




University  
of Glasgow

# Modelling Search and Stopping in Interactive Information Retrieval

**David Martin Maxwell**

School of Computing Science  
College of Science and Engineering  
University of Glasgow  
Scotland 

A thesis submitted for the degree of  
*Doctor of Philosophy (PhD)*

© David Maxwell



## Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this doctoral thesis are original and have not been submitted in whole (or in part) for consideration for any other degree or qualification in this (or any other) university.

This doctoral thesis is the result of my own work, under the supervision of Dr. Leif Azzopardi (*University of Strathclyde*) and Professor Roderick Murray-Smith (*University of Glasgow*). Nothing included is the outcome of work done in collaboration, except where specifically indicated within the text.

Permission to copy without fee all or part of this doctoral thesis is granted, provided that copies are not made or distributed for commercial purposes, and that the name of the author, the title of the thesis and date of submission are clearly visible on the copy.

A handwritten signature in black ink, appearing to read 'D. Maxwell', with a long horizontal flourish extending to the left.

David Martin Maxwell

Glasgow, Scotland 

31 May 2018



## Abstract

*Interactive Information Retrieval (IIR)* is a complex, non-trivial process (Ingwersen and Järvelin, 2005). During the course of a search session, a searcher may issue multiple queries, and for each query, examine a varying number of result summaries (textual snippets) and documents. That is to say, the interactions and behaviours exhibited by a searcher are *complex*, and he or she will inherently adapt their behaviour based upon a number of factors – perhaps most notably due to the perceived relevance of the list of results presented to them (Moffat et al., 2013). A searcher, for example, would be wise to abandon a list of results if they are perceived to be of low quality.

Despite the findings of a large number of studies in the field of IIR, many of the models and measures still in use today within the *Information Retrieval (IR)* community do not consider these complex factors regarding user behaviours. Central to these models and measures is the so-called *Cranfield Paradigm*, the *de facto* model used for IR evaluation, developed in the early 1960's. While this approach has been adapted over the years to suit ever more complex evaluation tasks (Harman, 2010), the model is still largely systems-focused, rather than user-focused. Such an evaluation approach considers a single query, and the consideration of a large number of doc-

uments returned from that query to be relevant – a wholly unrealistic approach.

As such, many researchers have proposed different models and frameworks for the purposes of IIR modelling and evaluation. In this thesis, we propose the *Complex Searcher Model (CSM)*, a high-level model that attempts to capture the various complex interactions and decisions that take place during *ad-hoc topic retrieval*. Central to the CSM is the inclusion of several key *decision points*, which provide a means for modelling the *stopping behaviour* of a searcher. Stopping behaviour is a fundamental aspect of human behaviour (Nickles, 1995), and during search, this is no different. Research into the stopping behaviours has however until recently been sparse, with work suggesting that searchers stop when they feel that what they have found is simply “good enough” to stop (Wu et al., 2014).

Considering the CSM, this thesis attempts to ascertain a more precise definition of what is “good enough”. We operationalise a variety of different stopping *heuristics* defined within the literature over a number of years, considering heuristics from both IR research and those based upon ecology, where researchers examined the stopping behaviour of animals when foraging for food. These operationalised heuristics are examined in varying search contexts. We compare the behaviour of real-world subjects who partook in a series of user studies against simulated behaviours utilising the CSM.

From the work undertaken as part of this thesis, we are able to show that developing searcher models incorporating some form of stopping behaviour is

able to offer more realistic, and credible simulations of the complex interactions that take place during a search session.





## Acknowledgements and Reflections

A good friend of mine – and a fellow PhD student – once said to me that when the time came to write his PhD thesis, he would avoid an acknowledgements section where *everyone and their dog* would be thanked for being part of his life. I on the other hand take a very different light on this matter; there are a lot of people who have helped me get to where I am today in some form. *To show my appreciation, I want to dedicate this work to each and every one of them* – regardless of whether they have a dog or not.

So, where do I begin? I guess day one of the PhD is a good place to start – October 1<sup>st</sup>, 2013. Sitting in my new office with (relatively) new faces around me, I remember thinking something along the lines of *what have I just let myself in for?* But despite the occasional set of disappointing results, the existential questioning of *what am I doing*, along with the challenges I've faced outside my PhD, I've got through to the other side complete with a much better understanding of my subject and life in general – as well as countless good memories to go with all of that. I did not for one minute think that within the intermediary four years, I would drive along the Pacific Highway in New South Wales, wander the streets of New York City with my good friend Horațiu, or have drinks with friends in downtown Tokyo. During my

PhD, I have been incredibly fortunate to undertake interesting experiments (that *did* work), and as a perk of getting them published, travel the world. And in doing so, I have met some amazing people with whom I have had the opportunity to share what I have been working on – and indeed, learn from. As Professor Ian Ruthven pointed out in his own 2001 thesis, there may only be one name on the front page of a PhD, but countless numbers of people behind the scenes who were involved in the making of it, “*cajoling*” the author (as Ian put it) towards the finishing line.

To the people in SAWB Rooms 220 and 221 – my friends, Colin Wilkie, Jarana Manotumruksa, Stuart Mackie, Stewart Whiting, Fatma Elsafoury, David Paule, James McMinn, Xi Yang, Rami Alkhawaldeh, Fajie Yuan, Phil McParlane, Jesus Perez and his brother Felix, and Shawki Al-Dubae. Thank you all for your friendship and the good times we have shared. Thank you also to Frances Cooper, Gözel Shakeri, Craig Reilly, James Trimble, Blair Archibald, Laura Voinea and Natascha Harth for your friendship. To Sean McKeown, it has been an absolute pleasure – and I know you are doing Edinburgh Napier proud. Most of all though, my thanks to Horațiu Bota for your continued friendship as we progressed through our respective PhDs – *we made it!*

To the Head of School, Professor Chris Johnson – and Dr. Simon Rogers – thank you both for the support and trust you bestowed upon me throughout my PhD. To Dr. Yashar Moshfeghi, thank you for your friendship, wisdom and advice throughout the years – even as my tutor when I was a second year undergraduate! To Professor Rod Murray-Smith, thank you for your feedback and assistance in shaping this thesis into what it is now. Your support after Leif moved to the University of Strathclyde was invaluable. To Helen

Border, Gail Reat and Teresa Bonner in the teaching office, I am so glad I could help out with my tutoring and exam collection efforts throughout the years. I enjoyed every second of my tutoring jobs – I hope that over the eight years I tutored, I got *something* through to a student who was struggling. Indeed, thank you to everyone else in the School of Computing Science who made my time there such an enjoyable and rewarding experience.

To those in the College of Science and Engineering – especially Heather Lambie – thank you for letting me disappear several times to go do internships elsewhere. I know it is a pain seeing students disappear for months at a time, so it makes me appreciate your efforts even more. Generally, thank you to the University of Glasgow for providing me with the opportunity to show the world what I can do, and to the *EPSRC* for providing me with a PhD scholarship to get this work done without any financial headaches.<sup>1</sup>

To my PhD examiners – both my internal examiner, *Dr. Some Guy*, and my external examiner, *Dr. Another Guy* at *Some University, Some Country*. I know just how much effort you both put into reading this thesis, and I hope I was able to provide you with the clarity, detail and presentation that you both expected. Thanks to you both for the work you put into assessing this thesis. My thanks also goes to my viva convenor, *Dr. Convenor Person*, for the time and effort in ensuring the examination process was swift and (relatively) painless.

In September 2014, I spent several weeks in Tampere, Finland. Here, I had the opportunity to work with world leaders regarding user modelling and

---

<sup>1</sup>I acknowledge the financial support offered by the UK Government (through the EPSRC), under grant number 1367507.

simulation of the Interactive Information Retrieval process. Thank you to Professor Kalervo Järvelin, Dr. Jaana Kekäläinen, Dr. Heikki Keskustalo, Teemu Pääkkönen and Dr. Feza Baskaya for a fantastic, rewarding and memorable time. I learnt so much from you all, and this reflected in the simulation software that was developed during my PhD – including use of your prototypical querying strategies.

To my good friends Vu Tran and Ioannis Karatassis, thanks for the wonderful time I spent at the University of Duisburg-Essen in Germany in late November to early December 2015. Thank you also to Professor Norbert Fuhr for allowing me to visit, for his counsel, and allowing me to participate on developing the interesting work that we were able to present at the third *International Conference on the Theory of Information Retrieval* in Amsterdam.

To Professor Jaap Kamps at the University of Amsterdam – thank you for your Doctoral Consortium mentorship at the first *Conference on Human Information Interaction and Retrieval* in North Carolina. The feedback you provided on my work allowed me to formulate some new ideas which subsequently made it into this thesis.

To Johanne and Penny, thank you both for everything you did for me when I visited in Melbourne. To Mostafa and Samira, thank you both for everything you did for me when I visited in Amsterdam. Mostafa, the moped ride from Sloterdijk station was *awesome!* The kindness you all showed me was very much appreciated, and I am happy to repay this at any time.

My thanks also to Dr. Peter Bailey, Dr. Paul Thomas and Professor Dave Hawking for the memorable time at Microsoft Australia. The kindness and

support shown by you all was very much appreciated. I learnt so much from my time there – and this has undoubtedly equipped me to become a better scientist. Thank you so much. To Gabrielle, James, Alan and Az – it was a pleasure to meet you all, and thank you for your continued friendships. I have nothing but fond memories of my times in Canberra.

I was also fortunate enough to become an intern at the *Alan Turing Institute* in London, during the Summer of 2017. While this internship was not directly applicable to my PhD (nor was my time at Microsoft, for that matter), I did learn lots about a new area – and of course, had the opportunity to meet and work with some wonderful people. To Emily Neilson, Professor Terry Lyons, Dr. Hao Ni, Dr. Jeremy Reizenstein, Alexandru Cioba, Radosław Kowalski, Tim King, James Bell, Haichen Shi and William Kayat – to name but a few – thank you for your friendship and support throughout my time as an intern, and to the Alan Turing Institute for a wonderful experience. Indeed, the Institute is fantastic place to work, with so many bright minds working together. I hope to be able to visit again in the future.

To my mother, Denise, my father, William, and my brother, Alastair – thank you all for your unwavering support, love and encouragement. It was tough at times. But all three of you were always there for me. I am so happy I have been able to do you all proud. *There is another doctor in the family, now!* To Ian Phillips – you were the person who taught me everything I needed to know about computing at Standard Grade and Advanced Higher.<sup>2</sup> As my high school teacher, you were one of the people most conducive in helping

---

<sup>2</sup>Standard Grades and Advanced Highers were the Scottish qualifications that you obtained from school and/or college, although it seems as though everything has changed since I was a high school student.

me get to where I am today. Thank you so much for everything you did for me at Mearns Castle High School, and I hope I have made you and the School proud. I sometimes look back at my Advanced Higher project and question what on earth I was doing – I *think* this shows that I have come a long way with everything that I have learnt since 2008. I remember thinking the software I wrote back then as being so complex; it now seems so simple.

Finally, I wish to thank my supervisor, Dr. Leif Azzopardi. Leif has seen me through so much – from my undergraduate Honours and Masters projects, and now as my PhD supervisor. His advice, wisdom, expertise and friendship have been instrumental in getting me to the stage that I could actually write this thesis up. Leif gave me so many opportunities, bestowed faith and trust in me, and provided support and encouragement when I was feeling down. I cannot express thanks enough for everything Leif has done for me. *Thank you so much, Leif!* I hope we will work together again in the future.

Yes, the past four years have been an incredible journey. At times everything felt overwhelming – like what I was doing was simply impossible. But with the love and support from everyone mentioned above, I got through it, and have managed to produce an original piece of research. With that has come a newfound confidence in myself that I will be able to achieve my future goals, whatever they may be. I think producing this thesis is testament to that, and it is something that I am very proud of.

But do you know what?

*This is just the beginning!*

“Essentially, all models are wrong, *but some are useful.*”

(George E.P. Box, 1919–2013)

At least my model might be useful...





## Presentational Conventions

A number of different presentational conventions have been employed in this thesis. These are listed below for reference.

- Spelling is according to the *Oxford English Dictionary*; the version referred to is searchable online at <https://en.oxforddictionaries.com/>.
- *Italicised text* is used to define a term, but not thereafter. This applies to acronyms, where the full expansion is presented initially; associated abbreviations are used thereafter.
- Research questions and other key shorthand descriptions for components of this research are presented inline within a shaded box.
- Pseudo-code that is presented within this thesis uses the *HAGGIS* high-level reference programming language, as per Cutts et al. (2014). This is particularly relevant for a Scottish PhD!

This thesis is typeset in 12-point Palatino (body) and Foundry Sterling (headers) using *XeTeX* version 0.99998, using a custom style complying with University of Glasgow thesis regulations.



# Table of Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements and Reflections</b>	<b>vii</b>
<b>Presentational Conventions</b>	<b>xv</b>
<b>List of Figures</b>	<b>xxvii</b>
<b>List of Tables</b>	<b>xxix</b>
<b>Glossary and Acronyms</b>	<b>xxxiii</b>
<b>I Introduction and Background</b>	<b>2</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Motivation and Context . . . . .	5
1.1.1 Why Simulation? . . . . .	8
1.1.2 Why Stopping? . . . . .	11

1.2	Thesis Statement . . . . .	12
1.3	Overarching Research Questions . . . . .	12
1.4	Origins of the Material. . . . .	13
1.5	Thesis Contributions. . . . .	14
1.6	Remaining Thesis Outline . . . . .	15
<b>2</b>	<b>Information Retrieval</b>	<b>17</b>
2.1	A (Brief) History . . . . .	18
2.1.1	Libraries, Indexing and Punchcards . . . . .	18
2.1.2	The Rise of Computers . . . . .	19
2.1.3	The World Wide Web . . . . .	19
2.2	Information Retrieval Basics . . . . .	19
2.2.1	The Cranfield Paradigm . . . . .	19
2.3	Retrieval Models . . . . .	19
2.3.1	Boolean Retrieval . . . . .	19
2.3.2	Probabilistic Retrieval . . . . .	19
2.4	System/User Evaluation . . . . .	19
2.4.1	Precision-at-k . . . . .	19

2.4.2 Rank-Biased Precision . . . . .	19
2.4.3 INST . . . . .	19
2.5 Chapter Summary . . . . .	19
<b>3 Searching and Stopping</b>	<b>21</b>
3.1 User Studies . . . . .	22
3.2 Stopping Heuristics . . . . .	22
3.2.1 The Fixed Depth . . . . .	23
3.2.2 Cognitive Heuristics . . . . .	24
3.2.2.1 Tolerance to Non-Relevance . . . . .	24
3.2.2.2 Satiation Rules . . . . .	25
3.2.2.3 Difference . . . . .	25
3.2.2.4 The Mental List . . . . .	25
3.2.2.5 Representational Stability . . . . .	25
3.2.2.6 Magnitude Threshold. . . . .	25
3.2.2.7 The Single Criterion . . . . .	25
3.2.3 Time-Based . . . . .	25
3.2.4 Patch Rules. . . . .	25

3.3	Formal Theories of Interaction . . . . .	25
3.3.1	From Conceptual to Formal . . . . .	26
3.3.2	Information Foraging Theory . . . . .	28
3.3.3	Considering Costs and Benefits . . . . .	31
3.3.3.1	The Interactive Probability Ranking Principle. . . . .	32
3.3.3.2	Search Economic Theory . . . . .	34
3.4	Chapter Summary. . . . .	34
<b>II</b>	<b>Improving User Modelling in IIR</b>	<b>36</b>
<b>4</b>	<b>Existing Approaches to User Modelling</b>	<b>37</b>
4.1	Individual Components . . . . .	37
4.1.1	Query Generation. . . . .	37
4.1.2	Judging Relevancy . . . . .	38
4.1.3	Document Examination . . . . .	38
4.2	Entire Search Session . . . . .	38
4.2.1	TREC User . . . . .	38
4.2.2	Baskaya et al.. . . . .	38
4.2.3	Thomas et al.. . . . .	38

4.3 Chapter Summary . . . . .	38
<b>5 Advancing User Modelling in IIR</b>	<b>39</b>
5.1 The Basic User Model . . . . .	39
5.2 The Complex Searcher Model . . . . .	40
5.2.1 CSM, Mark I . . . . .	40
5.2.2 CSM, Mark II . . . . .	40
5.2.3 Considering State and Agency. . . . .	40
5.3 Evaluating Model Effectiveness . . . . .	41
5.3.1 Research Questions. . . . .	41
5.3.2 Experimental Design . . . . .	41
5.3.3 Results . . . . .	41
<b>III Stopping Behaviours and Context</b>	<b>42</b>
<b>6 The Effect of Temporal Delays on Stopping Behaviour</b>	<b>43</b>
6.1 Introduction. . . . .	44
6.2 The User Study . . . . .	44
6.2.1 Study Motivation and Design . . . . .	44
6.2.2 Results . . . . .	44

6.2.3	Conclusions . . . . .	44
6.3	Simulation Experiments . . . . .	44
6.3.1	Simulated Experimental Design . . . . .	44
6.3.2	Comparisons . . . . .	44
6.3.3	Conclusions . . . . .	44
6.4	Conclusions . . . . .	44
6.5	Chapter Summary . . . . .	44
<b>7</b>	<b>The Effects of Snippet Lengths on Stopping Behaviour</b>	<b>45</b>
7.1	Introduction. . . . .	45
7.2	The User Study . . . . .	46
7.2.1	SERP Layouts and Presentation . . . . .	47
7.2.2	Generating Snippet Text . . . . .	48
7.2.3	Results per Page . . . . .	49
7.2.4	Snippet Lengths: Longer or Shorter? . . . . .	50
7.2.5	Corpus, Search Topics and System. . . . .	53
7.2.6	Snippet Generation . . . . .	54
7.2.7	Behaviours Logged . . . . .	55



7.2.8	Capturing User Experiences . . . . .	56
7.2.9	Crowdsourced Subjects & Quality Control . . . . .	57
7.2.10	Search Behaviours. . . . .	60
7.2.11	User Experience. . . . .	63
7.2.12	Study Motivation and Design . . . . .	67
7.2.13	Results . . . . .	67
7.2.14	Conclusions . . . . .	67
7.3	Simulation Experiments . . . . .	67
7.3.1	Simulated Experimental Design . . . . .	67
7.3.2	Comparisons . . . . .	67
7.3.3	Conclusions . . . . .	67
7.4	Conclusions . . . . .	67
7.5	Chapter Summary. . . . .	67
<b>8</b>	<b>The Effect of Diversifying Results on Stopping Behaviour</b>	<b>69</b>
8.1	Introduction. . . . .	69
8.2	The User Study . . . . .	69
8.2.1	Study Motivation and Design . . . . .	71

8.2.2	Corpus and Search Topics . . . . .	73
8.2.3	Tasks: Aspectual and Ad-Hoc Retrieval . . . . .	74
8.2.4	Relevance Judgments and Aspects . . . . .	74
8.2.5	Systems: Non-Diversified and Diversified . . . . .	76
8.2.6	Experimental Procedure. . . . .	77
8.2.7	Recruitment and Controls. . . . .	77
8.2.8	Logging and Measures . . . . .	79
8.2.9	User Experience . . . . .	80
8.3	Results . . . . .	81
8.3.1	Observed Behaviours . . . . .	83
8.3.2	Gain over Time . . . . .	87
8.3.3	Results . . . . .	91
8.3.4	Conclusions . . . . .	91
8.4	Simulation Experiments . . . . .	91
8.4.1	Simulated Experimental Design . . . . .	91
8.4.2	Comparisons . . . . .	91
8.4.3	Conclusions . . . . .	91

8.5	Conclusions . . . . .	91
8.6	Chapter Summary . . . . .	91
<b>IV</b>	<b>Conclusions</b>	<b>92</b>
<b>9</b>	<b>Conclusions and Future Work</b>	<b>93</b>
9.1	Discussion and Contributions . . . . .	93
9.1.1	User Modelling . . . . .	93
9.1.2	Examining Stopping Behaviours . . . . .	93
9.2	Conclusions . . . . .	93
9.3	Future Research Directions . . . . .	93
	<b>Appendices</b>	<b>95</b>
<b>A</b>	<b>The SimIIR Framework</b>	<b>97</b>
A.1	Architecture Overview . . . . .	98
A.1.1	The Searcher Model . . . . .	98
A.1.2	Searcher Contexts . . . . .	98
A.1.3	Topics . . . . .	98
A.1.4	Search Interface/Engine . . . . .	98
A.1.5	Output Controller. . . . .	98

