Automatically Identifying Coherent Web Sessions from Browser Logs



Chaoyu Ye

Department of Computer Science

University of Nottingham

This dissertation is submitted for the degree of

Doctor of Philosophy

Mixed Reality Lab September 2019

I would like to dedicate this thesis to:

my wife

my son

my parents

my brother

my parents-in-law

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation 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 dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains appendices, bibliography, footnotes, tables, equations, pseudo-code and figures.

Chaoyu Ye

September 2019

Acknowledgements

Firstly, I would like mostly to thank my primary supervisor, Dr. Max L.Wilson, for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Also, I would like to thank Emily (Max's wife) for the beautiful planet choloclate on my viva day.

Besides my primary supervisor, I would like to thank examiners: Prof. Natasa Milic-Frayling and Prof. Paul Clough for their time, effort and insightful feedback. I would like to thank Dr. Michel Valstar for his time and effort for my first two years annual review. I would like to thank Dr. Zhou Ke for applying my GT dataset as one of his student's final year projects, and also some guidance on my first touch in Machine Learning. I would like to thank Prof Twidale, Dr. Sophie Rutter, and Verena for the exciting Search Literacy Project. I would like to thank my secondary supervisor Prof. Tom Rodden.

I would like to thank my labmates and friends Matthew, Horia, Wenchao, Neha, Yitong, Leigh and Shashank for all the fun we have had in the last few years.

I am indebted to the participants in this research. Without their contribution, I would never complete this thesis.

Last but not the least, I would like to thank my family: my wife, my 3-year old boy, my parents, my brother and my parents-in-law for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

Due to the increasing diversity in both user's behaviour and the types of web tasks performed, many studies in information retrieval (IR) are turning towards session-based retrieval rather than single URL-query pairs. However, extracting the meaningful session data from the raw discrete logs is still a significant challenge. Most prior studies have been based on datasets where the logs of each user's web history were simply divided by fixed periods of inactivity, such as 5, 15, or 30 minutes [52, 31]. There have also been some attempts beyond these simplistic fixed timeouts [91]. Rather than covering all web activities, they focus on search-related activities only. Consequently, it is necessary to finding a meaningful way to cluster all activities including both searching and browsing on a web browser. The goal of this study is to find a way to better automatically segment users' web activity into sessions. There are three research stages: 1) how people understand their mental model in the session segmentation, 2) how these self-identified sessions look in practically implemented web logs, and 3) how we can algorithmically identify these sessions from browser activity, and how each algorithm performs. To answer these questions, firstly a qualitative study was conducted and a taxonomy of six factors related to the user-identified sessions was generated. Then a Chrome Extension was built that provided the practical reflection of user-identified sessions with comprehensive sets of web logs including both user interaction and visit details. This helped to gather a ground truth dataset to support further evaluation. Finally, several algorithmic approaches to automatically clustering web activities closer to user-identified sessions were evaluated.

Table of contents

List of figures **xvii**

List of tables **xix**

1 Introduction **1**

1.1 Background 1

1.2 Goal and Research Questions 4

1.3 Overview of the Research Stages 5

1.4 Thesis Structure 7

1.5 Publications and Related Projects 8

1.5.1 PhD Relevant Publication 8

1.5.2 Other Publication 8

2 Literature Review **11**

2.1 The Information-Seeking Process 11

2.2 User Interaction on Web 14

2.2.1 Factors That Influence User Interaction 14

2.2.2 Successive Versus. Interleaving 16

2.2.3 Searching: Query Reformulation 18

2.2.4 Navigation: Page Revisitation 19

2.3 Session Definition 20

2.3.1 Historical Definition 20

2.4 Existing Session Determination Techniques 23

2.4.1 Automatic Segmentation: Content-Ignorant 24
2.4.2 Automatic Segmentation: Lexical 25
2.4.3 Automatic Segmentation: Combined 25
2.5 Practical Applications 26

3 Methodology **29**
3.1 Data Collection 30
3.1.1 Logging Tool 30
3.1.2 Interview 31
3.2 User Study 31
3.3 Evaluation 32
3.3.1 Cranfield Paradigm 33
3.4 Combining Approaches 33
3.5 Summary 35

4 Study 1 **37**
4.1 Goal 37
4.2 Participants and Protocol 38
4.3 Findings 41
4.3.1 Taxonomy 41
4.3.2 Session Categorisation via Card Sorting 50
4.3.3 Session Categories: Sellen's Six Categories and Elserwiler's Three Categories 54
4.3.4 Other Findings 55
4.4 Summary 56
4.4.1 Applying the Taxonomy 57

5 Chrome Extension **59**
5.1 Design Detail 60
5.1.1 Single Log 60

5.1.2 Data Storage 60
5.1.3 Data Submission 61
5.1.4 Pre-Processing Dummy Logs 62
5.1.5 Test 63
5.2 Software Architecture 63

6 Study 2: Ground Truth Dataset 67
6.1 Goal 67
6.2 Participants and Protocol 68
6.3 Ground Truth Dataset: Session 69
6.3.1 Overview 69
6.3.2 User-Level 70
6.3.3 Session Level 71
6.4 Ground Truth Dataset: Session Boundary 82
6.4.1 Interactions at Each Boundary 83
6.5 Summary 84

7 Study 3: Five Versions of Algorithms and Fixed Timeout 87
7.1 Whole Flow 89
7.1.1 Atomic Decision Making 92
7.2 Five Versions Algorithm 94
7.2.1 A_1 (Baseline): Solo CC 94
7.2.2 A_2: Solo TC 94
7.2.3 A_3: Hybrid - Trusts A_2 94
7.2.4 A_4: Hybrid - Checks A_2 95
7.2.5 A_5: Hybrid - Half-Trusts A_2 95
7.2.6 Algorithmic Baselines and Alternatives 96
7.3 Design Concerns for Practical Implementation 97
7.3.1 Session Lifetime 97
7.3.2 Text Similarity Measurement 98

7.4 Algorithm Evaluation 100
7.4.1 Overview 100
7.4.2 Boundary Investigation: All Logs 101
7.4.3 Boundary Investigation: All Decision-Needed Logs 104
7.4.4 Boundary Investigation - All Decision-Needed Logs (Three-Classes Classification) 106
7.4.5 Session Content Overlapping 115
7.5 Summary 119

8 Discussion **121**
8.1 Discussion of Algorithms 121
8.2 Evaluation of Interactions 122
8.3 Finding Summary 125
8.3.1 Search VS. Browse 125
8.3.2 Fixed Timeout is not helpful 126
8.3.3 Domain-based or tab-based is not helpful 126
8.3.4 User Interaction did better than Content Relevance 126
8.3.5 Combined Versions perform better than Solo Versions 127
8.3.6 User Interaction 127
8.4 Algorithm's Compatibility 128
8.5 The Potential Application 128
8.5.1 Researchers 129
8.5.2 General Users 130
8.5.3 Developers 131
8.6 Future Work 131

9 Conclusion **133**
9.0.1 Answering the Research Questions 134
9.1 Contribution 135
9.1.1 Taxonomy 135

9.1.2 Tool Developed 136
9.1.3 Dataset 136
9.1.4 Algorithm Recommendations 137
9.2 Final Remarks 137

References **139**

Appendix A Search Literacy Project **151**

Appendix B Instruction for using Chrome Extension **153**

List of figures

1.1 Topic-Interleave in Conversation 3
1.2 Google Conversational Search 4
1.3 Elements Utilised to Understand Query 4
1.4 Three Stages 6
1.5 Other Projects 9

2.1 The Berry-Picking Model of the Information Seeking Process (adapted from Bates [10]) 13
2.2 The Standard Model of the Search Process (adapted from Broder [16]) . . . 13
2.3 Norman's Cognitive Model, Sketch by Hearst [41] 14
2.4 Wilson's Model of Information Behaviour 14
2.5 Session Example 23

3.1 HCI and IR 29

4.1 Study 1 Procedure 38
4.2 Study 1: Paper-Based Web History 40
4.3 Study 1: Card Sorting - Single Card 40
4.4 Study 1: Card Sorting - Clustering Cards 41
4.5 Study 1: Card Sorting - Groups 41
4.6 Happening Time, by Activity Goal 55
4.7 Happening Time, by Activity Context 56

5.1 Sample Log 60

5.2 Chrome Extension Structure 64
5.3 Interview Interface for Data Input 65
5.4 Each Log Entry during the Interview 65
5.5 Session Selection 65
5.6 Validation from User 66

6.1 Number of Search vs. Browse Sessions for Each Participant 70
6.2 Number of Long vs. Medium vs. Short Sessions for Each Participant 71
6.3 Mean Session Measures for Each Participant 72
6.4 Distribution of Measures for All Sessions 73
6.5 Correlation Overview Table 74
6.6 Overall Frequency of Each Interaction 77

7.1 Decision Flow of Sample Log Sequence 88
7.2 Flow Chart of a Log Decision 89
7.3 Pick the Right Similarity Measurement 98
7.4 The Relationship Between Logs and Decision-Needed Logs 101
7.5 A_1 vs. A_2: A_1 Interaction Distribution 108
7.6 A_1 vs. A_2: A_2 Interaction Distribution 108
7.7 A_1 vs. A_3: A_1 Interaction Distribution 109
7.8 A_1 vs. A_3: A_3 Interaction Distribution 109
7.9 A_2 vs. A_3: A_2 Interaction Distribution 111
7.10 A_2 vs. A_3: A_3 Interaction Distribution 111
7.11 A_4 vs. A_5: A_4 Interaction Distribution 112
7.12 A_4 vs. A_5: A_5 Interaction Distribution 112
7.13 (A_4,A_5) vs. A_3: A_3 Interaction Distribution 114
7.14 (A_4,A_5) vs. A_3: A_4 Interaction Distribution 114
7.15 (A_4,A_5) vs. A_3: A_5 Interaction Distribution 114
7.16 Measure the Session Overlap Between Two Versions of Datasets 115

List of tables

2.1 Session Definition Table 22
2.2 Practical Extension List 27

3.1 Session Features 32

4.1 Taxonomy of Six Factors 42
4.2 Dimensions of Open-Card Sorting 51
4.3 Analysis of Card Sorted Sessions, by Importance, Frequency, Difficulty, and Perceived Duration. 52
4.4 Confusion Matrix of Actual and Perceived Session Duration 53
4.5 Perceived VS. Actual Duration by Activity Goal 54
4.6 Perceived vs. Actual Duration by Activity Context 55

