) and long snippets for ITs resulted in the most attention (6.68 s in Table 2). The pattern for navigational tasks looked similar but the difference observed between short and long snippets was only about 1.4 s, which is not as large as the effect with ITs. Considering the effect due to task type, this result also suggests that the subjects did not need to read the NT snippets as much as they needed to read IT snippets.

Although there was a little difference in user attention on the title (less than 1 s), the above user-attention results on each element suggested that participants basically read the title and URL with some duration across all snippet lengths; then, they paid more attention to the snippet if its length was longer. This pattern was especially strong for ITs.

Aggregating the search-performance results, ITs caused a longer search time than NTs, and participants with long snippets for the ITs required more time, whereas they exhibited no difference by snippet length for NTs. In addition, they exhibited very high accuracy rates without the effect of snippet length. When we considered user attention, we confirmed that the different reading time for snippets is an important component of the biggest cost for search time with long snippets for ITs.

Considering several different conditions (e.g., tasks, participants, and measurements for search time) between our experiment and Cutrell and Guan's experiment [4], we cannot compare the results directly. However, our findings suggest at least that the long snippet does not improve search speed for ITs, unlike the result on a desktop monitor [4]. Therefore, our result does not support H1. In addition, we expected that we could test H2 (user preference among title, URL, and snippet for both task types) with the results of the proportion of fixation duration on each element. However, because the proportions varied according to snippet length, we could not judge whether our results supported H2 or not. This test is revisited in Section 4.3 with the user-preference results in the post-experiment questionnaire.

4.2 Search behaviour

We confirmed that fixation duration is one main contributor to the difference in search time. However, the common slope for the relation between search time and total fixation duration is 0.73, which indicates that the difference in search time is not entirely explained by the time spent reading. In this section, we explore scanning behaviours such as scrolling and the scanpath, and investigate some relations between the behaviour and the search time to determine what else caused the longer search time with long snippets for ITs.

Scroll action. Once users scroll a page, they tend to read a few more results over the page fold [8, 10], and this may cause an additional cost beyond the page fold. Using a

GLMM, we found significant effects due to task type and snippet length on scroll rate ($\sigma_s^2 = 0.387$, $\chi^2 = 7.36$, df = 1, $p < 0.01$ and $\chi^2 = 4.85$, df = 2, $p < 0.01$, respectively). As can be seen in Table 2, participants exhibited different scrolling habits and tended to scroll more often for ITs (14% higher).

The main reason for the difference due to snippet length was the high chance of scrolling with long snippets for both task types. Although snippet length caused a similar pattern in both task types, the difference between the medium and long snippets for ITs appears larger (25% and 12% for ITs and NTs, respectively) and this frequent chance of scrolling seems to be related to the searching and reading time.

Using an LMM, we confirmed the effect of scrolling (scrolled vs. non-scrolled) on search time: scrolling significantly affected search time ($\sigma_s^2 = 0.011$, $\chi^2 = 112.57$, df = 1, $p < 0.001$) without any interaction with task type or snippet length. This result means that participant scrolling led to an increased search time (mean: from 12.93 s to 26.11 s).

With the result of how often users scrolled, we also wondered how far the participants moved the viewport to look at further results beyond the page fold once they had scrolled, because participants with long snippets needed to scroll more to see another link. Based on an LMM and using a log-transformation for the normality assumption, a significant snippet length effect could be observed in viewport movement ($\sigma_s^2 = 0.078$, $\chi^2 = 5.71$, df = 2, $p < 0.01$). As shown in Table 2, the long snippet brought more viewport movement for both task types.

Both results regarding scrolling indicate that the higher chance of scrolling (46%) with bigger viewport movements is another contributor for longer search times with long snippets for ITs.

Scanpath. Scanpath presents the sequence of movements in user attention. Based on a GLMM, we analysed some scanpath measurements, e.g., compressed sequence and minimal scanpath lengths. As introduced in Section 2.1, compressed sequence (how many links a user looked at, including repeat visits) and minimal scanpath (how many different links a user looked at, removing repeated visits from the compressed sequence) were commonly adopted [6, 14, 16, 18] to investigate users' scanning strategy.

Only task type effect was observed on compressed sequence length ($\sigma_s^2 = 0.088$, $\chi^2 = 10.73$, df = 1, $p < 0.001$). This indicates that the subjects tended to visit more links (including revisits) for ITs. Because this length consists of the actual number of links the users looked at and revisits into the links, we could investigate this further by analysing minimal scanpath length.