A.1.6	Querying Strategies/Generators . . . . .	98
A.1.7	SERP Level Stopping . . . . .	98
A.1.8	Snippet/Document Classifiers. . . . .	98
A.1.9	Snippet Level Stopping . . . . .	98
A.1.10	Loggers. . . . .	98
A.2	Example Input. . . . .	98
A.3	Example Output. . . . .	98
<b>B</b>	<b>Original Publications</b>	<b>99</b>
B.1	How Temporal Delays Affect Search Behaviour . . . . .	100
B.2	Fixed and Adaptive Stopping Strategies . . . . .	101
B.3	An Analysis of Stopping Rules and Strategies . . . . .	102
B.4	Building Realistic Simulations for Interactive Information Retrieval	103
B.5	SimIIR: A Framework for the Simulation of Interaction. . . . .	104
B.6	Agents, Simulated Users and Humans . . . . .	105
B.7	A Study of Snippet Length and Informativeness. . . . .	106
	<b>Bibliography</b>	<b>107</b>

# List of Figures

1.1 Racing Simulator Example . . . . .	10
2.1 Search Integration within Windows 10 . . . . .	19
3.1 Berry Picking. . . . .	27
3.2 Illustration of the Patch Model . . . . .	28
3.3 Plots of optimal patch stopping points for Information Foraging Theory. . . . .	31
4.1 Interaction model by Thomas et al. . . . .	38



## List of Tables

7.1 Characters, words and <i>Information Gain (IG)</i> across each of the four interface conditions. An ANOVA test reveals significant differences, with follow-up tests (refer to Section 8.3) showing that each condition is significantly different to others. There are clearly diminishing returns in information gain as snippet length increases. An IG value closer to zero denotes a higher level of IG. In the table, <i>IG/W.</i> denotes <i>IG per word</i> . . . . .	59
7.2 Summary table of both interaction and performance measures over each of the four interfaces evaluated. For each measure examined, no significant differences are reported across the four interfaces. . . . .	60
7.3 Summary table of times over each of the four interfaces evaluated. Significant differences exist between T0 and T4 (identified by the *, where $\alpha = 0.05$ ) on a follow-up Bonferroni test. . . . .	60
7.4 Table illustrating a summary of interaction probabilities over each of the four interfaces evaluated. Note the increasing trends for each probability from <b>T0</b> → <b>T4</b> (short to long snippets). Refer to Section 7.2.10 for an explanation of what each probability represents. . . . .	62

7.5 Summary table of the recorded observations for the post-task survey, indicating the preferences of subjects over six criteria and the four interfaces, where * indicates that <b>T0</b> was significantly different from the other conditions. In the table, <i>Conf.</i> represents <i>Confidence</i> , <i>Read.</i> represents <i>Readability</i> , <i>Inform.</i> represents <i>Informativeness</i> , and <i>Rel.</i> represents <i>Relevancy</i> . . . . .	63
--	----

7.6 Table presenting responses from the exit survey completed by subjects. The survey is discussed in Section 7.2.8. . . . .	64
--	----

8.1 Query statistics and performance measures across both of the experimental systems trialled, <b>ND</b> (Non-Diversified) and <b>D</b> (Diversified). Note the significant differences between the diversity-centric measures, $\alpha DCG$ (where $\alpha = 0.5$ ) and Aspectual Recall ( <b><i>Asp.R.</i></b> ), highlighting that the diversification algorithm did indeed provide a more diverse set of results to the subjects.. . . .	82
---	----

8.2 Behavioural and performance measures across each condition, system and task.. . . .	83
---	----

8.3 Interaction times across each condition, system and task. Included is: the mean total session time; the per query time ( <i>Per Q.</i> ); the per document time ( <i>Per D.</i> ); and the per result summary (snippet) time ( <i>Per Snip.</i> ). Results presented in seconds. . . . .	84
--	----



8.4 Interaction probabilities, as observed over the four experimental conditions. Refer to Section 8.2.8 for an explanation of each probability's meaning. . . . .	85
8.5 Table highlighting the fitting parameters for the gain curves illustrated in Figure ?? over each experimental condition. Also included are the estimations from the model for the time to examine a document, and the depth to which subjects should go -- as well as the observed number of documents examined, and stopping depth (on average). . . . .	90



# Glossary and Acronyms

**IIR**

The study of Interactive Information Retrieval

**IR**

The study of Information Retrieval

**World Wide Web**

**(WWW)** An information space in which documents and other resources, linked together via hypertext links, can be accessed via the Internet.



# Glossary and Acronyms

# Introduction and Background

*In this part, we provide an introduction to the thesis, present the overarching research questions, and provide an overview of the thesis structure. We also provide background to the problem, with a detailed literature review of the various techniques and components commonly used in Information Retrieval, with an emphasis on stopping.*

# Chapter 1

## Introduction

We live today in the so-called *Information Age*, an era of human history characterised by the rapid development of technology allowing for the creation, transmission and retrieval of large volumes of information. Two key technological developments that have permitted such an increase in information generation are the computer and the associated technologies that allow for near-instantaneous communications with devices all around the planet, including the *Internet* and World Wide Web (Berners-Lee et al., 1994). Indeed, with such technologies being essentially ubiquitous in today's society, humankind today generates something in the range of 2.5 *quintillion bytes* of information *per day*<sup>1</sup>, according to a recent *IBM* technical report<sup>2</sup>. To use a different unit of measurement, 250,000 times the amount of information held at the *US Library of Congress*<sup>3</sup>.

Sifting through such large volumes of information online to find the elusive and proverbial *needle in the haystack* has been an area of active research and development over a number of decades. Since the early 1990's, the World Wide Web has emerged as the dominant means of

---

<sup>1</sup>2.5 quintillion bytes = 2,500,000,000,000,000,000 bytes, or 2,500,000 terabytes.

<sup>2</sup><https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN> – last accessed November 23<sup>rd</sup>, 2017.

<sup>3</sup>This is calculated using the approximation on the US Library of Congress Blog, that says the Library holds the equivalent of 10 terabytes of information.

publishing information online over the Internet, replacing obsolete technologies such as the *Gopher* protocol. As the amount of information available on the World Wide Web grew<sup>4</sup>, so too did the paradigms employed by those wishing to seek information on it. Over the course of the first decade of the 21<sup>st</sup> century, our approach to exploring the offerings on the World Wide Web changed from the concept of *surfing* a particular domain by following a series of hyperlinks, to the *searching* of the ever-expanding universe of hyperlinked documents<sup>5</sup>. Indeed, developing effective *search engines* can be considered as the *raison d'être* of the study of the study of *Information Retrieval*.

*"...but perhaps the key technology that took the web from a useful supplement of current information practice to become the default communication medium is search."*

(Wilson et al. 2010)

Contemporary search engines such as *Google* and *Bing* are considered to offer an effective means of finding the proverbial needle in the haystack (Wilson et al., 2010), where near perfect accuracy is regularly attained for popular *queries* (Vaughan, 2004). These search engines, along with the many others in existence today under a variety of different contexts, are the product of the collective work undertaken in the field of IR – from early research in the 1970's (Cleverdon, 1962; Rijsbergen, 1979); to the development of the various retrieval models that we use (Robertson and Zaragoza, 2009); to the setting up of evaluation forums (Harman, 1993); to the development of the large-scale retrieval systems that we all use today (Baeza-Yates and Ribeiro-Neto, 1999; Wang et al., 2010).

The work that the field of IR collectively undertakes is all done in the interests of making it easier for potential users of search engines to satisfy their underlying *information need*. Arriving with an *anomalous state of knowledge* (Belkin, 1980), a searcher then formulates a *query* – an expression of what they are looking for (Borlund, 2003) – before being presented

---

<sup>4</sup>According to <http://www.internetlivestats.com/total-number-of-websites/>, there are over 800 million individual websites online today.

<sup>5</sup>Indeed, (Mcbryan, 1994) considered a search engine as a means of *taming* the large number of online documents.



with a potentially relevant set of documents. However, the interactions that take place between a searcher and a search engine are complex (Ingwersen and Järvelin, 2005). This process, where the searcher engages in *dialogue* with the search system, is considered the study of *Interactive Information Retrieval* (Borlund, 2003).

## 1.1 Motivation and Context

Central to much of the work undertaken in the field of IR over the past 50 years is the so-called *Cranfield Paradigm*, which denotes a standardised approach for the evaluation of IR systems. The Cranfield paradigm descends directly from *Cranfield II* (Cleverdon et al., 1966), a set of experiments designed to evaluate the efficiency of indexing systems. At the time, searching for information on computer systems was achieved through the issuance of queries to a *boolean retrieval system*, with terms matched against a small set of manually indexed documents (Harman, 2010). Included is a set of *relevance judgements* for each of the documents, as judged by humans, allowing one to then ascertain the performance of a given search system.

While the basic principles of the Cranfield Paradigm have remained in place since it was established in the 1960's, components of the approach have evolved over the years to cater for the ever increasing complexity of the tasks at hand (Harman, 2010). Indeed, the approach is still widely used in evaluation forums, such as the NIST-sponsored *Text REtrieval Conference (TREC)* – the first of which was held in 1993 (Harman, 1993). Indeed, many of the *relevance assessments* and *topics* provided as part of *TREC Tracks* over the years are used as ground truths throughout the work discussed in this thesis.

The approach however can be argued to remain simplistic in terms of when considering the complex user interactions that take place during the search process (Borlund, 2000; Ingwersen and Järvelin, 2005). In other words, the Cranfield Paradigm broadly fails to consider the complexities of the IIR process, where, for example, searchers can issue multiple queries

## 1.1 Motivation and Context

during the course of a search session, and adapt their interactions based upon the perceived quality of the presented rank list of results for each associated query (Moffat et al., 2013). Selecting good terms to use within a query is difficult yet important (Efthimiadis, 2000); as such, the initial query posed in a search session often acts as an entry to the search system, followed by phases of browsing and query reformulations (Marchionini et al., 1993). Therefore, the first query formulation of the user often acts as an entry to the search system followed by subsequent phases of browsing and query reformulations (Marchionini et al. 1993). Searchers also will typically abide by the principle of least effort – striving to minimise the probable average rate of work expenditure over time (Zipf, 1949). Cranfield considers that a user will (i) issue a single query; (ii) examine documents to a large depth (typically 1,000 documents); and (iii) consider all documents to be relevant. This is largely unrealistic, and numerous researchers have proposed alternatives to the Cranfield Paradigm, such as Borlund, 2003.

Keskustalo et al., 2008 categorised IIR research into four different approaches that allow for the consideration of the complex interactions that take place between a human and the search engine being used. These are:

- 1 the observation of real-world searchers, in real world scenarios (e.g. general web search), through the use of interaction logs;
- 2 the observation of real-world searchers undertaking simulated work tasks in a lab-based environment;
- 3 performing *simulations* of interaction in a lab-based environment, sans real-world searchers; and
- 4 the undertaking of *traditional* lab-based experimentation.

Obtaining interaction data from studies utilising real-world users undertaking search tasks (categories 1 and 2) will of course always be the preferred option. However, there are

pitfalls with both approaches that must be considered, primarily in terms of *availability* and *cost*. Obtaining data for studies conducted in category **1** is difficult if the researchers do not work for an organisation offering a large-scale, search engine. Working within an academic setting for example may greatly restrict what data can be obtained. Indeed, working with real-world interaction data also leads to major ethical and privacy concerns (Korolova et al., 2009). The release of the *AOL Query Log* (and subsequent fallout) in August 2006 is testament to that, although this has not stopped researchers from utilising the data in their work (e.g. Brenes and Gayo-Avello, 2009).

This will leave many researchers with category **2**. While this approach also leads to the capturing of real-world interaction data, pitfalls of this approach primarily are the significant costs involved in such an approach – both financially and in terms of time. Considerations must also be placed into study design to mitigate potential biases as much as possible. In recent years however, the concept of *crowdsourcing* may alleviate some of these concerns, and open up a potential study to a larger potential audience. A recent study has shown that using crowdsourcing to capture interaction data is no worse than a carefully controlled lab-based user study (Zuccon et al., 2013), although quality control measures must be taken.

While categories **1** and **2** provide real-world interaction data, options **3** and **4** do not. Such approaches however, if executed correctly, can potentially lead to insights that would not otherwise be possible when involving real-world users. As an example, covering an extensive set of test cases may simply not be possible (Keskustalo et al., 2008) – perhaps due to financial constraints, or a lack of suitable subjects. Category **4** can be considered as a means of conducting *traditional, TREC-style* IR lab experimentation that is, as previously mentioned, largely naïve of the user’s complex interactions. This leaves category **3** as a means of conducting and evaluating user-sided experimentation without the explicit need for real-world users to be present. This approach uses *simulation* as a means to conducting such experiments.

## 1.1 Motivation and Context

### 1.1.1 Why Simulation?

Simulation is defined as the *imitation of the operation of a real-world process or system over time* (Banks et al., 1996). Such an approach allows one to gain insight into the functioning of some real-world phenomenon. Simulation has been used in a wide range of areas, including, for example, examining physical processes (Haessig and Friedland, 1991), psychology (Hastie, 1988), road traffic (Mahmud and Town, 2016) and training for various activities, such as piloting an aeroplane (Sparko et al., 2010). Central to contemporary uses of simulation is the idea *computerised simulation* (Heermann, 1990), thanks to ever increasing computational power available for such tasks – such as the simulation of racing cars for the purposes of driver development, as shown in Figure 1.1. In summary, employing simulation provides a rapid means of exploring the components, all at a low cost – all while permitting repeatable, and therefore reproducible, results (Maxwell and Azzopardi, 2016a).

One of the key components of any simulation – regardless of whether it is executed on a computer or not – is that of an underlying *model* of the real-world phenomenon being simulated (Tocher, 1963). This model defines the various stages, rules and other descriptive components of the phenomenon that simulations must consider. These rules are often high level, and as such, a number of *assumptions* are made (Tocher, 1963). For example, the TREC Paradigm assumes that a single query is issued by a searcher within a search session, and searchers examine to great depths per query. These assumptions should be made in the face of supporting evidence; evidence in IIR studies strongly suggests that searchers do not follow this rigid approach, but rather adapt their behaviour based upon the proximal cues presented to them.

Why though, use simulation when other alternatives to modelling the search process are available? Alternatives include some form of closed-form system, such as a series of linear equations to describe the complex interactions that take place. However, according to Fishwick, 1995, this approach is not flexible enough, and lists a number of reasons as to why

simulation is essential in complex, dynamic systems:

- the model is very complex, with many variables and interacting components;
- the underlying variables and relationships are non-linear;
- the underlying models contains random variates; and
- the model output is to be visual, as in a three-dimensional computer animation.

In the context of IIR, the first three reasons can be considered as acceptable reasons for why simulation is an advantageous methodology to pursue. For example, many state-of-the-art IIR models consider a stochastic component when determining the relevancy of a document to a particular topic.

Simulation provides a means of using a uniform model execution technique that can be used to solve a large variety of systems, without resorting to a “bag of tricks”, where one must choose special-purpose and sometimes arcane solutions to avoid simulation (Fishwick, 1995). Simulation provides the freedom and flexibility to permit the implementation of a model that better represents the real-world phenomenon that is being considered. In contrast, with a more closed-form approach, the underlying model that is created is often twisted and altered to suite the closed-form approach, rather than to actually represent the real-world system. This leads to a larger gap between the model and reality, and as such leads to a greater number of assumptions within the model than what would otherwise be required. In other words, the technique used to develop the model constrains just how realistic is can be.

As a means to providing better and more flexible models, simulation has been used extensively within IR, such as the simulation of work tasks and tracks, simulated and synthetic data collections, and of course *simulated interaction*, which is what this thesis considers. While this encapsulated category **3** of the four categories as outlined by Keskustalo et al.,

## 1.1 Motivation and Context



**Figure 1.1:** An example of a *video game simulation*, *Assetto Corsa*. In this figure is a model of the McLaren-Mercedes MP4/13 race car, driving around the Autodromo Enzo e Dino Ferrari race track in the Emilia-Romagna region of Italy. In the past ten years, the increase in computing power in CPUs and GPUs has enabled the development of increasingly complex (and *more realistic*) video game simulations, which, for example, are now used in the training of racing drivers.

2008, we also argue that in order to run simulations of interaction, they must be *grounded* using real-world observations, which in turn means the assumptions that are made in the simulated model can be considered to be a credible abstraction of the real-world phenomenon. As such, this necessitates access to data obtained through either categories **1** or **2** – and without access to a large-scale search engine, this leaves category **2** as means for acquiring such data. As such, this thesis presents the *Complex Searcher Model (CSM)* as a means of modelling the IIR process, with each component grounded using interaction data from a number of user studies, also conducted as part of this thesis. Of particular interest within this model is the concept of *stopping behaviour* (e.g. *how far down this ranked list of results should I go?*).



### 1.1.2 Why Stopping?

Knowing when to stop is a fundamental aspect of human thinking, with individuals commonly employing some form of *stopping criterion* to decide when they should stop with their interactions in the world around them (Nickles, 1995). As an example, a shopper looking to buy a new smartphone will stop shopping around once he or she has obtained sufficient information on what new device to purchase. A doctor, once their case notes about a patient's condition have been finished, will then diagnose their ailment. In the context of search, stopping may be considered at a variety of different points during the search process. The commonly used example of *search stopping behaviour* is the point at which a searcher should stop examining a list of ranked results, or, in other words, *how far down the ranked list the searcher should go*, for example.

The decision of when to stop is not necessarily due to external factors, but from a series of *internal factors* of the decision maker's thinking process. In the context of informational search, knowing when to stop requires that the individual makes a judgement regarding the sufficiency of the information obtained, and whether or not additional information is required to be obtained (Browne and Pitts, 2004). This is normally characterised by both the completeness and correctness of the information obtained thus far (Smith et al., 1991). These claims can be mirrored by qualitative studies on examining stopping behaviour, where researchers have found that searchers stop examining a ranked list of results simply because what they have found previously is "*good enough*" (Wu et al., 2014).

Considering the above, is there any means by which we can quantify what this feeling of "*good enough*" actually is? Researchers have devised a series of different *stopping heuristics* as a means to try and encapsulate the differing stopping behaviours exhibited by searchers. However, the literature in examining which of these heuristics offers the best approximations to what searchers actually do is somewhat limited. By focusing on stopping behaviour within this thesis, our model provides additional points at which a simulated user can stop

## 1.2 Thesis Statement

interactions with a search engine, and thus save time and effort that might otherwise have been wasted if they continued to examine content.

## 1.2 Thesis Statement

The statement of this thesis is that by considering various *stopping decision points* within model representing a user's interactions during the search process, one can run simulations of interaction that offer a greater degree of realism – and thus a better approximation of the actual behaviours exhibited by real-world searchers – than currently employed models.

In particular, incorporating such stopping decision points within a representation of the search process will allow the simulated user following such a model a greater degree of flexibility – allowing them, for example, to abandon a set of results that is judged to be of low relevancy to the given query. In addition, these decision points can be instantiated by operationalising a series of different *stopping heuristics* that attempt to encapsulate the stopping behaviours exhibited by real-world searchers. Furthermore, by taking this knowledge forward, we can then apply grounded simulations over a variety of different search contexts, examining how stopping behaviour of real-world searchers varies under each context – and through simulation, we can then deduce what particular heuristic(s) offer the best approximation to searcher behaviours.

## 1.3 Overarching Research Questions

From the introductory remarks, motivation and thesis statement provided previously in this chapter, we can now formulate the three main research questions that this thesis addresses.

**HL-RQ1** Considering stopping behaviours, can we improve upon and make *more realistic models* of the IIR search process – as a whole?



**HL-RQ2** How can we operationalise and subsequently implement more realistic stopping strategies for use within many of the commonly used IR and IIR models and measures that are in use today?

**HL-RQ3** How do the real-world stopping behaviours of searchers vary under different contexts?

## 1.4 Origins of the Material

Material presented in this thesis has appeared previously in several conference papers. All of these papers were published throughout the duration of this PhD programme (2013-2018). The list below details each of the publications in chronological order.

- Maxwell, D. and Azzopardi, L. (2014). Stuck in traffic: How temporal delays affect search behaviour. In *Proc. 5<sup>th</sup> IliX*, pages 155–164
- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015a). An initial investigation into fixed and adaptive stopping strategies. In *Proc. 38<sup>th</sup> ACM SIGIR*, pages 903–906
- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015b). Searching and stopping: An analysis of stopping rules and strategies. In *Proc. 24<sup>th</sup> ACM CIKM*, pages 313–322
- Maxwell, D. (2016). Building realistic simulations for interactive information retrieval. In *Proc. 1<sup>st</sup> ACM CHIIR*, pages 357–359
- Maxwell, D. and Azzopardi, L. (2016b). Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proc. 39<sup>th</sup> ACM SIGIR*, pages 1141–1144

## 1.5 Thesis Contributions

- Maxwell, D. and Azzopardi, L. (2016a). Agents, simulated users and humans: An analysis of performance and behaviour. In *Proc. 25<sup>th</sup> ACM CIKM*, pages 731–740
- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proc. 40<sup>th</sup> ACM SIGIR*, pages 135–144

## 1.5 Thesis Contributions

This thesis offers four main contributions to the research community, revolving around the concepts of user modelling and stopping behaviours during search.

