Understanding search behaviour on mobile devices

Jaewon Kim

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

November 2017
Except where otherwise indicated, this thesis is my own original work.

Jaewon Kim
November 2017
To wife, daughter, son, parents, parents-in-law, and my sisters
for all your love, understanding and encouragement
Acknowledgments

This work would not be completed without the help of the people listed below,

- **Supervisors:** I would like to thank my supervisors, Dr. Ramesh Sankaranarayana (ANU, the panel chair) and Dr. Paul Thomas (Microsoft, the primary supervisor) for continuous encouragement throughout the research, unsparing guidance, and valuable feedback. I have been heartily happy to work with them. My PhD research could be on the right track due to their supervision.

- **Advisors:** I am thankful for Prof. Tom Gedeon (ANU, head of the Human Centred Computing (Hcc) group), and Dr. Hwan-Jin Yoon (ANU, statistical advisor) for providing their guidance and support regarding research areas of Human-computer Interaction and Statistics, and the life in Canberra.

- **People in Research School of Computer Science:** I am also grateful to administrative and technical staff in Research School of Computer Science. During my PhD program, I could concentrate on my research without any problem under their assistance and support.

- **ANU:** Thanks to ANU for the financial support and grants for a Vice chancellor project and a travel fund.

- **Colleagues:** Thanks to my colleagues of PhD students in HCI group, research scientists in Information Retrieval at CSIRO, and researchers in Microsoft (Drs. David Hawking, Peter Bailey, and Nick Craswell) for their friendly advice and kindness.

- **People in RMIT University:** Thanks to my current manager and mentor, Prof. Mark Sanderson and Assoc. Prof. Falk Scholer in RMIT University, for their support in completing my thesis.
List of Publications

The work reported in this thesis is based on materials from following publications. During my PhD, one of my contributions had been accepted for publication by a journal, the other three have been published in conferences, and another has been presented as a poster.


Abstract

Web search on hand-held devices has become enormously common and popular. Although a number of studies have revealed how users interact with search engine result pages (SERPs) on desktop monitors, there are still only few studies related to user interaction in mobile web search, and search results are shown in a similar way whether on a mobile phone or a desktop. Therefore, it is still difficult to know what happens between users and SERPs while searching on small screens, and this means that the current presentation of SERPs on mobile devices may not be the best.

According to the findings from previous studies, including our earlier work, we can confirm that search behaviour on touch-enabled mobile devices is different from behaviour with desktop screens, and so we need to consider a different SERP presentation design for mobile devices. In this thesis, we explore several user interactions during search with the aim of improving search experience on smartphones.

First, one remarkable trend of mobile devices is their enlargement of screen sizes during the last few years. This leads us to look for differences in search behaviour on different sized small screens, and if there are any, to suggest better presentation of search results for each screen size. In the first study, we investigated search performance, behaviour, and user satisfaction on three small screens (3.6 inches for early smartphones, 4.7 inches for recent smart-phones and 5.5 inches for phablets). We found no significant differences with respect to the efficiency of carrying out tasks. However, participants exhibited different search behaviours on the small, medium, and large sizes of small screens, respectively: a higher chance of scrolling with the worst user satisfaction on the smallest screen; fast information extraction with some hesitation before selecting a link on the medium screen; and less eye movements on top links on the largest screen. These results suggest that the presentation of web search results for each screen size needs to take into account differences in search behaviour.

Second, although people are familiar with turning pages horizontally while reading books, vertical scrolling is the standard option that people have available while searching on mobile devices. So following a suggestion from the first study, in the second study we explored the effect of horizontal and vertical viewport control types (pagination versus scrolling) with various positions of a correct answer in mobile web search. Our findings suggest that although users are more familiar with scrolling, participants spent less time to find the correct answer with pagination, especially
when the relevant result is located beyond the page fold. In addition, participants using scrolling exhibited less interest in lower-ranked results even if the documents were relevant. The overall result indicates that it is worthwhile providing different viewport controls for better search experiences in mobile web search.

Third, snippets occupy the biggest space in each search result. Results from a previous study suggested that snippet length affects search performance on a desktop monitor. Due to the smaller screen, the effect seems to be much larger on smartphones. As one possible idea for a SERP presentation design from the first study, we investigated appropriate snippet lengths on mobile devices in the third study. We compared search behaviour with three different snippet lengths, that is, one line, two to three lines, and six or more lines of snippets on mobile SERPs. We found that with long snippets, participants needed longer search time for a particular task type, and the longer time consumption provided no better search accuracy. Our findings suggest that this search performance is related to viewport movements and user attention.

We expect that our proposed approaches provide ways to understand mobile web search behaviour, and that the findings can be applied to a wide range of research areas such as human-computer integration, information retrieval, and even social science for a better presentation design of SERP on mobile devices.
Contents

Acknowledgments vii
List of Publications ix
Abstract xi

1 Introduction 1
   1.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
   1.2 Research Purpose . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
   1.3 Thesis outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2 Background and Related Work 5
   2.1 Evaluating SERPs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
      2.1.1 System measures: IR approaches . . . . . . . . . . . . . . . . . . . . 5
         2.1.1.1 Online metrics . . . . . . . . . . . . . . . . . . . . . . . . . 5
         Search time . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5
         Click . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
         2.1.1.2 Offline metrics . . . . . . . . . . . . . . . . . . . . . . . . . . 6
            Precision, recall, and F-measure . . . . . . . . . . . . . . . . . . . . 6
            MAP and NDCG . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
   2.1.2 Behaviour and physiological methods: HCI approaches . . . . . 8
      ECG . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
      GSR . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
      EEG . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
      Eye tracker . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
   2.2 Reading and scanning SERPs . . . . . . . . . . . . . . . . . . . . . . . . . 10
      2.2.1 Standard SERPs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
      2.2.2 Manipulated SERPs . . . . . . . . . . . . . . . . . . . . . . . . . . 14
      2.2.3 Search strategies . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
   2.3 Mobile web search . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
   2.4 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

xiii
3 Measurements

3.1 Search performance
   3.1.1 Search time
   3.1.2 Search accuracy

3.2 Search behaviour
   3.2.1 Fixation duration
   3.2.2 Click pattern
   3.2.3 Scanpaths
   3.2.4 Scanning direction
   3.2.5 Skip and regression
   3.2.6 Scroll
   3.2.7 Trackback

3.3 Questionnaire measures

3.4 Summary

4 Study One: Three Different Small Screens

4.1 Introduction

4.2 User study and data collection
   4.2.1 Participants
   4.2.2 Tasks
   4.2.3 Design and procedure
   4.2.4 Apparatus
   4.2.5 Data collection and post processing

4.3 Results and discussion
   4.3.1 Power analysis
   4.3.2 Search performance
      Search time
      Search accuracy
   4.3.3 Search behaviour
      Fixation duration
      Click pattern
      Scanpath
      Scanning direction
      Skip and regression
      Scroll
      Trackback
   4.3.4 Questionnaire measures
   4.3.5 General Discussion
Contents

4.4 Conclusions and future work .................................................. 46

5 Study Two: Pagination versus scrolling .................................... 49
  5.1 Introduction ................................................................. 49
  5.2 User study ........................................................................ 51
    5.2.1 Participants .......................................................... 52
    5.2.2 Tasks ....................................................................... 52
    5.2.3 Design and procedure .............................................. 53
    5.2.4 Apparatus .................................................................. 54
  5.3 Results ................................................................................ 55
    5.3.1 Search performance and user satisfaction ..................... 55
    5.3.2 Scroll action and duration ....................................... 58
    5.3.3 User attention ......................................................... 60
    5.3.4 User preference ....................................................... 65
  5.4 Discussion ........................................................................... 65
    5.4.1 Limitations .............................................................. 66
  5.5 Conclusions and future work ................................................. 67

6 Study Three: Appropriate snippet length .................................... 69
  6.1 Introduction ........................................................................ 69
  6.2 User study ........................................................................... 71
    6.2.1 Participants .......................................................... 72
    6.2.2 Tasks ....................................................................... 72
    6.2.3 Design and procedure .............................................. 73
    6.2.4 Apparatus .................................................................. 74
  6.3 Results ................................................................................ 74
    6.3.1 Search performance and user attention ..................... 75
    6.3.2 Search behaviour .................................................. 81
      6.3.2.1 Scroll action ................................................... 81
      6.3.2.2 Scanpath ......................................................... 82
      6.3.2.3 Relations between search behaviours ............... 84
    6.3.3 User preference and post-experiment questionnaire .... 85
  6.4 Discussion ........................................................................... 87
    6.4.1 Limitations .............................................................. 88
  6.5 Conclusions ........................................................................ 89
7 Conclusions and future directions 91
  7.1 Summary of Contributions ............................... 91
  7.2 Future Directions ........................................ 93
  7.3 Closing Remarks ......................................... 95

A Appendix A: Experiment forms 97

B Appendix B: Task descriptions and queries 101

C Appendix C: The detailed post-experiment questionnaire 105

Bibliography 113
# List of Figures

2.1 An eye tracker ................................................................. 9
2.2 The heat-map of Google's golden triangle. .......................... 12
2.3 The heat-map of F-shaped pattern ................................. 13
2.4 Examples of three different snippet lengths for a single search result ......................................................... 16
2.5 Examples of the depth-first strategy and mixed strategy 17
2.6 An example of the breadth-first strategy .......................... 18
2.7 Examples of evaluation styles ......................................... 19
2.8 An example of the search results page showing Knowledge Graph result .................................................. 20
2.9 Search results as shown on the large and small screens. ........ 21
3.1 Definitions of four kinds of search time .......................... 24
3.2 A example of a scanpath on the search results page .............. 26
4.1 Examples of search engine result pages on three different screens ................................................................. 31
4.2 Microsoft Surface Pro 3 .................................................. 33
4.3 Theeyetribe ................................................................. 34
4.4 Mean fixation duration on each AOI .............................. 39
4.5 Relationship between search time and fixation duration per link ................................................................. 41
5.1 Example of two subsequent pages of a SERP with the horizontal control type ......................................................... 50
5.2 An Example of the experimental environment .......................... 54
5.3 Scroll rate with each target position, with standard error of mean (SEM) ................................................................. 58
5.4 Fixation duration on each AOI along target positions .............. 63
6.1 Examples of constructed SERPs with short, medium, and long snippets ................................................................. 70
6.2 Search time ................................................................. 77
6.3 Total fixation duration ................................................................. 78
6.4 Proportion of total fixation duration ................................................................. 79
6.5 Fixation duration on snippets ................................................................. 80
6.6 Relationship between minimal scanpath and search time ................................................................. 83
6.7 Fixation duration per link ................................................................. 84
6.8 User satisfaction ................................................................. 86
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Examples of task descriptions and queries from our earlier work.</td>
<td>14</td>
</tr>
<tr>
<td>4.1</td>
<td>Examples of task descriptions and queries.</td>
<td>32</td>
</tr>
<tr>
<td>4.2</td>
<td>Search performance and behaviour.</td>
<td>38</td>
</tr>
<tr>
<td>4.3</td>
<td>Effects of scrolling on search time.</td>
<td>44</td>
</tr>
<tr>
<td>5.1</td>
<td>Examples of task descriptions and queries.</td>
<td>53</td>
</tr>
<tr>
<td>5.2</td>
<td>Search performance and user satisfaction.</td>
<td>56</td>
</tr>
<tr>
<td>5.3</td>
<td>Scroll actions and search time excluding the scroll duration.</td>
<td>56</td>
</tr>
<tr>
<td>5.4</td>
<td>Search performance, user satisfaction, and scroll rate with each of the</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>six target positions.</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Fixation duration: relations between control types and others.</td>
<td>60</td>
</tr>
<tr>
<td>5.6</td>
<td>Fixation duration.</td>
<td>62</td>
</tr>
<tr>
<td>6.1</td>
<td>Examples of task descriptions and queries.</td>
<td>72</td>
</tr>
<tr>
<td>6.2</td>
<td>Search performance, behaviour, and satisfaction.</td>
<td>76</td>
</tr>
<tr>
<td>B.1</td>
<td>Full task descriptions and queries (Study One).</td>
<td>102</td>
</tr>
<tr>
<td>B.2</td>
<td>Full task descriptions and queries (Study Two).</td>
<td>103</td>
</tr>
<tr>
<td>B.3</td>
<td>Full task descriptions and queries (Study Three).</td>
<td>104</td>
</tr>
</tbody>
</table>
This chapter introduces the motivation, research purpose and outline of the thesis as why this study was worth conducting, what the purpose of this study is, and what we address in this thesis.

## 1.1 Motivation

With the expansion in the volume of information on the Internet, users have been able to search appropriate information and to retrieve them, and this has affected web usage. According to recent reports, web usage has increased by over 900% from 2000 to mid-2016 globally [Internet World Stats, 2016].

Meanwhile, 91% of all people on earth have a mobile device, the growth of smartphone ownership has been increasing to 56% at the middle of 2013 [Digital Buzz Blog, 2013], and web search is one major activity on mobile devices [Adwords, 2015]. Web search on mobile devices have become common due to their convenience. A report from 2014 indicates that mobile internet usage has soared by 67% from Sept., 2013 to Aug., 2014, and the use of hand-held devices (i.e. mobile phones and tablets) has grown rapidly from 21.9% to 35.3% worldwide, while accessing the web by using desktops decreased to 64.6% [Statcounter Global Stats, 2014]. More recently, Google, a popular search engine company, announced that more than half of its searches occur on mobile devices globally [Search Engine Land, 2015].

In the interest of enhancing the user search experience, a number of studies have investigated user interactions with elements of the search engine results pages (SERPs) on desktop monitors by evaluating search performance, behaviour, and user satisfaction. Researchers have studied user interaction under the standard conditions provided by search engines [Dumais et al., 2010; Granka et al., 2004; Lorigo et al., 2006], manipulated elements of SERPs (e.g., length of snippets and rank order) [Cutrell and Guan, 2007; Guan and Cutrell, 2007; Joachims et al., 2005; Kelly and Azzopardi, 2015], and classified task types according to the search goals [Broder, 2002; Lorigo}
et al., 2006]. The results of these studies have led to better SERP interfaces, and a number of search engines have incorporated some of these suggestions into their design.

There have been recent efforts to improve this design by understanding user interaction with small devices [Guo et al., 2013; Lagun et al., 2014, 2016; Raptis et al., 2013]. However, it seems that current search engines do not provide different contents for small devices, instead essentially simplifying their results pages, and compared to the amount of investigation in user interaction on desktop screens, there is not enough research on small screens to suggest optimised presentation of SERPs for mobile devices.

Our earlier work [Kim et al., 2015] compared search performance and behaviour on a desktop monitor versus a mobile device screen, and revealed that users have some difficulty in extracting information and exhibit less eye-movement, and are slower in completing tasks. The study suggested that the interface design for web searches on mobile devices needs to be different from that on a desktop monitor.

As an extension of this earlier work, this research focused on understanding user interaction on various small screen sizes and developing some function and design for better mobile search experience. The research includes three main issues as follows:

- Search behaviour among different small screen sizes.
- Effects of horizontal pagination for viewport controlling in mobile web search.
- Finding an appropriate snippet size on mobile SERPs.