5.1 Interaction Types 62

6.1 Session Characteristics 71
6.2 Characteristics of Search Session and Browse Session 80
6.3 Characteristics of Short, Medium, and Long Session 80
6.4 Comparison of Search vs. Browse sessions for each length group. 81
6.5 Log Types 82
6.6 Probability of a Log Being a Create Boundary or a Boundary in General . . 83

7.1 Sample Log Sequence with Users' Self-Annotated Session Boundaries . . . 88
7.2 Interaction in TC 93

7.3 Be Clustered to One Existing Session 95
7.4 Runtime in Milliseconds (ms): Efficiency 97
7.5 Accuracy of Using Similarity Thresholds 0.1 - 0.9 in Content Check 99
7.6 Session Characteristic 100
7.7 Binary Classification Confusion Matrix - All Logs 102
7.8 Binary Classification Boundary Investigation Performance - All Logs . . . 102
7.9 Binary Classification Boundary Investigation Performance: Decision Needed Logs 104
7.10 Binary Classification Boundary Investigation Performance Summary - Decision-Needed Logs 104
7.11 Three-Classes Confusion Matrix: Decision-Needed Logs 106
7.12 Three-Classes Detail 106
7.13 A_1: Three-Classes 106
7.14 A_2: Three-Classes 106
7.15 A_3: Three-Classes 107
7.16 A_4: Three-Classes 107
7.17 A_5: Three-Classes 107
7.18 Three-Classes: Create-Log 107
7.19 Three-Classes: Resume-Log 107
7.20 Three-Classes: Continue-Log 107
7.21 Session Content Overlapping Performance 117
7.22 Part 1: GT as Target 117
7.23 Part 2: GT as Base 117
7.24 Performance Summary Based on F1 Measures 119

8.1 Performance Summary of Each Interaction (Based on F1 Measures) 122
8.2 Performance Summary of Each Interaction (Based on Precision) 122
8.3 Performance Summary of Each Interaction (Based on Recall) 123
8.4 Possible Improvement for Each Interaction 127

A.1 Interaction Types in Search Literacy Project 151
A.2 SERP Component 152

List of Glossary:

Session Characteristic	Description
Session-Duration	The sum of active dwell time of all web pages belongs to one session
Page-Duration	The active dwell time spent on each web page when that page is active and onfocus
#Queries	Number of unique queries submitted to major search engines during one session
#Pages	Number of unique web page visited during one session
#Domains	Number of unique web domain visited during one session
#Tabs (What about #Windows)	Number of browser tab during one session
#Boundaries	Number of boundary for one session
#Interactions	Number of user interaction during one session

Tab Type	Description
Blank_Tab	Brand new Tab, for example, chrome://newtab in chrome browser
NonBlank_Tab	Any Other Tabs

User Interaction Type	Description
CNT	Create New Blank Tab
CT	Close Tab / Window
SAB_Blank_Tab	Type query / URL in Address Bar in Blank Tab
SAB_NonBlank_Tab	Type query / URL in Address Bar in NonBlank Tab
TULC_Blank_Tab	Update Tab via Clicking HyperLink on Blank Tab
TULC_NonBlank_Tab	Update Tab via Clicking HyperLink on NonBlank Tab
OLNT_Blank_Tab (*not found in this study)	Create a tab / window via clicking hyperlink in blank tab
OLNT_NonBlank_Tab (OLNT)	Create a tab / window via clicking hyperlink in non-blank tab
SW	Switch attention between tab / windows
BOF	Click Back or Forward Button
FEPNT	Create a tab via clicking Hyperlink from an external application (MS Word)
LBIW	Switch attention off from the browser or Go to Incognito window
SBFE	Switch your attention back from the external application

Session Type (by contain query or not)	Description
Search-Session	If there are any web search queries submitted to one of several major search engines (e.g. Google, Yahoo, Bing, Baidu) inside the session
Browse-Session	If there are no web search queries submitted to one of the several major search engines inside the session.

Session Type (by duration)	Description
Short-Session	When the session duration is less than 5 minutes.
Medium-Session	When the session duration is equal to or more than 5 minutes, and it is also equal to or less than 30 minutes.
Long-Session	When the session duration is equal to or more than 30 minutes.

Log Type	Description
NoDecision-Needed Logs	the logs without any content update to the tab. For example, tab-switching to a tab whose session-id has already been determined
Decision-Needed Logs	The other logs

Transition	Description
Transition-connected	When page 1 is opened/updated via the link in page 2, then we assume page 1 and page 2 are transition-connected. (i.e., TULC_NonBlank, OLNT, BOF), it will pass the transition-check (TC)
Transition-break	Other interactions

Chapter 1

Introduction

1.1 Background

Using web browser is one of the main activities comprising people's daily life on web. With the enhanced functionality of web browsers, rapidly improving network performance, and increased variety of web applications, users' behaviour on the internet has changed significantly. For example, some users' online patterns are not simply successive task-by-task but frequently interleave tasks. In addition, the types of web activities they perform are diverse, ranging from simple look-for to complex exploration. As a result, many studies in IR are turning towards series of web activities rather than atomic URL-query pairs. However, determining the relevance between visits is still ambiguous and extracting out the correctly relevant group of visits is still challenging. In last decade there have been many studies in session-based retrieval via either temporal (e.g. 5, 15, or 30 minutes [52, 31]) or lexical segmentation (e.g. query similarity used to decide whether users are changing focus [91]), and sometimes even a combination of the two [37, 66]. However, most of these prior studies were based on either 1) existing server-side datasets (e.g. query logs from EXCITE [90] or AOL [11]) or 2) browser data gathered from a controlled laboratory setting with a pre-defined task (i.e. the task is not natural). Both approaches lacked insight into real natural interactions from internet users. In addition, most of these segmentation methods were only driven by researchers without any direct input from users themselves. Moreover, most of these

studies focused more on the search-related activities only on web, but sometime query-free navigation could happen more frequently than searching/querying.

The comprehensive view of sessions could provide a great deal of context to help determine the ambiguious cases. Shifting from the atomic URL-query pair to session-based study could create benefits for both researchers and general users.

For Researchers: With the session-based study, researchers can explore more data characteristics such as 1) completion time on a session (i.e., a set of logs rather than a single log) and 2) the starting or ending interaction (e.g., a query) on a session. Allan et al. [3] demonstrated that time on task has a statistically significant negative relationship with system retrieval accuracy as measured by bpref. Su [114] identified time as the most frequently mentioned reason users give contributing to their rating of the overall success of the IR system. Xu and Mease [135] demonstrated that total task completion time is correlated with user satisfaction for difficult tasks. Jung et al. [56] argued that considering the last click of a session may be the most important piece of information in relating user clicks to document relevance and they have found that using click data from the entire search session could be valuable. Research also evaluated sessions by focusing on user experience [122, 106] and how users interact between the browser and other programmes on their computers. For example, Huffman and Hochster [48] applied the Pearson Correlation to calculate relevance between user satisfaction and query-result relevance, and they found a surprisingly strong relationship between the relevance of the first query and the user's ultimate satisfaction with the session. White and Drucker [130] focused on web trails with the goal of identifying where users typically end up from different initial queries in order to recommend these typical destinations sooner.

For General Users: It could be beneficial for general users to know more information about their online life such as 'how have I spent my time online this afternoon?' or 'which webpages have I visited to plan my holiday?'. There are some Chrome extensions that allow users to manage their own web visits, tabs, and windows in sessions; these extensions have been widely used and well reviewed by a large group of users. For example, Session Buddy[1],

[1] https://chrome.google.com/webstore/detail/session-buddy/edacconmaakjimmfgnblocblbcdcpbko

which allows the user to save open tabs as collections that are quick to find and restore later, has more than half million of users with nearly perfect reviews. More examples were presented in practical application section in Chapter 2.

However, both academic research and practical application development have encountered some difficulties, and determining the session boundary is one of them. For example, Google's Conversational Search [117], which integrates the natural language process (NLP) and a semantic search, can keep antecedents as shown in Figure 1.1 and 1.2. In semantic search techniques, determining the right starting and ending boundaries of the antecedent information is a significant challenge due to the frequent occurrence of multitasking [128]. Similarly, another feature on Google Search called 'Previous Query' [116], which has been practically applied for awhile, returns the customised results page based not only on the current query but also on some previous queries. Besides learning from previous search queries, Google utilises other context information by obtaining the user's geolocation [115] and scanning the user's email content [76], as illustrated in the top portion of Figure 1.3. Both of these innovations have one same barrier: clustering the user's online activities to determine the boundary where task switching occurred. For example, if the session of relevant previous visits can be extracted correctly, the direct impact from this session would be more helpful than the entire query history as illustrated by the bottom part of Figure 1.3.