**C1** We provide a series of contributions on user modelling within the IIR process. Specifically, we focus on the development of accepted user models by incorporating additional **stopping decision points**, and incorporating the ability for a simulated user to remember what has been examined through the inclusion of some basic **form of state**.

**C2** The second main contribution is the **SimIIR framework**, developed to allow for the running of **simulations of interaction**. Within the framework, our proposed user model is encoded, along with the ability to specify and configure various components of the search process. An explanation of the framework is provided in Appendix A.

**C3** We provide a comprehensive survey on various **stopping heuristics** that have been proposed over the years in the literature, before providing analysis on how performance and behavioural characteristics vary when considering:

- Snippet-Level stopping and associated strategies;
- SERP-Level stopping; and
- Session-Level stopping.

- C4** Finally, we provide analysis examining how the **stopping behaviour of searchers varies under different search contexts**, such as when we vary the overall search goal, or vary presentational aspects of the presented results.

## 1.6 Remaining Thesis Outline



## Chapter 2

# Information Retrieval: History and Background

*“Information Retrieval deals with the representation, storage, organisation of and access to information items.”*  
(Baeza-Yates and Ribeiro-Neto, 1999)

Without the availability of retrieval systems today, finding information would be a much more difficult task. Considering how easy the systems that we use on a daily basis are so straightforward to use, one might be forgiven in thinking that they were easy to develop.

In reality, this could not be further from the truth. Central to how these retrieval systems operate is the work that has been undertaken in the field of IR over a number of decades. In this chapter, we provide a provide a brief summary of the field of IR – from the first experiments up to the present day (Section 2.1) – and then examine the key concepts that are assumed within IR systems and experimentation today. From then, we begin to discuss the different models and measures that are commonly used within the field to evaluate the effectiveness of systems and users, all with an emphasis on how one can consider search stopping within them.

## 2.1 A (Brief) History

## 2.1 A (Brief) History of Information Retrieval

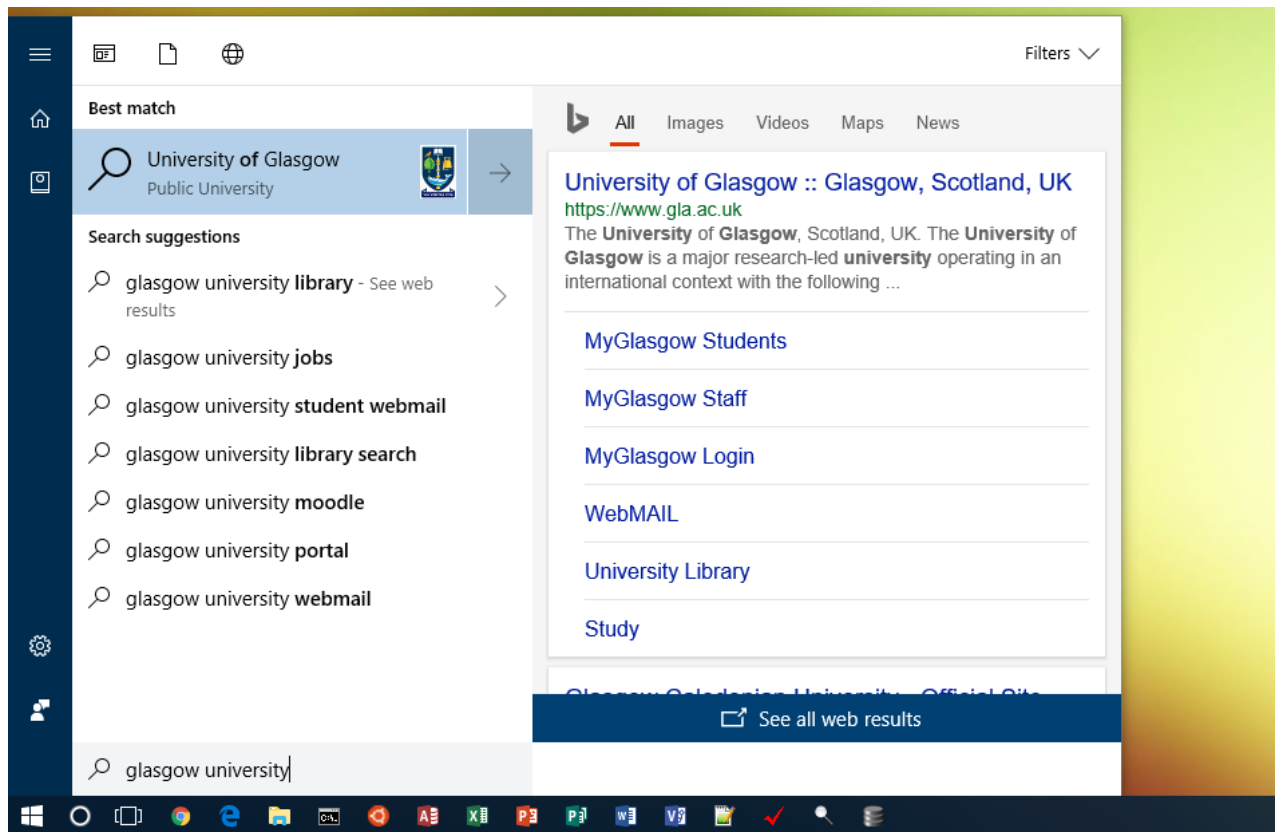
Given the abundance of contemporary, computer-based technologies readily available in our lives today, one might be forgiven in thinking that the humble *book* is in itself a technological development. The book provides us with the means of storing knowledge and information, with a book's ability to preserve information through the ages not going unnoticed (Hedstrom, 1997). With approximately 130 million unique titles in existence as of 2010<sup>1</sup>, individuals would access a *library* to acquire the book(s) – and subsequent knowledge that they possess.

### 2.1.1 Libraries, Indexing and Punchcards

The field of IR has transformed the way in which we access information. Back in the early 20<sup>th</sup> century, an individual would access books in their local library, with books organised by means of a classification system. A popular example of such a system is the *Dewey Decimal System* (Dewey, 1891). By examining the classification system, an individual would then be able to navigate themselves to the relevant area of the library floor to find their book. Such an approach – while acceptable at the time – was an inconvenience. Due to the comparatively high costs involved, information seekers would limit themselves to perhaps a small number of questions (Sanderson and Croft, 2012).

---

<sup>1</sup>This figure is estimated and presented on the *Google Books Search blog*, now archived, at <http://booksearch.blogspot.co.uk/>.



**Figure 2.1:** An illustrative example of how search has become integrated within computer operating systems. In this screenshot, the query `glasgow university` has been issued to *Bing*, through the *Microsoft's Windows 10* desktop interface. This search interface allows both searching locally for files on disk, and for external web search queries.

### 2.1.2 The Rise of Computers

### 2.1.3 The World Wide Web

## 2.2 Information Retrieval Basics

### 2.2.1 The Cranfield Paradigm

## 2.3 Retrieval Models

### 2.3.1 Boolean Retrieval

### 2.3.2 Probabilistic Retrieval





## Chapter 3

# Searching and Stopping: A Background

Having established the main focus of this thesis, this chapter provides a detailed overview of the current state of the literature in the context of stopping during search, and the varying ways in which it has been examined and subsequently modelled. We examine this from two main perspectives, considering:

- the **conceptual and descriptive** approaches that have been undertaken; and
- the **formalised** approaches that have been examined.

Conceptual and descriptive approaches generally consider a high-level approach of stopping. Approaches taken, as described in this chapter, include a series of *stopping heuristics and rules* (Section 3.2) and a series of user studies examining the phenomenon and the wider *Information Seeking and Retrieval (ISR)* process (Kelly and Sugimoto, 2013), as detailed in Section 3.1. These approaches generally provide a picture of the domain, with this information subsequently allowing us to develop more formalised approaches (Azzopardi and Zuccon, 2015).

### 3.1 User Studies

In turn, formalised approaches are more precise, building upon the knowledge and understanding obtained from the conceptual and descriptive approaches. Generally, a formalised approach enumerates each aspect of the phenomena as a variable, allowing one to explore how varying each variable functionally relates to others. In the context of examining one's stopping behaviour, a number of different formalised ISR models have been proposed, which we discuss in relation to stopping in Section 3.3.

### 3.1 User Studies

what have studies shown that examine stopping behaviour?

Wu et al. (2014)

Dostert and Kelly (2009)

Toms and Freund (2009)

Zach (2005)

Wu and Kelly (2014)

Marchionini (1995)

### 3.2 Stopping Heuristics

Despite the inherent difficulties that have been observed when ascertaining how and when people stop searching, researchers have over the years proposed a number of different stopping heuristics which are *believed* to provide a more concrete means of quantifying and / or explaining when searchers decide to stop, and thus quantifying the sense of what is “*good enough*” (Wu et al., 2014).

A majority of these heuristics can be found within IR research, with the earliest having been defined as far back as the early 1970's. These heuristics are generally high level, *conceptual*

approaches that describe when searching should stop, at different levels (e.g. the session level, or the query level). To describe each of the heuristics, we break this section up into a number of different subsections, which provides a loose classification of the different approaches that have been described in the literature. We consider first:

- a *fixed-depth* approach, considered as the de facto stopping heuristic used in much IR research today.

From this, a series of more *adaptive* heuristics are defined in the literature which we consider in turn. These are:

- tolerance to non-relevance;
- satiation rules;
- difference rules;
- time-based rules; and
- patch-based rules.

### 3.2.1 The Fixed Depth

Precision-at-k is the classic example here. You will go to a fixed depth always. Assumed by the Cranfield paradigm, for example.

unrealistic. so what other approaches are there in the literature? a number of different rules are explained in the literature.

## 3.2 Stopping Heuristics

### 3.2.2 Cognitive Heuristics

By far the largest category of stopping heuristics defined within the literature are *cognitive based* heuristics. That is to say, heuristics that attempt to model in some form of criterion or criteria that are met in the mind of the searcher as he or she examines content to determine what is and what is not relevant.

#### 3.2.2.1 Tolerance to Non-Relevance

Cooper worked on two studies (Cooper, 1973a,b) that argued that the best way in which to evaluate a retrieval system would be to elicit subjective estimates of the system's utility to its users.

It was argued in Part I (see JASIS, March-April 1973 p. 87) that the best way to evaluate a retrieval system is, in principle at least, to elicit subjective estimates of the system's utility to its users, quantified in terms of the numbers of utiles (e.g. dollars) they would have been willing to give up in exchange for the privilege of using the system; and a naive methodology was outlined for evaluating retrieval systems on this basis. But the impracticality of the naive evaluation procedure as it stands raises the questions: How can one decide which practical measure is likely to yield results most closely resembling those of the naive methodology? And how can one tell whether the resemblance is close enough to make applying the measure worth while? In the present paper two kinds of solution to these problems are taken up. The first answers the questions in terms of the reasonableness of the simplifying assumptions needed to get from the naive measure to the proposed substitute. The second answers it by experimentation.

Cooper (1973a)

Cooper (1973b)

cooper. pretty much the same rule was defined by nickles

considers a searcher's tolerance to non-relevance.

#### 3.2.2.2 Satiation Rules

#### 3.2.2.3 Difference

#### 3.2.2.4 The Mental List

#### 3.2.2.5 Representational Stability

#### 3.2.2.6 Magnitude Threshold

#### 3.2.2.7 The Single Criterion

#### 3.2.3 Time-Based

#### 3.2.4 Patch Rules

### 3.3 Formal Theories of Interaction

Understanding the complex behaviours exhibited by searchers during the ISR process has been a longstanding problem (Azzopardi and Zuccon, 2015). With most approaches considering the direct observation of searcher behaviours through interaction log data as described previously in Section 3.1, several attempts have been made to provide mathematically grounded, formalised theories that describe this process. From these theories, one can

### 3.3 Formal Theories of Interaction

then begin to deduce a series of testable hypotheses when undertaking interaction studies, such as the hypothesis provided in Chapter 7, and as published in Maxwell et al. (2017).

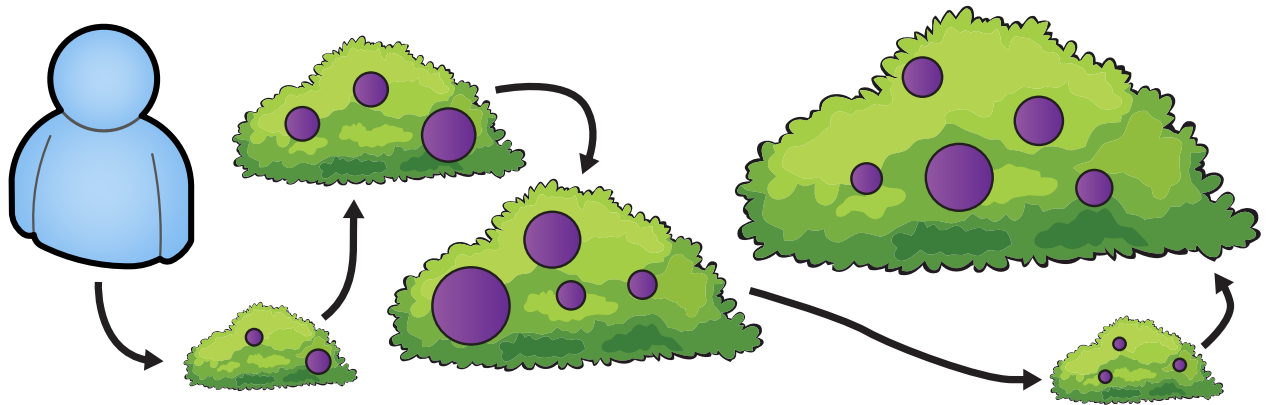
We discuss in this section three main competing ISR theories, all of which provide a rational explanation of *when searchers should stop*.

- **Information Foraging Theory (IFT)** (Pirolli and Card, 1999), which considers a searcher *foraging* for information to be analogous to an animal foraging for food in the wild – allowing for the inclusion of many theoretical approaches used in ecology to be applied to search (Section 3.3.2);
- the **Interactive Probability Ranking Principle (iPRP)** (Fuhr, 2008), allowing for the modelling of a searcher in making decisions, or *choices*, during an interactive search session (Section 3.3.3.1); and
- **Search Economic Theory (SET)** (Azzopardi, 2011), where the search interactions that take place between a searcher and the computer is modelled as an economics problem (Section 3.3.3.2).

Paradoxically, an astute way in which to introduce these formal models is to begin by discussing a *conceptual model* – the *Berry Picking Model*.

#### 3.3.1 From Conceptual to Formal

As illustrated in Figure 3.1, the Berry Picking Model (Bates, 1989) draws an analogy between a searcher and a *forager* – in this case, a forager looking for berries. The forager attempts to collect the ripest berries within a series of different *patches* (refer to Section 3.3.2 for more information on the *patch model*). This is crudely illustrated in Figure 3.1, where the forager moves between bushes (patches). Once the ripe berries have been collected from a patch, the forager moves to the next patch, until all patches have been exhausted. The analogy

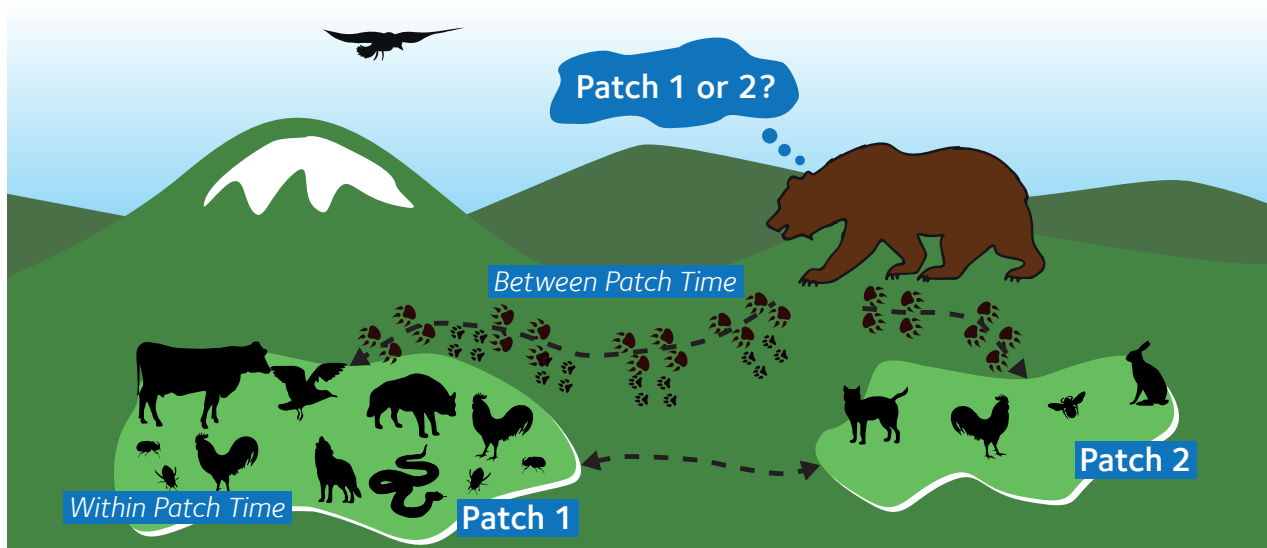


**Figure 3.1:** A crude illustration depicting *Bates' Berry Picking model* (Bates, 1989). In this conceptual model, a berry picker forages through a series of *patches* (refer to Section 3.3.2 for more information; patches are represented in the illustration as bushes) in search of the ripest berries (a simple assumption would be to assume larger berries are riper). This model mirrors how an individual searches with an evolving information need.

here regarding search considers a patch as a SERP, with a forager here attempting to look for the most relevant documents to their information need. As stated by Bates, the forager's information need evolves, and so the type of information they attempt to find valuable at a given point also changes. Switching back to the berry picking analogy, this may mean a switch between blueberries and strawberries, for example.

While this conceptual model is intuitive and subsequently easy to understand, the model lacks the ability to describe how long a forager will spend in an individual patch, or how long the time taken to reach a patch in the first place will affect their behaviour (Azzopardi and Zuccon, 2015). The lack of explanation is where more formalised approaches attempt to provide an answer. For example, researchers took forward the concept of a searcher as a forager (e.g. Russell et al. (1993); Sandstrom (1994)), with their results suggesting that *Optimal Foraging Theory (OFT)* (Stephens and Krebs, 1986) could be used to model the search process. This subsequently gave rise to the theory of Information Foraging Theory (Pirolli and Card, 1999).

### 3.3 Formal Theories of Interaction



**Figure 3.2:** A graphical representation of various models that form Information Foraging Theory. Should the forager expend effort navigating to *Patch 1*, or *Patch 2*? *Patch 1* is further away, with a stronger scent, and offers more potential gain (food). However, *Patch 2* is closer. IFT provides a framework for addressing these issues. Also shown within the illustration are the *between patch* and *within patch* times that are spent by the forager. Refer to Section 3.3.2 for an explanation on what these times represent. Silhouettes acquired from *freepik.com*.

#### 3.3.2 Information Foraging Theory

Analogous to an organism foraging for food in the wild, Information Foraging Theory considers a searcher as an *informavore*<sup>1</sup>, an organism that consumes *information*. IFT is comprised of a *ternion* of underlying models Pirolli and Card (1999), which are explained below. A graphical illustration of the three models can also be seen in Figure 3.2.

- The **Information Scent model** considers how informavores rely upon various *proximal cues* (Chi et al., 2001) (e.g. bolded terms, other textual snippets, or graphics within *information cards* – see Bota et al. (2016), for example) to indicate how promising a particular patch (e.g. a document) looks to be in terms of satisfying their information need. This is analogous to an organism foraging for food; organisms rely upon various cues in the surrounding environment (e.g. paw prints, smells) to guide them to a

<sup>1</sup>The term *informavore* was originally coined by Miller. “Just as the body survives by ingesting negative entropy, so the mind survives by ingesting information. In a very general sense, all higher organisms are informavores.”



promising patch. Both informavores and herbivores / carnivores estimate how beneficial following a particular path, or *scent trail*, will be<sup>2</sup>. Once the scent starts to weaken (i.e. when no more additional information is expected to be found), a informavore will stop and progress to a different scent trail (Piorkowski et al., 2012).

- The *Information Diet model* – when considering a webpage as a patch and information as the prey – provides a rationale for determining which information a forager will consume. Leaving a particular website may be straightforward, but finding a better one is not necessarily so – although it can be argued that in recent years, the advancement of search engine technology has made this issue to be less of a challenge (Vaughan, 2004). If it for example is easier for a forager to find lots of potentially useful webpages, there is less incentive for them to stay on a single page; technological developments today encourage the consumption of small chunks of information from a high number of sources, and, as such, a change in our behavioural characteristics.<sup>3</sup>
- Finally, *Information Patch model* concerns how long a forager will stay in a particular *patch* (e.g. an area of land, or, in terms of an informavore, a list of ranked documents, for example) before deciding to move to a new patch.