In this thesis, we measure search performance, search behaviour, and user satisfaction to investigate effects of the research variables, i.e., screen sizes, control types and snippet sizes, and considering the relationships among the three measurements. First, we record search speed and accuracy as the main variables to evaluate search performance. This data represents explicit search performance as broadly used in previous studies (e.g., Dumais et al., 2010; Granka et al., 2004; Raptis et al., 2013). Second, eye-tracking is a useful technology to investigate where users are interested in, how users interest moves, and how much users exhibit the interest [Just and Carpenter, 1976; Rayner, 1998]. We use eye-trackers in the experiments, since several previous works looked into search behaviours with such equipment [Cutrell and Guan, 2007; Guan and Cutrell, 2007; Lorigo et al., 2006]. Third, recording users search experience can provide helpful data in evaluating usability of search interface [Kelly and Azzopardi, 2015; Lagun et al., 2014]. We measure user satisfaction using post-task and post-experiment questionnaires to discover how user’s scores are different by


1.2 Research Purpose

The research purposes come from the discovery of the improved presentation of SERP by understanding search behaviour on small screens. Although a number of studies have been performed to investigate user behaviour in web search for better presentations designs of SERP on conventional desktop screens, it appears that only a few studies have attempted to comprehend the search behaviour on small screens separately. Therefore, the research purposes are as follow:

**Search behaviour among different small screen sizes**

- To investigate the effects of different small screens in search performance and behaviour.
- To provide appropriate presentation designs for each small screen size by analyzing the search performance and behaviour, if there is a difference according to the screen sizes.

**Effects of horizontal pagination for viewport controlling in mobile web search**

- To study the effects of the viewport control types (horizontal pagination versus vertical scrolling) on SERPs.
- To test if both control types have some interaction with the position of the correct answer.
- To investigate user preference and satisfaction between the two viewport control types.

**Finding an appropriate snippet size on mobile SERPs**

- To observe the effects of different snippet lengths (short, medium, long) in mobile web search.
- To investigate if the purpose of searching (task type) is related to the effect of the snippet length.
- To survey which snippet length users prefer for each purpose of the task, and the reason for the preference.
1.3 Thesis outline

This thesis is organised as follows:

- **Chapter 2. Background and Related Work** comprises the related background works regarding implications of eye-movements, search behaviour and search strategies on the web, and user interaction on small screens.

- **Chapter 3. Measurements** focuses on how to measure search performance, behaviour and user preference, and what each measurement means.

- **Chapter 4. Study One (Three different small screens)** describes the first experimental study which was conducted to investigate the effects of different small screen sizes in mobile web search, and suggests appropriate SERP presentation designs for each screen size.

- **Chapter 5. Study Two (Pagination versus scrolling)** presents the second experiment to study the effects of horizontal and vertical viewport control types along the positions of a relevant answer on SERPs.

- **Chapter 6. Study Three (Length of snippets)** describes the third study that investigated appropriate snippet length for mobile SERPs with different task types.

- **Chapter 7. Conclusion and Future direction** concludes this thesis by summarizing the main finding, considering the contribution and the limitations, and discussing the future directions.
Chapter 2

Background and Related Work

In this chapter, we introduce three necessary general lines of background knowledge for better understanding of this research. The first describes evaluating SERPs (Section 2.1) with several measurements from information retrieval (IR) and Human-computer interaction (HCI) research fields, the second addresses the potential outcomes from an reading and scanning SERPs (Section 2.2), and the third concerns mobile web search (Section 2.3).

2.1 Evaluating SERPs

There are several ways to evaluate SERPs. In this section, we explore IR approaches such as search time, click, precision and recall, and HCI approaches like Electrocardiogram (ECG), Galvanic skin response (GSR) and eye tracking.

2.1.1 System measures: IR approaches

We first address some evaluation methods such as search time and click-through as online metrics, which may be obtain from user studies (both online and laboratory). These evaluating metrics can be more powerful by combining with physiological behaviour. We then explore the other IR approaches (also known as offline metrics) which are also broadly used to evaluate SERP, although they are less related to user studies.

2.1.1.1 Online metrics

Search time  Search time can be defined as the time spent on SERPs and web documents in this thesis, and has been broadly used to measure the time on SERP and dwell time as the most common variable for search performance in user studies. For examples, [Lorigo et al. 2006] compared two task types, [Guan and Cutchell 2007]
investigated effect of target position, [Paek et al., 2004], Cutrell and Guan [2007] studied different snippet lengths, and Kelly and Teevan [2003], Yilmaz et al. [2014] explored relations between dwell time and relevance of SERPs. These previous studies adopted search time as one main measurement, most of them are dealt with in Section 2.2 in details. In this thesis, the search time is used as one major evaluation method and is addressed in detail in Section 3.1.1.

Click  Click-through is also commonly used to evaluate SERPs as how many clicks a result link in a SERP received, and a click decision on SERPs mostly indicates that users expect a relevant answer in the result [Hofmann et al., 2016]. The click-through also has been used in several previous studies, for example, to investigate the relationship between click and the relevance of search results [Joachims et al., 2005], to explore the effect of number of results in a SERP [Kelly and Azzopardi, 2015], and to compare user attention to click decision [Granka et al., 2004] (addressed more in Section 2.2). In this thesis, we explore the click behaviour by connecting it to other measurements such as search time and user attention.

2.1.1.2 Offline metrics

**Precision, recall, and F-measure**  For unranked sets of results, precision and recall were firstly introduced by [Kent et al., 1955], and they are basic and common measurements to investigate effectiveness in IR research area. By the definition in [Manning et al., 2008], Precision (P) is the proportion of relevant documents retrieved amongst all the retrieved documents, and Recall (R) is the proportion of the relevant documents retrieved amongst all the relevant documents as shown below:

\[
\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant|retrieved}) \tag{2.1}
\]

\[
\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved|relevant}) \tag{2.2}
\]

Combining the above measures, a single measure which is the weighted harmonic mean of precision and recall can be defined, named F-measure as below:

\[
F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2P + R} \text{ where } \beta^2 = \frac{1 - \alpha}{\alpha} \tag{2.3}
\]

The default balanced F measure found by giving the same weights (\(\alpha = 0.5\), so
\[ F_{\beta=1} = \frac{2PR}{P + R} \quad (2.4) \]

**MAP and NDCG** Considering SERP from search engines, we address two measurements for ranked retrieval results. First, one of most common measurements is *mean average precision* (MAP): the mean of the average precision scores for a set of queries, assuming that users are looking for several relevant documents. The equation is as below [Manning et al., 2008]:

\[
\text{MAP} = \frac{1}{|Q|} \sum_{j=1}^{\left| Q \right|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk}) \quad (2.5)
\]

where \( Q \) is the set of queries, \( m_j \) is the number of documents relevant to the \( j \)-th query, and \( R_{jk} \) is the set of ranked retrieval results from the top result until you get to document \( d_k \).

Another measurement for ranked results is *Normalized discounted cumulative gain* (NDCG), which is one popular method to evaluate the effectiveness of search engine, especially with machine learning approaches [Manning et al., 2008]. This measurement has two assumptions: one is that higher ranked relevant documents are more useful than marginally (less) relevant documents located in higher ranks; and the other is that the lower ranked relevant document is less useful for users due to a lower chance of it being examined. NDCG considers the usefulness of a document with a graded relevance scale. The gain is accumulated from top to bottom of the SERP, and decreased as moving down to lower ranks [Järvelin and Kekäläinen, 2002]. The formulation is as below:

\[
\text{NDCG}(Q, i) = \frac{1}{|Q|} \sum_{j=1}^{\left| Q \right|} Z_{ij} \sum_{d=1}^{i} \frac{2^{R(j,d)}}{\log_2(1 + d)} \quad (2.6)
\]

where \( Q \) is set of queries, \( Z_{ij} \) (rank at \( i \)) is a normalization factor, and \( R(j,d) \) is the relevant score assessors gave to document \( d \) for query \( j \).

There are more numerous online and offline methods in IR approaches apart from above metrics. The addressed measurements are small samples commonly used, but those are worth to know as the background knowledge before conducting the studies related to IR research area. Considering our research as laboratory experiments, we adopt some online metrics to investigate user behaviour rather than evaluating the quality of SERPs using offline metrics. In addition to this, we use eye-tracking technology to observe where users are interested in as introducing in next subsection,
and combine both approaches (IR and HCI) for understanding mobile web search behaviour in details.

2.1.2 Behaviour and physiological methods: HCI approaches

Human-Computer Interaction (HCI) approaches are sometimes very helpful to understand users’ cognitive behaviour. We briefly describe several measures such as ECG, GSR, and EEG, commonly used in HCI research field. We then introduce eye-tracking, which is a popular method in both IR and HCI areas, as the main objective measurement in this thesis.

ECG  Electrocardiogram (ECG), known as elektrokardiogramm in the German term (EKG) is a graphical record of a wavelength pattern by analyzing electrical activity of the heart during a particular period. Using the ECG, we can obtain both heart rate (HR) and heart rate variability (HRV), and these signals correlate well with human emotion such as joy, sadness, fear, and anger [Sinha et al., 1992].

GSR  Galvanic skin response (GSR), also known as skin conductance (SC) and electrodermal response (EDR), is a measurement of electrical conductance of skin change on the skin. From the data of GSR, we can measure levels of happiness and fear [Nasoz et al., 2004], stress [Liao et al., 2005] and differentiate between different content of data presented to a reader [Sharma and Gedeon, 2011] which is cognate to measuring signals during an information retrieval task.

EEG  Electroencephalography (EEG) is a method in which electrical signals from neural activity in the brain are recorded. This method is a useful and common measurement for brain activity due to high temporal resolution and low cost [Sharma, 2014]. The signal has been used to recognize human emotion [Horlings et al., 2008], reading behaviour [Vo and Gedeon, 2011] and even stress level during playing a computer game [Dharmawan and Rothkrantz, 2007].

In addition to the above measurements, human bodies produce other physiological signals such as blood pressure (BP), electromyography (EMG): electrical activity from active muscles, and skin temperature (ST). These signals also can provide indications of human cognition particularly regarding emotions. For the next thing, we introduce eye tracking which is commonly used for user studies in information retrieval research.
Eye tracker As the main method in this research, eye-tracking is a useful technology to investigate user cognition and user interaction in various research fields of computer science [Copeland et al., 2014; Jacob and Karn, 2003; Rayner, 1998]. In particular, a number of previous works regarding web search have adopted eye tracking for better understanding of users’ attention, because the gaze provides the information about which elements of SERPs attract attention and that how the attention moves (e.g. [Aula et al., 2005]; [Buscher et al., 2010]; [Cutrell and Guan, 2007]; [Dumais et al., 2010]; [Granka et al., 2004]). Therefore, eye-tracking is an appropriate method to investigate search behaviour on mobile devices.

![Figure 2.1: An eye tracker to record users’ gaze data (the image from Seeing-machines).](image)

We describe a few major eye-movement metrics provided by eye trackers which are relevant to our experiments, because eye-tracking is a large research area and it cannot be covered in details in this thesis. The main information source with eye-tracking are fixations and saccades. We can define fixations as the moments which that the eyes are relatively static in order to extract some information: fixation durations can be between 50–75ms and 500–600ms and depending on the reading material. We consider saccades to be the rapid eye movements occurring between fixation points, lasting about 20–35ms. Fixations can have a number of meanings regarding user cognition; saccades, even though they may provide nothing about users’ perception, may provide scanpaths (i.e., direction of fixations) by connecting to fixation
These eye-movements provide several implications in understanding search behaviour. A study by Goldberg and Kotval [1999] suggested that more effective search exhibits less fixations, and that the optimal scanpath in searching displays a short fixation duration with less hesitation. In addition, the number of fixations on a particular area of interest (AOI) represents the importance of the information: more fixations means more importance [Poole et al., 2005], and a longer fixation duration indicates complexity and difficulty of tasks [Just and Carpenter, 1976; Rayner, 1998].

2.2 Reading and scanning SERPs

In this section, we survey the previous works regarding how users interact with SERPs. Although these studies were conducted for desktop search, their results may provide a general background knowledge regarding web search behaviour.

2.2.1 Standard SERPs

Several approaches have been adopted to investigate users’ web search behaviour. One method, analysing transaction log files, has been used for a long time. Log files contain data about clicks, queries, and scrolling events that users made on search engines [Jansen and Spink, 2006; Silverstein et al., 1998]. Silverstein et al. [1998] analysed a large query log file to investigate the interaction between users and a commercial search engine. They found that searchers tend to scan only the first 10 search results, while rarely modifying their query. Similarly, Jansen and Spink [2006] found that the number of scanned results became fewer than in the past, and that users spent more time on SERPs than the time on web documents. They suggested that this may be caused by users’ higher familiarity with current search engines than in the past and the improvement of web search engines. More recently, Buscher et al. [2012] analyzed large query logs from one commercial web search engine, which contains cursor movements and text highlighting information. The results indicated that shorter search time causes users to inspect just a few results, scroll less, and use fast mouse movements.

Another approach uses diary studies and interviews to understand search behaviours. For example, the effect of task complexity [Byström and Järvelin, 1995], orienteering behaviour [Teevan et al., 2004] (performing directed situated navigation), and context in online information seeking [Kelly, 2006]. Byström and Järvelin analyzed the relationship among task complexity, necessary information types, information channels, and sources. They collected data using a combination of diaries.
and questionnaires. The results suggested that the relationships among these factors are significantly logical and systematic. For example, higher task complexity results in more information from various sources, but decreases the success rate of finding the required information. Teevan et al. [2004] investigated how users look for information on the web, adopting a modified diary study supplemented with direct observations and hour-long semi-structured interviews, in order to find an optimized search tool design. They found that users often did not use keyword-based search engines as part of the orienteering strategy. Users seek information using small steps with no specific information, and they suggested that we should consider the orienteering behaviour for web search tools. Using a diary study approach, Kelly [2006] investigated data about information seeking context, the aspects of this context, and relationships among these aspects. She observed the behaviour of seven participants over a three-month period, and suggested that the task and topic significantly affect the perception of usefulness of documents in completing tasks.

The method of analysing transaction log files can provide information on how users interact with SERPs, using the information in the files (e.g. mouse clicks, queries or cursor movement). Diary studies and interviews are useful in information interaction studies. However, these methods seem to have a limitation as they do not provide detailed information about where users are looking and why users interact with different elements of SERPs moment-by-moment. Thus, it is worthwhile to consider studies that have adopted eye-tracking.

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. [2005] found that users make a “golden triangle” pattern: the most popular area in first time visits to a SERP (see Figure 2.2), 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 [Nielsen, 2006] indicates that users exhibit an “F-shaped pattern” while searching, i.e., they scan one vertical stripe followed by two horizontal stripes (see Figure 2.3). 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, that is, which link in SERPs is firstly selected. Granka et al. [2004] 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 et al. [2005] and Nielsen [2006]. They also found that users tend to scan the links above the selected link. However users often use different patterns near the page fold (the end of initial results with no scrolling) with
Background and Related Work

Figure 2.2: The heat-map of Google’s golden triangle from Hotchkiss et al. [2005]: users’ eye fixations exhibits a triangle shape around the top ranks.

A study by Joachims et al. [2005] presented results similar to the findings of Granka et al. [2004] and additionally suggested that a user’s click decisions are affected by the relevance of search results. The results of both studies broadly indicate that users read the search results with a top-to-bottom scanning pattern, although Thomas et al. [2013] found that some users began their exploration from a different position rather than the top rank on SERPs.
Reading and scanning SERPs

There was a study regarding children's search behaviour. Bilal and Gwizdka [2016] investigated the effect of grade level (grades 6 and 8, respectively) and task type on children's reading behaviour on SERPs. Their preliminary findings indicate that children show a significant difference in reading behaviour, fixation count, and the first visit according to the grade level or age. Participants in grade 8 tended to read top result first, whereas children in grade 6 were more likely to look at lower rank for their first eye visit.

A few studies examined search behaviours according to the user's goals in web search. Broder [2002] 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. [2006] conducted an experiment to investigate user behaviours with the informational and navigational tasks (see Table 2.1 for examples of the task types in our earlier work [Kim et al., 2015]). Their findings suggest that users take more time to complete informational tasks. Both informational and navigational tasks have been commonly used (e.g., Granka et al. [2004], Joachims et al. [2005], Kim et al. [2015]), and the task types are one main consideration in this thesis.