We also found significant task type and snippet length effects on minimal scanpath length ($\sigma_s^2 = 0.046$, $\chi^2 = 15.53$, df = 1, $p < 0.001$ and $\chi^2 = 16.70$, df = 2, $p < 0.001$, respectively). This means that participants looked at more links with short snippets for both task types (maximum about one link), whereas they exhibited no difference between medium and long snippets (3.1–3.3 links for ITs, 2.5–2.8 links for NTs).

However, when we considered the numbers of result links displayed on the initial SERPs with each snippet length (averages: 5.0, 3.7, and 2.4 links; see Figure 1 for examples of short and long snippets), the mean minimal scanpath lengths for long snippets (3.13 and 2.5) appeared somewhat

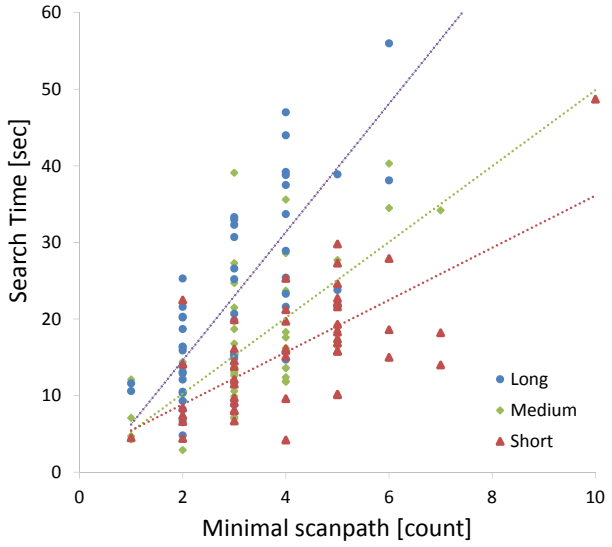

Figure 6: Relationship between minimal scanpath (scanned links) [count] and search time (elapsed time to first click) [sec]. When a participant looks at three links, this leads to about 12.3, 15.2, and 23.0 s spent with short, medium, and long snippets, respectively. The slopes (SEs) are 8.38 (0.66), 4.95 (0.78), and 3.41 (0.62) with p < 0.001 for short, medium, and long snippets, respectively.

high to explain the effect without scrolling, especially for ITs. As we confirmed the relationship between scrolling and search time, we also wondered how the number of scanned links affected search time. Therefore, we investigated the relationship between minimal scanpath (how many links a participant looked at) and the search time (time to first click) for ITs.

Figure 6 uses linear regression to show the relationship. When the minimal scanpath was one, users spent the same amount of time with each snippet length. However, as the length of minimal scanpath increased, users with longer snippets needed to spend more time. This suggests that participants with long snippets scanned slightly less/similar numbers of links compared to those with short and medium snippets; however, the number of scanned links with long snippets required a larger time cost, possibly because of the higher scroll rate.

Relations between search behaviours. With the results of compressed sequences and minimal scanpaths, we can measure two interesting search behaviours. First, using the difference between minimal scanpath and compressed sequence, we can extract the number of links the participants revisited. Revisit count exhibited a significant difference according to snippet length ($\sigma_s^2 = 0.144$, $\chi^2 = 7.85$, df = 1, $p < 0.01$). Although the revisit counts differed between task types (about 1.5 links more on average for ITs), the participants exhibited no different revisit patterns across the snippet length for either task types; even the number of revisited links with long snippets in ITs (5.15) is larger than the count for medium snippets (4.27). If we take the revisit counts as a proxy for users' hesitating, the participant tended to similarly hesitate with the different snippet lengths before

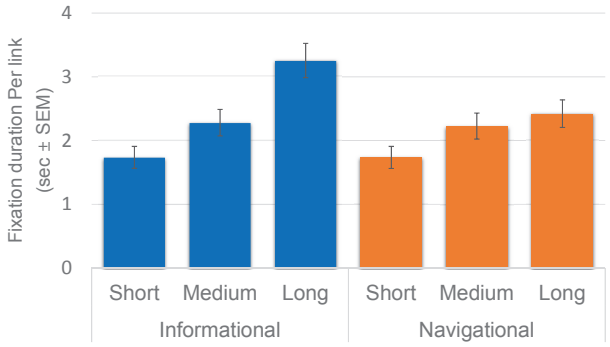

Figure 7: Fixation duration per link. Note that numbers on each y-axis are the values after back-transformation.