We consider the Information Patch model as of particular relevance to the work in this thesis, as it concerns when a forager will *stop* examining a given patch. Given the illustration in Figure 3.3, the analogy for an information seeker – or informavore – is that:

- *moving between a patch* is like *issuing a new query*, and the subsequent query formulation cost that must be expended; and

---

<sup>2</sup>A discussion of this explanation can be found by Jakob Nielsen at <https://www.nngroup.com/articles/information-scent/> – last accessed December 7<sup>th</sup>, 2017.

<sup>3</sup>The idea of technology changing our behavioural characteristics is not new; an article in the *New York Times* suggests that the development of technology has made us more impatient, expecting instant answers – and more forgetful, too, in the sense that the relative cost of accessing information is now so low, we readily discard information. The article is available at <http://www.nytimes.com/2010/06/07/technology/07brainside.html>, last accessed December 7<sup>th</sup>, 2017. Refer also to Carr (2008).

### 3.3 Formal Theories of Interaction

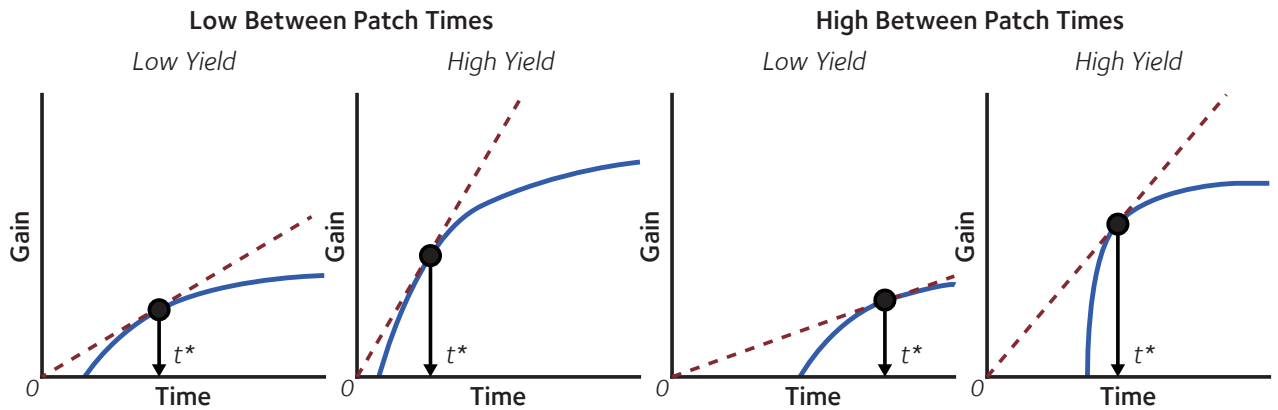
- *staying within a patch* is analogous to *assessing a series of documents*, where each document incurs an examination cost.

With these assumptions in hand, the Information Patch model allows for the prediction of how long a forager should stay in a patch before moving to the next patch. As IFT is based upon Optimal Foraging Theory (OFT) (Stephens and Krebs, 1986), two main assumptions about the forager are made, in that, unsurprisingly, they will act in an optimal fashion:

- the forager will *always enter the patch with the highest potential yield first*; and
- the forager will *maximise their gain per unit of time* spent in a patch (Pirolli and Card, 1999; Stephens and Krebs, 1986).

When considering the sample illustration in Figure 3.2, this therefore means that the forager will enter *Patch 1* first, as this patch offers the largest return for the energy the forager initially expends reaching the patch. When considering the stopping behaviour exhibited by a forager, one needs to calculate the gain attained at a given point in time, or  $g(t)$ . From this, the moment at which a forager should optimally stop examining a patch is the point in time at which the maximal value of gain per unit of time is reached. This is dependent upon a number of factors, including aspects such as the *between patch* and *within patch* times – the times spent getting to a patch, and spent within a patch, examining content, respectively.

Figure 3.3 provides two plots to demonstrate the theory in action, demonstrating on the left a gain. The illustration shows four gain curves that are for individual patches, each under different conditions. We show the difference in stopping times between patches with both a low and high between-patch time (i.e. a longer time to enter the patch from zero time), and patches which are fruitful, giving a high yield, and those with a low yield. By taking the gain curves and drawing the tangent to the curve through the origin, we can then see that the optimal stopping point as dictated by IFT is the point on the gain curve where its tangent intersects with it. This is the point of the *maximal rate of gain*; foragers spending time



**Figure 3.3:** Four plots, each denoting a different scenario for the optimal stopping point within a patch, as outlined by *Information Foraging Theory*. The two plots on the left denote a low between-patch time, with the two right plots illustrating a high between patch time. The *Low Yield* plots denote smaller levels of gain per unit of time compared to the *High Yield* patches. Note the optimal stopping point is denoted by  $t^*$ , the point at which the tangent of the gain curve intersects. Time spent within the patch after this point will result in steadily diminishing returns.

within the given patch after this point will experience continually diminishing returns. IFT suggests that after this point, a forager should abandon the patch, and then proceed to issue a new query, and subsequently enter a new patch.

### 3.3.3 Considering Costs and Benefits

The other means by which information seeking has been formally modelled is through the consideration of the *costs* and *benefits* that a searcher must consider during the search process. Although not discussed in the original Berry Picking model, Bates does discuss the idea that these should be weighed up in subsequent work (Bates, 1979).

The idea of using a cost/benefit analysis is not new; a number of other formal models have been created using this underlying approach. In IR research for example, purchasing decisions and ranking have been considered using this approach. A more user-centric approach was undertaken by Cooper, who modelled the trade-off between how long a searcher should spend searching, and how much time the system should itself spend searching. More related to the work in this thesis is the *Probability Ranking Principle (PRP)* by Robert-

### 3.3 Formal Theories of Interaction

son, who proposed a formalised model utilising decision theory to the so-called *ranking problem* (Robertson, 1977). The PRP essentially formed a theoretical foundation for optimising the results of ad-hoc retrieval. The PRP itself has in past decade been extended to yield the *Interactive Probability Ranking Principle* (Fuhr, 2008), as discussed in Section 3.3.3.1 below.

#### 3.3.3.1 The Interactive Probability Ranking Principle

With the PRP considering only the ranking problem, the notion of user interactions during ad-hoc retrieval are largely ignored by this theory. Systems have been shown in previous studies to perform differently in in a standard retrieval setting, of a single query and ad-hoc retrieval (e.g. Voorhees and Harman (2000); Turpin and Scholer (2006)). Using the PRP, this approach is indistinguishable from an interactive setting. Studies have also shown that scanning through a list of ranked documents to identify potentially relevant entries may not be the most crucial activity in the IIR process (Turpin and Hersh, 2001). As such, the iPRP provides an extension to the classical PRP, providing a means of formalising various activities within a search session besides simply scanning a ranked list of results.

Based upon the preexisting PRP, the iPRP removes two key assumptions that the PRP makes to increase the flexibility of the underlying model. These changes include the removal of the assumptions that consider:

- a *fixed information need*, allowing a user modelled by the iPRP to have an information need that changes and adapts as documents are examined (e.g. as per the *Anomalous State of Knowledge (ASK)*, described by Belkin); and
- the assumption that the *relevance of documents is independent of other documents*.

While sensible assumptions to make for the purposes of modelling, it has been demonstrated that the assumptions do under certain circumstances break down (Gordon and

Lenk, 1991). From these updated assumptions, three main requirements were specified, namely that iPRP must:

1. consider the interaction process as a whole, rather than simply considering document ranking as was assumed in the classical PRP;
2. allow for different of costs and benefits to be invested and given respectively, meaning that, for example, a longer document will take a greater period of time to examine than a shorter document; and
3. allow for the changing of the information need throughout the course of the search session, à la Belkin (1980).

This last point is reminiscent of Bates' Berry Picking model. As outlined in Section 3.3.1, a forager would traverse through patches/bushes in order to find the ripest berries available. However, as they forage, they may find, for example, certain berries to be riper than others. As such, they adjust what they are foraging for as they acquire more berries. In terms of an information seeker, this is analogous of a searcher's mental model of a particular topic continually evolving as they are subjected to new information through each document they examine. As such, one's information need may change as new information is uncovered.

From these requirements, the iPRP was then created with four revised assumptions, namely that:

1. there should be a focus on the function level of interaction, meaning that there are a variety of different activities one can undertake (e.g. query formulation, document examination), with each activity having a cost and benefit associated with it;
2. *decisions* forming the basis of interaction;
3. the searcher evaluating the *choices* laid before him/her in a *linear order* (i.e. when examining a list of ranked results, starting from the top and working your way down); and

### 3.4 Chapter Summary

4. only decisions that are positive and correct are of benefit to the searcher.

#### 3.3.3.2 Search Economic Theory

The origins of SET can be traced back to work by Varian (1999), who outlined three directions in which

### 3.4 Chapter Summary

- there has been lots of conceptual work on trying to understand stopping behaviours, but most boil down to the fact that people stop because what they have found is “good enough”.
- despite this, we have all these different heuristics that researchers define — primarily mental rule that searchers tick off as they go through results. And once a criterion/some criteria have been satisfied, they stop.
- and now we have theories of information seeking behaviour which also provide us with a mathematical means for trying to deduce when people stop. These are themselves stopping rules.
- this goes back to what I am saying in chapter 2, which states that most models and measures that we use all encode within them some form of stopping model — they are just not particularly realistic.
- and from here, we can go: we have these heuristics and rules, but there is nothing in the literature to show how well these rules perform when compared to real behaviours. We know that behaviours change under different search contexts; it therefore follows that stopping behaviour will also change. And so, in order to look at this, we now introduce our conceptual search model (next chapter), before using this

model in a series of simulations to ascertain exactly how stopping behaviour changes under different conditions / different search goals (remaining chapters).

## Improving User Modelling in IIR

*In this part of the thesis, we explore the various approaches that have been used to model all or parts of the IIR process, before introducing the Complex Searcher Model (CSM).*



## Chapter 4

# Existing Approaches to User Modelling

### 4.1 Individual Components

#### 4.1.1 Query Generation

Azzopardi, de Rijke Jordan Keskustalo, Baskaya...

## 4.2 Entire Search Session

### 4.1.2 Judging Relevancy

### 4.1.3 Document Examination

## 4.2 Entire Search Session

### 4.2.1 TREC User

### 4.2.2 Baskaya et al.

### 4.2.3 Thomas et al.

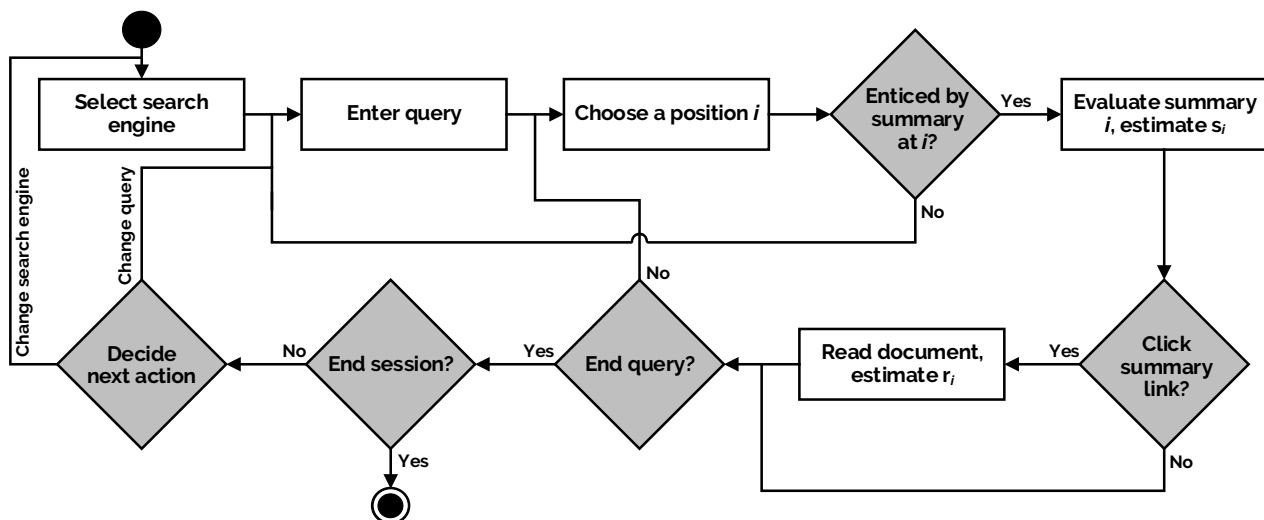


Figure 4.1: Model by Thomas et al.

## 4.3 Chapter Summary

- have all these different models

## Chapter 5

# Advancing User Modelling in IIR

“Empirical studies of complete systems mostly focus on variations of single components” (Fuhr, 2008)

In previous chapter we have seen a number of different entire search session models. In this chapter, we introduce a number of models that have been considered throughout this PhD, with a focus on stopping decision points.

- Why do we consider this as a flow chart? It's a high-level, conceptual model, where different states/activities are represented as components. Totally up to how one implements each component – i.e. with stopping components, we can instantiate them in a number of different ways.

- these models are based on stochastic approaches.

### 5.1 The Basic User Model

Query, Document, Mark

## 5.2 The Complex Searcher Model

### 5.2 The Complex Searcher Model

introduce the model, explaining and motivating it

#### 5.2.1 CSM, Mark I

Query, Snippet, Document, Mark

#### 5.2.2 CSM, Mark II

- motivation for including the SERP - Query, SERP, Snippet, Document, Mark - ECIR 2018 paper

#### 5.2.3 Considering State and Agency

- as an aside, we can also incorporate state and agency into the model. - to a degree, the mark I and mark II models already do consider some form of state, by virtue of having a record of what documents have been previously examined, for example. - however, we can go further here, introducing more advanced state, leading to some form of agency. for example, an agent can determine a new set of queries based upon what it has previously observed.

- while not central to the thesis, we can at least consider it, and demonstrate that it is indeed useful.

### 5.3 Evaluating Model Effectiveness

Core to developing a new model is to basically improve our representation of a real world phenomenon, such as, in this case, the search process.

#### 5.3.1 Research Questions

Allowing us to address HL-RQ1.

#### 5.3.2 Experimental Design

#### 5.3.3 Results

## Examining and Simulating Searcher Stopping Behaviours

*In this part of the thesis, we explore various stopping heuristics that have been defined in the literature, examine stopping behaviour under different search contexts, and simulate these behaviours.*

## Chapter 6

# The Effect of Temporal Delays on Stopping Behaviour

## 6.1 Introduction

### 6.1 Introduction

## 6.2 The User Study

### 6.2.1 Study Motivation and Design

### 6.2.2 Results

### 6.2.3 Conclusions

## 6.3 Simulation Experiments

### 6.3.1 Simulated Experimental Design

### 6.3.2 Comparisons

### 6.3.3 Conclusions

## 6.4 Conclusions

## 6.5 Chapter Summary



## Chapter 7

# The Effect of Snippet Lengths on Stopping Behaviour

### 7.1 Introduction

*Interactive Information Retrieval (IIR)* is a complex, non-trivial process where a searcher undertakes a variety of different actions during a search session Ingwersen and Järvelin (2005). Core to their experience and success is the *Search Engine Results Page (SERP)*, with its presentation and design over the years having been subject to much research. With more complex components now becoming commonplace in modern day Web search engines (such as the *information card* ?? or *social annotations* ?), much work however still remains on examining how more traditional SERP components (such as *result summaries*) are designed and presented to end users.

Result summaries have traditionally been viewed as the ‘*ten blue links*’ with the corresponding URL of the associated document, and one or more textual *snippets* of *keywords-in-context* from the document itself, approximately 130-150 characters (or two lines) in length ?. Numerous researchers have explored result summaries in a variety of different ways, such as: examining their length ???; the use of thumbnails ??; their attractiveness ??; and the gener-

## 7.2 The User Study

ation of *query-biased snippets* ?? . The performance of users has broadly been evaluated in a limited fashion (e.g. by examining task completion times). In this work, we are interested in how the length and information content of result summaries affects SERP interactions and a user's ability to select relevant over non-relevant items. Prior research has demonstrated that longer result summaries tend to lower completion times for informational tasks (where users need to find one relevant document) ?, but does this hold in other contexts, specifically for ad-hoc retrieval, where users need to find *several* relevant items? Furthermore, how does the length and information associated with longer result summaries affect the user's ability to discern the relevant from the non-relevant?

This work therefore serves as an investigation into the effects of search behaviour and search performance when we vary (i) result summary snippet lengths, and by doing so (ii) the information content within the summaries. To this end, a within-subjects crowdsourced experiment ( $n = 53$ ) was designed and conducted. Under ad-hoc topic retrieval, participants used four different search interfaces, each with a different size of result summary. Findings allow us to address the two main research questions of this study. **RQ1** How does the value of information gain represented as snippet length affect behaviour, performance and user experience? **RQ2** Does information gain – again represented as snippet length – affect the decision making ability and accuracy (identifying relevant documents) of users? We hypothesise that longer and more informative snippets will enable users to make better quality decisions (i.e. higher degrees of accurately identifying relevant content).

## 7.2 The User Study

As previously mentioned, the design and presentation of SERPs has been examined in depth. Researchers have examined various aspects of SERPs, and how the designs of such aspects influence the behaviour of users. Here, we provide a summary of the various aspects that have been investigated. Specifically, we focus upon: the layout of SERPs; the size

of SERPs; how snippet text is generated; and how much text should be presented within each result summary – the latter being the main focus of this work.

### 7.2.1 SERP Layouts and Presentation

Early works regarding the presentation of result summaries ?? examined approaches to automatically categorise result summaries for users, similar to the categorisation approach employed by early search engines. Chen and Dumais ? developed an experimental system that automatically categorised result summaries on-the-fly as they were generated. For a query, associated categories were then listed as verticals, with associated document titles provided underneath each category header. Traditional result summaries were then made available when hovering over a document title. Subjects of a user study found the interface easier to use than the traditional ‘ten blue links’ approach - they were 50% faster at finding information displayed in categories. This work was then extended by Dumais et al. ?, where they explored the use of hover text to present additional details about search results based upon user interaction. Searching was found to be slower with hover text, perhaps due to the fact that explicit decisions about when to seek additional information (or not to) were required.

Alternatives to the traditional, linear list of result summaries have also been trialled (like grid-based layouts ???). For example, Krammerer and Beinhaur ? examined differences in user behaviour when interacting with a standard list interface, compared against a tabular interface (title, snippet and URL stacked horizontally in three columns for each result), and a grid-based layout (result summaries placed in three columns). Users of the grid layout spent more time examining result summaries. The approach demonstrated promise in overcoming issues such as *position bias* ?, as observed by Joachims et al. ?.

Marcos et al. ? performed an eye-tracking user study examining the effect of user behaviour while interacting with SERPs – and whether the *richness* of result summaries provided on a

## 7.2 The User Study

SERP (i.e. result summaries enriched with metadata from corresponding pages) impacted upon the user's search experience. Enriched summaries were found to help capture a user's attention. Including both textual and visual representations of a document when presenting results could have a positive effect on relevance assessment and query reformulation ?. Enriched summaries were also examined by Ali et al. ? in the context of navigational tasks. Striking a good balance between textual and visual cues were shown to better support user tasks, and search completion time.

### 7.2.2 Generating Snippet Text

Users can be provided with an insight by result summaries as to whether a document is likely to be relevant or not ?. Consequently, research has gone into examining different kinds of snippets, and how long a snippet should be. Work initially focused upon how these summaries should be generated ??????. These early works proposed the idea of summarising documents with respect to the query (query-biased summaries) or keywords-in-context – as opposed to simply extracting the representative or lead sentences from the document ?. Tombros and Sanderson ? showed that subjects of their study were likely to identify relevant documents more accurately when using query-biased summaries, compared to summaries simply generated from the first few sentences of a given document. Query-biased summaries have also been recently shown to be preferred on mobile devices ?.

When constructing snippets using query-biased summaries, Rose et al. ? found that a user's perceptions of the result's quality were influenced by the snippets. If snippets contained truncated sentences or many fragmented sentences (*text choppiness*), users perceived the quality of the results more negatively, regardless of length. Kanungo and Orr ? found that poor readability also impacts upon how the resultant snippets are perceived. They maintain that readability is a crucial presentation attribute that needs to be considered when generating a query-biased summary. Clarke et al. ? analysed thousands of pairs of snippets where result *A* appeared before result *B*, but result *B* received more clicks than result

A. As an example, they found results with snippets which were very short (or missing entirely) had fewer query terms, were not as readable, and attracted fewer clicks. This led to the formulation of several heuristics relating to document surrogate features, designed to emphasise the relationship between the associated page and generated snippet. Heuristics included: (i) ensuring that all query terms in the generated snippet (where possible); (ii) withholding the repeating of query terms in the snippet if they were present in the page's title; and (iii) displaying (shortened) readable URLs.