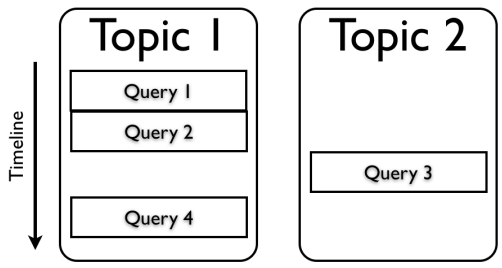


Fig. 1.1 Topic-Interleave in Conversation

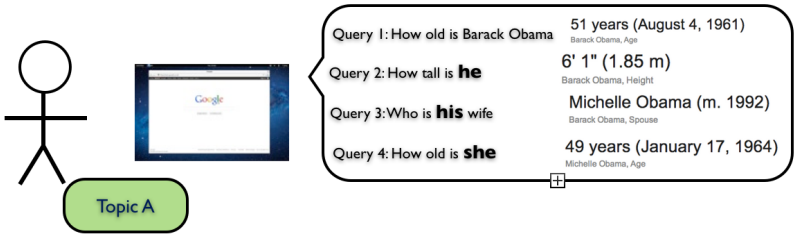


Fig. 1.2 Google Conversational Search

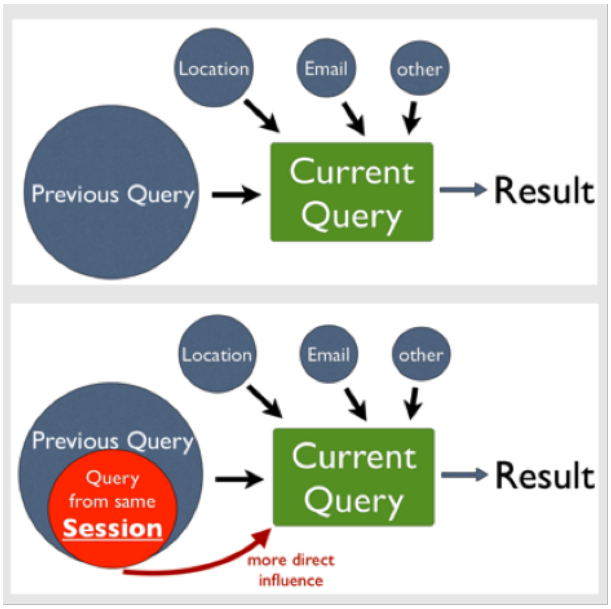


Fig. 1.3 Elements Utilised to Understand Query

1.2 Goal and Research Questions

The goal of my research is to work out an automatic segmentation on a web log to generate the session unit data. According to most of the prior research, there are several key limitations as listed below:

- a) Limited detail of each individual log (e.g., the user's interaction data between tabs is missing),
- b) Lack of users' direct involvement when studying their data (e.g., always having researchers annotate users' data without knowing much about the user's context);
- c) Lack of natural data (e.g., in order to get a clear copy of experiment data for a successive tasking pattern, many researcher have adopted a controlled laboratory setting with a pre-defined topic/task)

In order to corrrect the second limitation, several user studies were conducted in which all of the participants self-reviewed and identified their days-long web logs. In response to the other two limitations, a Chrome extension was built to gather user interaction logs with more detail from outside a controlled laboratory environment. It can provide more context information, such as how a user switches between tabs and how much time he/she actively spends on each page. In the end, give session segmentation attempts were proposed and evaluated. The discussion covered both boundary investigation and session content overlapping (i.e. merging different discrete logs).

My final outcomes were 1) a theoretical taxonomy about factors related to user's identification of sessions, 2) a ground truth dataset of user-identified sessions, 3) evaluations of and suggestions for algorithms including both content similarity and user's interaction, and 4) a chrome extension that is able to I) gather rich log data and convert the abstract system logs to a human-friendly log and II) determine the session based on different versions of algorithms in real-time.

As a result, the research questions are divided as indicated in Figure 1.4:

- RQ1: How do people understand and describe their own web sessions, and what factors create the divisions between them?
- RQ2: How are these self-identified sessions reflected in real-world logs of web activity, and what observable features represent their boundaries?
- RQ3: How can we algorithmically identify web sessions from detailed web activity logs?

1.3 Overview of the Research Stages

The whole study can be divided into three stages.

Stage 1: ***How do people understand and describe their own web sessions, and what factors create the divisions between them?*** The main goal of this stage was to get more information from users' perspectives. It focused on qualitative insights rather than quantitative analysis, so a smaller number of participants were recruited for a longer study period with

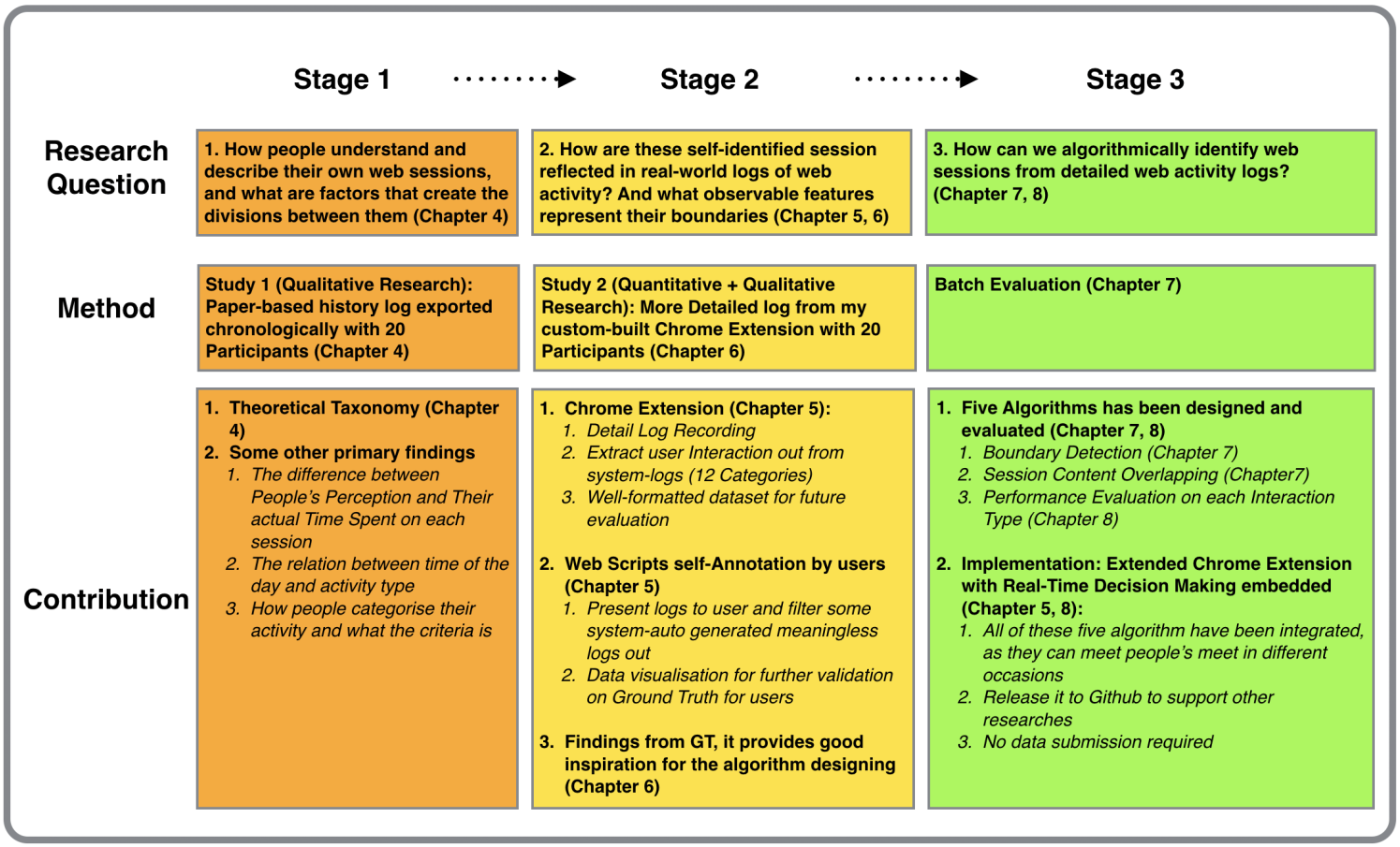


Fig. 1.4 Three Stages

each one. This stage included a weeklong data gathering process and a 60-minute semi-structured interview including card sorting in order to explore the factors the users consider during their annotation. It helped to generate a list of factors from users' perspectives during their segmentation.

Stage 2: ***How are these self-identified sessions reflected in real-world logs of web activity, and what observable features represent their boundaries?*** There was a difference between knowing the list of factors users consider in their session segmentation exercises and the practical reflections in logs. In order to create a more straightforward and rich presentation of these user-identified sessions, a custom-built logging tool was developed, and its detail was fully presented in Chapter 5. It is helpful to obtain more details about

user interactions in the browser in a field study rather than in a controlled laboratory setting. Therefore, it generates a copy of richer detail for each single log of a user's self-driven (i.e., not given or pre-defined by researchers) task from his/her daily life. A similar experiment protocol from Stage 1 was applied here and generated a ground truth dataset of user-identified sessions in a practical logging version. This provided some relevant implications for the algorithm in Stage 3.

Stage 3: ***How can we algorithmically identify web sessions from detailed web activity logs?*** As evidenced by the ground truth dataset in Stage 2, there were several algorithms that were mainly related to content relevance, users' interactions, or combination of them. Evaluations of their performance, including both boundary investigation and session content overlapping, were presented in Chapter 7. In addition, different levels (i.e. participant level, session level, and log level) of analysis were presented.

1.4 Thesis Structure

This thesis was divided into following chapters: first, Chapter 2 included background research about 1) information seeking models, 2) user's online behaviour, 3) the definition of session, 4) previous session determination techniques, and 5) existing relevant tools and applications. In Chapter 3, the methodology for data collection, experiment design and evluation were presented. Chapter 4 presented the first study, which was intended to answer Research Question 1, is presented. It ended with a taxonomy of six factors involved in participants' annotation of their own web data. Then, since a detailled logging tool was required, a custom-built Chrome Extension used for gathering users' comprehensive web data was described in Chapter 5. After that, the second study that employed an interview similar to that from Study 1 but with a richer features dataset, was presented in Chapter 6. It ended with a ground truth dataset of participant-identified sessions and also some implications for later implementation. It helped to further deliver the concept from previous taxonomy and provides a bridge between the initial theory and practical implementation. Chapter 7 involved

five algorithms proposed. Evaluations of their performance in both boundary investigation and session content overlapping were presented. Chapter 8 then presented a discussion of the performance of each interaction type as well as further suggestions for each algorithm based on the evaluation. At the same time, the compatibility of algorithm, potential application and a future plan were also presented. Finally, the findings of the entire study and its contributions were summarised in Chapter 9.

1.5 Publications and Related Projects

1.5.1 PhD Relevant Publication

A TOIS Journal Paper mainly covered the Chapter 5, 6, 7, 8 would be submitted within several weeks. (I am the primary-author).