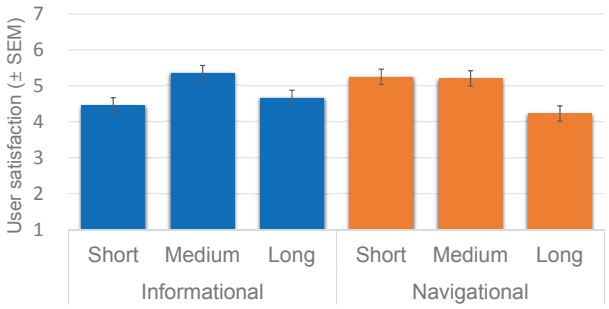

Figure 8: User satisfaction: 7-points Likert scale.

deciding where to click on SERPs, despite those with long snippets needing more effort to move their eye gaze to other links (and possibly scroll).

Second, connecting minimal scanpath to the fixation duration, we could extract fixation duration per link which means how much effort users made to read a link. We found significant effects due to task type, snippet length and their interaction on fixation duration per link ($F_{(1,259)} = 4.59$, $p < 0.05$, $F_{(2,259)} = 27.70$, $p < 0.001$ and $F_{(2,259)} = 3.68$, $p < 0.05$, respectively). As can be seen in Table 2, the participants exhibited slightly less reading time for each scanned link for NTs (2.65 s vs 2.36 s, for ITs and NTs, respectively), and they also spent different amounts of time reading with the three different snippet lengths. As shown in Figure 7, participants needed more time to extract information from one link as snippet length increased (1.85 s, 2,52 s, and 3.56 s from short to long snippets), this pattern was observed between short and medium snippets for NTs.

This explains the relation between the number of scanned links (minimal scanpath) and the reading time. The search time in Figure 6 includes both reading time and other actions (e.g. scrolling). Therefore, we could not examine whether the participants also needed to spend different times reading each link according to snippet length. With this figure, we can confirm that the long snippets for ITs caused both longer reading and search time per link.

In this sub-section, we found that participants with long snippets scrolled more frequently with bigger viewport movements, and that they needed more effort with long snippets to both search and read one link with similar or slightly fewer numbers of looks at links per task, and with similar hesitation in choosing a link. We confirmed that the above two results were related to search time, and these were other reasons for the greater time consumption with long snippets for ITs.

4.3 User preference and post-experiment questionnaire

In a qualitative analysis, user satisfaction is one of the most important factors in designing user interfaces, including SERPs design. As mentioned in Section 3.3, participants were required to score their satisfaction after each task using a 7-point Likert scale regarding the snippet length. Table 2 shows that there were significant effects due to snippet length and the interaction of task type and snippet length on user satisfaction ($\sigma_s^2 = 0.447$, $\chi^2 = 12.93$, df = 2, $p < 0.001$ and $\chi^2 = 7.68$, df = 2, $p < 0.001$, respectively). The overall satisfaction was affected by snippet length, however the patterns differed according to task type.

Figure 8 shows that participants with ITs marked the highest scores (5.35) for the medium snippets (no difference between short and long snippets, with scores of 4.46 and 4.67, respectively), and they expressed the worst satisfaction on the long snippets for NTs (4.23). In other words, participants preferred the medium snippet for ITs, and disliked long snippets for NTs. Unfortunately, it is complicated to find a relationship between user satisfaction and search performance.

After the experiment, participants were asked to choose the most important element among the title, URL, and snippet for deciding about each task type. They also needed to fill out their preference about snippet length for the task types, along with the reasons for their preference.

For ITs, sixteen participants replied that the snippet was the most important element. Six of the remainder replied that they mainly looked at the title when choosing (one of them also chose the snippet as well). Only three participants considered the URL as the most valuable element for ITs. Similar to the user satisfaction results, the participants' overall preferred IT snippet length was the medium snippet (medium: 18, long: 5, and short: 1). Participants' opinions, such as those below, reflected our observations of searching and reading time:

"6–7 lines [of snippet] were useful, but I could not concentrate."

"Long [snippet] takes more time to read and short [snippets] sometimes does not contain enough information."

For NTs, eleven participants replied that the title was the most important component, and 14 subjects said the URL was the most useful factor (one user chose both). However, no one chose 'snippet'. Participants preferred one or 2–3 (short or medium) line snippets for NTs to six or more lines (long snippets). Thirteen of the subjects expressed that short snippets were the best, the remainder (11) replied that medium snippets were suitable for NTs. No one chose the long snippets for NTs. This also matches the user satisfaction results: long snippets were worst and medium or short were better for NTs.