In addition, Lorigo et al. [2006] defined a compressed sequence and a minimal scan-path 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 fix-
Background and Related Work

Table 2.1: Examples of task descriptions and queries from our earlier work [Kim et al., 2015].

<table>
<thead>
<tr>
<th>Task description</th>
<th>Initial task query</th>
<th>Task type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find the official homepage of the Canberra casino and hotel in Canberra.</td>
<td>Canberra Casino</td>
<td>Nav</td>
</tr>
<tr>
<td>Go to the homepage of the Canberra Cavalry baseball team.</td>
<td>Canberra cavalry baseball</td>
<td>Nav</td>
</tr>
<tr>
<td>What is the standard length of a cue used for playing billiards?</td>
<td>billiard cue size</td>
<td>Info</td>
</tr>
<tr>
<td>How many spikes are in the crown of the Statue of Liberty?</td>
<td>statue of liberty crown spikes</td>
<td>Info</td>
</tr>
</tbody>
</table>

Note. Nav denotes navigational task and Info denotes informational task.

ations 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 2–2–2–1–1–2–3–2–4–4, 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.

With the scanpaths, Lorigo et al. [2006] investigated additional user behaviours such as 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) 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. In this thesis, these patterns are valuable measurements for analysing the sequence of movements in user attention.

2.2.2 Manipulated SERPs

Several studies have investigated the effects of rank order and the number of rank. Joachims et al. [2005] found that users are affected by rank order, after analysing the effects of three different manipulated rank orders (normal: the original ranking, swapped: the top two results were switched, and reversed: the ranking in reversed order).

A similar study was conducted by Guan and Cutrell [2007] to investigate the effects of target position: the rank position of the relevant result. They suggested that
the search speed and accuracy were much lower when the relevant links were located at lower ranks.

Wu et al. [2014] conducted a user study to investigate search behaviour based on Information Foraging Theory, introduced by Pirolli and Card [1999]. They showed different number and position of relevant results to see the effects of information scent level (ISL) and information scent pattern (ISP). They found that users tend to scan lower ranked document if a search engine provide more relevant results and they abandon their search when the top links in a SERP are not relevant.

Recently, Kelly and Azzopardi [2015] investigated the effects on search behaviour and user experience of the number of results on a SERP. By showing the participants in their experiments a different number of results (three, six, or ten), they examined the resulting different click patterns and found that participants shown three results viewed more SERPs than those shown ten results.

Some researchers studied the effects of the effects of different SERP elements (title, URL, snippets) on search behaviour by manipulating the contents. Paek et al. [2004] 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 [2007] focused on the effects of snippet length. They examined search behaviours with three different snippet lengths on a desktop screen (short, medium, and long, see Figure 2.4). Their findings indicated that users tend to spend less search time with the rich 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. [2008] 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 richer snippets may be more useful if the snippets are relevant for the query.
Several studies have classified user search behaviour according to gaze patterns. Klöckner et al. [2004] found that more than half of participants used a depth-first strategy: the subjects scanned only the links above the selected link, while the remaining exhibited a breadth-first: looking through all the links before making a decision, or mixed strategy: looking ahead a few results past the selected link (about 10% and 20%–30%, respectively). Figures 2.5 and 2.6 exhibited the examples of each strategy with small red dots which indicates the selected links.

<table>
<thead>
<tr>
<th>Short</th>
<th>Medium</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>The oldest zoo in the Southwest and one of the top in the nation, the Oklahoma ...</td>
<td>The oldest zoo in the Southwest and one of the top in the nation, the Oklahoma City Zoo's 110 acres are home to more than 2,800 of the world's most exotic animals.</td>
<td>The oldest zoo in the Southwest and one of the top in the nation, the Oklahoma City Zoo's 110 acres are home to more than 2,800 of the world's most exotic animals. &quot;The Cat Forest/Lion Overlook was completed in 1997. New in 1993 was the Great EscApe, a simulated tropical forest with gorillas, orangutans and chimpanzees. Also found at the zoo are the Noble Aquatic Center: Aquaticus, a Children's Zoo and Discovery Area, Herpetarium, Island Life Exhibit, Dan Moran Aviary and the Safari Tram. Open 9-5 (Oct-March), 9-6 (April-Sept). Rides additional (weather permitting and seasonal). 2101 N.E. 50th Street Oklahoma City, OK (405) 424-3344 (OKC Zoo Phone Directory)</td>
</tr>
</tbody>
</table>
§2.2  Reading and scanning SERPs

Figure 2.5: Examples of the depth-first strategy (top): a user follow a promising link immediately, and mixed strategy (bottom): a user reads ahead, but to a small extent.
Aula et al. [2005] defined two kinds of search strategy: economic and exhaustive. They divided the patterns on the basis of whether a user scanned less than or more than half of the visible results before making a selection (see Figure 2.7). Their findings suggested that the 54% of subjects were “economic” evaluators, and that the others had an “exhaustive” evaluation style.

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 proportionally more time on the top three links than the
exhaustive users did. In addition, the exhaustive group showed a slower scanning pattern of reading links.

### 2.3 Mobile web search

In the above section, we surveyed numerous previous works to investigate how users read the SERPs, under standard and manipulated conditions to find better SERP presentation designs for desktop screens. Although the user interactions are worth understanding before we start this research, we need to concentrate more on the following few studies that focused on search behaviour on small screens.

Several studies adopted eye-trackers to analyse user interaction on small screens. [Drewes et al., 2007] investigated gaze interaction for controlling applications on a handheld device using dwell time and gaze gestures and [Biedert et al., 2012] investigated text interaction and reading on a mobile phone screen, although their research was not about web search tasks.

[Lagun et al., 2014] studied the effect of relevance in Knowledge Graph (KG, see Figure 2.8) 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 of interest, we focus on the organic results (i.e. titles, snippets, URLs), because SERP with the additional result types is 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).

![Example of the search results page showing Knowledge Graph result](image)

Figure 2.8: An example of the search results page showing Knowledge Graph result from Lagun et al. [2014].

A few studies evaluated the search performance and behaviour between mobile devices and desktop monitors. Jones et al. [2003] 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 lower search speed and accuracy.
In our previous work [Kim et al., 2015], we studied the differences of user performance and behaviour for web search tasks based on large and small screens (for a desktop and mobile device, see Figures 2.9, respectively). Although we adopted an emulator with a mouse for the mobile-sized screen, this study was able to compare user interaction according to screen size. We 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 (the gap between the clicked link and the farthest link looked at) were exhibited on the smaller screen.

Figure 2.9: Search results as shown on the large and small 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 (reading behaviour) are significantly correlated to the most predictive signals of document relevance.

Recently, Ong et al. [2017] investigated search behaviour on a desktop and mobile device based on Information Foraging Theory (IFT) [Pirolli and Card, 1999] with a similar experiment design to Wu et al. [2014], as mentioned in Section 2.2.2. Their findings indicate that participants with a mobile device exhibited higher search ac-
accuracy when more relevant results were presented, and they were less accurate when the relevant results were distributed across the SERPs.

Some studies focused on search behaviour with small screens. [Raptis et al., 2013] 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 than those with other screen sizes, although no effect was found related to the perceived usability and task completion rate.

The above previous works 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.

2.4 Summary

In this chapter, we explored the background knowledge for conducting the research of user interaction in mobile web search. We surveyed several methods for evaluating SERPs from the system measures (e.g., search time and accuracy) and HCI cognition (e.g., eye tracking, EEG, and GSR) points of view. We then addressed general search behaviours in desktop screens with both standard and manipulated conditions. Most importantly, we examined a few user interaction on small screens at the end. From examining the related works, we can see that eye-tracking is a very useful technology to analyse user behaviour in web searches. The previous works on general search behaviours and search strategy suggests what we need to consider while conducting the experiment, and SERPs for mobile devices need to be different from the SEPRs for desktops. However, we also noticed that current research regarding mobile web search could not recommend the best presentation design of mobile SERPs. Therefore, we expect that this study regarding understanding search behaviour on mobile devices would contribute to finding a better SERPs design and supplementing the limited current knowledge of mobile web search. Based on this background and related work, we explain what user interaction we measured, and how we extracted the data, in Chapter 3.
measurements

In this chapter, we explain what methods we consider for extracting results from raw data, how the measures operate, and the reasons of adopting these particular measures for the analysis. We took several measurements to investigate search performance and behaviour, and user satisfaction. This chapter is comprised of three sections: search performance (Section 3.1), search behaviour (Section 3.2), and questionnaire measures (Section 3.3).

### 3.1 Search performance

We measured user search performance by search time and accuracy. The search time was measured as four kinds of time spent according to several stages from the start to the end of the tasks shown in Figure 3.1. The search accuracy was calculated according to whether a participant’s answer was right or not.

#### 3.1.1 Search time

Participants’ time on each task could be divided into the four stages shown (see Figure 3.1). If a user clicked (tapped) the relevant link at the first attempt, the time spent on the initial SERP (stage 1) was the first stage, and the time spent on the correct web document (stage 4) was the last stage needed to reach the correct answer. In addition, we investigated instances in which users made an incorrect choice on the SERP. Therefore, we considered the time spent on the wrong web document(s) (stage 2), and the time spent re-reading SERP(s) (stage 3) while searching for a different choice.

Using these different stages, we could measure four search times: Time to first click denotes the time taken to make the first decision, Time on SERPs combines stages 1 and 3 to give the total time spent on SERPs, Time to correct click combines stages 1–3 to investigate the time taken to make the correct decision, and Task completion duration is the total time required to complete a task.
In this thesis, we considered the elapsed time to the first click on a SERP as the primary search speed, because our participants were presented with the same SERP content whereas the web documents presented after a click could vary, and the time spent on the linked web pages varied considerably according to the design for either a full web site or a mobile-friendly version. The effect of different resolutions among web pages is beyond the search engine’s control. We calculated the other indicators of search time to be supplementary search times.

### 3.1.2 Search accuracy

Participants were given only one chance to provide 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, the search accuracy was assigned a score of ‘0’.

We could define the search accuracy as whether a participant selected the ‘best’ answer like Cutrell and Guan [2007] measured. However we had a difficulty in deciding the ‘best’ answer, because each task had at least two or more relevant links on the SERPs. In addition, most of the tasks were designed for one target answer, except for some questions which may have a changeable answer, such as weather.

The details of search accuracy definitions for each study were varied according to tasks and experimental conditions, and they are addressed in each study in more details.
3.2 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.

3.2.1 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 and Carpenter, 1976; Rayner, 1998]. The tasks in the experiments had the same components (10 ranks including titles, snippets, and URLs as well as the periphery such as the query box and Google logo), so we initially assigned 10 areas of interest (AOIs) to each SERP to investigate user attention. For some study, we additionally divided each rank into three smaller AOIs, in order to measure the fixation duration on each component (the title, snippet, and URL).

We adopted two different eye-trackers for the experiments and the methods (algorithms) of recording fixation were varied according to types eye-tracker. The details of gaze recording are explained in each study.

3.2.2 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 to determine how much participants were biased toward the rank order by the search engine [Guan and Cutrell, 2007; Joachims et al., 2005]. In addition, the click point is also used to determine the trackback value, as explained in the subsection on Trackback Section 3.2.7.

3.2.3 Scanpaths

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 as shown in Figure 3.2, 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 length of minimal scanpath can be interpreted as how many different links a user looked at, removing repeated visits from the compressed sequence.

![Figure 3.2: A example of a scanpath on the search results page.](image)

### 3.2.4 Scanning direction

Given the above two definitions, we can analyse five types of scan patterns. In our previous work [Kim et al., 2015], we refined the measurement methods from previous studies [Dumais et al., 2010; Lorigo et al., 2006] 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.

For example, when the raw scanpath is 1-1-2-3-2-5, the compressed sequence is 1-2-3-2-5 (not monotonically increasing), and the minimal scanpath is 1-2-3-5 (increases monotonically). We can consider this scanpath as a linear pattern although it includes a skip (3 to 5) and a regression (3 to 2), whereas it cannot be a strictly linear pattern.

The previous study did not explain the cases when users selected a link after looking at only the link in the definition of linear and strictly linear patterns. In our previous work [Kim et al., 2015], we considered that immediate decision with only one link as linear and strictly linear patterns, because our participants often made
such a decision during the experiment. Therefore, as defined in our previous work, 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. As noted in our previous study, any pattern of linear or/ID is also considered as strictly linear or/ID (and vice versa), and the length of compressed sequence will have length 1. For example, if a user only scans rank 3, and then selects the link, the search behaviour is considered neither linear nor strictly linear, but is measured as both linear/ID and strictly linear/ID.

### 3.2.5 Skip and regression

We investigated skip and regression patterns using the compressed sequence and minimal scanpaths across the screens. As defined in previous studies [Lorigo et al., 2006; Kim et al., 2015], 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.

### 3.2.6 Scroll

Due to the small screen sizes, the initial viewport displayed visible ranks from two and half to six results, depending on the tasks, and users needed to use the scroll function to see the lower links over the page fold. We measured the proportion of scrolling, the scrolled distance, and when scrolling took place, in order to investigate its usability, and compared the results to other search behaviours. Like the fixation data, exploring scrolling behaviour can be one measurement that is used to investigate users’ attention on SERPs [Lagun et al., 2014].

### 3.2.7 Trackback

We used trackback, introduced in our earlier work [Kim et al., 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 clicked links in previous studies [Granka et al., 2004; Joachims et al., 2005]. Participants in our studies displayed the same pattern. We measure trackback as the distance between the selected link and farthest link visited. For example, if a subject looked as far as rank 8 and then clicked the rank 2, we recorded a trackback value of 6. This means that the participant made some additional effort (e.g., search
time and scrolling) to scan the links between rank 3 and 8. In the rare cases that a user does not exhibit top-to-bottom scanning pattern, we consider this to be zero trackback value.

### 3.3 Questionnaire measures

Questionnaires are a useful supplement to determine a participant’s personal information and experience regarding the experiments. We asked participants to fill in a post-task questionnaire after completing each task and a post-experiment questionnaire when they ended all tasks.

To measure the user satisfaction, we adopted a 7-point Likert scale that ranged from ‘Extremely disagree’ to ‘Extremely agree’ (valued as 1 and 7, respectively) for the post-task questionnaire.

The post-experiment 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. For some studies, participants were also asked required to score their overall preference with the reasons.

### 3.4 Summary

In this chapter, we explored the measurements we adopted in the analysis with the reason for the adoptions. We broadly explained the measurements in terms of search performance, behaviour and user satisfaction, because each study was conducted under various experimental conditions with different research variables. The further explanation of each measurement is addressed in each study in details.
Chapter 4

Study One: Three Different Small Screens

This chapter describes the first study regarding search performance and behaviour on three different small screens, and suggests possible better presentation designs for each screen size.

4.1 Introduction

One notable 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 (e.g. Samsung Galaxy S1 and Apple iPhone 3 or 4). In contrast, recent smart phones are equipped with a screen of 4.5 inches (e.g. Apple iPhone 6) and have a higher resolution. Another recent device is a phablet (a portmanteau word combining 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 the needs of 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 it came with some disadvantages such as heavier weight and lower mobility. We cannot say whether or not this need was 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. We need to investigate the effects of the change to determine if there is any difference in user interaction. When improving the search engine interface design for mobile devices, in contrast to conventional monitors, we need to consider different designs for each
mobile device that has a different screen size. This investigation could enhance search engine result pages (SERPs) on small devices for better search experience. In this chapter, 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 4.1) using eye-tracking technology. We adopt search time, search accuracy, and user satisfaction as the user performance metrics, similarly to previous studies [Kim et al., 2015; Lagun et al., 2014] and employ implicit data such as fixation and scanning patterns on SERPs to understand user behaviour.