- Ye C, and Wilson ML (2014): **A user defined taxonomy of factors that divide online information retrieval sessions**. – *In Proceedings of the 5th Information Interaction in Context Symposium*, IIiX2014
- Ye C, and Wilson ML (2014): **The characteristics of casual sessions in search behaviour logs**. – *Searching4Fun*, (pp. 23-26)
- Ye C, Porcheron M, and Wilson ML (2013): **Studying Extended Session Histories**. – *Euro-HCIR* (pp. 23-26)

1.5.2 Other Publication

- Rutter S, Blinzler V, Ye C, Twidale MB, and Wilson ML (2017): **Complex search task: how to make a phone safe for a child**. – *CEUR Workshop Proceedings 1798, 34-36*
- Wilson ML, Ye C, Twidale MB, Grasse H, Rosenthal J, and McKittrick M (2016): **Search literacy: learning to search to learn**. –

Besides the main research conducted for this thesis, I had other connected projects as listed in Figure 1.5. The search literacy project investigated how search literacy affects information seeking in complex search tasks. Thirty-nine participants had been recruited. My

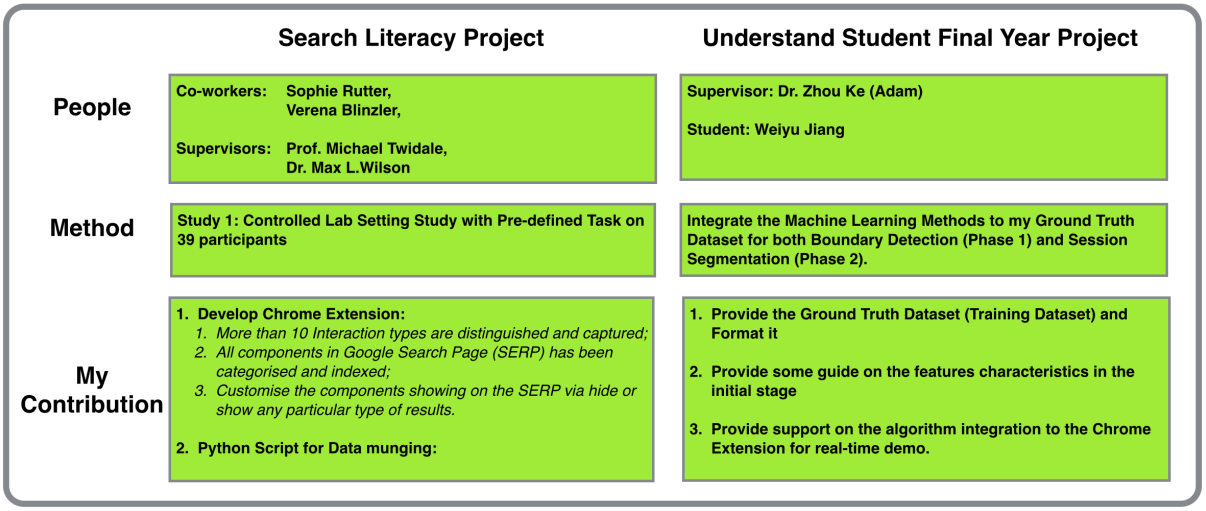


Fig. 1.5 Other Projects

responsibility was to develop the Chrome extension to capture users' logging data during the study. Rather than simply logging all of these direct system logs from the Chrome browser, all of the system logs were recorded and converted to user's interaction data. The details were described in Appendix A. The second project was about integrating the machine learning methods based on my ground truth dataset. It was the final project of an undergraduate student supervised by Dr. Zhou Ke.

Chapter 2

Literature Review

With better network support, improved website design, and web browser multi-functionality, users' web behaviour has changed from single successive tasking patterns to diverse approaches such as interleaving. The types of web activity have also changed: for example, users query search engines in order to accomplish tasks at a variety of granularities by issuing multiple queries, meaning that a series of interactions between user and search engine may be necessary to fulfil a a single information need [109]. Meanwhile, in order to extract more insights, many studies in IR are also turning towards series of web activities instead of atomic URL-query pairs.

2.1 The Information-Seeking Process

Research on humans' information-seeking process has been available for decades. According to these studies [120, 102, 104], the atomic cycle of the information-seeking process could be summarised as follows:

- 1) Problem identification,
- 2) Query planning,
- 3) Action (query input or browsing), and
- 4) Result evaluation (assessing whether or not one is satisfied with the results)

Many information needs cannot be satisfied by a single query or action, and they may consist of query reformulation or problem re-identification. Sometimes the actual information seeking may involve iterations like the expanded version from Marchionini and White [75] shown below:

- 1) Recognising a need for information,
- 2) Accepting the challenge to take action to fulfil the need,
- 3) Formulating the problem,
- 4) Expressing the information need in a search system,
- 5) Examining of the results,
- 6) Reformulating of the problem and its expression, and
- 7) Using the results.

The berry-picking model illustrated in Figure 2.1, which was first proposed by Bates [10], indicates that continuous changes to the understanding of questions can occur as a result of reading and learning during the whole procedure. Broder [16] presented a standard model of the search process as shown in Figure 2.2 and emphasised the potentially iterative loops of query refinement in search process. Hearst [41] provided a good summary of these information-seeking models.

With the involvement of human beings, the real execution of these actions is not always perfectly matched with these theoretical processes. Norman's [86] cognitive model provided a good explanation of the gap between execution and evaluation as illustrated in Figure 2.3. It indicates that what people intend to do (intention) can be different from what they actually do (action). The iteration of cycle and gulf in human beings' cognitive model makes the real information-seeking processes in people's daily lives more dynamic.

Besides the intrinsic characteristic of users themselves, there are also some external elements that affect their information-seeking processes, because users never interact with information in vaccum, but rather with some distractions [34], such as physcial distractions like television [14] or other online tasks. Actually, Wilson's model [133] provided a view with 1) multiple online and offline (e.g., a library) information resources and 2) the information exchange activities between people.

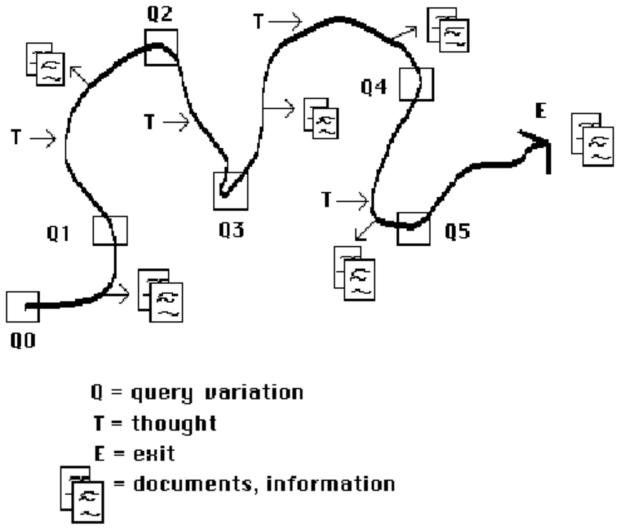


Fig. 2.1 The Berry-Picking Model of the Information Seeking Process (adapted from Bates [10])

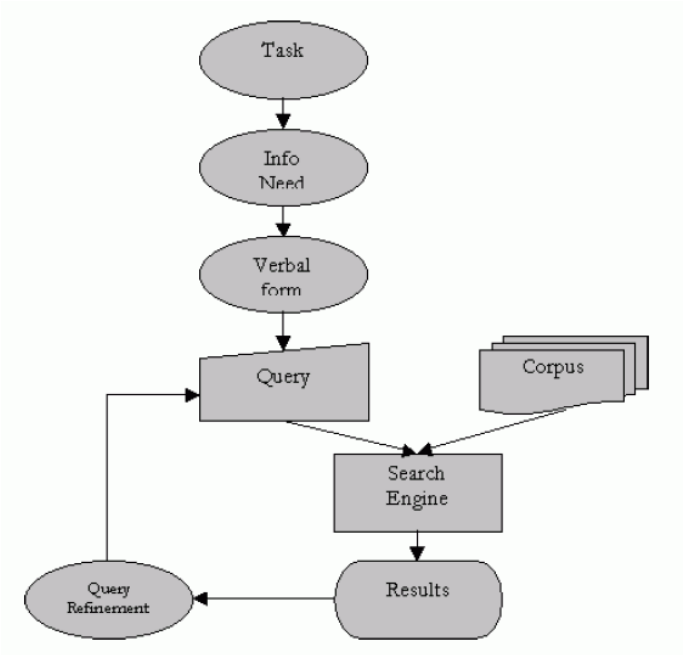


Fig. 2.2 The Standard Model of the Search Process (adapted from Broder [16])

As a result, the information seeking process may end with diverse possibilities, which makes it more difficult to predict. Rather than being stuck with the initial question, user may 1) investigate further with unchanged question and query, 2) shift to something relevant but different, or 3) give up the previous question and switch to a completely different one.

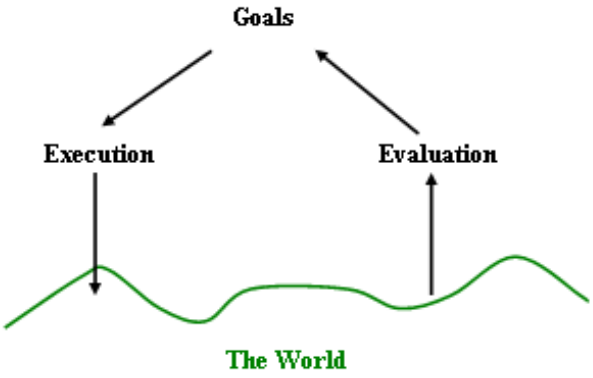


Fig. 2.3 Norman's Cognitive Model, Sketch by Hearst [41]

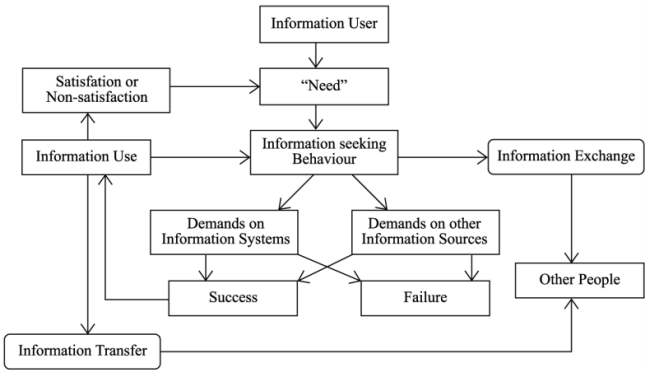


Fig. 2.4 Wilson's Model of Information Behaviour

2.2 User Interaction on Web

The intrinsic and external elements that affect users' interactions with these information systems have made the whole procedure of human information seeking beyond the atomic interaction cycle, and have led to behaviour diversity. A number of factors had been investigated in many prior studies.