Recent work has examined the generation of snippets from more complex angles – from manipulating underlying indexes ?? to language modelling ??, as well as using user search data to improve the generation process ?. Previous generation approaches also may not consider what parts of a document searchers actually find useful. Ageev et al. ? incorporated into a new model post-click searcher behaviour data, such as mouse cursor movements and scrolling over documents, producing *behaviour-biased snippets*. Results showed a marked improvement over a strong text-based snippet generation baseline. Temporal aspects have also been considered – Svore et al. ? conducted a user study, showing that users preferred snippet text with *trending* content in snippets when searching for trending queries, but not so for general queries.

### 7.2.3 Results per Page

Today, a multitude of devices are capable of accessing the *World Wide Web* (WWW) – along with a multitude of different screen resolutions and aspect ratios. The question of how many result summaries should be displayed per page – or *results per page* (RPP) – therefore becomes hugely important, yet increasingly difficult to answer. Examining behavioural effects on mobile devices when interacting with SERPs has attracted much research as of late (e.g. ???), and with each device capable of displaying a different number of results *above-the-fold*, recent research has shown that the RPP value can influence the behaviour of searchers ?. Understanding this behaviour can help guide and inform those charged with

## 7.2 The User Study

designing contemporary user interfaces.

In a Google industry report, Linden ? however stated that users desired more than 10RPP, despite the fact that increasing the RPP yielded a 20% drop in traffic; it was hypothesised that this was due to the extra time required to dispatch the longer SERPs. This drop however be attributed to other reasons. Oulasvirta et al. ? discusses the *paradox of choice* ? in the context of search, where more options (results) – particularly if highly relevant – will lead to poorer choice and degrade user satisfaction. In terms of user satisfaction, modern search engines can therefore be a victim of their own success, presenting users with *choice overload*. Oulasvirta et al. ? found that presenting users with a six-item search result list was associated with higher degrees of satisfaction, confidence with choices and perceived carefulness than an a list of 24 items.

Kelly and Azzopardi ? broadly agreed with the findings by Oulasvirta et al. ?. Here, the authors conducted a between-subjects study with three conditions, where subjects were assigned to one of three interfaces - the baseline interface, showing 10RPP (the ‘ten blue links’), and two interfaces displaying 3RPP and 6RPP respectively. Their findings showed that individuals using the 3RPP and 6RPP interfaces spent significantly longer examining top-ranking results and were more likely to click on higher ranked documents than those on the 10RPP interface. Findings also suggested that subjects using the interfaces showing fewer RPP found it comparatively easier to find relevant content than those using the 10RPP interface. However, no significant difference was found between the number of relevant items found across the interfaces. Currently, 10RPP is still considered the *de-facto* standard ?.

### 7.2.4 Snippet Lengths: Longer or Shorter?

Snippet lengths have been examined in a variety of ways. A user study by Paek et al. ? compared a user’s preference and usability against three different interfaces for displaying result summaries. With question answering tasks, the interfaces: displayed a *normal*

SERP (i.e. a two line snippet for each summary, with a clickable link); an *instant* interface, where an expanded snippet was displayed upon clicking it; and a *dynamic* interface, where hovering the cursor would trigger the expanded snippet. The instant view was shown to allow users to complete the given tasks in less time than the normal baseline, with half of participants preferring this approach.

Seminal work by Cutrell and Guan [1] explored the effect of different snippet lengths (*short*: 1 line, *medium*: 2-3 lines; and *long*: 6-7 lines). They found that longer snippets significantly improved performance for *informational tasks* (e.g. 'Find the address for Newark Airport.'). Users performed better for informational queries as snippet length increased. This work was followed up by Kaiser et al. [2]. They conducted two experiments that estimated the preferred snippet length according to answer type (e.g. finding a person, time, or place), and comparing the results of the preferred snippet lengths to users' preferences to see if this could be predicted. The preferred snippet length was shown to depend upon the type of answer expected, with greater user satisfaction shown for the snippet length predicted by their technique.

More contemporary work has begun to examine what snippet sizes are appropriate for mobile devices. Given smaller screen sizes, this is important – snippet text considered acceptable on a computer screen may involve considerable scrolling/swiping on a smaller screen. Kim et al. [3] found that subjects using longer snippets on mobile devices exhibited longer search times and similar search accuracy under informational tasks<sup>1</sup>. Longer reading times and frequent scrolling/swiping (with more viewport movements) were exhibited. Longer snippets did not therefore appear to be very useful on a small screen – an *instant* or *dynamic* snippet approach (as per Paek et al. [4]) may be useful for mobile search, too.

The presentation of result summaries has a strong effect on the ability of a user to judge relevancy [5]. Relevant documents may be overlooked due to uninformative summaries – but conversely, non-relevant documents may be examined due to a misleading sum-

---

<sup>1</sup>The tasks considered by Kim et al. [3] were similar to those defined by Cutrell and Guan [1], where a single relevant document was sought.

## 7.2 The User Study

mary. However, longer summaries also increase the examination cost, so there is likely a trade-off between informativeness/accuracy and length/cost. The current, widely accepted standard for result summaries are two query-based snippets/lines <sup>2</sup>. This work examines whether increasing and decreasing the length (and consequently the informativeness) of result summary snippets affects user accuracy and costs of relevance decisions in the context of ad-hoc topic search, where multiple relevant documents are sought.

To address our two key research questions outlined in Section ??, we conducted a within-subjects experiment. This allowed us to explore the influence of snippet length and snippet informativeness on search behaviours, performance and user experience. Subjects used four different search interfaces, each of which varied the way in which result summaries were presented to them.

To decide the length and informativeness of the result summaries, we performed a preliminary analysis to determine the average length (in words) and informativeness (as calculated by the *Kullback-Leibler distance* <sup>2</sup> to measure *information gain*, or *relative entropy*) of result summaries with the title and varying numbers of snippet fragments (0–10). The closer the entropy value is to zero, the more information gained. Figure ?? plots the number of words, the information gain, and the information gain per word<sup>2</sup>. It is clear from the plot that a higher level of information gain was present in longer snippets. However, as the length increases with each additional snippet fragment added, the informativeness per word decreased. Consequently, for this study, we selected the four different interface conditions in the region where informativeness had the highest change, i.e. from zero to four. The conditions we selected for the study were therefore:

**T0** where only the title for each result summary were presented;

**T1** where for each result summary, a title and one query-biased snippet fragment were presented;

---

<sup>2</sup>To obtain these values, we submitted over 300 queries from a previous study (refer to Azzopardi et al. <sup>2</sup>) conducted on similar topics and on the same collection to the search system that we used.



*T2* where a title and two query-biased snippet fragments were presented; and

*T4* where a title and four query-biased snippet fragments were presented,

where our independent variable is snippet informativeness, controlled by the length. Figure ?? provides an example of the different result summaries in each condition. The remainder of this section details our methodology for this experiment, including a discussion of: the corpus, topics and system used (Subsection 8.2.2); how we generated snippets (Subsection 7.2.6); the behaviours we logged (Subsection 8.2.8); how we obtained the opinions of subjects regarding their experience (Subsection 7.2.8); and further details on our study, including measures taken for quality control (Subsection 7.2.9).

### 7.2.5 Corpus, Search Topics and System

For this experiment, we used the TREC AQUAINT test collection. Using a traditional test collection provided us with the ability to easily evaluate the performance of subjects. The collection contains over one million newspaper articles from the period 1996-2000. Articles were gathered from three newswires: the *Associated Press* (AP); the *New York Times* (NYT); and *Xinhua*.

We then selected a total of five topics from the *TREC 2005 Robust Track*, as detailed by Voorhees ?. The topics selected were: № 341 (*Airport Security*); № 347 (*Wildlife Extinction*); № 367 (*Piracy*); № 408 (*Tropical Storms*); and № 435 (*Curbing Population Growth*). We selected topic № 367 as a practice topic so that subjects could familiarise themselves with the system. These topics were chosen based upon evidence from a previous user study with a similar setup, where it was shown that the topics were of similar difficulty ?. For each subject, the remaining four topics were assigned to an interface (one of *T0*, *T1*, *T2* or *T4*) using a Latin-square rotation.

To ground the search tasks, subjects of the experiment were instructed to imagine that they

## 7.2 The User Study

were newspaper reporters, and were required to gather documents to write stories about the provided topics. Subjects were told to find as many relevant documents as they could during the allotted time, which was 10 minutes per topic – hereafter referred to as a *search session*. With the traditional components of a SERP, such as the query box and result summaries present (refer to Figure ??), subjects were instructed to mark documents they considered relevant by clicking on the 'Mark as Relevant' button within the document view – accessed by clicking on a result summary he or she thought was relevant. Coupled with a two minute period to familiarise themselves with the system (using topic № 367), subjects spent approximately 45-50 minutes undertaking the complete experiment when pre- and post-task surveys were accounted for.

For the underlying search engine, we used the *Whoosh Information Retrieval (IR)* toolkit <sup>3</sup>. We used BM25 as the retrieval algorithm ( $b = 0.75$ ), but with an implicit ANDing of query terms to restrict the set of retrieved documents to only those that contained all the query terms provided. This was chosen as most search systems implicitly AND terms together.

### 7.2.6 Snippet Generation

For interfaces *T2* and *T4*, each result summary presented to the subjects required one or more textual snippets from the corresponding document. These snippet fragments were query-biased ?, and were generated by scoring sentences according to BM25 and selecting fragments from those sentences. Fragments were then extracted from the ordered series of sentences, by identifying query terms within those sentences with a window of 40 characters from either side of the term. Figure ?? provides a complete, rendered example of the result summaries generated by each of the four interfaces. Each result summary contains a document title, a newswire source (acting as a replacement for a document URL), and, if required, one or more textual snippets.

---

<sup>3</sup>Whoosh can be accessed at <https://pypi.python.org/pypi/Whoosh/>.

### 7.2.7 Behaviours Logged

In order for us to address our research questions, our experimental system was required to log a variety of behavioural attributes for each subject as they performed the variety of actions that take place during a search session. Search behaviours were operationalised over three types of measures: (i) interaction, (ii) performance, and (iii) the time spent undertaking various search activities. All behavioural data was extracted from the log data produced by our system, and from the TREC 2005 Robust Track QRELS<sup>4</sup>. All data was recorded with the interface and topic combination used by the subject at the given time.

**Interaction measures** included the number of queries issued, the number of documents viewed, the number of SERPs viewed, and the greatest depths in the SERPs to which subjects clicked on – and hovered over – result summaries.

**Performance measures** included a count of the documents marked as relevant by the subject, the number of documents marked that were also TREC relevant – as well as TREC non-relevant, and  $P@k$  measurements for the performance of the subject's issued queries for a range of rankings.

**Time-Based measures** included the time spent issuing queries, examining SERPs – as well as examining result summaries<sup>4</sup> – and the time spent examining documents. All of these times added together yielded the total search session time, which elapsed once 10 minutes had been reached.

From this raw data, we could then produce summaries of a search session, producing summarising measures such as the number of documents examined by searchers per query that they issued. We could also calculate from the log data probabilities of interaction, including a given subject's probability of clicking a result summary link, given that it was

---

<sup>4</sup>Result summary times were approximated by dividing the total recorded SERP time by the number of snippets hovered over with the mouse cursor. We believe this is a reasonable assumption to make – the timings of hover events proved to be unreliable due to occasional network latency issues beyond our control.

## 7.2 The User Study

TREC relevant ( $P(C|R)$ ) or TREC non-relevant ( $P(C|N)$ ) – or the probability of marking a document that was clicked, given it was either TREC relevant ( $P(M|R)$ ) or TREC non-relevant ( $P(M|N)$ ). Actions such as hover depth over result summaries were inferred from the movement of the mouse cursor, which in prior studies has been shown to correlate strongly with the user's gaze on the screen ??.

### 7.2.8 Capturing User Experiences

To capture user experiences, we asked subjects to complete both pre- and post-task surveys for each of the four interface conditions. Pre-task surveys consisted of five questions, each of which was on a seven-point Likert scale (7 – *strongly agree* to 1 – *strongly disagree*). Subjects were sought for their opinions on their: (i) prior knowledge of the topic; (ii) the relevancy of the topic to their lives; (iii) their desire to learn about the topic; (iv) whether they had searched on this topic before; and (v) the perceived difficulty to search for information on the topic.

The same Likert scale was used for post-task surveys, where subjects were asked to judge the following statements: (*clarity*) – the result summaries were clear and concise; (*confidence*) – the result summaries increased my confidence in my decisions; (*informativeness*) – the result summaries were informative; (*relevance*) – the results summaries help me judge the relevance of the document; (*readable*) – the result summaries were readable; and (*size*) – the result summaries were an appropriate size and length.

At the end of the experiment, subjects completed an exit survey. From five questions, they were asked to pick which of the four interfaces was the closest fit to their experience. We sought opinions on what interface: (*most informative*) – yielded the most informative result summaries; (*least helpful*) – provided the most unhelpful summaries; (*easiest*) – provided the easiest to understand summaries; (*least useful*) – provided the least useful result summaries; and (*most preferred*) – the subject's preferred choice for the tasks that they un-

dertook.

### 7.2.9 Crowdsourced Subjects & Quality Control

As highlighted by Zuccon et al. [?], crowdsourcing provides an alternative means for capturing user interactions and search behaviours from traditional lab-based user studies. Greater volumes of data can be obtained from more heterogeneous workers at a lower cost – all within a shorter timeframe. Of course, pitfalls of a crowdsourced approach include the possibility of workers completing tasks as efficiently as possible, or submitting their tasks without performing the requested operations [?]. Despite these issues, it has been shown that there is little difference in the quality between crowdsourced and lab-based studies [?]. Nevertheless, quality control is a major component of a well-executed crowdsourced experiment [?]. Here, we detail our subjects and precautions taken.

The study was run over the *Amazon Mechanical Turk (MTurk)* platform. Workers from the platform performed a single *Human Intelligence Task (HIT)*, which corresponded to the entire experiment. Due to the expected length of completion for the study (45-50 minutes), subjects who completed the study in full were reimbursed for their time with US\$9; a typically larger sum (and HIT duration) than most crowdsourced experiments. A total of 60 subjects took part in the experiment, which was run between July and August, 2016. However, seven subjects were omitted due to quality control constraints (see below). In all, of the 53 subjects who satisfied the expected conditions of the experiment, 28 were male, with 25 female. The average age of our subjects was 33.8 years ( $min = 22$ ;  $max = 48$ ;  $stdev = 7.0$ ), with 19 of the subjects possessing a bachelor's degree or higher, and all expressing a high degree of search literacy, with all subjects stating that they conducted at least five searches for information online per week. With 53 subjects, each searching over four topics, this meant a total of 212 search sessions were logged.

We examined extra precautionary measures to ensure the integrity of the log data that was

## 7.2 The User Study

recorded. Precautions were taken from several angles. First, workers were only permitted to begin the experiment on the MTurk platform that: (i) were from the United States, and were native English speakers; (ii) had a HIT acceptance rate of at least 95%; and (iii) had at least 1000 HITs approved. Requiring (ii) and (iii) reduced the likelihood of recruiting individuals who would not complete the study in a satisfactory manner. Recruits were forewarned about the length of the HIT, which was considerably longer than other crowd-sourced experiments.

We also ensured that the computer the subject was attempting the experiment on had a sufficiently large screen resolution (1024x768 or greater) so as to display all of the experimental interface on screen. With the experiment being conducted in a Web browser popup window of a fixed size, we wanted to ensure that all subjects would be able to see the same number of results on a SERP within the popup window's viewport. As the experiment was conducted via a Web browser, we wanted to ensure that only the controls provided by the experimental apparatus were used, meaning that the popup window had all other browser controls disabled to the best of our ability (i.e. history navigation, etc.). The experimental system was tested on several major Web browsers, across different operating systems. This gave us confidence that a similar experience would be had across different system configurations.

We also implemented a series of log post-processing scripts after completion of the study to further identify and capture individuals who did not perform the tasks as instructed. It was from here that we identified the seven subjects that did not complete the search tasks in a satisfactory way – spending less than three of the ten minutes searching. These subjects were excluded from the study, reducing the number of subjects reported from 60 to 53. Finally, results are reported based upon the first 360 seconds as some of the remaining subjects didn't fully use all 600 seconds.

Both search behaviour and user experience measures were analysed by each interface. To evaluate these data, ANOVAs were conducted using the interfaces as factors; main effects were examined with  $\alpha = 0.05$ . Bonferroni tests were used for post-hoc analysis. It should

**Table 7.1:** Characters, words and *Information Gain (IG)* across each of the four interface conditions. An ANOVA test reveals significant differences, with follow-up tests (refer to Section 8.3) showing that each condition is significantly different to others. There are clearly diminishing returns in information gain as snippet length increases. An IG value closer to zero denotes a higher level of IG. In the table, *IG/W.* denotes *IG per word*.

	T0	T1	T2	T4
<b>Words</b>	6.58±0.01	12.52±0.06*	16.29±0.10*	17.06±0.13*
<b>Chars.</b>	37.37±0.15	153.29±0.16*	168.36±0.28*	234.78±0.31*
<b>IG</b>	-	-	-	-
	6.35±0.01	3.59±0.00*	3.00±0.00*	2.67±0.00*
<b>IG/W.</b>	-	-	-	-
	1.17±0.00	0.18±0.00*	0.08±0.00*	0.04±0.00*

be noted that the error bars as shown in the plots for Figures ?? and ?? refer to the *standard error*.

To check whether the interfaces were different with respect to snippet length and information gain, we performed an analysis of the observed result summaries. Table 7.1 summarises the number of words and characters that result summaries contained on average. As expected, the table shows an increasing trend in words and characters as snippet lengths increase. Information gain for each snippet was then calculated using the Kullback-Leibler distance ? to measure information gain (e.g. relative entropy). Statistical testing showed that the differences between snippet length ( $F(3, 208) = 1.2 \times 10^5, p < 0.001$ ) and information gain ( $F(3, 208) = 2.6 \times 10^5, p < 0.001$ ) were significant. Follow up tests revealed that this was the case over all four interfaces, indicating that our conditions were different on these dimensions. These findings provide some justification for our choices for the number of snippet fragments present for each interface – a diminishing increase in information gain after four snippets suggested that there wouldn't be much point generating anything longer.

## 7.2 The User Study

**Table 7.2:** Summary table of both interaction and performance measures over each of the four interfaces evaluated. For each measure examined, no significant differences are reported across the four interfaces.

	<b>T0</b>	<b>T1</b>	<b>T2</b>	<b>T4</b>
<b>Number of Queries</b>	3.72± 0.34	3.19± 0.35	3.30± 0.35	3.28± 0.31
<b>Number of SERP Pages per Query</b>	2.87± 0.29	2.69± 0.23	2.43± 0.13	2.40± 0.20
<b>Number of Docs Clicked per Query</b>	4.23± 0.55	4.83± 0.54	5.14± 0.66	4.76± 0.62
<b>Depth per Query</b>	24.47± 2.96	22.87± 2.47	20.02± 1.46	19.40± 2.04
<b>P@10</b>	0.25± 0.02	0.23± 0.02	0.27± 0.02	0.25± 0.03
<b>Number of Documents Marked Relevant</b>	6.68± 0.66	7.00± 0.63	6.49± 0.58	7.60± 0.79
<b>Number of TREC Rels Found</b>	2.58± 0.34	2.28± 0.25	2.47± 0.28	2.66± 0.32
<b>Number of Unjudged Docs Marked Relevant</b>	1.85± 0.32	2.08± 0.29	1.98± 0.24	1.68± 0.32

**Table 7.3:** Summary table of times over each of the four interfaces evaluated. Significant differences exist between T0 and T4 (identified by the \*, where  $\alpha = 0.05$ ) on a follow-up Bonferroni test.

	<b>T0</b>	<b>T1</b>	<b>T2</b>	<b>T4</b>
<b>Time per Query</b>	8.29± 0.57	7.99± 0.57	9.42± 0.79	8.12± 0.48
<b>Time per Document</b>	17.31± 2.12	22.82± 6.03	17.19± 1.86	18.99± 2.13
<b>Time per Result Summary*</b>	1.63 ± 0.13*	2.21± 0.21	2.35± 0.23	2.60 ± 0.27*

### 7.2.10 Search Behaviours

**Interactions.** Table 8.2 presents the mean (and standard deviations) of the number of queries issued, the number of SERPs viewed per query, documents clicked per query, and the click depth per query over each of the four interfaces examined. Across the four different interfaces, there were no significant differences reported between any of these measures. The number of queries issued follows a slight downward trend as the length of result summaries increases (3.72 ± 0.34 for **T0** to 3.28 ± 0.31 for **T4**), as too does the number of SERPs examined, and the number of documents examined per query. The depth to which subjects went



to per query however follows a downward trend – as the length of snippets increases, subjects were likely to go to shallower depths when examining result summaries ( $24.47 \pm 2.96$  for *T0* to  $19.4 \pm 2.04$  for *T4*).