"Firstly, I use the URL to specify the website and I use snippets to support my thought. 2-3 lines [of snippet] are enough to check some basic information."

"One line is too short, and I don't need long sentences [for NTs]."

"[I need one line because,] the title and URL are enough to reach the particular web page."

In addition, we asked participants about the quality of the query and the task difficulty; i.e., if the cached queries were relevant and how difficult the tasks were. With a 7-point Likert scale (1: completely different/difficult, 7: completely same/easy), the participants scored 6.4 (SD: 0.6) for similarity between the cached queries and their own if they could make them up (they could re-check all descriptions and queries before they marked the scores), and gave 6.25 (SD: 0.8) points for task difficulty. Therefore we could confirm that we prepared appropriate queries for the tasks, and the tasks were simple.

We could not confirm whether our results supported H2 by analysing the proportion of fixation duration on each element, because it varied by snippet length. According to the user preference results for SERP elements, almost consistent with H2, users considered that the most valuable component was the snippet for ITs, and the title and URL for NTs. Regarding H3, although it was not 100% consistent, participants were satisfied with the medium snippets for ITs, and with short and medium snippets for NTs. Therefore our results broadly supported H3.

5. DISCUSSION

For each result, we have discussed the meanings of explicit and implicit data for search performance, behaviour, and user preference. In this section, we summarize our discussion and address several limitations in this experiment that we should consider.

First, our search performance results show that navigational tasks are easier than informational tasks, similar to the results from previous studies (e.g. [4, 14]). We also found a snippet-length effect: i.e., participants with long snippets for ITs exhibited longer search time with no difference in search accuracy. This result on a mobile device differed from the results on a desktop [4]: long snippets reduced search time, although we could not compare directly.

When we investigated user attention, we found one reason for the longer time consumption because of the long snippets: participants' attention on SERPs formed a pattern very similar to search time across the snippet length and task types. Furthermore, we finally confirmed that the longer reading time with the long snippets mainly came from the longer user attention on the snippets. Further analysis is required to investigate the reason why users read the long snippets.

Second, we analysed search behaviour to investigate how participants reacted differently to SERPs with different snippet sizes, and to see if there was another factor that affected search time. Our findings suggested that the higher scrolling frequency with more viewport movements to see additional links beyond the page fold was another reason for the worse search time for long IT snippets, even when participants could reach a relevant link without scrolling.

We also confirmed that participants with long snippets scanned similar (or slightly different) numbers of links with similar hesitation before deciding on the SERPs, although the long snippets clearly required more time for searching and reading each link. The combination of scrolling and scanning behaviour was another reason for the long search times with long snippets for ITs. In addition, the difference in search time between both task types seemed to come from obvious differences in scroll rates and scanpath: less chance of scrolling and fewer scanned links. For users' better search experience, horizontal pagination [15] may help reduce the effect of scrolling, and highlighting the query words in the snippets may reduce hesitation behavior on SERPs.

Third, our findings displayed similar results in both post-task and experiment questionnaires. For ITs, our participants expressed that the snippet was the most important element, but they preferred two or three line snippets. For NTs, they considered the URL and title as more valuable than the snippet, and no one wanted long snippets (over six lines). In addition, our participants expressed the view that the prepared queries were very appropriate for each tasks and the tasks were quite easy to solve. This results seems that they might be "trained" through their everyday searches with the snippet length provided by current search engines.

5.1 Limitations

We acknowledge that our experiment had several limitations, which might affect our results. First, we recruited participants from a particular pool, although the age range was wide and they had varying backgrounds. Therefore, our results cannot represent the search behaviour of all users in the world.

Second, one disadvantage of the laboratory study was that our participants could not move while conducting the experiment. They were asked to sit on a chair and not make big movements to ensure accuracy for our gaze recording. We know this condition is different from actual searching on mobile devices.

Third, the tasks had similar difficulties including at least two relevant results within the top three links, and we counter-balanced the task distribution across the participants. Moreover our participants confirmed that the quality of tasks was very appropriate. However, we recognize that different task difficulties and/or lower quality of queries might lead to different user interaction.

Fourth, even though the mobile device in our experiment had a popular screen size [19], the screen size is a major factor in displaying different snippet lengths. The number of links shown in SERPs should differ according to screen size [16, 22], and this could cause different results.