2.2.1 Factors That Influence User Interaction

Search expertise, which refers to how familiar users are with the information system's features (e.g. the 'back' button on a browser [45]), has some influence on user behaviour.

Some studies demonstrated that different levels of search expertise can lead to different performance. In 2000, Hölscher and Strube's research [45] found some interaction differences between novices and experts. For example, the back-oriented behaviour of clicking on the back button in a browser is more common among novices. Lazonder's 2000 study [64] also found that novice searchers who were less familiar with system features (e.g., clicking hyperlinks or scrolling a page) needed more time to get used to these features and were less likely to end with successful search strategy execution. Meanwhile, a 2003 study by Vakkari et al. [124] about observing users' search behaviour during a long-term thesis writing task demonstrated that the novices' search performance could be improved by training (e.g., increasing search system feature knowledge).

Domain Knowledge refers to how familiar users are with a specific topic. White et al.'s 2009 study [131] found that there were some differences in the online behaviour of domain experts and novice web users, and the differences were reflected in query terms, the diversity of sites visited, dwell time and also search strategies. Similiarly, Hembrooke et al in 2005 [42] found that domain knowledge influenced user's term selection and search pattern. Hirsh's 1997 research [44] proved that higher domain knowledge was helpful for child searchers when making relevance judgments.

Task types A user may exhibit different behaviours when performing different types of tasks. Kellar et al [58] investigated four types of tasks - fact finding, information gathering, browsing and transaction - and they found different behaviours across these different tasks. For example, users tended to spend more time, view more pages, and use more browser system features in information-gathering tasks. Liu et al. [68], whose objective was to predict the task type from user behaviour logging, have proven the association between task type and user behaviour. They covered four task types - 1) background information collection (BIC), 2) interview preparation (INT), 3) advance obituary (OBI), 4) copy editing (CPE) - and measured the differences in 1) completion time, 2) number of pages visited, 3) number of sources visited, 4) number of queries, and 5) number of search sources. In the end, they found that BIC took longer and utilised more pages, sources, queries, and search sources than the other tasks.

Gender: A 2005 study by Lorigo et al [70] investigated the relationship between gender and online behviour. They found that males had greater average fixation durations on web documents than females, but males submitted shorter queries.

2.2.2 Successive Versus. Interleaving

Currently, many online information needs cannot be satisfied with a single query but rather require a series of interactions. Therefore, learning from successive patterns on one topic and problem solving have been widely adopted by many researchers. However, the reality of users' daily lives online is not always the same as what happens in a controlled laboratory environment (i.e., simulating successive task-by-task).

Successive: refers to a series of consecutive interactions concerning one information need or problem. As explained in the information-seeking model above, users may decide to continue searching or browsing after evaluating the found results [107]. Successive tasking, especially successive searching, is conducted during a user's interaction with IR systems over time when seeking information to solve an information problem [99, 108]. Spink et al [108] provided the following definition of successive searching: *"The process of repeated, related searches over time in relation to a given, possibly evolving, information problem (including changes or shifts in beliefs, cognitive, affective and/or situational states)"*. They found that more than half (56.5%) of their end-users conducted more than one search. It was also identified in Robertson and Hancock-Beaulieu's study [99].

Interleaving: Since users never interact with information in a vaccum but rather are always confronted by some distraction, the reality of users' information seeking may not always be that they stick to a single task. Sellen et al. [103] found that some users often keep the web pages spread out over time, especially in the information-gathering tasks. Other research also demonstrated that people interleave web episodes, such as email, social networking, and information gathering [103, 112] as shown in Figure 1.1. Waller provided the following definition of multitasking in 1997 [127]: 'individual-level multitasking process involve a person's allocation of his or her own scarce cognitive resources among several tasks and the moderating impact of task elements, task processes, and task resources on individual

multiple-task performance'. Multitasking is the ability of human beings to deal with more than one task simultaneously via task switching [79]. Multitasking is also related to 'parallel browsing' in Huang et al's 2000 study [47], which described behaviour where users visit webpages in multiple concurrent threads. People engage in parallel browsing 57.4% of the time.

In general, there are several types of causes for interleaving:

- 1) Physical distraction (interference): the ubiquity of digital device such as mobile phones or television [129], and
- 2) Other online activities [77, 128].

In the other online activities, for example, Spink and Park [111] worked out a model of multitasking including the interplay between information and non-information-related task switching. A 2003 study by Ozmutlu et al. [92] reported a high frequency of multitasking in people's information seeking online and found that 32% of FAST[1] users submitted multitasking web search sessions. Some research indicated that people typically engage in a range of multitasking activities [63, 93, 110, 73] and proved that queries or activities from the same task are not always consecutive. Jones and Klinker's 2008 study [55] reported on properties of this search task hierarchy as seen in a random sample of user interactions from a major web search engine's log annotated by human editors. The result indicated that 17% of tasks are interleaved and 20% are hierarchically organised.

Even Lee and Taatgen [65] argued that multitasking behaviours could be understood as a product of skill acquisition. This important online behaviour has been regarded as 'noise' in many prior web studies because it always brings more complexity to the data feature and may necessitate heavy manual processing to extract the group of successive activities focused on one information need. As a result, many researchers recruited participants and conducted their studies in a controlled laboratorry with pre-defined topics or tasks in order to simulate a single successive tasking pattern. However, in my research, instead of seeing multitasking as negatively impacting on the study and trying to eliminate this 'noise', I consider it one

[1]FAST is the search engine for Microsoft's SharePoint collaboration platform.

of the most important and valuable parts, and I believe that this behaviour truly reflects the diversity of human behaviour.

2.2.3 Searching: Query Reformulation

Jansen defined query reformulation as *'the process of altering a given query to improve search or retrieval performance'* [49]. In Lau and Horvitz's 1999 study [62], there were seven classes of query refinement: 1) new, 2) generalisation, 3) specialisatio, 4) reformulation, 5) interruption, 6) request for additional results, and 7) blank queries. This classification of query refinement was applied in Jansen et al.'s research [51], and high occurrences of query reformulation after generalisation (17%) and specification (32%) were found. Other researchers came up with similar taxonomy. For example, a 2016 study from Rha et al. [97] explored query refinement types and suggested the following query reformulation taxonomy: generalisation, sepcialisation, word substitution, repetition, new, spelling correction, and stem identical. Huang and Efthimiadis (2009) [46] created taxonomy of query reformulation strategies with nine categories: word reordering, whitespace and punctuation, removing words, adding words, URL stripping, stemming acronyms, substring abbreviation, word substitution, and spelling correction. This taxonomy focused more on interaction rather than content difference. Rieh and Xie (2006) [98] put forward a facet-based query reformulation structure with three main facets (content, format, and resource) and nine sub-facets. They also categorised modification sequences of query reformulation into eight distinct patterns: specified, generalised, parallel, building-block, dynamic, multi-tasking, recurrent, and formatting. Ozmutlu [91] reported that 28% of queries are reformulations of previous queries. In addition, Liu et al.'s [67] 2010 study in query analysis also reported a high frequency of query reformulation occurrence, and they found that users were more likely to submit query reformulations in more complex search tasks. Some researchers found that the pattern of query reformulation could be related to other factors such as the user's expertise on web [45] and location (i.e., regional differences) [53]. In addition, Hölscher and Strube's [45] 2000 study comparing novices and experts found that web novices more frequently reiterate their search queries with smaller and less effective changes to their queries. Some other

research also demonstrated that effective query reformulation can improve the outcome of user searches [30, 98].

2.2.4 Navigation: Page Revisitation

Revisitation is defined as *'returning to a site that the user has visited in the past'* [78]. Tauscher and Greenberg [121] found that 58% of pages visited by users had been accessed before. A higher frequency (81%) of page revisitation was reported in Cockburn and MacKenzie's 2001 study [22]. Adar et al's study [2] identified four main revisitation patterns: fast (revisited many times within one hour but rarely revisited over longer intervals), medium (revisited primarily at intervals between an hour and a day), slow (revisiting mostly at intervals of a week or more), and hybrid (combination of fast and slow revisitations).

There were some applications related to revisitation. For example, the history list [20], bookmark [21], and back buttons [57, 43, 45] of web browsers were significantly related to page revisitation. Herder's 2005 study [43] found that the majority of page revisitation happened via these backtracking activities, and they were significantly related to browser back buttons. Some research [5, 20] demonstrated that the history list was not used quite as much due to the difficulty (e.g. page name inconsistency) of extracting information from this chronologically ordered list. The better management of bookmark compared to history list nakes them more useful; however, the effort of manual management created some drawbacks [1]. SmartBack, presented by Milic-Frayling et al. [78], is a feature that supports users in revisiting pages, and it marks some visited pages (e.g., hubs[2], results of form submissions, or beginnings of navigation trials) for quick access. These authors have confirmed a high frequency of page revisitation and the benefit of SmartBack for user navigation when revisiting. A study by Hölscher and Strube [45] found that novice users were more likely to engage in backward-oriented behaviour through clicking the back button than were experts. Catledge and Pitkow [18] demonstrated that page visitation via back buttons accounts for 41% of all navigation acts, which makes it the second most frequent way

[2][33]: when navigating through a hyperlinked environment with some possible revisitations of pages, hub-and-spoke is a tree strucutre for describing the navigation map. For example, 'hub': created when Back button clicked, 'spoke': the branching

of accessing pages after link navigation (52%). Tauscher and Greenberg's 1997 study [121] confirmed that link navigation is the most frequent activity, comprising 50% of navigation acts. They also found that pressing the back button comprised over 30% of all navigational acts and occurred far more frequently than other revisitation operations like bookmark (< 3%) and history list (< 1%).

2.3 Session Definition

The notion of a 'session' dates back decades to when early work on DIALOG [118] kept track of searchers' local queries and allowed them to reuse those queries, especially to re-find information. Following the advent of 'Internet Web', the research on web session began in the mid-1990s [18]. The IR moving towards focusing on the success of a session rather than just single URL-query pair analysis led to the development of relevance measurements that consider change in relevance over time [114]. Up to this point there has been increasing focus on web sessions, where search engines are keen to better support searchers who continue to search for more than a few queries or minutes such as query reformulations [91, 96, 130].