After first describing the user study and data collection including the experimental procedure. We then present our results and a discussion about the findings 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.

4.2 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 method and post processing.

4.2.1 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 rated themselves expert or good at finding information using web search engines and did so frequently, and most (16) of them had experience using mobile devices for web searches (see Appendix C for the question).

4.2.2 Tasks

Each participant performed a total of nine search tasks (see the descriptions and queries in Table 4.1; see Appendix B.1 for further details) for given initial (controlled) 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
Figure 4.1: Examples of search engine result 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.
the relevant links were located in the top links. Therefore, it is not interesting because the ranks which included relevant answers are equally prominent on all three screen sizes. 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 4.1: 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 4.1: Examples of task descriptions and queries.

<table>
<thead>
<tr>
<th>Task description</th>
<th>Initial task query</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPhone6 is recently out. In what memory sizes can you get it? (3 kinds)</td>
<td>iPhone 6 specs</td>
</tr>
<tr>
<td>Which two countries will play for the first match in the cricket world cup 2015?</td>
<td>cricket world cup 2015</td>
</tr>
<tr>
<td>When does daylight-saving time end in Australia? (any applied states such as NSW, ACT, or VIC)?</td>
<td>2015 daylight savings</td>
</tr>
<tr>
<td>How many seats are there in the Australian parliament for MPs (elected by the Australian people)?</td>
<td>Australian parliamentary seats</td>
</tr>
</tbody>
</table>

### 4.2.3 Design and procedure

A total of nine tasks were shown to each participant, three on each of the three screen sizes. 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 (see Appendix A). 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, which is about 40 pixels. 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 right 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 (see Appendix C). A time notice was verbally 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 run time for the experiment was less than 20 minutes to complete everything from the instruction to post-questionnaire. Participants were not paid.

4.2.4 Apparatus

All SERPs were obtained from the Google mobile search engine and shown using Internet Explorer 11. We conducted the experiment on a MS Surface Pro 3 as shown in Figure 4.2 [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 as shown in Figure 4.3 [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, which was emulated by window size, were $378 \times 672$ (3.57 inches), $495 \times 880$ (4.67 inches), and $585 \times 1040$ (5.52 inches) for small-, medium-, and large-size screens, respectively. As noted 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. These are discussed in detail in the Conclusions and future work section.
4.2.5 Data collection and post processing

We collected data from three sources, i.e., gaze, window events, and questionnaire data. 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. Another 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 doing as well as the usability score for each screen with system times. We embedded custom Javascript 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 employed a commonly used algorithm; dispersion-threshold identification (I-DT, see Salvucci and 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. We also collected user satisfaction and users’ general information using post-experiment questionnaire.

The participants were asked several questions (see Appendix C). 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.

4.3 Results and discussion

We first analysed the data-set of 162 tasks (54 tasks for each screen) using the measures introduced in the previous section. We also investigated relationships between search time 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) [Cochran and Cox, 1957] with block structure (participants) for search time 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 and 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 and Clayton, 1993] for binary data and count data with subjects as a random term, \(\sigma_s^2\). GLMM is an extension to GLM in which the linear predictor has the fixed effects as well as random effects. Thus, if the random term is higher than the standard error, we used GLM rather than GLMM. In these analyses, we used the GenStat version 17 statistical package [VSN International, 2014].

In addition, when we found a significant effect due to screen size, we ran a post-hoc test to confirm the difference among the three different screen sizes using a standard error of the difference (SED), and we noted labels; e.g., ‘A’ and ‘B’: B is significantly different (e.g., bigger, higher, or longer) from ‘A’, but neither ‘A’ nor ‘B’ is significantly different from ‘AB’.

4.3.1 Power analysis

We carried out post-hoc analyses with the significance level \(\alpha = 0.05\) to confirm the power of our design [Chow et al., 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 analyses below.

4.3.2 Search performance

As noted in the Measurement chapter [3], we analysed two kinds of explicit data for search performance: search time, search accuracy. There was no significant screen size effect on search time and search accuracy.
§4.3  Results and discussion

Search time  We analysed two kinds of search time: 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 4.2, there is no significant screen size effect on the two kinds of search time, 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 is believed that task completion duration decreases as screen size is reduced [Jones et al., 2003; Raptis et al., 2013; Kim et al., 2015]. These results, however, indicate that participants exhibited similar search time on SERPs as well as on completing tasks. This result might be caused by locations of the relevant links and task difficulty. If the relevant links were located beyond the page fold, and/or tasks were more difficult, our result might be similar results to the previous studies.

Search accuracy  For the accuracy data, we used a GLM with a binomial distribution. Similar to a result from our previous study [Kim et al., 2015], 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 4.2, 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 and 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.

4.3.3 Search behaviour

We can observe 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 time results.

Fixation duration  We applied ANOVA with using a log-transformation to mean fixation duration on SERPs. Although Table 4.2 shows that the effect is near the significant level, there is no significant difference with respect to screen size. As fixation duration represent users’ effort in search [Just and Carpenter, 1976; Rayner 1998; Dumais et al., 2010], this means that subjects expended similar effort to obtain
<table>
<thead>
<tr>
<th></th>
<th>Mean values Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value L M S</td>
</tr>
<tr>
<td>Search performance</td>
<td></td>
</tr>
<tr>
<td>Search time Time to first click [s]</td>
<td>7.70 10.47 9.12</td>
</tr>
<tr>
<td>Task completion duration [s]</td>
<td>20.89 24.79 23.08</td>
</tr>
<tr>
<td>Search accuracy Correct answer rate [%]</td>
<td>94.44 98.15 94.44</td>
</tr>
<tr>
<td>Search behaviour</td>
<td></td>
</tr>
<tr>
<td>Fixation duration on SERP Per task [s]</td>
<td>3.97 5.60 5.53</td>
</tr>
<tr>
<td>Fixation duration on SERP Per link [s]</td>
<td>2.16 1.89 2.49</td>
</tr>
<tr>
<td>Clicks Ranks 1.39 1.52 1.46</td>
<td>0.697</td>
</tr>
<tr>
<td>Scanpath Minimal scanpath</td>
<td>2.06 2.76 2.26</td>
</tr>
<tr>
<td>Scanpath Compressed sequence</td>
<td>3.33 5.50 4.35</td>
</tr>
<tr>
<td>Scanpath Compressed minus minimal</td>
<td>3.33 5.50 4.35</td>
</tr>
<tr>
<td>Scanpath Clicks</td>
<td>0.09 0.19 0.12</td>
</tr>
<tr>
<td>Scanpath Trackback 1.07 1.91 1.28</td>
<td>0.087</td>
</tr>
<tr>
<td>Search satisfaction</td>
<td>5.24 4.91 4.20</td>
</tr>
</tbody>
</table>

Note: Labels A and B indicate the type of result, A type is significantly different from B, but not different to AB. Labels A and B denote search engine result page large, medium and small, respectively.

Note: SERP denotes search engine result page and L, M and S denote large, medium and small, respectively.

\[ 100 \% > d \] 
\[ 0.05 > d \] 
\[ 0.001 > d \]
enough information to make a selection on the small screens, unlike the difference between a monitor and small screen in our earlier work [Kim et al., 2015]. 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.

Figure 4.4: Mean fixation duration on each AOI (second/user/task), the standard error of the difference (SED) = 0.2.

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 4.4 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 and 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 on ranks 1–3). 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. 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 (the standard error of the difference (SED) = 0.2). The above findings may be connected to scrolling, therefore this is discussed in detail in the subsection below on scroll.

**Click pattern** Using a GLMM with a Poisson distribution for the analysis, we found that a participant’s click patterns were mostly distributed across the top three links with no significant differences. The chance of clicking one of 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., 2015].

**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 4.2, we determined that there are significant differences for the minimal scanpath length ($\sigma^2_s = 0.055$, $\chi^2 = 12.21$, df = 2, $p < 0.01$) and compressed sequence length ($\sigma^2_s = 0.134$, $\chi^2 = 11.39$, 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 by calculating that compressed sequence length minus minimal scanpath length. 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 ($\sigma^2_s = 0.288$, $\chi^2 = 9.23$, 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, meaning that participants had some hesitation before making a click decision.

As mentioned in the subsection on fixation duration, we investigated how much effort participants made 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 time (time to first click) and the fixation duration per link. Figure 4.5 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 4.5: Relationship between search time (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), which is the same as the approach adopted in previous studies to investigate the scanning direction [Dumais et al. 2010; Lorigo et al. 2006]. We adopted a GLMM with a binomial distribution for the scanning direction analysis. Table 4.2 indicates that there is a significantly different effect on proportion of people using the complete pattern ($\sigma^2 = 4.585, \chi^2 = 11.86, df = 2, p < 0.01$). 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, as addressed in the next subsection. We next looked into the linear and strictly linear patterns. There is a significantly different effect on the linear pattern ($\sigma^2 = 0.629, \chi^2 = 8.29, df = 2, p < 0.05$), 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: $\sigma^2 = 1.059, \chi^2 = 13.62, df = 2, p < 0.01$; strictly linear/ID: $\sigma^2 = 1.732, \chi^2 = 7.61, df = 2, p < 0.05$). 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 using the large screen. The participants using 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, difference in ID pattern between the large and medium screens produces significant effects on the linear/ID and strictly linear/ID, although there is not such a different effect on linear and strictly linear. We believe that this might be caused by users being more confident with the large screen, though they are biased in the rank order.
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 ($\sigma^2_s = 1.732$, $\chi^2 = 7.61$, df = 2, $p < 0.05$) by a GLMM with a binomial distribution. On the other hand, we determined that participants exhibited a higher regression rate on the medium screen than on the large one (about 20% higher). 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 complete rate, although it was only near the significance 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 in scroll rates can be observed ($\sigma^2_s = 1.969$, $\chi^2 = 15.90$, 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). 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 time 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 4.3, there are significant effects of scrolling on search time (elapsed time to first click) for the medium and small screens ($\sigma^2_s = 0.061$, $\chi^2 = 12.67$, df = 1, $p < 0.001$, and $\sigma^2_s = 0.019$, $\chi^2 = 27.91$, 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 by around 9 seconds.

Trackback  We measured the trackback value to investigate how much additional effort user make before making a click decision. A significantly different effect on trackback was observed using a GLMM with a Poisson distribution ($\sigma^2_s = 0.191$, $\chi^2 = 16.24$, df = 2, $p < 0.001$). Participants using the medium screen recorded a higher trackback value, which is related to scanpath. As we determined from previous behaviours, users 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
### Table 4.3: Effects of scrolling on search time (elapsed time to first click) (s)

<table>
<thead>
<tr>
<th>Screen size</th>
<th>Non-scrolled</th>
<th>Scrolled</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>7.73</td>
<td>6.95</td>
<td>0.530</td>
</tr>
<tr>
<td>Medium</td>
<td>8.66</td>
<td>17.54</td>
<td>*</td>
</tr>
<tr>
<td>Small</td>
<td>6.01</td>
<td>14.84</td>
<td>***</td>
</tr>
</tbody>
</table>

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

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. This is an important result, suggesting that we need to consider a better SERP presentation design to reduce the extra effort.

#### 4.3.4 Questionnaire measures

At the end of the experiment, participants were asked several questions (see Appendix C). 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 ($\sigma_s^2 = 0.051, \chi^2 = 28.07, df = 2, p < 0.001$). By comparing mean values among the screens with the standard error of the difference (SED), a difference is shown for the small screen with the lower score. 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.

#### 4.3.5 General Discussion

In this section, we summarize the implications for the screen sizes by reviewing 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 (i.e., fixation duration) 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 each link; however, they visited more links, with frequent revisits, with higher trackback values, and with higher regression rates than on the large screen. With the above results, it seems that they exhibited 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 be hard to read.

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, we can say that people using small screens—such as those of early smartphones—tend to hesitate before choosing a link to follow. They spend a long time reading each link, and use the scroll function a lot.

**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, though 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
4.4 Conclusions and future work

In this chapter, we observed search performance and behaviour across three different sizes of screen. Our findings indicate that search behaviours and usability vary across screens, although participants exhibited a similar search time and accuracy. Glancing through the results reported in this chapter, one might conclude that there is no need to change anything in SERPs because there is no significant difference for search time and accuracy. However, if we could create better design for SERPs for each type of screen by understanding search behaviour, this 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 and Guan 2007]. However, we need to consider the effects 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. 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 displays 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, our users reported a low usability for the small screen and had a slow reading performance for each link. The best way to improve satisfaction, and reduce scrolling, would be to widen the screen. 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 et al. 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.

Third, 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 and 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. We focused on this in study three, see Chapter 6.

In addition, for general mobile devices, we recommend 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 in our previous work [Kim et al., 2015], and as initially suggested by Jones et al. [2003]. 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 all sizes of small screens.

We believe that the above recommendations from this study may contribute to better SERP presentation designs for each mobile screen size.
Study Two: Pagination versus scrolling

One possible idea for the SERP design for small screens in the previous chapter (Study One) indicates that ‘embedding page up and down button on the interface or horizontal page changes instead of a vertical scroll function’ would provide better user experience in mobile web search. Horizontal pagination is used as one viewport control type on several mobile applications. However, mobile search engines still provide only vertical scrolling to show the contents beyond the page fold. In this chapter, we address the second study about the effect of horizontal pagination as one of the future work from the study one.

5.1 Introduction

Although several studies have indicated that the use of a scroll function is closely related to search performance and behaviour (e.g., increasing search time and some special interests on links around a page fold) [Cutrell and Guan, 2007; Granka et al., 2004; Joachims et al., 2005; Kelly and Azzopardi, 2015], people must still scroll through SERPs to see beyond the page folds.

Compared to scrolling on desktop monitors, screens for smartphones need to be scrolled more because they have less visible space, and the effect of scrolling may be more important because users need to cover part of the screen with their fingers while scrolling.

To initiate a scroll event on a touch screen, users drag their fingers vertically to produce a similar effect to spinning a mouse wheel, holding and dragging the scroll bar, or using the page up/down keys on desktops. To the best of our knowledge, vertical scrolling is the only option provided by all search engines to control SERPs on touch-enabled mobile devices. However, people are familiar with turning pages
Study Two: Pagination versus scrolling

Figure 5.1: Example of two subsequent pages of a SERP with the horizontal control type.

horizontally when reading a book, either on paper or via an eReader, with discarding unwanted applications or web pages on smartphones [Warr and Chi, 2013], with the start screens for smartphones (pages of icons or tiles), and with controlling applications which provide a switching interface using horizontal swiping gestures (e.g., weather, stock price and shopping). Therefore, horizontal swiping for mobile web search could also have been considered by interface designers, and we wondered what would happen, in terms of search performance and behaviour, if users could swipe horizontally (paginate), as shown in Figure 5.1. Although most SERPs have buttons for moving to the next page, this function is somewhat different from a horizontal swiping interface. With new HTML and CSS standards, it is now relatively simple to load multiple SERPs, and the pre-loaded pages can be hidden before being called up. This means that there is no additional loading time to display the same
number of search results as with a vertical scroll function.

In this study, we investigate the effect of two control types (horizontal pagination and vertical scrolling) for mobile web search, with relevant results at different positions, as shown in Figure 5.1. We form a series of tasks that have only one target result among ten items to examine the effect of the control type.

We formulated the following four hypotheses prior to conducting the experiment. First, we imagined that users will be affected by the position of relevant links, similar to the results in previous studies [Cutrell and Guan, 2007; Joachims et al., 2005]:

- H1. When the relevant result is located after the page fold rather than before, users take longer to make their first decision and complete the task.

Second, although people are familiar with turning the pages of a book horizontally, and swiping horizontally between screens, for example, weather apps, they are used to scrolling with a vertical control type on touch-enabled devices, as this is the only mechanism provided by search engines.