Interaction probabilities all showed an increasing trend as snippet length increased over the four interfaces, as shown in Table 7.4. Although no significant differences were observed over the four interfaces and the different probabilities examined, trends across all probabilities show an increase as the snippet length increases. An increase of both the probability of clicking result summaries on the SERP ( $P(C)$ ) and marking the associated documents ( $P(M)$ ) as relevant were observed. When these probabilities are examined in more detail by separating the result summaries clicked and documents marked by their TREC relevancy (through use of TREC QREs), we see increasing trends for clicking and marking – both for TREC relevant ( $P(C|R)$  and  $P(M|R)$  for clicking and marking, respectively) and TREC non-relevant documents ( $P(C|N)$  and  $P(M|N)$ ). This interesting finding shows that an increase in snippet length does not necessarily improve the accuracy of subjects – simply the likelihood that they would consider documents as relevant.

**Performance.** Table 8.2 also reports a series of performance measures over the four conditions, averaged over the four topics examined. We report the mean performance of the queries issued with  $P@10$ , the number of documents marked relevant, and the number of documents marked relevant that were TREC relevant. Like the interaction measures above, no significant differences were observed over the four interfaces for each of the performance measures examined. The performance of queries issued by subjects was very similar across all four conditions ( $P@10 \approx 0.25$ ), along with the number of documents identified by subjects as relevant ( $6.49 \pm 0.58$  for *T2* to  $7.6 \pm 0.79$  for *T4*), and the count of documents marked that were actually TREC relevant ( $2.28 \pm 0.25$  for *T1* to  $2.66 \pm 0.32$  for *T4*). We also examined the number of documents marked that were not assessed (unjudged) by the TREC assessors, in case one interface surfaced more novel documents. On average, subjects marked two such documents, but again there was no significant differences between interfaces.

## 7.2 The User Study

**Table 7.4:** Table illustrating a summary of interaction probabilities over each of the four interfaces evaluated. Note the increasing trends for each probability from **T0** → **T4** (short to long snippets). Refer to Section 7.2.10 for an explanation of what each probability represents.

	<b>T0</b>	<b>T1</b>	<b>T2</b>	<b>T4</b>
$P(C)$	$0.20 \pm 0.02$	$0.25 \pm 0.02$	$0.26 \pm 0.03$	$0.28 \pm 0.03$
$P(C R)$	$0.28 \pm 0.03$	$0.34 \pm 0.03$	$0.35 \pm 0.03$	$0.40 \pm 0.04$
$P(C N)$	$0.18 \pm 0.02$	$0.23 \pm 0.02$	$0.25 \pm 0.03$	$0.24 \pm 0.03$
$P(M)$	$0.61 \pm 0.04$	$0.68 \pm 0.04$	$0.65 \pm 0.03$	$0.71 \pm 0.03$
$P(M R)$	$0.66 \pm 0.06$	$0.69 \pm 0.05$	$0.67 \pm 0.05$	$0.66 \pm 0.05$
$P(M N)$	$0.55 \pm 0.04$	$0.65 \pm 0.04$	$0.58 \pm 0.04$	$0.67 \pm 0.04$

**Time-Based Measures.** Table 8.3 reports a series of selected interaction times over each of the four evaluated interfaces. We include: the mean total query time per subject, per interface; the mean time per query; the mean time spent examining documents per query; and the mean time spent examining result summaries per query. No significant differences were found between the mean total query time, the time per query and the time per document. However, a significant difference did exist for the time spent per result summary. A clear upward trend in the time spent examining snippets can be seen in Figure ?? as result summaries progressively got longer, from  $1.63 \pm 0.13$  for **T0** to  $2.6 \pm 0.27$  for **T4**, which was significantly different ( $F(3, 208) = 3.6, p = 0.014$ ). A follow-up Bonferroni test showed that the significant difference existed between **T0** and **T4**. This suggests that as result summary length increases, the amount of time spent examining result summaries also increases (an intuitive result). This also complies with trends observed regarding examination depths. When the length of result summaries increased, subjects were likely to examine result summaries to shallower depths.

**Table 7.5:** Summary table of the recorded observations for the post-task survey, indicating the preferences of subjects over six criteria and the four interfaces, where \* indicates that **T0** was significantly different from the other conditions. In the table, *Conf.* represents *Confidence*, *Read.* represents *Readability*, *Inform.* represents *Informativeness*, and *Rel.* represents *Relevancy*.

	<b>T0</b>	<b>T1</b>	<b>T2</b>	<b>T4</b>
<b>Clarity</b>	4.16± 0.27*	5.00± 0.21	5.06± 0.24	5.40± 0.20
<b>Conf.</b>	3.71± 0.26*	4.66± 0.26	4.75± 0.24	5.06± 0.25
<b>Read.</b>	5.18± 0.31*	6.32± 0.17	6.46± 0.14	6.36± 0.14
<b>Inform.</b>	4.20± 0.30*	5.38± 0.24	5.27± 0.24	5.62± 0.20
<b>Rel.</b>	3.84± 0.28*	4.89± 0.25	5.08± 0.24	5.36± 0.20
<b>Size</b>	4.00± 0.31*	4.94± 0.25	5.21± 0.22	5.36± 0.19

### 7.2.11 User Experience

**Task Evaluations.** Table 7.5 presents the mean set of results from subjects across the four interfaces, which were answered upon completion of each search task. The survey questions are detailed in Section 7.2.8. Using the seven-point Likert scale for their responses (with 7 indicating *strongly agree*, and 1 indicating *strongly disagree*), significant differences were found in all question responses (**clarity**  $F(3, 208) = 5.22, p = 0.001$ , **confidence**  $F(3, 208) = 5.3, p = 0.001$ , **readable**  $F(3, 208) = 9.25, p < 0.001$ , **informative**  $F(3, 208) = 5.22, p = 0.001$ , **relevance**  $F(3, 208) = 6.44, p < 0.001$ , and **size**  $F(3, 208) = 7.28, p < 0.001$ ). Follow-up Bonferroni tests however showed that the significant difference existed only between **T0** and the remaining three interfaces, **T1**, **T2** and **T4**. A series of discernible trends can be observed throughout the responses, with subjects regarding longer snippets as more concise, and a higher degree of clarity ( $4.16 \pm 0.27$  for **T0** to  $5.4 \pm 0.2$  for **T4**). This perceived clarity also made subjects feel more confident that the longer result summaries helped them

## 7.2 The User Study

**Table 7.6:** Table presenting responses from the exit survey completed by subjects. The survey is discussed in Section 7.2.8.

	T0	T1	T2	T4
Most Informative	1	4	20	29
Least helpful	46	5	1	2
Easiest	4	4	24	22
Least Useful	49	4	0	1
Most Preferred	3	5	20	26

make better decisions as to whether they were relevant to the given topic – interaction results presented above however differ from this, where the overall probability of marking documents increased, regardless of the document/topic TREC relevancy judgement. Other notable trends observed from the results included an increase in how informative subjects perceived the result summaries to be – again, with longer summaries proving more informative. Subjects also reported a general increase in satisfaction of the length of the presented result summaries/snippets – although, as mentioned, no significant difference existed between the three interfaces that generated snippets (*T1*, *T2* and *T4*).

**System Evaluations.** Upon completion of the study, subjects completed the exit survey as detailed in Section 7.2.8. Responses from the subjects are presented in Table 7.6. From the results, subjects found result summaries of longer lengths (i.e. those generated by interfaces *T2* and *T4*) to be the most informative, and those generated by *T0* – without snippets – to be the least helpful and useful. The longer result summaries were also consistently favoured by subjects, who preferred them over the result summaries generated by interfaces *T0* and *T1*. Subjects also found the result summaries of longer length easier to use to satisfy the given information need.

From the results, it is therefore clear that a majority of subjects preferred longer result summaries to be presented on SERPs, generated by interfaces *T2* and *T4*. Figure ?? provides summary plots, showing general trends across the four interfaces, examining observed in-

teractions and reported experiences.

In this paper, we investigated the influence of result summary length on search behaviour and performance. Using the Kullback-Leibler distance  $\mathcal{D}_{KL}$  as a measure of information gain, we examined result summaries of different lengths, selected a series of snippet lengths where there was a significant difference in information gain between them, which yielded the configurations for our four experimental conditions,  $T0$ ,  $T1$ ,  $T2$  and  $T4$ . We conducted a crowdsourced user study comprising of 53 subjects, each of whom undertook four search tasks, using each of the four interfaces.

Our work was focused around addressing our two research questions, which explored **RQ1** how the value of information gain (represented by snippet length) affected search behaviour and user experience; and **RQ2** whether information gain affected the decision making ability and accuracy of users. Addressing **RQ1** first in terms of search behaviour, there was little difference – but we did observe the following trends: as summary length increases, participants: issued fewer queries; examined fewer pages; but clicked more documents, i.e. they spent more of their time assessing documents at higher ranks. Second, our results show that in terms of experience, subjects broadly preferred longer summaries. The participants felt that longer summaries were more clear, informative, readable – and interestingly – gave them more confidence in their relevance decisions. With respect to **RQ2**, we again observed little difference in subjects' decision making abilities and accuracy between the four interfaces. While subjects perceived longer snippets to help them infer relevance more accurately, our empirical evidence shows otherwise. In fact, it would appear that longer result summaries were more attractive, increasing the information scent of the SERP  $\mathcal{I}$ . This may account for the increase in clicks on the early results, without the benefits, however: accuracy of our subjects did not improve with longer snippets; nor did they find more relevant documents. Increased confidence in the result summaries (from  $T0 \rightarrow T4$ ) may have led to a more relaxed approach at marking content as relevant – as can be seen by increasing click and mark probabilities for both relevant and non-relevant content. It is also possible that the *paradox of choice*  $\mathcal{P}$  could play a role in shaping a user's preferences. For example, in

## 7.2 The User Study

the condition with longer result summaries (*T4*), users viewed fewer results/choices than on other conditions. This may have contributed to their feelings of greater satisfaction and increased confidence in their decisions.

These novel findings provide new insights into how users interact with result summaries in terms of their experiences and search behaviours. Previous work had only focused upon task completion times and accuracy of the first result while not considering their experiences (e.g. ??). Furthermore, these past works were performed in the context of Web search where the goal was to find one document. However, we acknowledge that our work also has limitations. Here, we examined out research questions – with respect to topic search within a news collection – to explore how behaviour and performance changes when searching for multiple relevant documents. It would be interesting to examine this in other search contexts, such as product search, for example. News article titles also can be crafted differently from documents in other domains. Summaries in this domain may perhaps be more important than in other domains, and so the effects and influences are likely to be larger. Furthermore, we only considered how behaviours changed on the desktop, rather than on other devices where users are more likely to be sensitive to such changes (e.g. ??). For example, during casual leisure search, multiple relevant documents on tablet devices are often found, and so it would be interesting to perform a follow up study in this area.

7.2.12 Study Motivation and Design

7.2.13 Results

7.2.14 Conclusions

7.3 Simulation Experiments

7.3.1 Simulated Experimental Design

7.3.2 Comparisons

7.3.3 Conclusions

7.4 Conclusions

7.5 Chapter Summary





## Chapter 8

# The Effect of Diversifying Results on Stopping Behaviour

Perhaps for this final study, examine things on an individual level. Are individuals more likely to follow a particular stopping strategy than others?

### 8.1 Introduction

### 8.2 The User Study

*Interactive Information Retrieval (IIR)* is a complex (and often exploratory) process Ingwersen and Järvelin (2005) in which a searcher issues a variety of queries as a means to explore the topic space ?. Often, such tasks are *aspectual* in nature, where an underlying goal is to find out about the different facets, dimensions or aspects of the topic. This is often referred to as (*aspectual retrieval*). While aspectual retrieval has been heavily studied in the past (during the TREC Interactive Tracks ?), there has been renewed interest in the search task as it represents a novel context to explore the idea of “*search as learning*” ?. In this context, the goal of the system is to help the searcher learn about a topic ? – and in doing so, the number

## 8.2 The User Study

of aspects the searcher finds provides an indication of how much they learned during the process ?. If the goal is to help people learn about a topic, then by returning results that are more diverse in nature and presenting a broader view on the topic, *should* help searchers learn more about said topic. This reasoning suggests that employing diversification will lead to an improved search and learning experience ?.

However, while there have been numerous diversification algorithms developed and proposed over the years ?????, the focus here has been on addressing the problem of intents, rather than how diversification affects complex search tasks, such as *ad-hoc* or aspectual retrieval. Thus, in this paper, we perform one of the first investigations into the influence and impact of how diversifying the results (or not) affects the search behaviour and search performance of users when performing different search task (ad-hoc or aspectual). Our focus is on understanding how behaviours – in particular, how searching and stopping behaviours – change under the different conditions. We ground our study by drawing upon *Information Foraging Theory (IFT)* Pirolli and Card (1999). However, somewhat counter to intuition, IFT leads to the following hypotheses: (i) that when searching for aspects, diversification will lead to searchers examining fewer documents per query; and, either, (ii) issuing more queries or (iii) lower task completion times.

Yet intuitively we reason that if a system provides a more diversified set of results, searchers then *should* be able to exploit the diversification, and find more varied aspects by examining more documents per query – and thus issue fewer queries. To explore these hypotheses and test our intuitions, we designed a  $2 \times 2$  within-subjects user study, where participants were tasked to learn about four different topics under the following conditions, using: (i) a non-diversified system (BM25); versus (ii) a diversified system (BM25+ $x$ QuAD ?), and when the search task is either: (a) ad-hoc retrieval, where they need to only find relevant documents; or (b) aspectual retrieval, where they need to find documents that are both relevant, and different – i.e. covering new aspects of the topic. We perform our experiments in the context of learning about a topic in order to write a report where participants use a standard search interface to search the *TREC AQUAINT* news collection.

### 8.2.1 Study Motivation and Design

When searching for information, searchers pose a varying number of queries, examine *Search Engine Result Pages (SERPs)*, and examine a number of documents (if any) before issuing a new query, or stopping their search altogether. This may be because they have found enough information, have run out of time, were dissatisfied, or simply gave up their search ??????. Prior work has shown that there are a variety of different factors that can influence people's search behaviours. Of particular relevance to this paper, it has been shown that different search tasks influence the search behaviour of users ?.

An interesting task that has not received much attention as of late is aspectual retrieval. Aspectual retrieval is a type of search task that concerns the identification of different *aspects* of a given topic. This task type differs from traditional ad-hoc retrieval in the sense that ad-hoc retrieval is concerned only with what constitutes a *relevant* document to a given topic, rather than identifying relevant documents, and whether they are *different* to what has been seen previously. A relevant and different document will contain unseen *aspects* associated with the topic in question. As an example, take the topic *wildlife extinction*, one of the topics in the *TREC 2005 Robust Track* ?. In an ad-hoc search task, if the searcher finds several documents concerning 'Pandas in China', then these would all be considered relevant. However, for the aspectual retrieval task, where *different* examples must be found, then the first document concerning 'Pandas in China' is considered relevant/useful, and other aspects (in this case, species of endangered animals) would need to be found, such as 'Sumatran Rhinos in Malaysia', 'Crested Ibis in Japan', etc.

Aspectual retrieval found significant traction in the *TREC Interactive Tracks* from 1997–2002. The overarching, high-level goal of the TREC Interactive Tracks was to investigate searching, as an interactive task, by examining the process of searching, as well as the outcome ?. Historically, interaction was considered from the inaugural *TREC-1* in 1993 Harman (1993), where one group investigated interactive searching under "*interactive query mode*" within

## 8.2 The User Study

the ad-hoc task. From TREC-6 to TREC 2002, a substantial volume of research was directed towards the development of systems and search interfaces that: (i) assisted users in exploring and retrieving various aspects of a topic, such as cluster-based and faceted interfaces that explicitly showed different aspects ??; (ii) tiles and stacks to organise documents ????; and (iii) mechanisms to provide query suggestions that lead to different search paths ??. However, a disappointing conclusion from this initiative was that little difference was observed between such systems, and the standard control systems (*ten blue links*), both in terms of behaviour and performance ?.

As work on aspectual retrieval subsided, work related to determining the intent of a searcher's query began to take hold, where the goal of this problem is to diversify the results retrieved with respect to the original query ?. Thus, this addresses the problem of *ambiguity* for short, impoverished queries. This led to a series of diversification algorithms (and intent-aware evaluation measures) being proposed, changing focus from the interface to the underlying algorithms and their evaluation measures (e.g. ?????????). However, while there have been numerous studies investigating the effectiveness of diversification algorithms for the problem of *intents* (e.g. one query, several interpretations), little work has looked at studying how such algorithms apply in the context of aspectual retrieval (e.g. one topic, many aspects). This is mainly due to the fact that most of these algorithms were developed after the TREC Interactive Track finished in 2002.

Recently however, a growing interest in new, more complex and exploratory search tasks has taken hold – especially in the aforementioned context of “*search as learning*” ?. Syed and Collins-Thompson ? hypothesised that diversifying the results presented to users would improve their learning efficiency, and that this would be observed by the change in vocabulary expressed in user queries. This study motivates our interest examining the effects of diversification (or not) when considering the task of aspectual retrieval (where a user needs to learn about different aspects) This in this paper, our aim is to better understand how search performance and search behaviour changes when people undertake different types of search task, using search systems that diversify the ranked results, and those that

don't. To ground this study, we first consider how search behaviour is likely to change by generating hypotheses from Information Foraging Theory.

To address our research questions and examine the hypotheses as outlined in Section ??, we conducted a within-subjects experiment with two factors: system and task. For the system factor, our baseline control system was based on BM25 (no diversification) and a diversified system based on BM25+xQuAD ?. For the task factor, we used the standard ad-hoc retrieval task, and compared against the aspectual retrieval task. This resulted in a  $2 \times 2$  factorial design. Each participant, therefore, completed four different search tasks, one in each of the four conditions (see below). Conditions were assigned using a Latin square rotation to minimise any ordering effects.

**D.As:** *A diversified system, with an aspectual retrieval task.*

**ND.As:** *A non-diversified system, with an aspectual retrieval task.*

**D.Ad:** *A diversified system, with an ad-hoc retrieval task.*

**ND.Ad:** *A non-diversified system, with an ad-hoc retrieval task.*

### 8.2.2 Corpus and Search Topics

For this experiment, we used the *TREC AQUAINT* test collection that contains over one million articles from three newswires, collected over the period 1996-2000. The three newswires were: the *Associated Press* (AP); the *New York Times* (NYT); and *Xinhua*. From the *TREC 2005 Robust Track* ?, we selected five contemporary topics that have been used in prior works ???. These were: № 341 (*Airport Security*); № 347 (*Wildlife Extinction*); № 367 (*Piracy*); № 408 (*Tropical Storms*); and № 435 (*Curbing Population Growth*). These topics were chosen based upon evidence from a previous user study with a similar setup, where it was shown that the topics were of similar difficulty and interest ?. Topic, № 367 was used as a practice topic, while the others topics were used as part of the experimental study.

## 8.2 The User Study

### 8.2.3 Tasks: Aspectual and Ad-Hoc Retrieval

Subjects were asked to imagine that they need to learn about a number of topics on which they need write a report on. Then, given the topic, they were further instructed on whether to focus on finding *relevant* articles in the case of ad-hoc retrieval or *relevant* articles that discussed *different* aspects of the topic in the case of aspectual retrieval. For example, for Airport Security, in the ad-hoc retrieval conditions, subjects were required to learn about the efforts taken by international airports to better screen passengers and their carry-on luggage. While in the aspectual retrieval condition, they were also asked to find relevant documents that are different and mention *new* airports. Thus, there were explicitly instructed to find a number of examples from different airports, as opposed to a similar or the same example based in the same airport multiple times. Subjects were instructed to find and save at least four useful documents (useful being relevant, or relevant and different, depending on the task).

### 8.2.4 Relevance Judgments and Aspects

For each topic, we used the corresponding TREC QREs from the Robust Track, to provide the relevance judgements for the study. However, to assess how many aspects were retrieved, we needed to commission additional labels as existing labels were not available for all the selected topics. First, for each topic, we examined the topic descriptions to identify what dimensions could be considered aspects of the topic. We noted that for each topic there was at least two ways this could be achieved: entity or narrative based. For example, in the topic on population growth, for a document to be relevant it could state the country (entity based) or the measure taken to reduce population growth (narrative based).