Fifth, we extracted snippets using Nutch and Solr/Lucene. Therefore the quality of snippets may differ from snippets extracted by other search engines.

6. CONCLUSIONS AND FUTURE WORK

The purpose of this study was to investigate user interaction according to different snippet lengths with both informational and navigational task types. Considering the limitations, we concluded that if we provide users with long snippets for mobile searches, instead of the typical two or three snippet lines, it will take longer because they will read the snippets.

In addition, even if a relevant link is on the top page, users will frequently want to read more to check further links over the page fold, and the cost of scrolling and reading far outweighs the benefit (no better chance of reaching a correct answer), especially for finding a particular piece of

information. Most importantly, users would not be satisfied with the long snippets and would finally want to read two–three lines for ITs and one–three lines for NTs.

Unlike the effect of long snippets for desktop screens, the long snippets did not seem to be useful for mobile devices. Overall, although users might become accustomed to searching with the typical snippet size, our results suggested that mobile users are best served by snippets of two to three lines.

The limitations indicated that the screen size and task difficulty might cause other effects, and the function for expanding snippets (e.g., Yandex mobile search engine) appears to improve search performance and user satisfaction. Therefore, our future work will include extending this investigation with various levels of task difficulty on several different screen sizes of mobile devices, and evaluating the usability of expandable snippets.

7. REFERENCES

[1] Adwords. Building for the next moment, 2015. Retrieved from http://adwords.blogspot.com.au/2015/05/building-for-next-moment.html.

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

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

[4] E. Cutrell and Z. Guan. What are you looking for?: An eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference*, pages 407–416. ACM, 2007.

[5] Devguru. Twomon USB. Retrieved from http://devguru.co.kr/easynlight-en/.

[6] S. Dumais, G. Buscher, and E. Cutrell. Individual differences in gaze patterns for web search. In *Proceedings of the third symposium on Information interaction in context*, pages 185–194. ACM, 2010.

[7] Eyetracking. Eyeworks. Retrieved from http://www.eyetracking.com/Software/EyeWorks.

[8] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference*, pages 478–479. ACM, 2004.

[9] G. Hotchkiss, S. Alston, and G. Edwards. Eye tracking study. *Research white paper, Enquiro Search Solutions Inc*, 2005.

[10] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference*, pages 154–161. ACM, 2005.

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

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

[13] M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving search results quality by customizing summary lengths. In *ACL*, pages 701–709, 2008.

[14] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. 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*, 66(3):526–544, 2015.

[15] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. Pagination versus scrolling in mobile web search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 751–760. ACM, 2016.

[16] J. Kim, P. Thomas, R. Sankaranarayana, T. Gedeon, and H.-J. Yoon. Understanding eye movements on mobile devices for better presentation of search results. *Journal of the Association for Information Science and Technology*, 2016.

[17] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference*, pages 113–122. ACM, 2014.

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

[19] Mobile Marketing. Majority of mobile web browsing on smartphones bigger than 5 inches, 2015. Retrieved from http://mobilemarketingmagazine.com/54708-2/.

[20] J. Nielsen. F-shaped pattern for reading web content. *Jakob Nielsen's Alertbox*, 17, 2006.

[21] T. Paek, S. Dumais, and R. Logan. Wavelens: A new view onto internet search results. In *Proceedings of the SIGCHI conference*, pages 727–734. ACM, 2004.

[22] D. Raptis, N. Tselios, J. Kjeldskov, and M. B. Skov. 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*, pages 127–136. ACM, 2013.

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

[24] Search Engine Land. Worldwide, more than half of Google's searches happen on mobile, 2015. Retrieved from http://searchengineland.com/half-of-google-search-is-mobile-232994.

[25] S. Searle. *Linear Models*. Wiley Classics Library. Wiley, 1997.

[26] Seeingmachines. Facelab5. Retrieved from http://www.seeingmachines.com/.

[27] The Apache Software Foundation. Nutch and Solr/Lucene. Retrieved from http://www.apache.org/.

[28] P. Thomas, F. Scholer, and A. Moffat. What users do: The eyes have it. In *Asia Information Retrieval Symposium*, pages 416–427. Springer, 2013.

[29] VSN International. Genstat for Windows 17th edition., 2014. VSN International, Hemel Hempstead, UK. Web page: www.vsni.co.uk.

[30] B. T. West, K. B. Welch, and A. T. Galecki. *Linear mixed models: A practical guide using statistical software*. CRC Press, 2014.