- H2. Users need less search time with the vertical control type than horizontal pagination.

Third, under horizontal swiping, users bring a set of new links over the page fold in one swipe without scrolling. This might lead to users spending more time on each SERP, and paying attention to more results, which could allow them to carefully consider the most relevant link.

- H3. Users exhibit higher search accuracy with the horizontal control type.

Fourth, similar to our reasoning for H2, we expect that the first attempt of using horizontal pagination in our experiment would not change user preferences.

- H4. Users are more satisfied with the vertical control type.

In the following sections, we describe the design and procedure of our user study. We then present our findings regarding the effects of both control types and the position of relevant results. We discuss the observed effects, and describe some limitations of our study. Finally, we conclude and suggest ideas for future work.

5.2 User study

In this section, we describe the experimental design and procedure, and introduce the participants, tasks, and apparatus.
5.2.1 Participants

We recruited 27 subjects from both inside and outside of the university campus. The data from three participants were excluded because of low eye gaze calibration accuracy, leaving us 24 participants (14 male) aged 22–41 years (mean: 28.3, standard deviation: 5.0). Two-thirds were studying computer science, and the remaining third had varied backgrounds: accounting, business, law, biochemistry, mathematics, and international studies. Most subjects (22) classified themselves as heavy users of web search engines, usually submitting multiple queries every day. These subjects considered themselves good at using search engines, although one participant thought that he/she was neither good nor bad at searching. Over half of the subjects (17 of 24) regarded themselves as good or expert at using mobile devices; two subjects were not familiar with using mobile phones, and the remaining five claimed to be neither good nor bad at using mobile devices. The subjects participated voluntarily in the experiment, and were compensated for attending with a meal voucher.

5.2.2 Tasks

Each participant completed twelve tasks, which were based on the tasks used in a study by Dumais et al. [2010], and modified for the local participants. Samples of the task descriptions and initial queries are listed in Table 5.1 (see Appendix B for further details). We varied the task category, including categories such as tourism, sports, science, weather, and transport. All initial SERPs were extracted from the Google mobile search engine. Each task had 10 search results (five results before the page fold), as shown in Figure 5.1 and we ensured that each task showed the same elements (title, snippet, and URL) by excluding advertisements, related links, and stars for popularity as in previous studies [Dumais et al., 2010; Kelly and Azzopardi, 2015; Kim et al., 2015]. All of the tasks were informational in that they required a particular piece of information to be found [Lorigo et al., 2006]. To control some variability, we excluded navigational tasks, which are much easier to complete [Guan and Cutrell, 2007; Lorigo et al., 2006], and testing other task types remains as a future study. The tasks were very simple, but were made more difficult by changing the target position. Inspired by Guan and Cutrell [2007], we ensured that each task had six variant SERPs, each with a single target result (by checking if other links have relevant information) at a different position: ranks 1, 3, and 5 before the fold, and ranks 6, 8, and 10 after (see Figure 5.1). The target result was not marked, and it looked the same as the other nine links. This let us investigate the effect of target position.
Table 5.1: Examples of task descriptions and queries.

<table>
<thead>
<tr>
<th>Task description</th>
<th>Initial task query</th>
</tr>
</thead>
<tbody>
<tr>
<td>You are interested in membership of Questacon.</td>
<td>Questacon membership benefits</td>
</tr>
<tr>
<td>What are the benefits?</td>
<td>Cricket world cup 1996</td>
</tr>
<tr>
<td>Which two countries played the third match in the 1996</td>
<td>Cricket world cup 1996</td>
</tr>
<tr>
<td>Cricket world cup?</td>
<td>Cricket world cup 1996</td>
</tr>
<tr>
<td>How many demerit points can you collect before your</td>
<td>demerit point ACT suspend</td>
</tr>
<tr>
<td>driving license is suspended in ACT?</td>
<td>Jervis bay weather 14 days</td>
</tr>
<tr>
<td>You plan to visit Jervis Bay in a couple of weeks.</td>
<td>Jervis bay weather 14 days</td>
</tr>
<tr>
<td>Check the weather there 14 days from now.</td>
<td>Jervis bay weather 14 days</td>
</tr>
</tbody>
</table>

5.2.3 Design and procedure

We used a within-subject design (two control types × six target positions). Participants were assigned a total of 12 tasks across two task sets (tasks 1–6: from task set 1; tasks 7–12: from task set 2), i.e., six tasks from each task set had a different control type. Under this condition, each task in a task set was shown with the target at a different position, so only two of twelve tasks had the same target position. To minimize the carry-over effect in this design, we randomized the task order in each task set. Furthermore, the orders of task set and control type were counter-balanced, and every task was presented with all six target positions across the participants.

Before starting the experiment, we attempted to relax the participants by emphasizing that they were not being tested. They were then given a consent form explaining the privacy of individual data and their rights, such as terminating their participation at any time or withdrawing their data. After signing the form, subjects were instructed about the number of tasks, the control types, and the experimental procedure (see Appendix A). To familiarize themselves with both methods of controlling the viewport, participants conducted two sample tasks with each control type, although they were already very familiar with the vertical control type. We then calibrated their gaze recordings using a 9-point procedure provided by the eye-tracker software, and presented the task list page. When the participants clicked the first task on the list, a description and initial query were shown. Once they clearly understood the descriptions and the query, they pressed the start button, and the initial SERPs were presented. When the participants announced their own answer regardless of the correct answer, or gave up on a task (did not happen during the experiment), the task was considered complete. There was no time limit on reaching the correct answer. After each task, they were asked to score the difficulty of the task and usability of the control type. This cycle was repeated until all twelve tasks had been
completed. The task/task set order was controlled by Javascript in the cached HTML files. After completing all tasks, participants were asked to fill out a post-experiment questionnaire (see Appendix C) that asked about their age, preferred scrolling type, and familiarity with search engines and mobile devices. The run-time from sitting on the chair to leaving the laboratory was about 35–40 min.

5.2.4 Apparatus

All tasks were displayed with Internet Explorer 9 on an iPhone 6 Plus (5.5 inches, which is one of the most popular screen sizes [Mobile Marketing, 2015]) connected as a secondary monitor by a USB cable using the Twomon software [Dev guru], as shown in Figure 5.2. We obtained eye gaze data using Facelab 5 [Seeingmachines], which records gaze data at 60 Hz, and used the Eyeworks software [Eyetracking] for our analysis. With the mobile device, the SERPs were displayed at full resolution (1920 \times 1080), and the zoom level was adjusted in Internet Explorer to display five links on the initial screen, similar to the Google mobile search engine with most browsers.

We implemented the pagination function using the smoothscroll jQuery function [Swedberg, 2014]. The touchable screen recognized a horizontal swipe if the distance
between the initial point pushed by the finger and the point at which the finger left
the screen was over 100 pixels along the x-axis (about 0.63 cm).

5.3 Results

We analyzed the gaze data from 288 tasks: 144 tasks for each control type, 48 tasks
for each target position. As mentioned earlier, our focus is on the effect of the control
type and the target position. First, we divided each SERP into the front and back
pages (before and after the page fold) given by the horizontal swiping interface (see
Figure 5.1), and analyzed the target position effect by aggregating target positions
1, 3, 5 as target-front and 6, 8, 10 as target-back. For further analysis, we investigated
the effect of each target position with both control types. To test the hypotheses,
we measured search time, search accuracy, and user satisfaction. We then examined
other data such as time consumed for scroll actions and user attention in the area of
interest (AOI).

For data analysis, we adopted analysis of variance (ANOVA), generalized linear
(mixed) models (GLMs and GLMMs) as similar to the previous study: Chapter 4. For
the score data from the 7-point Likert scale, we used a linear mixed model (LMM)
[West et al., 2014]. Analysis was conducted using the GenStat statistical package
[VSN International, 2014] and R [R Core Team, 2013].

5.3.1 Search performance and user satisfaction

We applied ANOVA to the search time measurements. First, only the target position
significantly affected time to first click ($F_{(1,261)} = 7.39, p < 0.01$) and task completion
duration ($F_{(1,261)} = 10.13, p < 0.01$). Table 5.2 indicates that subjects spent more time
finding targets at lower ranks, where it was necessary to perform a scroll (mean time
to first click: 29.46 s vs. 31.93 s, and task completion duration: 60.84 s vs. 68.12 s).

For the remaining two search time, we found significant effects due to target
position and control type on time to correct click ($F_{(1,261)} = 20.67, p < 0.001$, and
$F_{(1,261)} = 4.81, p < 0.05$, respectively) and time on SERPs ($F_{(1,261)} = 18.01, p < 0.001$,
and $F_{(1,261)} = 5.05, p < 0.05$, respectively). With vertical scrolling, participants needed
more time to identify the correct result (about 7.96 s longer) and spent longer on the
SERP (7.05 s). This may be connected to scrolling behaviour, so we deal with this
inference in detail in the next Subsection 5.3.2.

The search accuracy was assigned a value of 1 if a participant clicked the target
result and found the answer at the first attempt; otherwise, the search accuracy was
set to 0. Using a GLM with a binomial distribution, significant differences were
Table 5.2: Search performance and user satisfaction.

<table>
<thead>
<tr>
<th>Control type</th>
<th>Target position</th>
<th>Interaction</th>
<th>Search time</th>
<th>Task completion duration</th>
<th>Time to correct click</th>
<th>Time on SERPs</th>
<th>Search accuracy</th>
<th>User satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.91</td>
<td>0.06</td>
<td>25.83</td>
<td>58.00</td>
<td>36.09</td>
<td>31.13</td>
<td>79.20</td>
<td>5.60</td>
</tr>
<tr>
<td>V</td>
<td>0.83</td>
<td>0.08</td>
<td>33.10</td>
<td>63.69</td>
<td>40.95</td>
<td>37.40</td>
<td>81.94</td>
<td>5.74</td>
</tr>
</tbody>
</table>

Note: H and V denote horizontal and vertical control types, respectively. Target position means the location of a relevant result.

Table 5.3: Scroll actions and search time excluding the scroll duration.

<table>
<thead>
<tr>
<th>Control type</th>
<th>Target position</th>
<th>Interaction</th>
<th>Scroll Rate (%)</th>
<th>Finger actions</th>
<th>Search time</th>
<th>Task completion duration</th>
<th>Time to correct click</th>
<th>Time on SERPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.67</td>
<td>0.06</td>
<td>41.67</td>
<td>0.82</td>
<td>25.42</td>
<td>57.56</td>
<td>35.65</td>
<td>30.69</td>
</tr>
<tr>
<td>V</td>
<td>0.83</td>
<td>0.08</td>
<td>54.17</td>
<td>3.57</td>
<td>30.22</td>
<td>60.64</td>
<td>37.84</td>
<td>34.30</td>
</tr>
</tbody>
</table>

Note: Finger action and scroll duration measured until first click. Target position means the location of a relevant result.
observed in search accuracy according to target position ($\chi^2 = 5.99$, df = 1, $p < 0.05$) and the interaction of target position and control type ($\chi^2 = 3.83$, df = 1, $p < 0.05$). Table 5.2 displays a small difference (2.74%) in the mean values for both control types when the target was on the front of the page, but this difference was much larger when the target was on the back page (19.45%). With these results, we can say that horizontal pagination results in better search accuracy when the correct answer is located on the back page. Considering the results of search time and accuracy, it seems that pagination helps compensate when rankings are poor, but does not degrade performance regardless of rankings, even when it is better.

We employed an LMM to examine user satisfaction. User satisfaction exhibited a significant difference due to target position ($\sigma^2 = 0.530$, $\chi^2 = 4.69$, df = 1, $p < 0.05$). This means that users were more satisfied (about 0.21 points higher) when a relevant link was located on the front page. However, participants expressed no preference for either control type, despite being more familiar with vertical scrolling on current search engines.

Table 5.4: Search performance, user satisfaction, and scroll rate with each of the six target positions.

<table>
<thead>
<tr>
<th>p-value</th>
<th>Target Position</th>
<th>Control Type</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to first click **</td>
<td>0.172</td>
<td>0.610</td>
<td></td>
</tr>
<tr>
<td>Task completion duration **</td>
<td>0.056</td>
<td>0.432</td>
<td></td>
</tr>
<tr>
<td>Time to correct click ***</td>
<td>*</td>
<td>0.322</td>
<td></td>
</tr>
<tr>
<td>Time on SERPs ***</td>
<td>*</td>
<td>0.593</td>
<td></td>
</tr>
<tr>
<td>Search accuracy 0.105</td>
<td>0.199</td>
<td>0.071</td>
<td></td>
</tr>
<tr>
<td>User satisfaction *</td>
<td>0.599</td>
<td>0.342</td>
<td></td>
</tr>
<tr>
<td>Scroll rate ***</td>
<td>0.381</td>
<td>0.181</td>
<td></td>
</tr>
</tbody>
</table>

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

We also investigated the effect of each (of six) target position and both control types using the same statistical methods. Across all search performance results, user satisfaction and scroll rate, we found very similar results compared to those obtained when the target positions were divided into target-front and -back, as shown in Table 5.4.

However, when we looked at control type effects by each target position, there was an interesting result for scroll rates. Figure 5.3 displays the scroll rates for the six target positions. The scroll rates with both control types on each target position were not different except for one target position—when the correct answer was at rank 5, subjects recorded only about 46% scroll rate using the horizontal swiping,
whereas that with the vertical control type was as high as 81%. The line in Figure 5.3 for vertical scroll suddenly increases from target position 5 (from about 40% to 80%), whereas the other line for horizontal swiping starts to soar from target position 6. This suggests that participants performed a scroll event to either look ahead a few links beyond rank 5, or used scrolling to position rank 5 in the middle of the screen with the vertical control type. This is addressed further when we examine the fixation duration with target position 5 in Subsection 5.3.3 regarding user attention.

Across the results for search performance and user satisfaction, there were clear target position effects, and we found several task type effects and interactions of the variables. Consistent with H1, when the target was on the front of the page, users tended to require less time, show better accuracy, and exhibit higher satisfaction. However, participants were not faster or more satisfied with vertical scrolling, and in fact were faster to find the target and spent less time on SERPs when given the horizontal controls. Therefore, our results do not support H2 or H4. In addition, consistent with H3, users recorded better search accuracy with pagination when the target was on the back page.

### 5.3.2 Scroll action and duration

Most people are more familiar with vertical scrolling in mobile web search, which is generally the only type provided by search engines. The results for search performance and user satisfaction indicated that horizontal pagination was similar or
better for mobile web search. As a main research variable in this study, we inferred that the scrolling itself may cause the longer search time with worse user satisfaction. Therefore, we looked more closely at the relationship between scrolling effect and search time.

First, using a GLMM with a binomial distribution, only the target position effect was observed on scroll rate ($\sigma_s^2 = 0.494$, $\chi^2 = 36.07$, df = 1, $p < 0.001$), and we found no control type effect (Table 5.3). We can explain the target position effect as follows. When the correct link was on the front page, neither control type required the subject to scroll to reach the relevant answer: overall scroll rate on the front page was less than 50%. However, the scroll rate does not seem to explain the difference in speed between the control types. Therefore, we decided to investigate how much effort the user made to use the scroll function.

In our experiment, minimal effort was required (only one flip) to see a whole SERP with pagination. Even if users wanted to see a few links on the back page, once swiped, they could see all of the results 6–10. Besides the time taken to change the cached pages (maximum 0.5 s), users could read the SERP immediately after swiping, whereas it may have been difficult to read while scrolling vertically. Thus, it is interesting to examine how vertical scrolling affected the search. Unfortunately, we were unable to find a previous study regarding scroll duration.