For the purposes of this study, it was decided that we should focus on entity based aspects. This was because “different narratives” were subject to greater interpretation than “different entities”. For each relevant document, two assessors extracted out the different

aspects, and we found that there was substantially higher agreement (95% vs 67%) between assessors across the entity based aspects: (341) airports, (347) species, (367) vessels, (408) storms, and (435) countries, as opposed to the more narrative based aspects: (341) the security measures taken, (347) the protection and conservation efforts, (367) the acts of piracy, (408) the death and destruction, and (435) the population control methods. Entity based aspects which we considered for each topic are listed below.

№ 341 (***Airport Security***) Different *airports* in which additional security measures were taken, e.g. *John F Kennedy International Airport, Boston Logan International Airport, or Leonardo Da Vinci International Airport.*

№ 347 (***Wildlife Extinction***) Different *species of endangered animals* under protection by states, e.g. *golden monkey, Javan rhino, or Manchurian tiger.*

№ 367 (***Piracy***) Different *vessels* that were boarded or hijacked, e.g. *Petro Ranger, Achille Lauro, or Global Mars.*

№ 408 (***Tropical Storms***) Different *tropical storms* where people were killed or there was major damage, e.g. *Hurricane Mitch, Typhoon Linda or Tropical Storm Frances.*

№ 435 (***Curbing Population Growth***) Different *countries* where population control methods were employed, e.g. *China, India or Zimbabwe.*

The total number of aspects identified for each topic were: 14 for № 341, 168 for № 347, 18 for № 367, 43 for № 408, and 26 for № 435. Judgements were put into the TREC Diversity format, as used by the `ndeval` application.<sup>1</sup>

---

<sup>1</sup>In the interests of promoting reproducibility and repeatability, the aspectual judgements will be made available for download at

## 8.2 The User Study

### 8.2.5 Systems: Non-Diversified and Diversified

Two experimental search systems were developed. These were identical except in terms of branding/logo and retrieval algorithm. First, in terms of branding, we created two fictional search engine names, *YoYo Search* and *Hula Search*, for which different colour schemes were used. The names were chosen as they were not associated with any major search engine (that we were aware of), nor did they imply that one of the systems performed better than the other. The colour schemes were chosen to provide the greatest difference in visual appearance to those with colourblindness (two variants of colourblindness, *protanopia* and *deutanopia*, were both considered). This was to ensure that subjects could later on indicate which system that they preferred, etc. Screenshots of the two systems in action are provided in Figure ???. Note, a generic *NewsSearch* system, which had a blue header, was used for the practice task, so that subjects could familiarise themselves with how to mark and save documents, and how the search functionality worked.

For the underlying search engine, we used the *Whoosh Information Retrieval (IR)* toolkit.<sup>2</sup> We used BM25 as the retrieval algorithm ( $b = 0.75$ ), but with an implicit ANDing of query terms to restrict the set of retrieved documents to only those that contained all query terms provided. This was chosen as most search systems implicitly AND terms together. BM25 thus served as the baseline, control for the non-diversified system condition. For the diversified system condition, we used BM25 to provide the initial ranking and then used xQuAD<sup>?</sup> to diversify the ranking. xQuAD has been shown to provide excellent, if not, state of the art performance, for web intent based diversification. To select the parameters for xQuAD, i.e.  $k$ , how many documents to re-rank, and  $\lambda$  how much focus on diversification, we performed a parameter sweep using a set of training queries from a prior user study. We explored a range of  $k$  and  $\lambda$  values, with 10–50 trialled for  $k$ , and 0.1–1.0 for  $\lambda$ . We selected  $k = 30$ ,  $\lambda = 0.7$  as is provided the best results ( $P@10 = 0.36$ ,  $\alpha DCG@10 = 0.075$ , aspectual recall@10 = 6.61) in terms of performance and efficiency – i.e. a higher  $k$  only slightly

---

<sup>2</sup>Whoosh can be accessed at <https://pypi.python.org/pypi/Whoosh/>.



increased performance, but took longer to compute.

### 8.2.6 Experimental Procedure

Subjects were provided a link to an online experimental system, that first presented the information sheet regarding the experiment followed by the consent form which they needed to agree to, in order to proceed. Note that ethics approval was sought before the experiment from Dept. of Computer and Information Sciences, The University of Strathclyde institution (ethics approval no. 622). Subjects were then asked to fill in a brief demographics survey, before undertaking a practice task to familiarise themselves with the interface. Once comfortable with the system, subjects could then proceed to undertake the four search tasks. Depending upon the Latin square rotation, subjects would then be provided with one of the four conditions on one of the four topics. For each task, first completed a pre task questionnaire, and after they had completed their search task, they were asked to fill in a post task questionnaire. At the end, of the experiment they were asked to fill in an exit questionnaire regarding which system they preferred.

### 8.2.7 Recruitment and Controls

Subjects for the experiment were recruited via the crowd sourcing platform *Amazon Mechanical Turk (MTurk)*. Previous work has shown that crowd sourced studies provide similar results as traditional lab-based user studies ???. That is, if sufficient controls are in place, otherwise workers may take try to take advantage and complete the task poorly ???. Therefore, it is important to ensure that quality control mechanisms are in place.

First we ensured that the browser/device and screen resolution used was desktop based (i.e. *Chrome, Firefox, Safari*, etc.) and 1024x768 or greater in size. As the experiment was conducted via a web browser, we wanted to ensure that only the controls provided by the experimental apparatus were used. So the experimental system launched a pop up size

## 8.2 The User Study

1024x768, which had all other browser controls disabled (to the best of our abilities), i.e. no history, back buttons, etc. The experimental system was tested on several major Web browsers, across different operating systems. This gave us confidence that a similar experience would be had across different system configurations.

Based on the suggestions from prior work ???, workers were only permitted to begin the experiment on the MTurk platform that: (i) were from the United States, and were native English speakers; (ii) had a HIT acceptance rate of at least 95%; and (iii) had at least 100 HITs approved. Requiring (ii) and (iii) increased the likelihood of recruiting individuals who wanted to maintain their reputation and would be more likely to complete the study in a satisfactory manner.

Subjects were informed that from our pilot study, it would take approximately 7-10 minutes to find at least four relevant documents per task - and the duration of the entire experiment would be approximately 40-50 minutes. Since we did not impose any time constraints on how long they searched for, we imposed an accuracy based control. We informed participants that their accuracy in identifying relevant material would be examined, and that they should aim to find four useful documents with at least 50% accuracy (based on TREC relevance judgments as the gold standard). Note that from a previous lab based study for this set of topics, the accuracy of participants was between 25% and 40% on average, depending on the topic, and so while we stipulated a higher accuracy, this was to motivate subjects to work diligently. Since we expected the experiment to take just under an hour, participants were compensated seven dollars (USD). In all, 64 subjects performed the experiment. However, 13 subjects were omitted either because they failed to complete all search tasks (five subjects were removed), failed to mark at least four documents (two subjects were removed), or spent less than two minutes per task and failed to retrieve any relevant documents (six subjects were removed).

Of the 51 subjects who successfully completed the experiment, 26 females and 25 males participated. The average age of the subjects was 38.66 years ( $min = 20$ ;  $max = 71$ ;  $stdev =$

11.43). 22 of the subjects reported having a bachelor's degree or higher, with the remaining 29 possessing an associate degree or lower. All subjects bar one expressed *Google* as their everyday search engine of choice. All subjects indicated that they conducted many searches for information via a search engine per week. Nearly three quarters of the subjects (i.e. 38 subjects) reported using a mouse for the experiment, with the remaining 13 using some form of trackpad.

### 8.2.8 Logging and Measures

Below we note the interactions logs and the measures taken while participants used the systems.

**Interaction Measures** included the number of queries issued by subjects, the number of documents that were examined, the number of different SERPs viewed, and the depths to which subjects clicked on – and hovered over – result summaries. It should be noted that components recorded such as hover depths over result summaries were inferred from the movement of the mouse cursor – eye-tracking equipment was not used in this study. In prior studies, the position of the mouse cursor on the screen has correlated strongly with the user's gaze on the screen ??.

**Performance Measures** included the number of documents that were saved by subjects, denoting that they were either relevant (for ad-hoc retrieval), or relevant and contain new information (for aspectual retrieval). From this, we could also break this number down into the number of documents that were saved and TREC relevant – as well as TREC non-relevant – and  $P@k$  measures at varying depths for the performance of the queries issued by the subjects. In addition, using the diversity QREs (generated as per the description in Section 8.2.4), we were able to determine how well the query performed in terms of how many new entities were in the top  $k$  results, and the  $\alpha$ DCG scores for each query. In addition, using the list of saved documents, we could identify how many entities that subjects

## 8.2 The User Study

had found, and how many documents contained one or more unseen entities – both in the context of the query results, and the overall search session.

From the log data, we could also compute additional performance measures, such as the accuracy that searchers reached during each session, as well as the probabilities of interaction. In the context of this study, accuracy referred to the ratio of documents that were TREC relevant, versus the total numbers of documents saved. For example, if a searcher saved four documents during a search session, with three of them being TREC relevant, the searcher's accuracy was 0.75. The interaction probabilities that we considered included: the probabilities of clicking on a result summary link ( $P(C)$ ) – given that it was either TREC relevant ( $P(C|R)$ ) or TREC non-relevant ( $P(C|N)$ ), and the probabilities of marking a document that was clicked ( $P(M)$ ) – given that it was either TREC relevant ( $P(M|R)$ ) or TREC non-relevant ( $P(M|N)$ ).

**Time-Based measures** included the time spent issuing queries (from query focus to query issue), the time spent on a SERP — as well as examining result summaries<sup>3</sup> – and the time spent examining documents. These times allowed us to then compute the total amount of time spent during the search session.

### 8.2.9 User Experience

To capture their perceived experiences, we asked subjects to complete both pre- and post-task surveys for each of the four experimental conditions they undertook.

Pre-task surveys consisted of five questions, each of which was on a seven-point Likert scale (7 – strongly agree to 1 – strongly disagree). Subjects were sought for their opinions on their: (i) prior knowledge of the topic; (ii) the relevancy of the topic to their lives; (iii) their desire

---

<sup>3</sup>Result summary times were approximated by dividing the total recorded SERP time by the number of snippets hovered over with the mouse cursor. We believe this is a reasonable assumption to make – network latency issues beyond our control ensured that the mouse hover events occasionally were delivered at the wrong times, and in the wrong order.

to learn about the topic; (iv) whether they had searched on this topic before; and (v) the perceived difficulty to search for information on the topic.

Following the completion of each search task, subjects were provided with a post-task survey, again using a seven-point Likert scale. The survey considered aspects on (i) their behaviour, and (ii) how they felt the system performed. Considering their behaviours, subjects were asked for their opinions on: how successful they thought they were at completing the task (*success*); how quickly they felt they completed the task (*subject speed*); whether they issued different queries to explore the topic (*queries*); if they only examined a few documents per query (*documents*); whether they checked each document carefully before saving (*checks*); and whether they saved more documents than was required, with a minimum of four being required (*more*). Subjects were also asked for their opinions on: whether they thought the system helped them complete the task quickly (*system speed*); whether they felt the system made it difficult to find useful information (*difficulty*); if the system made it easy to complete the task (*ease*); if they were happy with how the system performed (*happiness*); whether the system was cumbersome or not (*cumbersome*); and whether they were confident in the decisions they made (*confident*). Upon completion of the experiment, subjects were provided with an exit survey consisting of several questions. Here, we wanted to ascertain which of the two search system offered the better experience and which one they preferred.

## 8.3 Results

We now address our research questions and hypotheses as addressed in Section ???. Both the behaviour and performance of each subject were analysed across each of the four experimental conditions, *D.As*, *ND.As*, *D.Ad* and *ND.Ad*. Task (*As.* vs *Ad.*) and System (*ND.* vs *D.*) effects were also examined. To evaluate these data, ANOVAs were conducted using the conditions, systems and tasks each as factors; main effects were examined with  $\alpha = 0.05$ .

### 8.3 Results

**Table 8.1:** Query statistics and performance measures across both of the experimental systems trialled, **ND** (Non-Diversified) and **D** (Diversified). Note the significant differences between the diversity-centric measures,  $\alpha DCG$  (where  $\alpha = 0.5$ ) and Aspectual Recall (**Asp.R.**), highlighting that the diversification algorithm did indeed provide a more diverse set of results to the subjects.

		ND	D
	Queries Is-sued	718	555
	Terms per Query	3.59	3.80
Unique Terms		345	292
<i>Prec.</i>	<i>P@5</i>	$0.25 \pm 0.01^*$	$0.29 \pm 0.01^*$
	<i>P@10</i>	$0.22 \pm 0.01$	$0.24 \pm 0.01$
$\alpha DCG$	$\alpha DCG@5$	$0.02 \pm 0.00^*$	$0.04 \pm 0.00^*$
	$\alpha DCG@10$	$0.03 \pm 0.00^*$	$0.04 \pm 0.00^*$
<i>Asp.R.</i>	<i>Asp.R.@5</i>	$1.40 \pm 0.11^*$	$3.39 \pm 0.21^*$
	<i>Asp.R.@10</i>	$2.11 \pm 0.14^*$	$4.07 \pm 0.24^*$

Bonferroni tests were then used for post-hoc analysis. It should be noted, however, that where  $\alpha DCG$  is reported, we compute the values using  $\alpha = 0.5$ .

To begin with our analysis, we first examined whether the performance experienced by participants on the two systems was in fact different (as indicated by our pilot study). We took the queries participants issued to each system, and measured the performance according to  $\alpha DCG$ , aspectual recall and precision (see Table 8.1). Statistical testing confirms that the two systems were significantly different in terms of diversity (i.e.  $\alpha DCG@10$ :  $F(1, 1272) = 28.74, p < 0.001$ ), and aspectual recall@10:  $F(1, 1272) = 55.43, p < 0.001$ ). However,  $P@10$  was not significantly different - suggesting that the re-ranking promoted relevant and diverse documents, but only in the top 10 on average.

**Table 8.2:** Behavioural and performance measures across each condition, system and task.

Measure	Experimental Conditions				Systems		Tasks	
	D.As	ND.As	D.Ad	ND.Ad	ND	D	Ad.	As.
#Queries	5.92± 0.88	5.25± 0.80	4.96± 0.74	5.20± 0.69	5.23± 0.53	5.44± 0.58	5.08± 0.51	5.59± 0.59
#SERPs/Q.	1.78± 0.14	2.42± 0.24	2.28± 0.31	2.28± 0.20	2.35± 0.16	2.03± 0.17	2.28± 0.18	2.10± 0.14
Doc./Q.	3.02± 0.39	3.65± 0.46	3.48± 0.51	3.23± 0.37	3.44± 0.29	3.25± 0.32	3.36± 0.31	3.34± 0.30
Depth/Q.	12.85± 1.49	15.73± 2.53	16.19± 2.14	13.94± 1.93	14.84± 1.58	14.52± 1.31	15.07± 1.44	14.29± 1.47
#Saved	5.80± 0.26	5.96± 0.25	5.92± 0.25	5.78± 0.20	5.87± 0.16	5.86± 0.18	5.85± 0.16	5.88± 0.18
#TREC Saved	2.63± 0.22	2.18± 0.23	2.51± 0.23	2.22± 0.22	2.20± 0.16	2.57± 0.16	2.36± 0.16	2.40± 0.16
#TREC Non.	1.75± 0.22	1.96± 0.23	1.37± 0.22	1.82± 0.23	1.89± 0.16	1.56± 0.16	1.60± 0.16	1.85± 0.16
#Ent. Found	7.22± 0.94*	4.31± 0.60*	5.82± 0.77	4.37± 0.59*	4.34± 0.42*	6.52± 0.61*	5.10± 0.49	5.76± 0.57
#Docs. Ent.	3.20± 0.21*	2.35± 0.20*	2.63± 0.23	2.02± 0.18*	2.19± 0.13*	2.91± 0.16*	2.32± 0.15*	2.77± 0.15*

Aside from showing query performance, Table 8.1 also reports the number of terms issued per query over each systems *ND* and *D*; of the 1273 queries issued, those issued to *ND* were shorter on average, with 3.59 terms compared to 3.80 terms for *D*. However, the vocabulary used by subjects issuing queries to *ND* was greater than *D* – queries issued to *ND* contained 345 unique terms, compared to 292 for *D*. This provides our first finding of note. When using *ND*, participants issued more queries – but were slightly shorter and more varied – in order to accomplish their tasks.

### 8.3.1 Observed Behaviours

**Interactions.** Table 8.2 presents the mean (and standard deviations) of (i) the number of queries issued, (ii) the number of SERPs that were examined by subjects per query, (iii)

### 8.3 Results

**Table 8.3:** Interaction times across each condition, system and task. Included is: the mean total session time; the per query time (*Per Q.*); the per document time (*Per D.*); and the per result summary (snippet) time (*Per Snip.*). Results presented in seconds.

Time	Experimental Conditions				Systems		Tasks	
	D.As	ND.As	D.Ad	ND.Ad	ND	D	Ad.	As.
<b>Total Ses- sion</b>	443.65± 45.05	430.50± 38.39	432.18± 49.87	447.55± 47.82	439.02± 30.52	437.91± 33.44	439.86± 34.38	437.08± 29.45
<b>Per Q.</b>	8.80± 0.89	9.99± 1.21	9.69± 0.79	8.69± 0.57	9.34± 0.67	9.25± 0.59	9.19± 0.49	9.39± 0.75
<b>Per D.</b>	15.97± 1.96	13.03± 1.01	13.66± 1.02	15.09± 2.20	14.06± 1.21	14.81± 1.10	14.37± 1.21	14.50± 1.11
<b>Per Snip.</b>	1.59± 0.09	1.75± 0.15	1.71± 0.11	1.71± 0.13	1.73± 0.10	1.65± 0.07	1.71± 0.08	1.67± 0.09

the number of documents examined (clicked) per query, and (*iv*) the click depth (or search stopping depth) per query. Statistical tests reveal no effects across conditions, systems or tasks. However, there are several trends that are worth mentioning. Firstly, we notice that when participants used the diversified system to complete the aspectual retrieval task, they examined fewer documents per query than when completing the same task on the non-diversified system (12.85 vs. 15.73) – which is in line with *H1*. We also observed that participants issued slightly more queries on the diversified system compared to the non-diversified system with the aspectual retrieval task (5.92 vs. 5.25) – which is in line with *H2a* — but these results were not significantly significant.

Turning our attention to the ad-hoc retrieval tasks, while our hypotheses suggested that there would be no differences in terms of the number of documents examined (*H3*) or in the number of queries issued (*H4*) – which was the case – however we note that participants on the diversity system inspected more results than when on the non-diversified system (16.19 vs. 13.94), and they issued slightly fewer queries (4.96 vs 5.20). We can see the trade-offs between queries and the number of results inspected per query, where more queries tend to lead to fewer results being examined, and vice versa. This trend suggests that participants, when searching on the diversified system, for relevance, may have had to



**Table 8.4:** Interaction probabilities, as observed over the four experimental conditions. Refer to Section 8.2.8 for an explanation of each probability's meaning.

Prob.	D.As	ND.As	D.Ad	ND.Ad
$P(M)$	$0.67 \pm 0.03$	$0.66 \pm 0.03$	$0.70 \pm 0.03$	$0.71 \pm 0.04$
$P(M R)$	$0.78 \pm 0.04$	$0.63 \pm 0.05$	$0.74 \pm 0.04$	$0.67 \pm 0.05$
$P(M N)$	$0.59 \pm 0.04$	$0.61 \pm 0.04$	$0.65 \pm 0.04$	$0.65 \pm 0.04$
$P(C)$	$0.16 \pm 0.01^*$	$0.21 \pm 0.02^*$	$0.16 \pm 0.01^*$	$0.20 \pm 0.01^*$
$P(C R)$	$0.27 \pm 0.03$	$0.30 \pm 0.04$	$0.25 \pm 0.03$	$0.31 \pm 0.04$
$P(C N)$	$0.13 \pm 0.02^*$	$0.18 \pm 0.02^*$	$0.13 \pm 0.01^*$	$0.17 \pm 0.02^*$

dig deeper, to find more relevant material (due to system performance), or that the system encouraged participants to go deeper (which is what we intuitively inspected when they were searching for diversity). Either way, we find no conclusive evidence to support the studies main hypotheses – only trends.

Table 8.4 reports interaction probabilities associated with user interactions, i.e. the probability of marking a document saved  $P(M)$ , and the probability of clicking a document  $P(C)$  along with the conditional probabilities for each based on whether the document saved or clicked was TREC (R)levant or (N)on-Relevant. From the table, we can see that there was a significant difference between conditions (and systems, not shown) for the probability of a click, and the probability of clicking on non-relevant items. Comparing systems indicated that participants clicked more when using the non-diversified system, and clicked on more non-relevant documents. However, we did not observe any task effects. This suggests that the non-diversified system affected led to examining more documents, but often more non-relevant documents. This is reflected by the fact that across all the performance measures (see below), participants on the non-diversified system performed worse.