2.3.1 Historical Definition

Jansen, Spink, Blakely and Koshman [50] mentioned the session as *'a series of interactions by the user towards addressing a single information need'*. Jones and Klinkner [55] focused on searching behaviour research and introduced the different scales, namely search goal, search session, and search mission. Search goal was defined as *'an atomic information need, resulting in one or more queries, and it can be thought of as a group of related queries to accomplish a single discrete task, the queries need not be contiguous, but may be interleaved with queries from other goals'*. Search session was defined as *'all user activity within a fixed time window'*, and search mission as *'a related set of information needs, resulting in one or more goals, which would also include the hierarchical relevant tasks that get interleaved'*. Sessions contain missions, which in turn contain goals. A goal may have multiple queries, and a mission may have multiple goals. Temporally, these levels can

all be interleaved. In Wolfram et al.'s 2009 study [134], they mentioned that a session was defined as *'a set of one or more consecutive queries attributed to the same identifier where the time between adjacent queries does not exceed a cutoff value. A cutoff value is a threshold for a query interval or time between two temporally adjacent queries by the same identifier. Two temporally adjacent queries belong to the same session if the query interval value is less than the cutoff value; otherwise, the two queries belong to two adjacent sessions. One way to identify a suitable cutoff is to examine the characteristics of the query intervals to study how different cutoffs affect resulting session lengths'*. Wolfram et al. studied the characteristics of sessions according to the list of session variables from Chen and Cooper's research [Chen and Cooper]. Hassan et al.'s 2010 study [38] used the same definition from Jones and Klinkner's in 2008. A study by Hagen et al. [37] introduced some other terms: physical session (*'characterised by time based between queries'*), logical session (*'characteristed by consecutive queries for same information need within a physical session')*, and mission (similar to Jones and Klinkner's definition [55]). Additionally, the term 'episode' from Belkin [12] in 1996 was highly relevant to the session concept in our research: *'An information seeking episode consists of a series of kinds of interactions (slices in time), structured according to some plan associated with the person's overall goals, problem, experience, according to the person's specific goals, etc. at any one time, and according to what has happened during the course of the interaction'*. The example given in this study, which described a set of relevant documents exploration and query determine through one information need, provided a comprehensive view of a series of consecutive interactions required to fulfil one information need. Hagen, Gomoll, Beyer, and Setin [37] believed that in these cases the session with its typical flavour of some temporal connection should be replaced by the more general notion of a search mission (not just consecutive queries) established by Jones and Klinkner [55]. They argued that relying on series of consecutive queries alone will result in missing important connections, because behavioural patterns of search engine users include interleaving searches conducted in a multitasking manner.

Most of the prior studies mainly focused on the searching behaviour; however, there were many others about how web browser could be query-free. Kong et al.'s work [61]

Table 2.1 Session Definition Table

Term	Author	Year	Inter-leaving?	Cluster Factors	Definition	Log Detail
Session	Jansen, Spink, Blakely, Koshman [50]	2007	Yes	Cookie, User-Id, query, time-gap	a session is a series of interactions by the user towards addressing a single information need	[Query Log]: User-Id, Cookie, timestamp, query, location, source, feedback
Search Session	Jones, Klinkner [55]	2008	No	Time-gap	all user activity within a fixed time window, it is just a slice of user time	[Query Log]: Query, Timestamp, User-Id, URL, Page titles, relevant snippets, some other features (spelling and query suggestion), clicks
Search Goal	Jones, Klinkner [55]	2008	Yes	Time-gap, Content (Query, and other content-related info)	an atomic information need, resulting in one or more queries, and it can be thought of as a group of related queries to accomplish a single discrete task, the queries need not be contiguous, but may be interleaved with queries from other goals	Same as above
Search Mission	Jones, Klinkner [55]	2008	Yes	Content (Query, and other content-related info)	a related set of information needs, resulting in one or more goals, which would also inlcude the hierarchical relevant tasks that get interleaved, it is extended information need	Same as above
Session	Wolfram, Wang, Zhang [134]	2009	No	Cookie, IP, Time-gap	a set of one or more consecutive queries attributed to the same identifier where the time between adjacent queries does not exceed a cutoff value. A cutoff value is a threshold for a query interval or time between two temporally adjacent queries by the same identifier. Two temporally adjacent queries belong to the same session if the query interval value is less than the cutoff value; otherwise, the two queries belong to two adjacent sessions. One way to indentify a suitable cutoff is to examine the characteristics of the query intervals to study how different cutoffs affect resulting session lengths	[Query Log]: Query, Tiemstamp, IP Address, Cookie, User-Id, Result Pages, Click, URL
Physical Session	Hagen, Gomoll, Beyer, Setin [37]	2013	No	Time-gap	characterised by time based between queries	[Query Log]: Query, Timestamp, Wikipedia Articule for Queries
Logical Session	Hagen, Gomoll, Beyer, Setin [37]	2013	Yes	Lexical Similarity of Query and also semantic similiarty	characteristed by consecutive queries for same information need within a physical session	[Query Log]: Query, Timestamp
Mission	Hagen, Gomoll, Beyer, Setin (from Jones and Klinkner [55]) [37]	2013	Yes	Lexical Similarity of Query and **Semantic similiarty**	characterised by logical sessions, multitasking behvior, and hierarchical goals	[Query Log]: Query, Timestamp

on pre-search context indicates that there was a strong relationship between *searching* and *browsing* such that many queries were triggered by different types of contexts in which *browsing* was one of the key factors. In addition, Olston and Chi [89] stated that *browsing* and *searching* were the two predominant paradigms for finding information on web.

We have a similar view as Hagen et al. [37] and Jones and Klinkner [55], our session definition is similar to the mission definition, it covers all activities including both searching and browsing focused on one information need, and it considers the interleaving as pictured in Figure 2.5.

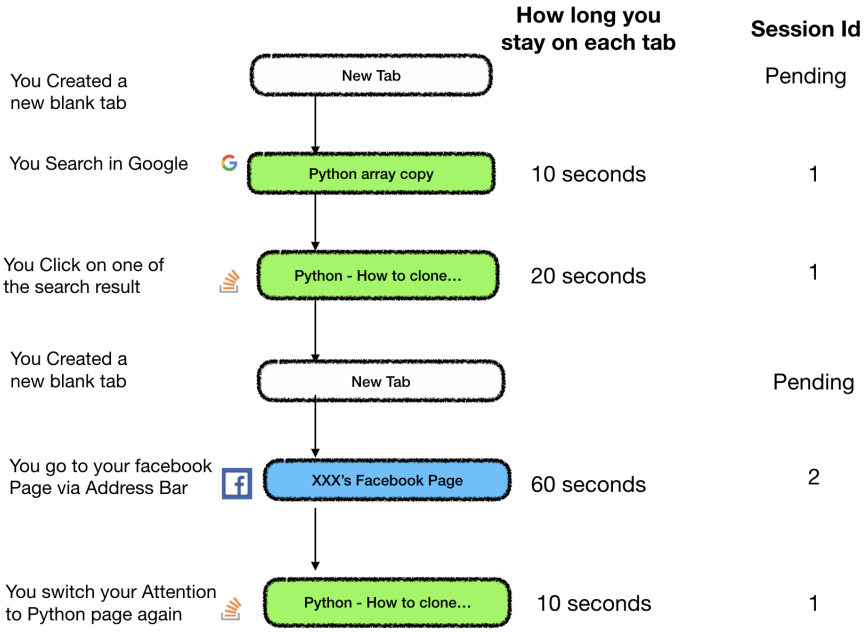


Fig. 2.5 Session Example

2.4 Existing Session Determination Techniques

Besides the conceptual defintions, there was a group of practice in session clustering. The key challenge in session clustering is to determine the boundarie, and this became a primary focus of web research in the 1990s [18]. Some researchers applied manual processing to data [39, 40, 100] to capture the relevance of the original raw dataset, which requires either researchers or users themselves to annotate the log to generate more accurate sessions. For

example, Rose and Levinson [100] manually labelled the log data from the AltaVista search engine and worked out a framework of users' goals consisting of three large categories. The ideal way to do this is to have the users annotate their own data, but this is not feasible in most studies. Even the manual subject analysis of logs can be performed more accurately to some extent, but this is impractical and difficult for large datasets. In the aforementioned automatic clustering, there are primarily three types: content-ignorant, lexical, and combined.

2.4.1 Automatic Segmentation: Content-Ignorant

Fixed time gap (temporal): In the early stages and even today, many researchers attempted to identify sessions from logs' applied simplistic fixed timeouts to separate users' visits into sessions [126, 80, 55]. These researchers were inspired by Catledge and Pitkow's use of 25.5 minutes [18] due to its simplicity. For example, in 2007 Beitzel et al. [11] investigated the properties of a large query log from AOL over a six-month period and examined topical trends over different time periods. However, some researchers found that fixed timeouts are not reliable. Montgomery and Faloutsos [80] tried several thresholds of timeout values and found that the choice of cutoff did not matter. Jones and Klinkner's study [55] also found that different thresholds of fixed timeouts did not create significantly different results. In addition, the frequency of interleaving described above in the user interaction section illustrates that the relevant visits stemming from one information need may not always occur in a consecutive period, so a fixed timeout will either group irrelevant logs into one session or tear one good session apart. The participants from our study feel that the roles of time gap vary depending on the size of the session and most of them do not really care about the timegaps, even including breaks for the night.

User's actions: Many studies found that there could be some correlation between relevance and user interaction such as document reading time [81], webpage dwell time in search [59], and scrolling behaviour [32]. Fox et al's study [29] found that document clickthrough is one of the strongest implicit signals for relevance feedback. Some other studies have indicated that mouse movement can be a good indicator for the relevance judge of search result [23, 35]. There have been some other attempts to identify indicators such as

eye tracking [54]. Hassan et al [38] used a behaviour-based approach. They focused on user action during the search and proved that models of using user behaviour were more predictive of goal success than is lexcial relevance. Downey et al. [25] provided a comprehensive table of events and parameterisation for log data analysis.