We needed to define the continuous scroll action from the scroll events to calculate the duration. First, the duration of a continuous scroll action was measured by considering whether the scroll event happened continuously over 100 ms. Second, we considered users to have stopped scrolling if no scroll event has been recorded for 100 ms, because, by definition, they could have made a fixation during this time (see Subsection 5.3.3). Finally, we neglected scroll distances of less than 30 pixels (the height of one snippet line) over 1 s to exclude slow scrolling in which the SERPs were being read and slight movements of the finger while holding the screen causes minor further scrolling. There was a clear limitation in this calculation, as certain participants might be able to fixate while scrolling relatively fast, although we excluded the case of reading while slowly scrolling. However, this could be considered as an effect of scrolling that made it more difficult to read the search results.

Using the above definition, we compared how many times the users performed a continuous scroll action with vertical scrolling against the number of swiping events with the horizontal control type. In addition, we investigated how long the participants spent scrolling vertically for each task. Table 5.3 lists the number of finger actions required to scroll with both control types until the first decision was made. Applying a GLMM with a Poisson distribution, we found a significant difference in the number of scrolling motions on the first SERP according to control type ($\sigma_s^2 =$
Table 5.5: Fixation duration: relations between control types and others.

<table>
<thead>
<tr>
<th></th>
<th>Target-front</th>
<th>Target-back</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>AOI-front [s]</td>
<td>8.91</td>
<td>11.73</td>
<td>0.345</td>
</tr>
<tr>
<td>AOI-back [s]</td>
<td>3.37</td>
<td>3.42</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>8.42</td>
<td>10.47</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>7.12</td>
<td>4.52</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>0.345</td>
<td>**</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>***</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOI-pages</td>
<td>***</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Control type</td>
<td>0.197</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>0.507</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

Note: H and V denote horizontal and vertical control types, respectively.
* p < 0.05, ** p < 0.01, *** p < 0.001.

0.262, χ² = 97.09, df = 1, p < 0.001). This result indicates that participants conducted about five times more finger actions with the vertical control type.

Finally, we recalculated the contents of Table 5.2 for the four kinds of search time by removing the scroll duration (mean 3.05 s and 4.18 s with target-front and -back) with the vertical scroll type. We recalculated the search time under horizontal swiping by assuming that participants could not read SERPs while changing pages (0.5 s per scroll event). Target position effects can still be found across the four measurements. However, the control type effects on the time to correct click and time on SERPs have disappeared, the mean values of both control types become closer, and the p-values of the control type effects are now out of the significance level.

Thus, we can be sure that the time needed for vertical scrolling is by itself the main factor in users spending more time on SERPs. It is not productive time, such as reading for example; but simply time needed to manage the interface. Therefore, pagination would be one possible viewport control to improve unnecessary time consumption.

5.3.3 User attention

While the search performance and user satisfaction results represent how efficiently users search and their impression of the search experience, fixation duration is believed to represent the degree and effort to which users try to extract information from SERPs and areas of interest (AOIs) [Just and Carpenter 1976, Rayner 1998].

In this experiment, we recorded fixations of over 100 ms within a region of 70 pixels diameter using algorithms in the Eyeworks software. We determined the duration and distance thresholds by considering the screen resolution (1920 × 1080), the distance between the user and the eye-tracker, and the calibration accuracy [Salvucci].
To investigate the effect on user attention with each control type, we applied ANOVA and assigned two AOIs (AOI-pages): contents on the front page (AOI-front), and the remaining content (AOI-back).

First, we analyzed the effects of AOI-pages, target position, and control type on fixation duration. Using a three-way ANOVA, we found no significant control type effect or interaction of the three variables, but there were significant differences in user attention due to AOI-pages ($F(1,545) = 246.46, p < 0.001$), target position ($F(1,545) = 21.45, p < 0.001$), interaction of AOI-pages and target position ($F(1,545) = 30.73, p < 0.001$), interaction of AOI-pages and control type ($F(1,545) = 7.20, p < 0.01$), and interaction of target position and control type ($F(1,545) = 4.12, p < 0.05$).

Although we confirmed the effects of target position, AOI-pages, and their interactions, it was difficult to explain the implication of these interactions. For further analysis, we investigated the effects of control type and target position and control type and AOI-pages, because the control type had no significant effect but did interact with the other factors.

First, Table 5.5 describes the attention paid to AOI-pages (AOI-front and -back), which can explain the interaction of control type and target position. When investigating the fixation duration on AOI-front, there was a significant control type effect ($F(1,261) = 6.84, p < 0.01$) on fixation duration, whereas no difference according to target position was found. In addition to this, we could see a target position effect ($F(1,261) = 33.95, p < 0.001$) and an interaction between control type and target position ($F(1,261) = 4.62, p < 0.05$) for AOI-back.

These results indicate that subjects spent more time (about 2.4 s) extracting information from AOI-front with the vertical control type. On AOI-back, participants spent a similar amount of time reading the target on the front page with both scrolling types, but paid more attention to targets on the back page while using the horizontal swiping function (about 2.6 s).

The interaction of target position and control type on AOI-back may be explained by the fact that all links on the back page were displayed with one action under the horizontal control type. This allowed subjects to read more links on the back page, whereas the vertical control type required continued scrolling to see links on the back page, which moved downwards with each vertical scrolling action, which users found it harder to read while scrolling.

Second, we investigated the effect of AOI-front and -back on fixation duration to estimate how much effort the users made before and after the page fold with each control type, by analyzing the interaction of AOI-pages and control type as shown in Table 5.5. When we considered a target on the front page, a significant AOI-pages effect could be observed in fixation duration ($F(1,261) = 215.70, p < 0.001$), whereas
Table 5.6: Fixation duration: relation between AOIs and control types.

<table>
<thead>
<tr>
<th>Target positions</th>
<th>AOI</th>
<th>Control type</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank 1</td>
<td>***</td>
<td>***</td>
<td>0.366</td>
</tr>
<tr>
<td>Rank 3</td>
<td>***</td>
<td>0.171</td>
<td>0.584</td>
</tr>
<tr>
<td>Rank 5</td>
<td>***</td>
<td>0.607</td>
<td>**</td>
</tr>
<tr>
<td>Rank 6</td>
<td>***</td>
<td>0.815</td>
<td>***</td>
</tr>
<tr>
<td>Rank 8</td>
<td>***</td>
<td>0.430</td>
<td>*</td>
</tr>
<tr>
<td>Rank 10</td>
<td>***</td>
<td>0.454</td>
<td>**</td>
</tr>
</tbody>
</table>

* p < 0.05, ** p < 0.01, *** p < 0.001.

no difference due to control type was found. Considering tasks where the target was on the back page, we found significant effects on fixation duration due to AOI-pages ($F_{(1,261)} = 60.01, p < 0.001$) and the interaction of AOI-pages and control type ($F_{(1,261)} = 11.29, p < 0.001$).

This indicates that when the target result was located on the front page, participants spent more time reading the front page with both control types: 5.54 s and 8.31 s more with horizontal swiping and vertical scrolling, respectively. When the target was on the back page, participants using the scrolling interface still spent 5.95 s longer on the front page (2.36 s less than before), but participants using horizontal swiping spent only 1.30 s longer on the front page than the back (4.24 s less). That is, when the target document was on the back page, people using the swiping interface shifted their attention much more than people using the scrolling interface. This result might be related to the difference in search accuracy on the back pages (77.78% with pagination and 58.33% with vertical scrolling), because the longer reading time on the back pages may allow users to make better decisions.

We also assigned ten AOIs, one to each link in the SERP, to investigate which results participants focused on. In particular, we were interested in the effect of target position and AOI with each control type.

When we considered the effects of AOI, control type and target position on fixation duration, the effects were very similar to the results with target-front and -back with AOI-pages: we could see effects due to target position, AOI and interactions between these two variables.

To investigate these interactions, we analyzed the fixation duration of each AOI along the target positions using ANOVA (see Table 5.6 and Figure 5.4). First, we examined the effects for target positions 1, 3, and 5, located on the front pages. With target position 1, there were significant AOI and control type effects ($F_{(1,437)} = 25.97$,}
Figure 5.4: Fixation duration on each AOI along target positions [s]: numbers on each y-axis are values after back-transformation. The numbers in each sub-figure denote target ranks, for example, target at rank 1 for the top-left sub-figure.
Study Two: Pagination versus scrolling

$p < 0.001$ and $F_{(1,437)} = 11.95, p < 0.001$, respectively) on fixation duration. Fixation duration with pagination and vertical scrolling followed a fairly similar pattern of decline across AOs. The fixation duration on each AOI was normally higher with vertical scrolling, which causes the control type effect, and the main difference could be seen on AOs 2, 3. When the target position was at rank 3, fixation duration was significantly affected by AOI ($F_{(1,437)} = 30.77, p < 0.001$). AOI 3 recorded the longest read times with both control types, which would appear to be due to the target result.

With target position 5, users exhibited significant differences in user attention due to AOI and the interaction of AOI and control type ($F_{(1,437)} = 21.27, p < 0.001$ and $F_{(1,437)} = 2.69, p < 0.01$, respectively). This interaction was apparently caused by the reversed pattern of reading effort between AOI 1 and AOs 4, 5. Subjects spent more time reading AOI 1, but less on AOs 4 and 5 with the horizontal control type. This can be connected to the scroll rate discussed in Subsection 5.3.2 participants only exhibited higher scroll rates for target position 5. This may be due to some special (higher) user interest near the page fold with the higher scroll rate, as a previous study suggested [Granka et al., 2004].

Across target positions 6, 8, and 10, we found significant effects on fixation duration due to AOI ($F_{(1,437)} = 24.62, p < 0.001$), $F_{(1,437)} = 11.54, p < 0.001$, and $F_{(1,437)} = 6.11, p < 0.001$) and due to the interactions of AOI and control type ($F_{(1,437)} = 3.40, p < 0.001$, $F_{(1,437)} = 2.17, p < 0.05$, $F_{(1,437)} = 2.95, p < 0.01$). With target position 6, the reason for this interaction may come from the reversed pattern for AOs 3, 4, 5, and AOI 6. Users took more time to read AOI 6, which contained the correct answer, when using the horizontal swiping interface, whereas subjects using the vertical control type did not exhibit any enhanced reading effort in this AOI. This might be explained by AOI 6 being the top link on the second page when using pagination. With the relevant result at rank 8, subjects using horizontal swiping concentrated more on AOI 8, which contained the correct answer. Similar to the result for target position 6, we could not find a relation between the target position and fixation duration on the AOI with the vertical control type. Finally, for target position 10, subjects displayed similar results to those for target positions 6 and 8. The main reason for this interaction seems to come from AOI 10. Subjects took more time reading AOI 5, but spent less time on AOI 10 with the vertical control type. This may be the reason for the different search accuracy (75% with horizontal swiping and 50% with vertical scrolling) with target position 10.

When the target results were on the front page, the fixation displayed a similar pattern with both control types: decreasing along the sequence of AOs and increasing on the AOs containing the correct answer. However, whereas targets on the back page received more attention than others with the horizontal swiping interface, there
are no such effect with vertical scrolling. This may indicate that the targets were simply not recognised as useful when scrolling. This, in turn, may explain why the paginated interface led to higher accuracy when targets were over the fold.

5.3.4 User preference

At the end of each session, we asked participants about their preferred control type, the reason for this preference, and their thoughts regarding the use of horizontal swiping on their phones.

Twelve of the 24 participants felt that the vertical scrolling was more convenient than horizontal pagination, and one reason from seven of the 12 was familiarity.

“That [the vertical scrolling] is common in most phones including mine”

“Phones provide vertical browsers, familiar”

The other reason from the remainder (5) of this 12 is that the vertical control type was easier/more convenient for web searches.

“I prefer to continuous scrolling to pagewise [one page at a time]”

One-third (8) replied that they preferred the pagination to the vertical scrolling, and they considered it to be more convenient and easier to see all of the results.

“I just needed one click to see two pages — it [the horizontal swiping] was fast”

“It [pagination] was improved visibility”

The remainder (4) considered both scroll types to have similar usability. If horizontal swipes were provided by commercial search engines, a quarter of participants said that they would definitely use it instead of vertical scrolling, and the majority (16 of 24) were willing to try using pagination as the main scroll type.

5.4 Discussion

In this section, we aggregate the previous discussions by focusing on the effects of both control types and consider the limitations of this study.

First, in dividing SERPs into front and back pages, we found interesting effects of control type on search performance and no difference in user satisfaction. Although it was their first time using pagination on SERPs with a touch-enabled mobile device, participants exhibited similar or better search performance across four speed measurements, and they were more accurate when relevant results were located after the page fold. Participants also expressed similar satisfaction with both control types.
Considering each target position, we found an interesting search behaviour: participants scrolled at different rates, across the two interfaces, when the correct answer was at link 5. Similar to the results of previous studies [Granka et al., 2004; Joachims et al., 2005], special scanning behaviour occurred at the page break. Although this may have several implications, one possible reason is that participants scrolled vertically to read a few links after the page fold or to locate the link from the bottom to the middle/top of the screen with higher user attention when a correct answer was in that AOI.

Second, since we have controlled for task and other effects, differences in speed were most likely due to control type: something about vertical scrolling made users slower. To investigate this, we measured the vertical scroll duration by defining a continuous scroll action and re-investigated the search time by excluding the scroll duration. From these results, we could confirm that the time consumed when using vertical scrolling is the main reason for the difference in search time.

Third, a major effect on user attention was that participants with the horizontal control type spent more time extracting information over the page fold when the correct answer was located on the back page. We inferred that the higher search accuracy with horizontal swiping when the relevant links were on the back page may be due to the higher user attention on lower ranks. We left this for a further study. With 10 AOIs, subjects using horizontal swiping exhibited a similar pattern with each target position: a decline in fixation duration along the rank order, but strong attention on AOIs containing the relevant links. However, we could find no such effect with the vertical control type when the relevant links were on the back page. Thus, pagination appears to encourage users to notice results further down the list.

5.4.1 Limitations

We acknowledge that this study had several limitations. First, user interaction on mobile web search may differ for various screen sizes as shown in the previous Chapter [4]. We considered the recent trend in screen size, and adopted one of the most popular smartphone sizes for this study [Mobile Marketing, 2015]. However, it is possible that the results may be different with other screen sizes. Second, although we recruited participants with various backgrounds and a wide range of ages, the results cannot cover all mobile users’ individual web search behaviours. Third, participants were seated while searching, and were not free to move the mobile device as it was held fixed in a holder as shown in Figure [5.2]. This restricted the participants’ freedom of movement with the smartphone, which could have affected their
behaviour. Finally, the results may be affected by the task type. Because we adopted only informational tasks to avoid complex interactions of many independent variables, the results may be different with other tasks.

5.5 Conclusions and future work

We conducted an experiment to determine the effect of horizontal and vertical control types when conducting web searches on mobile devices.

Despite participants having greater familiarity with vertical scrolling during search, our results suggested that horizontal swiping gave similar or better search speed. The main reason was the time taken to scroll when using vertical scrolling. In addition, participants using horizontal swiping tended to pay more attention to links beyond the fold when a correct answer was located there, and with better search accuracy. Furthermore, users using pagination exhibited special attention to the positions of the relevant link, whereas this trend was not observed with the vertical control type when the correct answer was beyond the page-break. Finally, although half of the participants expressed a preference for vertical scrolling, many stated that this was because of familiarity, and most subjects were willing to try horizontal swiping if provided.

Considering the limitations regarding users’ individual differences and the lab conditions of this study, we cannot conclusively say that the horizontal control type is better than vertical scrolling. We do suggest, at least, that it is worthwhile for search engines to provide both scrolling types to enhance the user search experience.
Study Three: Appropriate snippet length

According to results from a previous study [Cutrell and Guan, 2007], the length of snippet in a desktop search has a significant effect on search time and behaviour. Due to smaller screens on mobile devices, the effect could be larger. To investigate the ‘curiosity’ about the long snippet from Study one in Chapter 4 ‘enriching the content of top links by showing longer snippets’, this chapter describes the third study related to the effect of snippet length. We expect that this study would contribute to finding appropriate lengths of snippets for mobile web search.