### 8.3 Results

**Time-Based Measures.** Table 8.3 reports the time taken for various interactions, across each condition, system and task. We report: the mean total session time (from the first query focus to ending the task); the mean time spent entering queries; the mean per document examination time; and the mean time spent examining a result summary (or snippet). All values are reported in seconds. Surprisingly, no significant differences were found between any of the comparisons over the total session times, the per query times, the per document times, and the per snippet times. Results however do show a relatively constant mean session time over each of the four experimental conditions, at  $\approx 438.5$  seconds, which is about 7 minutes, on average – this was in line with the time taken to find four documents in our previous studies with similar workers ? and lab participants Maxwell and Azzopardi (2016a). Considering hypothesis *H2b*, no evidence was found to support that under the diversity system *D* with an aspectual task that completion times would be lower. Here, we can see that they were in fact slightly higher (443 seconds *D* vs. 430 seconds on *ND*, i.e. the difference of about examining one more document).

**Performance.** In Table 8.2, we also report a number of performance measures: the number of saved documents – also broken down into the number of TREC saved and TREC non-relevant and saved, along with the number of new entities found (within saved documents, with new being in the context of a search session) – and the number of documents containing at least one new entity. In terms of the documents saved, there were no significant differences between conditions, systems or tasks. On average, participants saved around 6 documents on average, which was two more than the goal set, 4 – suggesting that wanted to make sure that they found a few extra, just in case some were not relevant/useful.

However, when we look at the entity-related measures, we note that participants found more documents that contained new entities and found more entities overall when using the diversity system. This was significantly different ( $6.52 \pm 0.61$  compared to  $4.34 \pm 0.42$  respectively, where  $F(1, 203) = 8.70, p < 0.05$ ). When examining each condition, the Bonferroni follow-up test showed significant differences were observed between condition *D*. As

and conditions *D.Ad* and *ND.Ad*, where  $F(3, 203) = 3.49, p < 0.05$ . Also, we notice that participants also found more documents with entities, and more entities when using the ad-hoc retrieval task when using the diversity system than when they used the non-diversity system (docs with entities: 2.63 vs. 2.02, new entities: 5.82 vs 4.37). Though this was not significantly different, it does suggest that when participants used the diversity system, they did learn more about the different aspects of the topic (or at least encountered more aspects) than when using the non-diversity system.

**Post Task and Post System Questionnaires.** There were no notable significant differences between conditions, tasks, or system for any of the post task questions. For the post system questionnaires, participants were roughly evenly split between their preference for the diversified or non-diversified system – again with no significant differences. This finding suggests that despite the substantial (and significant) difference in aspectual recall and other system performance measures, between the systems, participants seemed largely ambivalent to the different system’s influence. Though, of course, their observed behaviours do suggest that the system (and task) did affect their performance.

### 8.3.2 Gain over Time

We motivated this study using IFT, where we constructed a number of gain curves that reflected our beliefs about the search performance experienced by users would look like on each system and task. This was done in order to generate the aforementioned hypotheses. Here, we examine how participants performed over time for each of the systems and conditions to infer the gain curves. We then compare that to our expectations (which are shown in Figure ??).

To create empirical gain curves, we plotted cumulative gain over time, where we defined gain to be the number of saved relevant documents (in the case of ad-hoc retrieval), and gain to be the number of saved relevant but different documents (in the case of aspectual

### 8.3 Results

retrieval). These definitions are what we said would constitute a useful document in these tasks. And as they are in the same units, we can plot the gain for both tasks on the same axes.

Figure ?? shows the corresponding empirical gain curves for: (a) the non-diversified system on both tasks, (b) the diversified System on both tasks, (c) the aspectual task for both systems, and (d) the ad-hoc task for both systems. Compared to our expectations in Figure ??, on visual inspection, we see that our predictions were roughly in line with the gain experienced. For example, in (a) we hypothesised that on the non-diversified system, participants would experience greater levels of gain, and the empirical gain curves show this. A critical difference though is for (b) – where we hypothesised that the gain curves would be similar on the diversified system, up until a point, before the aspectual gain would drop. From (b), it is clear that participants had a very different experience – and experienced lower gains from the beginning – motivating a revision of our expectations.

To do so, we first fit a logarithmic function to each of the gain curves given time, such that:  $gain = a \cdot b \log(time)$ . Table 8.5 shows the parameters and correlation co-efficients for fit ( $r^2$ ) for each condition. We then could calculate how many documents a participant would examine by drawing the tangent line to the estimated gain functions from the origin. This resulted in the predicted number of documents examined – which we see are in line with the actual documents examined. With respect to (b), we see that for the diversified system, the theory, given their performance, suggests that participants should examine more documents per query on the aspectual task than when undertaking the ad-hoc task (i.e. 4.98 to 3.36, respectively). We observed that they examined 3.48 and 3.02 documents per query – which follows the same trend, but not the same magnitude. Thus, the revising our expectations regarding how people would search differently between these tasks. With respect to *H1*, we see that the theory, given their performance, suggests that participants, when undertaking the aspectual task, would examine fewer documents per query when using the diversified system than on the non-diversified system (4.36 vs 4.92). Again, we see that they examined 3.02 and 3.65 documents per query respectively, again following the same trend

– but not the same magnitude. This post-hoc analysis has provided justification for some of our initial hypotheses regarding how search behaviour would change under the different conditions – but it has also led to us revising our expectations based on the observed, empirical data.

In this paper, we investigated the effects of diversifying search results when searchers undertook complex search tasks, requiring one to learn about different aspects of a topic. We inferred a number of hypotheses based upon Information Foraging Theory, in which diversification would lead to searchers examining fewer documents per query, and subsequently issuing more queries. We tested our hypotheses by conducting a within-subjects user study, using (i) a non-diversified system; versus (ii) a diversified system, when the search task was either: (a) ad-hoc; or (b) aspectual.

Our findings lend evidence to broadly support our hypotheses; however, our results were not statistically significant. This was despite the fact that there were significant differences in the two systems performance, i.e. the diversified system returned a ranked list of results with a greater number of documents containing new, unseen entities. Clearly, bigger differences need to be present before participants can subjectively report whether they had a different experience, or which one they preferred – as post task and system questions revealed no difference. However, in terms of performance, we found that participants on the diversified system did perform better – more relevant documents were found, and more new entities were found – suggesting they found out more about the topics on the diversified system. They also inspected few non-relevant documents. After conducting a post-hoc analysis, we showed that the hypotheses we posited given IFT were sound, but revised our expectations on how participants would behave when using the diversified system. That is, they would examine more documents per query, and thus issue fewer queries when undertaking the aspectual retrieval task, as opposed to there being no difference in performance. Again, we see a trend to support the hypothesis. Encouragingly, our application of Information Foraging Theory, before and after the study, led to new insights into how behaviours are affected under the different conditions – and is a useful tool in developing,

### 8.3 Results

**Table 8.5:** Table highlighting the fitting parameters for the gain curves illustrated in Figure ?? over each experimental condition. Also included are the estimations from the model for the time to examine a document, and the depth to which subjects should go -- as well as the observed number of documents examined, and stopping depth (on average).

Condition	Model Fitting			Pred.	Actual
	$a$	$b$	$r^2$	Docs.	Docs.
ND.Ad	-1.08	0.48	0.989	3.68	3.23
ND.As	-0.57	0.23	0.987	4.92	3.65
D.Ad	-1.22	0.52	0.959	4.98	3.48
D.As	-0.68	0.29	0.985	4.36	3.02

motivating and analysing search performance and behaviours. Counter to our intuition about how we *believed* people would behave in these conditions, the theory provided more informed and accurate hypotheses.

In past work, mainly interface based solutions were studied – where few significant differences in behaviour were found compared to a standard interface. Disappointingly, we also find that an algorithmic solution has very little influence either, though there were trends which indicated that diversifying the results does lead to better performance, greater awareness of the topic (even when not specifically instructed, i.e. *find relevant only*), and fewer examinations of non-relevant items. Thus, we suggest that diversification should be employed more widely – in particular in the context of news search – where bias is an issue and diversification algorithms can present a broader overview of the aspects within a topic.

8.3.3 Results

8.3.4 Conclusions

8.4 Simulation Experiments

8.4.1 Simulated Experimental Design

8.4.2 Comparisons

8.4.3 Conclusions

8.5 Conclusions

8.6 Chapter Summary

## Part IV

# Conclusions

*This final part of the thesis details the findings from this research, as well as exploring potential areas of exploration for future work.*



## Chapter 9

# Conclusions and Future Work

### 9.1 Discussion and Contributions

#### 9.1.1 User Modelling

#### 9.1.2 Examining Stopping Behaviours

### 9.2 Conclusions

### 9.3 Future Research Directions

When considering possible future directions, Apple's 1987 Knowledge Navigator vision of IR is still a strong exemplar of how search systems might develop. The short film showed a college professor pulling together a lecture presentation at the last minute. The professor used a form of tablet computer running an IR system presented as an agent capable of

### 9.3 Future Research Directions

impeccable speech recognition, natural dialogue management, a high level of semantic understanding of the searcher's information needs, as well as unbounded access to documents and federated databases. The Knowledge Navigator identified and connected the professor to a colleague who helped him with the lecture. The broader implications of finding people (rather than documents) to aid with information needs that we see facilitated in the vast growth of social media was not really addressed in the Apple vision. What it also did not encompass was the portability of computer devices opening the possibility of serving information needs pertinent to the particular local context of location, location type, route, the company one is in, or a combination of all these factors. **from the history of IR paper**

# Appendices



## Appendix A

# The SimIIR Framework

One of the key contributions of this thesis is the introduction of a framework enabling the simulation of interaction.

**N.B.** Portions of this chapter are based upon the demonstration paper *Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction* (Maxwell and Azzopardi, 2016b). This demonstration was presented at the 39<sup>th</sup> International ACM SIGIR Conference on Research and Development in Information Retrieval, held in Pisa, Italy. Additional components have however been included as the underlying searcher model was evolved during the course of this project.

## A.1 Architecture Overview

### A.1.1 The Searcher Model

### A.1.2 Searcher Contexts

### A.1.3 Topics

### A.1.4 Search Interface/Engine

### A.1.5 Output Controller

### A.1.6 Querying Strategies/Generators

### A.1.7 SERP Level Stopping

### A.1.8 Snippet/Document Classifiers

### A.1.9 Snippet Level Stopping

### A.1.10 Loggers

## A.2 Example Input

## A.3 Example Output

## Appendix B

# Original Publications

This section provides copies of all the publications that form the basis of this doctoral thesis.

## B.1 How Temporal Delays Affect Search Behaviour

### B.1 How Temporal Delays Affect Search Behaviour



## B.2 Fixed and Adaptive Stopping Strategies

## **B.3 An Analysis of Stopping Rules and Strategies**

## B.4 Building Realistic Simulations for Interactive Information Retrieval

## B.5 SimIIR: A Framework for the Simulation of Interaction

## B.6 Agents, Simulated Users and Humans

## **B.7 A Study of Snippet Length and Informativeness**

## Bibliography

- Azzopardi, L. (2011). The economics in interactive information retrieval. In *Proc. 34<sup>th</sup> ACM SIGIR*, pages 15–24.
- Azzopardi, L. and Zuccon, G. (2015). An analysis of theories of search and search behavior. In *Proc. 1<sup>st</sup> ACM ICTIR, ICTIR '15*, pages 81–90.
- Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Banks, J., Carson, J., and Nelson, B. (1996). *Discrete-event System Simulation*. Prentice-Hall international series in industrial and systems engineering.
- Bates, M. J. (1979). Information search tactics. *Journal of the American Society for Information Science*, 30(4):205–214.
- Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424.
- Belkin, N. (1980). Anomalous states of knowledge as a basis for information retrieval. 5:133–143.

- Berners-Lee, T., Dimitroyannis, D., Mallinckrodt, A. J., McKay, S., et al. (1994). World wide web. *Computers in Physics*, 8(3):298–299.
- Borlund, P. (2000). Evaluation of interactive information retrieval systems. Unpublished doctoral dissertation, Åbo Akademi University.
- Borlund, P. (2003). The iir evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8.
- Bota, H., Zhou, K., and Jose, J. M. (2016). Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proc. 1<sup>st</sup> ACM CHIIR*, pages 131–140.
- Brenes, D. J. and Gayo-Avello, D. (2009). Stratified analysis of aol query log. *Information Sciences*, 179(12):1844 – 1858.
- Browne, G. J. and Pitts, M. G. (2004). Stopping rule use during information search in design problems. *Organizational Behavior and Human Decision Processes*, 95(2):208 – 224.
- Carr, N. (2008). Is google making us stupid? *Yearbook of the National Society for the Study of Education*, 107(2):89–94.
- Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using information scent to model user information needs and actions and the web. In *Proc. CHI 2001*, pages 490–497.
- Cleverdon, C., Mills, J., and Keen, M. (1966). *Factors Determining the Perfor-*



*mance of Indexing Systems*, volume 1:2 of *Factors Determining the Performance of Indexing Systems*.

Cleverdon, C. W. (1962). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems / by*. Cranfield University.

Cooper, W. S. (1973a). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100.

Cooper, W. S. (1973b). On selecting a measure of retrieval effectiveness part ii. implementation of the philosophy. *Journal of the American Society for Information Science*, 24(6):413–424.

Cutts, Q., Connor, R., Michaelson, G., and Donaldson, P. (2014). Code or (not code): Separating formal and natural language in cs education. In *Proc. 9<sup>th</sup> WiPSCE*, pages 20–28.

Dewey, M. (1891). Decimal classification and relative index for libraries, clippings, notes, etc. 240(41):407–593.

Dostert, M. and Kelly, D. (2009). Users’ stopping behaviors and estimates of recall. In *Proc. 32<sup>nd</sup> ACM SIGIR*, pages 820–821.

Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(11):989–1003.

Fishwick, P. A. (1995). Computer simulation: The art and science of digital world construction. Technical report, University of Florida.

- Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265.
- Gordon, M. and Lenk, P. (1991). A utility theoretic examination of the probability ranking principle in information retrieval. *Journal of the Association for Information Science and Technology*, 42(10):703–714.
- Haessig, J. D. A. and Friedland, B. (1991). On the modeling and simulation of friction. *Journal of Dynamic Systems, Measurement, and Control*, 113(3):354–362.
- Harman, D. (1993). Overview of the first trec conference. In *Proc. 16<sup>th</sup> ACM SIGIR, SIGIR '93*, pages 36–47.
- Harman, D. (2010). Is the cranfield paradigm outdated? In *Proceedings of SIGIR 2010*, pages 1–1.
- Hastie, R. (1988). A computer simulation model of person memory. *Journal of Experimental Social Psychology*, 24(5):423 – 447.
- Hedstrom, M. (1997). Digital preservation: A time bomb for digital libraries. *Computers and the Humanities*, 31(3):189.
- Heermann, D. W. (1990). *Computer-Simulation Methods*, pages 8–12.
- Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*.

- Kelly, D. and Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770.
- Keskustalo, H., Järvelin, K., and Pirkola, A. (2008). Evaluating the effectiveness of relevance feedback based on a user simulation model: Effects of a user scenario on cumulated gain value. *Information Retrieval*, 11(3):209–228.
- Korolova, A., Kenthapadi, K., Mishra, N., and Ntoulas, A. (2009). Releasing search queries and clicks privately. In *Proc. 18<sup>th</sup> WWW*, pages 171–180.
- Mahmud, K. and Town, G. E. (2016). A review of computer tools for modeling electric vehicle energy requirements and their impact on power distribution networks. *Applied Energy*, 172(Supplement C):337 – 359.
- Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press.
- Marchionini, G., Dwiggins, S., Katz, A., and Lin, X. (1993). Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. *Library and Information Science Research*, 15(1):35–69.
- Maxwell, D. (2016). Building realistic simulations for interactive information retrieval. In *Proc. 1<sup>st</sup> ACM CHIIR*, pages 357–359.
- Maxwell, D. and Azzopardi, L. (2014). Stuck in traffic: How temporal delays affect search behaviour. In *Proc. 5<sup>th</sup> IliX*, pages 155–164.

- Maxwell, D. and Azzopardi, L. (2016a). Agents, simulated users and humans: An analysis of performance and behaviour. In *Proc. 25<sup>th</sup> ACM CIKM*, pages 731–740.
- Maxwell, D. and Azzopardi, L. (2016b). Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proc. 39<sup>th</sup> ACM SIGIR*, pages 1141–1144.
- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015a). An initial investigation into fixed and adaptive stopping strategies. In *Proc. 38<sup>th</sup> ACM SIGIR*, pages 903–906.
- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015b). Searching and stopping: An analysis of stopping rules and strategies. In *Proc. 24<sup>th</sup> ACM CIKM*, pages 313–322.
- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proc. 40<sup>th</sup> ACM SIGIR*, pages 135–144.
- Mcbryan, O. A. (1994). GENVL and WWW: Tools for taming the web. In *Proceedings of the first World Wide Web Conference*.
- Miller, G. A. (1983). Informavores. In Machlup, F. and Mansfield, U., editors, *The Study of information : interdisciplinary messages*, pages 111–113.
- Moffat, A., Thomas, P., and Scholer, F. (2013). Users versus models: What

- observation tells us about effectiveness metrics. In *Proc. 22<sup>nd</sup> ACM CIKM*, pages 659–668.
- Nickles, K. (1995). *Judgment-based and reasoning-based stopping rules in decision making under uncertainty*. PhD thesis, University of Minnesota.
- Piorkowski, D., Fleming, S., Scaffidi, C., Bogart, C., Burnett, M., John, B., Bellamy, R., and Swart, C. (2012). Reactive information foraging: An empirical investigation of theory-based recommender systems for programmers. In *Proc. CHI 2012*, pages 1471–1480.
- Pirolli, P. and Card, S. K. (1999). Information foraging. *Psychological Review*, 106:643–675.
- Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, 2nd edition.
- Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.
- Robertson, S. E. (1977). *Journal of Documentation*, 33(4):294–304.
- Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993). The cost structure of sensemaking. In *Proc. CHI 1993*, pages 269–276.
- Sanderson, M. and Croft, W. B. (2012). The history of information retrieval research. *Proceedings of the IEEE*, 100:1444–1451.
- Sandstrom, P. E. (1994). An optimal foraging approach to information

- seeking and use. *The Library Quarterly: Information, Community, Policy*, 64(4):414–449.
- Smith, G. F., Benson, P. G., and Curley, S. P. (1991). Belief, knowledge, and uncertainty: A cognitive perspective on subjective probability. *Organizational Behavior and Human Decision Processes*, 48(2):291–321.
- Sparko, A., Burki-Cohen, J., and Go, T. (2010). chapter Transfer of Training from a Full-Flight Simulator Vs. a High-Level Flight-Training Device with a Dynamic Seat. Guidance, Navigation, and Control and Co-located Conferences. American Institute of Aeronautics and Astronautics.
- Stephens, D. and Krebs, J. (1986). *Foraging Theory*.
- Tocher, K. (1963). *The art of simulation*. Electrical engineering series.
- Toms, E. G. and Freund, L. (2009). Predicting stopping behaviour: A preliminary analysis. In *Proc. 32<sup>nd</sup> ACM SIGIR*, pages 750–751.
- Turpin, A. and Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Proc. 29<sup>th</sup> ACM SIGIR*, pages 11–18.
- Turpin, A. H. and Hersh, W. (2001). Why batch and user evaluations do not give the same results. In *Proc. 24<sup>th</sup> ACM SIGIR*, pages 225–231.
- Varian, H. R. (1999). Economics and search. *SIGIR Forum*, 33(1):1–5.
- Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing & Management*, 40(4):677–691.

- Voorhees, E. and Harman, D. (2000). Overview of the eighth text retrieval conference (trec-8). In *Proc. TREC-8*, pages 1–24.
- Wang, K., Li, X., and Gao, J. (2010). Multi-style language model for web scale information retrieval. In *Proc. 33<sup>rd</sup> ACM SIGIR*, pages 467–474.
- Wilson, M. L., Kules, B., Schraefel, M., and Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Found. Trends Web Sci.*, 2(1):1–97.
- Wu, W. and Kelly, D. (2014). Online search stopping behaviors: An investigation of query abandonment and task stopping. Seattle, WA.
- Wu, W., Kelly, D., and Sud, A. (2014). Using information scent and need for cognition to understand online search behavior. In *Proc 37<sup>th</sup> ACM SIGIR*, pages 557–566.
- Zach, L. (2005). When is “enough” enough? modeling the information-seeking and stopping behavior of senior arts administrators: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(1):23–35.
- Zipf, G. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*.
- Zuccon, G., Leelanupab, T., Whiting, S., Yilmaz, E., Jose, J. M., and Azopardi, L. (2013). Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems. *Inf. Retr.*, 16(2):267–305.