2.4.2 Automatic Segmentation: Lexical

Query similarity: With the same awareness of the dynamic nature of sessions, other research considered query similarity as a means to determine when users switch between information needs (and thus sessions) [91, 95]. One of the most popular lexical approaches was to determine the similarity between queries based on common terms [50, 91] and check the addition and removal of terms between subsequent queries to determine topic change [39]. In 2005, Ozmutlu and Cavdur [90] replicated He, Goker, and Harper's [39] methods of automatic topic identification using an Excite query log dataset. They concluded that the method's performance was limited because of the log limit. Lucchese et al. [73, 71] proposed the clustering algorithm Query Clustering based on Head-Tail Components (QC-HTC), which was based on query content (common term overlap) and semantics (i.e., learning the term context from Wikipedia and Wiktionary).

2.4.3 Automatic Segmentation: Combined

Since there are both advantages and disadvantages of both types explored above, some researchers tried to combine some factors in their segmentation. Lucchese et al.'s [72] work introduced the task-based session discovery problem and proposed two similarity measurements: content-based (normalised levenshtein distance + Jaccard distance) and semantic-based (Wiktionary and Wikipedia), and also evaluated several clustering algorithms, but while using timeouts for the pre-processing breaking phase. Liao et al.'s work [66] focused on extracting the task trail from the chronological web search logs with a query task clustering (QTC) algorithm that measures the query similarity by time and word-based features with weights determined by an SVM clssifier. More recent work by Hagen et al. [37]

used perhaps the most comprehensive approach to identifying homogenous search sessions. They presented a cascading six-step method that includes temporal factors, lexical similarity, and semantic similarity using explicit semantic analysis (ESA) and Linked Open Data (LOD). The technique is tolerant of data from interleaved sessions. In our research, we also use the combined apporach. We consider the performance of content-based techniques in separating sessions but are also aware of the content-ignorant factors like temporal and users' action.

Some of these automatic approaches take assumption of consecutive loggings without any interleaving. As a result, some others have to involve other third-party sources to provide more information (e.g., Wikipedia articles for semantic similarity in Hagen et al's research [37]) because of the limited detail of each single log.

2.5 Practical Applications

Besides the session research and segmentation attempts, there are some session-related practical applications on the market. These include some features released by official browser teams and browser add-ons developed by some third-party teams.

Yahoo developed Searchpad [24, 83], while provided a note-taking facility for use during longer session. It was an online notebook that watched the user while the user was interacting with Yahoo. Rather than requiring explicit manual input from the user, it watched what the user was doing and saved the search results. Meanwhile, SearchBar [82] let people manipulate their search history as it related to certain topics or sessions. The web browser Vivaldi[3] was developed by the former Opera team. It incorporated an embedded tab management feature referred to as tab stacking [84] that helped to create folders to group relevant tabs together.

In addition, there were some existing Chrome extensions that allowed users to manage their own web visits, tabs, and windows into sessions to enhance some easy operations like revisitation or saving bookmarks, as shown in Table 2.2. Most of these extensions require manual input concerning tab and window management except Web Task Switcher[4], which

[3]Vivaldi Web Browser: https://vivaldi.com/features/

[4]Web Task Switcher: https://chrome.google.com/webstore/detail/web-task-switcher/hklklpilgegjkkkdbmhbkpapfcbjfdhk

tried to automatically organise tabs related to a single task into a separate window. Web Task Switcher created a new task when searching something online would lead to a new task (window) for that search query. The links in the search results that were opened would be added to the same task. The user could also create a new task out of an existing tab by taking that tab out of the current window.

Table 2.2 Practical Extension List

Extension Name	Rating	Users	Manually?
Session Manager [5]	4 out of 5 by 413 reviews	57k	Yes
FreshStart [6]	4.5 out of 5 by 707 reviews	72k	Yes
Session Buddy [7]	nearly 5 out of 5 from 21776 reviews	633k	Yes
Web Task Switcher [8]	5 out of 5 from 6 reviews	85	**No**
Task Timer Pro [9]	5 out of 5 from 35 reviews	1.4k	Yes
Simple Time Track [10]	4.5 out of 5 from 198 reviews	18k	Yes

Users seem to become tired easily when some explicit input is required from them. One study found that inexperienced search users prefer a system that uses implicit relevance feedback over explicit because it reduces the burden of providing feedback [132]. Analysis of search log data found that typically less than 10% of users make use of an available relevance feedback during searches even when results are often over as a result. The reluctance of this user in the subway of direct input in relevance feedback systems has motivated the development of implicit signals instead.

Besides commercial use, some researchers developed logging tools for academic purposes. One of the most popular of these tools was CrowdLogger [28, 27], which was developed by Feild. It was good software with much more detailed data recording from other web logging tools. It was able to record participants' interaction data (click and page focus) and basic information (URL, timestamp, title, and tabid). However, what I wanted to create is a way to generate a more detailed list of user interactions, and I tried to determine the correlation between these different types of interactions to session segmentation. As a result, rather than using this particular software, I chose to build my own software by myself.

[5] https://chrome.google.com/webstore/detail/session-manager/mghenlmbmjcpehccoangkdpagbcbkdpc
[6] https://chrome.google.com/webstore/detail/freshstart-cross-browser/nmidkjogcjnnlfimjcedenagjfacpobb
[7] https://chrome.google.com/webstore/detail/session-buddy/edacconmaakjimmfgnblocblbcdcpbko
[8] https://chrome.google.com/webstore/detail/web-task-switcher/hklklpilgegjkkkdbmhbkpapfcbjfdhk
[9] https://chrome.google.com/webstore/detail/task-timer-pro/dddgjgcemjgkhgeekbooigaiilbklkbm
[10] https://chrome.google.com/webstore/detail/simple-time-track/hbhcgdpbdenjjpekdajanfgchlgacdea

Chapter 3

Methodology

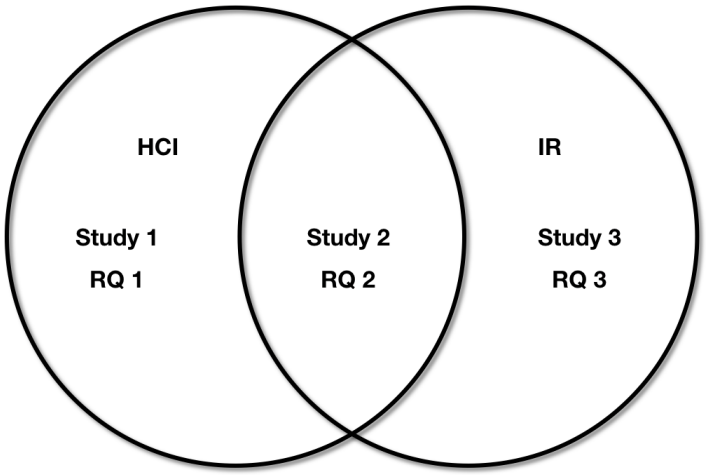


Fig. 3.1 HCI and IR

This research is a mixture of HCI (Human-Computer Interaction) and IR (Information Retrieval), as shown in the Figure 3.1. The goal is to create an automatic web logging segmentation to generate session unit data that consists of the **user's perspective** on session clustering and **more details of each individual log** gathered from **field study** of **self-driven tasks** rather than a controlled laboratory environment. It is a mixed methods research [125] that combines elements of qualitative and quantitative research approaches.

RQ1 is focusing on understanding user's real-world web behaviour, so HCI methods, for example, interviewing [137] is needed to collect data. RQ2 is connecting the finding

from the user study and the web log captured from custom-built chrome extension, which is a combination of HCI and IR. It helps to collect richer individual natural-data from the real-world (i.e., especially more interaction detail). RQ3 is focusing on IR, which proposed the algorithms based on finding from previous studies and follow the batch evaluation [8] to assess them. It is a full cycle from theoretical exploration to a practical implementation of findings involving both qualitative and quantitative study. It follows the User-Centered Design [87], and it emphasis the importance of understanding user's perspective during the session segmentation, which and consist of following phases:

- 1) Specify the context
- 2) Specify requirements: get the user's understanding about session segmentation
- 3) Create design solution: get the ground truth and propose the algorithm
- 4) Evaluatation: evaluate the algorithm performance.

During the research, a logging tool has been developed in order to obtain richer details from each individual log and more user interaction detail. In addition, the ground truth dataset of sessions annotated by participants themselves has been stored and can be provided in various formats to support further research. The gap between the theory and practical implementation has been looped through, and the balance between accuracy and efficiency during the algorithm design has been considered.

3.1 Data Collection

Collecting data is a key step in this research and there are several methods involved in this research:

3.1.1 Logging Tool

In order to characterise user's interaction data with more context information, a logging tool (Chrome Extension) has been developed. It runs quietly in the background while the user is interacting with the browser without causing any distraction or distruption, and it can get not

only the page content information (e.g. title, url, top10keywords), but also the interaction context between tabs/pages. It is therefore a useful observational method because it can capture user's natural interaction data without interrupting them. And it could also benefit providing objective dimension for further analysis. More detail of this tool was described in Chapter 5.

3.1.2 Interview

Interview [137] is appropriate when there are some complex or abstract questions to discuss and to extract some further insight. In this research, after the data collection via logging tool, the participants were invited to engage in an interview to review their own web data and annotate the sessions and boundaries.

3.2 User Study

In most prior user research, there are two primary approaches to experiment setup: natural settings (field study) and controlled laboratory settings. The comparison between these two by Sun and May [119] stated some differences, namely 1) the difficulty of experiment setup and 2) atmospheric differences. Sun and May have argued that field study experiments are more difficult to conduct due to the influence of distractions, which was noted by some other researchers like Kjeldskov et al. [60] and Baillie [9]. Because of the concentration on particular hypotheses and the 'noise-free' characteristic, the controlled laboratory settings had been widely applied in a large number of IR studies [69]. The tasks tested in laboratory settings are always pre-defined by the researchers, which creates ideal data that is removed from reality. A 2006 study by Gwizdka and Spence [36] found that the differences in task complexity greatly influenced users' behaviour, which led to differences in performance such as different numbers of pages visited and different dwell time on each page. As a result, designing the right type of task is another significant challenge in controlled laboratory studies. In addition, in controlled laboratory environments, participants act differently from the realistic situation sometime more politely during the study. For example, Thomas and

Kellogg [123] stated that there is a gap between the laboratory setting and users' real lives, and this leads to some bias in results and conclusions. Field study is more acceptable for participants in studies of some sensitive topics (e.g., sexual behaviour investigation [113, 6]. Some researchers extended the traditional paper-based diary study by using voice-mail paired with mobile phone to collect data [94], or a web-based diary [113, 6] like our custom-built chrome extension in this study. Distraction is a significant concern during studies conducted in natural settings. Different types of distraction have been investigated. For example, in 2011 Brasel and Gips [14] conducted a study about multitasking behaviours across devices and found an extremely high switching rate (i.e. around 120 switches in 27.5 minutes) between television and internet. In 2014, Yeykelis et al. [136] found an even higher frequency of switching. However, these distractions naturally exist in users' daily life, and they definitely influence users' perfomrnace [14, 136, 74, 4, 13]. As a result, a comprehensive web study should try to work with these natural distractions in a natural setting.