6.1 Introduction

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. Therefore, the snippet can be one of the most important elements that affects search performance, behaviour and user satisfaction.

To provide better SERP presentation designs, 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, and 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. However, 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 on desktop monitors.

Some studies have investigated the effect of snippet lengths on desktop screens [Cutrell and Guan, 2007; Kaisser et al., 2008; Paek et al., 2004]. Some of their results suggest that long snippets are more useful for finding a particular piece
Study Three: Appropriate snippet length

Figure 6.1: Examples of initial SERPs with short (left), medium (middle), and long (right) snippets.
of information [Cutrell and Guan 2007], and that snippets are currently too short to provide sufficient information in many cases [Kaisser et al. 2008].

Several studies [Jones et al. 2003; Raptis et al. 2013] and our previous work 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 6.1 for examples of short, medium, 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 due to the hint in snippets for the question, similar to the results in a previous study [Cutrell and Guan 2007]:

- 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 [Cutrell and Guan 2007]:

- 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.

After describing our experimental design and procedure. We then present our findings on the effects of task type and snippet length, and discuss the implications of the results. Finally, we conclude with some directions for future work.

6.2 User study

This section describes the experimental design and the procedure for how the experiment was conducted with the participants and tasks, and how the experiment was prepared and proceeded.
6.2.1 Participants

We recruited 30 participants\(^\text{1}\) 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 (see the post-experiment questionnaire in Appendix C). 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. The participants voluntarily attended and were compensated with a meal voucher.

6.2.2 Tasks

Table 6.1: Examples of task descriptions and queries.

<table>
<thead>
<tr>
<th>Informational</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Panadol is a brand of pain reliever. What are some side-effects of Panadol? Is a rash one of them? (panadol side effects)</td>
</tr>
<tr>
<td>• 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)</td>
</tr>
<tr>
<td>• Which two countries played for the 4th match in the Cricket World Cup 2011? (cricket world cup 2011 dates)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Navigational</th>
</tr>
</thead>
<tbody>
<tr>
<td>• You are interested in shoes from Adidas. Find the official Adidas homepage. (adidas shoes Australia)</td>
</tr>
<tr>
<td>• Find the web page where you can apply for a saving account on the Citibank website. (citibank new account)</td>
</tr>
<tr>
<td>• You bought a laptop from Sony and something doesn’t work as expected. Find the page for Sony technical support. (sony laptop technical support)</td>
</tr>
</tbody>
</table>

Note: The words in brackets are the queries.

Each participant completed twelve tasks (six each of informational and navigational tasks; see Table 6.1 for examples of the tasks; see Appendix B for further

\(^{1}\)Participants for each study were different people.
details) which were derived from those of Dumais et al. [2010]. The task categories varied, e.g., chemistry, sports, travel, history, and law, and the tasks were simple.

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 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. Each SERP had 10 search result links with titles, URLs and snippets and with no advertisement and the other features the same as the previous studies. With the rank order from the Google mobile search engine and the snippets we extracted, every task had at least two relevant links within the top three ranks, and included a correct answer/destination web page.

We prepared three different snippet lengths: short with one line, medium with two or three lines, and long with six to seven lines as shown in the short and long snippet examples in Figure 6.1. 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.

### 6.2.3 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 the participants listened to a short introduction to the experiment, they were asked to sign the consent form and to read the instructions regarding the procedure and tasks. Once they agreed and signed the form and had no further questions about the instructions, 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.

---

2 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.
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.

6.2.4 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 [Mobile Marketing, 2015]. The mobile phone was connected to the main system as a secondary monitor using the Twomon software [Dvrguru] and search results were displayed through Internet Explorer. To collect the gaze data, we used Facelab 5 [Seeingmachines], and we analyzed the data using Eyeworks [Eyetracking].

6.3 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, same as for the earlier chapters. We employed analysis of variance (ANOVA) for continuous data, generalized linear mixed models (GLMMs) for binary data and countable data, and a linear mixed
model (LMM) for score data. We also used linear regression [Searle, 1997] 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 \( \sigma^2 \) were greater than standard errors (SEs) in all variables.

As similar to the way we adopted in chapter 4, when we found a significant effect due to snippet length and/or the interaction of task type and snippet length, we ran a post-hoc test to confirm the difference among the three different snippet lengths using a standard error of the difference (SED), and we noted labels; e.g., ‘A’ and ‘B’: B is significantly different (e.g., bigger, higher, or longer) from ‘A’, but neither ‘A’ nor ‘B’ is significantly different from ‘AB’.

All analyses were conducted using the GenStat statistical package [VSN International, 2014].

### 6.3.1 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 a study of [Cutrell and Guan, 2007]. Only task type significantly affected task completion duration \( F_{(1,259)} = 172.22, p < 0.001 \). As can be seen in Table 6.2 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),
Table 6.2: Search performance, behaviour, and satisfaction for each task type, broken down by snippet length.

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Snippet Interaction</th>
<th>Informational tasks</th>
<th>Navigational tasks</th>
<th>Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>M</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Search performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task completion duration [s]</td>
<td>36.37</td>
<td>41.20</td>
<td>40.98</td>
<td>18.89</td>
</tr>
<tr>
<td>Time to first click [s]</td>
<td>16.24</td>
<td>16.35</td>
<td>24.03</td>
<td>12.90</td>
</tr>
<tr>
<td>Search accuracy</td>
<td>0.85</td>
<td>0.85</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>Fixation duration</td>
<td>7.71</td>
<td>7.98</td>
<td>11.69</td>
<td>6.71</td>
</tr>
<tr>
<td>Fixation on titles [s]</td>
<td>4.20</td>
<td>3.48</td>
<td>3.60</td>
<td>3.65</td>
</tr>
<tr>
<td>Fixation on URLs [s]</td>
<td>1.55</td>
<td>1.44</td>
<td>1.41</td>
<td>1.68</td>
</tr>
<tr>
<td>Fixation on snippets [s]</td>
<td>1.96</td>
<td>3.06</td>
<td>6.68</td>
<td>1.38</td>
</tr>
<tr>
<td>Search behaviour</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scroll rate</td>
<td>0.29</td>
<td>0.21</td>
<td>0.46</td>
<td>0.17</td>
</tr>
<tr>
<td>Viewport movement (pixel)</td>
<td>540</td>
<td>675</td>
<td>1148</td>
<td>640</td>
</tr>
<tr>
<td>Compressed sequence length</td>
<td>9.56</td>
<td>7.50</td>
<td>8.27</td>
<td>7.30</td>
</tr>
<tr>
<td>Minimal scanpath length</td>
<td>4.17</td>
<td>3.23</td>
<td>4.13</td>
<td>3.46</td>
</tr>
<tr>
<td>Next snippet [s]</td>
<td>3.63</td>
<td>3.65</td>
<td>3.60</td>
<td>2.79</td>
</tr>
<tr>
<td>Task completion duration [s]</td>
<td>36.37</td>
<td>41.20</td>
<td>40.98</td>
<td>18.89</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>4.46</td>
<td>5.35</td>
<td>4.67</td>
<td>5.25</td>
</tr>
</tbody>
</table>

*Significant at 0.05 level. ** Significant at 0.01 level. *** Significant at 0.001 level.

Note: S, M, and L denote short, medium, and long snippets, 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 6.2). As can be seen in Figure 6.2, participants with ITs exhibited the highest time consumption with long snippets (A-A-B), although the time spent for NTs (A-A-A) does not follow this pattern.

A significant difference can be observed in search accuracy between the task types ($\sigma^2 = 0.464, \chi^2 = 5.10, \text{df} = 1, p < 0.05$). This indicates that our 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 [Just and Carpenter 1976, Rayner 1998]. 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 6.2, 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 6.3 for each task type, and broken down by snippet length, is very similar to the pattern for search time (see Figure 6.2): both measurements displayed the patterns of A-A-B for ITs and A-A-A for NTs, with short-medium-long snippets, respectively. 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 better understand overall user attention and also make it possible...
§6.3 Results

Figure 6.4: Proportion of total fixation duration for each element of SERPs.

to test H2.

Figure 6.4 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 (B-BA-A with short-medium-long snippets, respectively), 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 6.2. 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 6.5 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 6.2) and long snippets for ITs resulted in the most attention (6.68 s in Table 6.2). 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, that means, if we provide more information, they will read them even more. 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 the experiment of Cutrell and Guan [2007], 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 [Cutrell and Guan, 2007]. 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.

### 6.3.2 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.

#### 6.3.2.1 Scroll action

Due to the small screen size, users needed to scroll to see the lower ranks beyond the page fold, which may increase search time. 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 6.2, 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 (AB-A-B with short-medium-long snippets, respectively). 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 ($\sigma^2_s = 0.011$, $\chi^2 = 112.57$, 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 ($\sigma^2_s = 0.078$, $\chi^2 = 5.71$, df = 2, $p < 0.01$). As shown in Table 6.2, the long snippet brought more viewport movement for both task types (A-A-B with short-medium-long snippets, respectively).

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.

6.3.2.2 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 Chapter 2, 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 [Dumais et al., 2010; Kim et al., 2015; Lorigo et al., 2006] to investigate users’ scanning strategy.

Only task type effect was observed on compressed sequence length ($\sigma^2_s = 0.088$, $\chi^2 = 10.73$, 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 ($\sigma^2_s = 0.046$, $\chi^2 = 15.53$, 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 more, B-A-A with short-medium-long snippets, respectively), whereas they exhibited no difference between medium and long snippets (3.1–3.3 links for ITs, 2.5–2.8 links for NTs).
§6.3 Results

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 6.1 for examples of short, medium, and long snippets), the mean minimal scanpath lengths for long snippets (3.13 and 2.5) appeared somewhat 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 6.6: 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.](image)

Figure 6.6 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.

6.3.2.3 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 ($\sigma^2 = 0.144, \chi^2 = 7.85, df = 1, p < 0.01$). 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. If we take the revisit counts as a proxy for users’ hesitating, the participant tended to similarly hesitate with the different snippet lengths before 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.
6.3 Results

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 6.2, 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 6.7, 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 scan-path) and the reading time. The search time in Figure 6.6 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. Considering the differences regarding word counts and reading time among the three snippet lengths, it appears that either users are not reading all of the snippet or they read faster with longer snippets, because the long snippet had more than twice the words that the medium snippet had, but the reading time per link does not show this amount of difference.

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.

6.3.3 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 6.2.3, participants were required to score their satisfaction after each task using a 7-point Likert scale regarding the snippet length. Table 6.2 shows that there were significant effects due to snippet length and the interaction of task type and snippet length on user satisfaction (\(\sigma^2_s = 0.447, \chi^2 = 12.93, df = 2, p < 0.001\) 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 6.8 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
Study Three: Appropriate snippet length

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, where were on average very easy. 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.

6.4 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 our navigational tasks are easier than informational tasks, similar to the results from previous studies (e.g. [Cutrell and Guan, 2007; Kim et al., 2015]). 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 [Cutrell and Guan 2007]: 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.

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, (we investigated the effect of horizontal pagination in the previous Chapter [5], and which may help reduce the effect of scrolling, and highlighting the query words in the snippets may reduce hesitation behaviour on SERPs [Iofciu et al. 2009]).

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. These results seem that they might be “trained” through their everyday searches with the snippet length provided by current search engines.

6.4.1 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 [Mobile Marketing, 2015], 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 as suggested in the previous Chapter 4 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, and hence
the results may be different with snippets extracted by other search engines.

Sixth, people’s search behaviours or preferences may change after a day, a week,
and a year, and the laboratory user study does not capture this. Therefore, the results
may be different by time and the environment.

6.5 Conclusions

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 more.

In addition, even if a relevant link is on the top page, users will 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 accus-
tomed to searching with the typical snippet size, our results suggested that mobile users are best served by snippets of two to three lines.
Conclusions and future directions

7.1 Summary of Contributions

Mobile web search has been getting very common [Statcounter Global Stats 2014] and more than half of the network traffic for web search comes from mobile devices [Search Engine Land 2015]. There have been a number of studies to improve presentation designs for SERPs by investigating user interaction. Although the results of the studies contributed to improving SERP interfaces, current mobile search engines do not display different contents from desktops display, and the studies for mobile web search are not enough to suggest optimised SERP presentation designs for the small screens.

From our earlier work [Kim et al. 2015] and the previous study [Jones et al. 2003], we confirmed that the differences of search behaviour between desktop screens and mobile devices need to be considered for the web search interface design. We conducted three user studies to understand mobile web search behaviour and to suggest better presentation designs for SERPs on small screens. These studies have considered how people read search results on various small screens; the effect of control types; and what snippet size is needed for mobile users.

The main features of each chapter and the contributions of each study presented in this thesis can be summarised as follows:

- **Chapter 1. Introduction** We addressed the research background and purpose. That is, why we are interested in this research, describing recent trends in web search in mobile devices: the increase of web usage, the growth of smartphone ownership, and the proportion of web traffic by mobile devices, and a brief introduction of previous work: user interaction under the standard and manipulated conditions of SERPs, and some studies for mobile web search. We also presented the goal of this research: a better presentation design for mobile devices.
• Chapter 2. Background and Related Work

We explored what we need to consider as the background knowledge for conducting the research. We introduced both online and offline evaluating metrics in aspects of IR and HCI research fields, and described evaluating SERPs with several methods such as eye-tracking technology, discussed general behaviour on reading and scanning SERPs, and strategies in web search. We also addressed previous studies related to mobile web search.

• Chapter 3. Measurements

We described what measurements we used in this thesis with the reason for adoption. We measured search time and accuracy as search performance, and considered fixation data (scanpath, direction, skip and regression), click and scroll events for search behaviour. In addition, we addressed post-task and post-experiment questionnaire, which are related to user preference and participants’ information.

• Chapter 4. Study One (Three different small screens)

We explored user search performance and behaviour on three different sizes of screens due to the remarkable trend of enlargement of screen sizes on mobile devices. Briefly, we found no significant difference with respect to the search performance, 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. With this result, we suggested several ideas for presentation design for each screen size: displaying a knowledge graph for phablets (5.5 inches or bigger); enriching the content of top links by showing longer snippets for recent smartphones (4.7 inches); embedding page up and down button on the interface or providing horizontal page changes functions for old smartphones (3.5 inches). Compared to the results from [Raptis et al. 2013], our study showed some different results: no difference in the task completion duration. This may be caused by a better quality of SERPs and some mobile-friendly web documents. The main contribution in this study is that we suggested better presentation designs for each screen size considering different search behaviours.

• Chapter 5. Study Two (Pagination versus scrolling)

Due to the one curiosity from the study one: what would happen, if users can use the horizontal pagination instead of the vertical scrolling, which all current search engines provide. We investigated the effect of horizontal and vertical viewport control types (pagination versus scrolling) in mobile web search. Our findings suggest that pagination improves search over scrolling, despite scrolling being more
future. The main reason for this is the time taken for the scroll itself. Participants using scrolling also spent less time reading lower-ranked results even if this is where relevant documents were found. We conclude that search engines need to provide different viewport controls to allow a better search experience on touch-enabled mobile devices. The contribution of this study is that we found that the pagination is worth being adopted for web search.

- **Chapter 6. Study Three (Length of snippets)*** We described the effect of three different snippet lengths on a mobile device, because snippets are the biggest element in SERPs and the effect of snippet length on smartphones can be much bigger than on desktop screens. We found that users with long snippets on mobile devices exhibit longer search times with similar search accuracy for informational tasks. This is caused by the long reading time and the frequent use of scroll function with bigger viewport movements. The over-all findings suggest that, unlike desktop users, mobile users are best served by snippets of two to three lines. This study contributes to finding appropriate snippet lengths for mobile devices.

### 7.2 Future Directions

As considering the limitations in each study, we may suggest future studies as further investigation of the interactions among screen size, effect of control type, and snippet length.

In the first study, we suggested several presentation designs for each small screen size by considering the search behaviour. Although some of the ideas for presentation designs has been tested in the chapter 5 and 6 (e.g., effects of viewport control type and snippet length), the others remain as future work.

- **Small screen: font size, page changing button, and rich information** It may 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 as suggested by Jones et al. [1999] or horizontal page changes instead of a vertical scroll function.

- **Medium screen: enriching top link contents** Enriching the content of top links by showing longer snippets could reduce the hesitation on the recent smart phones.

- **Large screen: displaying a knowledge graph** Display a knowledge graph that displays information regarding the keywords that may be helpful for search performance, as shown in a previous study [Lagun et al., 2014]
In the second study, we investigated the effect of viewport control type, following up on a suggestion from the first study. The results revealed that it is worth adopting horizontal pagination for mobile web search, although it left the following ideas for future study.

- **The effect of control type on various small screen sizes** Considering popularity of bigger screen size, we conducted the experiment adopting a phablet (i.e., iPhone 6 plus). As we confirmed the search behaviour can be different along screen sizes, we may have some different results on using the control types with different screen sizes like iPhone 4 (an old smartphone) and iPhone 8 (a recent smartphone).

- **The effect of control type on touchable monitors** Nowadays, we can use a touchable monitor on laptops and desktop and the vertical scrolling is also the main control type for the bigger screens. Therefore, investigating the effect of the control types on the large touchable screens would bring some interesting results.

- **Pagination with navigational tasks** In this study, we considered informational tasks for the experiment. Users with the other task types such as a navigational task may exhibit different result: no significant effect between both control types or even worse search performance with the pagination. This point also needs to be confirmed.

In the third study, we concluded that two-three lines of snippet is the best for mobile web search. However, this experiment was conducted with easy tasks as users indicated, and the users’ search experience might come from their familiarity of the current mobile search engines, which provides two or three lines of summary, typically. Therefore, we may consider the ideas as below for future study.

- **Task difficulty** For harder level of task difficulty (e.g., comparing information and finding complex information with visiting several web documents), the dependency of snippets may increase, and users may want to spend more time on snippets. Thus, conducting additional experiments with diverse task difficulties and topics are required.

- **Less experienced users in mobile web search** Most participants in our experiment were familiar with mobile web search, and the participants might be trained by the commercial search engines with two-three lines of snippet
provided. Therefore, an experiment with less-experienced participants regarding the use of long snippets would explain more about the reason of the user preference.

Additionally, SERPs on mobile device also provide images, advertisements, and other information such as a visited date and a popularity of the web document. Although we did not include those elements to the tasks in the studies in order to compare user attention on the main contents (i.e., title, URL, and snippet), investigating the effect of the additional information would be helpful for better understanding of search behaviour.

The studies proposed above are connected to each other. Thus, the goal of the future work is finding optimized presentation design along various screen sizes.

### 7.3 Closing Remarks

The work presented in this thesis helps us to understand mobile web search behaviour. We believe that our approach will contribute to understanding user behaviour in mobile web search. This will help us in improving the user experience in mobile web search.
Appendix A

Appendix A: Experiment forms

The consent form and instruction before the beginning of experiment.
1 The goal of experiment was varied for each study.
2 The average running time for the experiment was varied for each study.

*Note1:* The instruction is for the study three. The contents in other studies may be different from the instruction.

*Note2:* All experiments described in this thesis were approved by the human ethics committee of the university (Human ethics protocol: 2012/006).
User Interactions with Mobile Devices

Dear participant,

This experiment is being carried out as part of research at the Australian National University (ANU). This experiment investigates \textit{purpose of the experiment}\textsuperscript{1}. This should help researchers design future search engine design which better meet people’s needs.

Participation in this project will take about \textit{time}\textsuperscript{2} of your time. You will be asked some initial questions about yourself and your experience with web search on mobile devices, then asked to carry out some search tasks using the software we provide. We will record your interactions with the system, including data obtained from an eye-tracking system and clicks. These results will be analysed and used to guide our research.

Results from this project may be published in a research forum. No personal information will be published except in aggregate form (such as averages or totals). We will not publish any information which could link you to the experiment or to any particular search or webpage, and any information you provide will only be used for the purpose for which you have provided it. All information will be protected to the greatest extent allowed by law, and data will be kept secure on a password-protected computer during and after the project.

Your participation in this research is entirely voluntary. You are welcome to withdraw at any time, even after finishing the search tasks, and there will be no penalty whatsoever.

If you have any questions, comments please contact the researchers, Ramesh Sankaranarayana, at ramesh@cs.anu.edu.au or (02) 6125 4281, or Jaewon Kim, at jaewon.kim@anu.edu.au or on (02) 6125 9662. If you have any questions regarding your rights
as a research participant, please contact the Human Ethics officer at the ANU Office of Research Integrity at human.ethics.officer@anu.edu.au

Thank you for your participation.

__________________________
I, ________________________, (print name), hereby give my consent to participate in the “User Interaction with Mobile Devices” experiment. I understand that participation is completely voluntary, and that I may withdraw from the project at any time with absolutely no penalty.
Signature: __________________

Date: _____________________

If you would like to receive a copy of any publications based on this experiment, please provide your e-mail address (optional):

E-Mail: _____________________

If you have any questions, comments please contact the researcher, Ramesh Sankaranarayana, at ramesh@cs.anu.edu.au or (02) 6125 4281, or Jaewon Kim, at jaewon.kim@anu.edu.au or on (02) 6125 9662. If you have any questions regarding your rights as a research participant, please contact the Human Ethics officer at the ANU Office of Research Integrity at human.ethics.officer@anu.edu.au

Thank you for your participation.
**Instruction**

**In this experiment,**

- This is not a test, so feel free during the experiment.
- It takes about $time^2$ to the end.
- During the experiment, please avoid to move your head too much or to touch your face.

**Environment**

- Browser: Internet explorer
- Search Engine: Google Mobile

**Guide of experiment**

- You will do three practice tasks until you are familiar with carrying out a task.
- The below is the example

---

*For example)*
- Description: You are interested in some facts about the Telstra tower in Canberra. What is its height?
- Initial Query: Telstra tower fact

---

- If you can not understand the description, please ask the experimenter.
- When you find the answer on the search result page, please click the link then tell the experimenter, for example: this is the right answer!
- In this experiment, you will do three sample tasks on each snippet length. You then will do totally 12 tasks.
- Any questions are welcome!!

Thank you for your participation.
Appendix B: Task descriptions and queries

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

Table B.1: Full task descriptions and queries (Study One).
Table B.2: Full task descriptions and queries (Study Two).

<table>
<thead>
<tr>
<th>Task description</th>
<th>Initial task query</th>
</tr>
</thead>
<tbody>
<tr>
<td>• You are interested in membership of Questacon. What are the benefits?</td>
<td>Questacon membership benefits</td>
</tr>
<tr>
<td>• Which two countries played the third match in the 1996 Cricket world cup?</td>
<td>Cricket world cup 1996</td>
</tr>
<tr>
<td>• How many demerit points can you collect before your driving license is suspended in ACT?</td>
<td>demerit point ACT suspend</td>
</tr>
<tr>
<td>• You plan to visit Jervis Bay in a couple of weeks. Check the weather there 14 days from now.</td>
<td>Jervis bay weather 14 days</td>
</tr>
<tr>
<td>• Find a number of cesium (Cs) in the periodic table.</td>
<td>cesium</td>
</tr>
<tr>
<td>• You need to deliver your fridge from your old home to new home. Find a contact number of a delivery service for that.</td>
<td>fridge delivery Canberra</td>
</tr>
<tr>
<td>• You want to apply for roadside assistance and to compare the costs. Find a low price on a comparison site.</td>
<td>roadside assistance</td>
</tr>
<tr>
<td>• You need to buy a new 14W LED globe for your home. Check the price in the Bunnings website.</td>
<td>led globe Bunnings</td>
</tr>
<tr>
<td>• You plan to buy a car. Find the fixed interest rate for a car loan from ANZ.</td>
<td>ANZ loan</td>
</tr>
<tr>
<td>• You need to park at Sydney international airport 2 hours to pick your friend.</td>
<td>Sydney airport parking fee</td>
</tr>
<tr>
<td>• Find the cost on the airport official site.</td>
<td></td>
</tr>
<tr>
<td>• You are buying a property in ACT, which costs $350,000. Estimate the stamp duty using a calculator.</td>
<td>property stamp duty act</td>
</tr>
<tr>
<td>• Your weight is 60kg and you want to know how many calories you burn by swimming. Find the calorie consumption for you.</td>
<td>swimming 60kg</td>
</tr>
</tbody>
</table>
Appendix B: Task descriptions and queries

Table B.3: Full task descriptions and queries (Study Three).

Informational

- Panadol is a brand of pain reliever. What are some side-effects of Panadol? Is a rash one of them? (panadol side effects)
- The new iPhone SE is just out. In what colors can you get it (the color of the phone itself - not the color of additional cases for it)? (iphone se colour)
- 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)
- You want to buy alcoholic beverages for your friend when you come back from overseas. How many litres (maximum) can you bring? (Australian customs liquor)
- You forgot to bring your E-tag when you drove through the Harbour Bridge today. Will there be some additional charge? How much? (harbor bridge toll)
- Which two countries played for the 4th match in the Cricket World Cup 2011? (cricket world cup 2011 dates)

Navigational

- You bought a laptop from Sony and something doesn't work as expected. Find the page for Sony technical support. (sony laptop technical support)
- You need to apply for an E-tag account on the Citibank website. (citibank new account)
- Find the web page where you can apply for a saving account on the Citibank website. (citibank website)
- You are interested in Israeli electric cars. Find the official Tesla website for information on the Model 3. (tesla model3 australia)
- You are interested in shoes from Adidas. Find the official Adidas homepage for Adidas shoes Australia. (adidas shoes Australia)
- You are interested in the special offers page for Virgin flights. (virgin special offers flights)
- You are using Accor for electricity and now have a concession card. Find the website which has the concession form. (accor electricity concession)
- You are interested in Tesla (electric car) Model S. Find the official Tesla website for information on the Model S. (tesla model3)
- 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)
- You want to buy alcoholic beverages for your friend when you come back from overseas. How many litres (maximum) can you bring? (Australian customs liquor)
- You forgot to bring your E-tag when you drove through the Harbour Bridge today. Will there be some additional charge? How much? (harbor bridge toll)
- Which two countries played for the 4th match in the Cricket World Cup 2011? (cricket world cup 2011 dates)

Note: The words in brackets are the queries.
Appendix C: The detailed post-experiment questionnaire

Post-experiment questionnaire for each study.
Appendix C: The detailed post-experiment questionnaire

Post Experiment Questionnaire (Study One)

Q1: Age
Number (___)

Q2: Gender
1. Male ( ) 2. Female ( )

Q3: What year are you in? (eg., Undergraduate year 2, Master or Ph.D)

(__________________________________________________________)

Q4: What do you study? (eg., Computer science, Engineering, Finance or Law)

(__________________________________________________________)

Q5: Which of screen size was the most convenient for you?
1. Smallest size ( ) 2. Medium size ( )
3. Largest size ( ) 4. All similar ( )

Q6: How difficult were the tasks? (1: Entirely easy, 7 Entirely hard)

1 2 3 4 5 6 7

Q7: How often do you use a search engine like Google, Bing or Yahoo?
1. At least once a day ( ) 2. A few times per week ( )
3. Once a week ( ) 4. Rarely ( )
Q8: How good are you with search engine?
1. Expert ( ) 2. Good ( ) 3. Inexperienced ( ) 4. Terrible ( )

Q9: How good are you with using mobile device?
1. Expert ( ) 2. Good ( ) 3. Inexperienced ( ) 4. Terrible ( )

Thank you for your participation.
Appendix C: The detailed post-experiment questionnaire

Post Experiment Questionnaire (Study Two)

Q1: Age

Number (___)

Q2: Gender

1. Male ( )
2. Female ( )

Q3: What year are you in? (eg., Undergraduate year 2, Master or Ph.D)

______________________________

Q4: What do you study? (eg., Computer science, Engineering, Finance or Law)

______________________________

Q5: Which of controlling type (horizontal vs vertical) was the most convenient for you?

1. Horizontal ( )
2. Vertical ( )
3. Both similar ( )

Q5-1: Why? Please write the reason.

______________________________

Q6: Are you willing to use mainly the horizontal scroll (flip), if search engines provide the function?

1. Yes ( )
2. Give a try ( )
3. No ( )
Q7: How difficult were the tasks? (1: Entirely easy, 7 Entirely hard)

1 2 3 4 5 6 7

Q8: How often do you use a search engine like Google, Bing or Yahoo?

1. At least once a day ( )
2. A few times per week ( )
3. Once a week ( )
4. Rarely ( )

Q9: How good are you with search engine?

1. Expert ( )
2. Good ( )
3. Neither good nor bad ( )
4. Inexperienced ( )
5. Terrible ( )

Q10: How good are you with using mobile device?

1. Expert ( )
2. Good ( )
3. Neither good nor bad ( )
4. Inexperienced ( )
5. Terrible ( )

Thank you for your participation.
Post Experiment Questionnaire (Study Three)

Q1: Age
Number (___)

Q2: Gender
1. Male ( )
2. Female ( )

Q3: What year are you in? (e.g., Undergraduate year 2, Master or Ph.D)

(___________________________________________)

Q4: What do you study? (e.g., Computer science, Engineering, Finance or Law)

(___________________________________________)

Q5: For tasks to find a particular web page (e.g., find the official Adidas homepage.), what was the most important thing in choosing a link on a search result page?

1. Title ( )
2. URL ( )
3. Snippet(description) ( )
4. Others (_______________________)

Q6: For tasks to find a particular web page (e.g., find the official Adidas homepage.), which snippet length was suit for you?

1. Short (1 line) ( )
2. Medium (2-3 lines) ( )
3. Long (6-7 lines) ( )

Q6-1: Why? Please write the reason.

(___________________________________________)
Q7: For tasks to find a particular information (e.g., what is the height of Telstra tower), what was the most important thing in choosing a link on a search result page?

1. Title ( ) 2. URL ( )
3. Snippet(description) ( ) 4. Others (____________________)

Q8: For tasks to find a particular information (e.g., what is the height of Telstra tower), which snippet length was suit for you?

1. Short (1 line) ( ) 2. Medium (2-3 lines) ( ) 3. Long (6-7 lines) ( )

Q8-1: Why? Please write the reason.

(________________________________________)

Q9: The search terms (queries) automatically selected for each task were usually close to what I would have entered myself for that task. (1: completely disagree, 4: neither agree nor disagree, 7: completely agree)

1 2 3 4 5 6 7

Q10: The tasks were easy to complete for me (1: completely disagree, 4: neither agree nor disagree, 7: completely agree))

1 2 3 4 5 6 7

Q11: I am good with search engine. (1: completely disagree, 4: neither agree nor disagree, 7: completely agree)

1 2 3 4 5 6 7

Q12: I am good at using mobile device. (1: completely disagree, 4: neither agree nor disagree, 7: completely agree)

1 2 3 4 5 6 7
Thank you for your participation.
Bibliography


Proceedings of the fifth ACM international conference on Web search and data mining, 373–382. ACM. (cited on page 10)


Poole, A. and Ball, L. J., 2006. Eye tracking in HCI and usability research. (cited on page 10)