3.3 Evaluation

During the analysis of the ground truth dataset, these session characteristics are covered in Table 3.1, and they are similar to those from Wolfram et al.'s research [134].

Table 3.1 Session Features

Name	Description
Page-Duration	the active dwell time spent on each web page when that page is active and in-focus [1]
Session-Duration	The sum of active dwell time of all web pages belong to one session
#Query	number of unique queries submitted to major search engines during one session
#Pages	number of unique web pages visited during one session
#Domains	number of unique web domains visited during one session
#Tabs	number of browser tabs created during one session
#Boundaries	number of boundary [2] for one session.
#Interactions	number of user's interactions (as shown in Table 5.1) during one session

[1] It only include the time when this tab was focused.

[2] boundary refers to session-switch, Each session has a Create boundary, plus the number of times that it was resumed. For example, after being created (1 boundary), if a user switches to a different session and back again, it is resumed (now 2 boundaries)

Besides these, some extra information can be extracted from the dataset collected via the logging tool, and then it can be extended to get the answer to these following questions:

- How much time did the user spend on each page (active dwell time)?
- How did the user get to this page (e.g. via a hyperlink click on a page)?
- When did the user openclose this page?
- How many times did the user switch hisher attention offon this page, and when?
- If it is a search (e.g., in Google, Bing, Baidu, or Yahoo), what was the query?
- What did the user do on each page (e.g., click on some links)?

3.3.1 Cranfield Paradigm

During the evaluation, the Precision, Recall, and F1-Measure [7] have been adopted in the evaluation (more detail can be found in Chapter 7 Table 7.7). Precision (TP / (TP + FP)) is a measure of a classifier's exactness, which describes 'how many of the predicted boundaries are actual boundaries' in the boundary investigation. Then Recall (TP / (TP + FN)) is a measure of a classifier's completeness, which describes 'how good the algorithm is at detecting the actual boundaries out'. The F1 score (2 * Precision * Recall / (Precision + Recall)) conveys the trade-off balance between precision and recall.

3.4 Combining Approaches

The whole research flow was split into three stages in terms of the three research questions. The methodology involved in each stage was presented below, including experiment design, data gathering, and data analysis.

- RQ1: How do people understand and describe their own web sessions, and what factors create the divisions between them?
- RQ2: How are these self-identified sessions reflected in real-world logs of web activity, and what observable features represent their boundaries?
- RQ3: How can we algorithmically identify web sessions from detailed web activity logs?

For RQ1, the main goal was to obtain more information from users' perspectives during their session annotation. It focused on qualitative insight rather than quantitative analysis, so a smaller number of participants were recruited with a longer study period for each. In addition, it concentrated on the reality of users' daily web lives, and there was no given task during the weeklong data gathering period. After data collection, the participants were invited to a 60-minute semi-structured interview including card sorting to explore the factors users considered during their session annotation. The result was a list of session clustering factors from users' perspectives.

For RQ2, the main goal was to obtain a further reflection of user-identified session data in richer logging data, so a logging tool with richer detail was required. Most of the existing logging tools (e.g. the CrowdLogger [28, 27]) or datasets did not cover enough information. For example, the search datasets from AOL [11] and EXCITE [39] only contained the timestamp, query, clicked URL, and user ID. As a result, a custom-built Chrome extension was developed, and it can comprehensively log both pages visited and user interactions with the browser. More detail about this tool was presented in Chapter 5. The participants were asked to install the extension and browse the web as normal for one week while submitting data daily just like a dairy study [88] with automatic data gathering in a field setting, and it closely recorded users' interactions in a real 'noisy' environment without any pre-defined tasks. In the end, the final data submitted contained most of the users' daily web data except for the logs deleted by users themselves before the data submission.

For RQ3, five session clustering attempts were explored from different dimensions, namely user interaction and content similarity. Precision, recall, and F1 measures were then applied to evaluate the performance of these attempts in session boundary detection and session content overlapping accuracy compared to the ground truth dataset and some other fixed timeout versions.

3.5 Summary

I chose a natural setting with no pre-defined tasks for this study. Additionally, rather than using the manual user input required in dairy studies, I automatically recorded users' web data and then invited them back to perform the manual annotation following the data collection process. This helped to 1) reduce the side effects of the extra steps (i.e., manual data input) required from users during their interactions with the browser and 2) avoid missing some data. White et al. [132] found that some users preferred a system using implicit relevance feedback over explicit, because it reduced the burden of providing feedback.

In the end, several algorithms were proposed and evaluated. They were compared to the ground truth dataset and also the fixed timeout versions, and the performance in terms of both boundary investigation and session content overlapping were presented. Precision, recall and F1 measures were applied to evaluate the algorithms' performance.

Chapter 4

Study 1

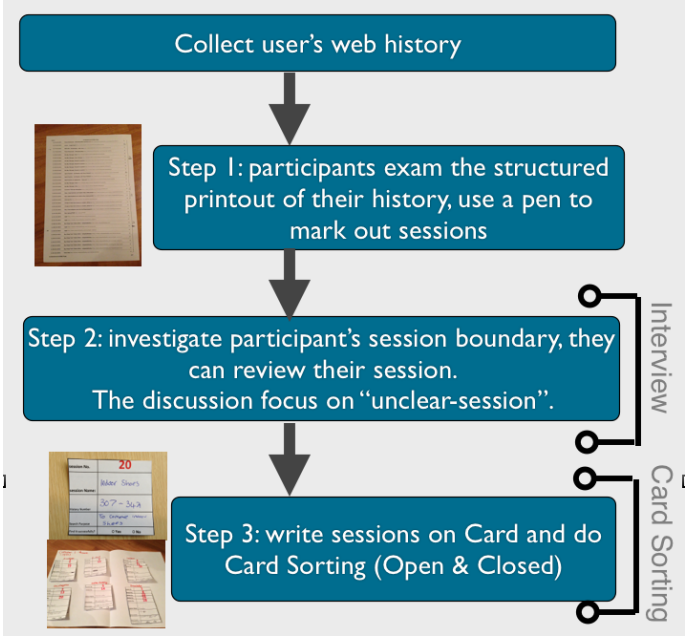
4.1 Goal

According to the first research question - 'How do people understand and describe their own web sessions, and what factors create the divisions between them?' - the goal of Study 1 was to understand how users want to cluster their own **natural** web data **by themselves** according to their information needs. This research question can be divided into the following sub-questions:

- 1) What factors affect the breakpoint (boundary) of a session?
- 2) What factors relate to the start of new sessions?
- 3) What factors divide apparently single sessions?
- 4) What factors resume the session?

Compared to some other research that has run the behaviour observation in controlled laboratory environments and provided the participants with pre-defined task, I concentrated on participants' natural web data withoutt any pre-defined tasks. This provided a direct reflection of participants' real online behaviour.

Fig. 4.1 Study 1 Procedure



4.2 Participants and Protocol

The whole procedure of this experiment is shown in Figure 4.1. A third-party browser add-on, History Export 0.4[1], was applied to gather users' web data. Twenty participants were recruited from the university. All of them were required to have a face-to-face meeting with me to ensure that they knew how to protect their own privacy in this study. After the data submission, the participants were invited to engage in a 60-minute interview to review their own web data and annotate the sessions and boundaries, which was similar to the interview process used by Sellen et al. [103]. During the interview, we actively avoided giving people a specific rule-based definition of 'session', and instead verbally provided an informal description of sessions by asking participants to 'categorise the logs serving the same information need or purpose into one session'. This is closer to Jansen's [50] conceptual view: *'a session is a series of interactions by the user towards addressing a single information need'*. No more specific definition was provided. The printout of web data shown in Figure 4.2 was presented to the participants. The log was then structured and pre-processed using a) automatic detection of search URLs and b) manual identification of

[1] addons.mozilla.org/en-us/firefox/addon/history-export/

periods of interest to discuss in the interview. After this preparation stage, the participants were invited to an interview to annotate their own web data (500 logs per participant) using a pen to mark sessions as shown in Figure 4.2. In the first 30 to 40 minutes the participants mainly focused on annotating their web data. After the annotation, the interview began by discussing the marked session boundaries. Participants were given the chance to review each session boundary and typically concentrated on the unclear or uncertain decisions that either the participant or the researcher found interesting or worth discussing. Therefore, the first iteration of self-annotation and the second iteration of review discussion 1) allowed the participants to deliver their first decision about the web data without any leading opinions from the researcher, 2) allowed the participants to review and revise some accidental mistakes, and 3) allowed the researcher to follow a participant's original thought to provide the corresponding terms for implicit criteria for each subsequent card sorting. To extract the corresponding criteria from the session and boundary, card sorting technique [101] was applied that consists of first open and then closed repeated single-criteria card sorting. As shown in Figure 4.5, the open card sorting allowed the participants to classify and group the sessions according to their own thoughts, and the closed card sorting allowed us to make sure the following dimensions were covered: **session duration**, **difficulty**, **importance**, and **frequency**. The interviews were audio recorded, and physical copies of the card sorts were kept for analysis. Although these methods do not allow me to analyse web sessions at a very large scale, they are conducive to building a better, richer understanding of web sessions and their boundaries, so the insights rather than the scale are emphasised. The procedure was approved and pilot tested by the school's ethics committee. Of course, in order to protect their privacy, all participants were asked to review their own data to exclude private and confidential information from this study. In the end, 10,000 history logs were manually analysed by 20 participants (i.e., each participant need to mark 500 logs that varied between 2 to 5 days depending on their individual usage).

This study focused on the qualitiative research exploring the factors users considered about their web data, so we focused on the deeper insights rather than a larger scale of user data. The twenty participants recruited from the university included students and staff from

Fig. 4.2 Study 1: Paper-Based Web History

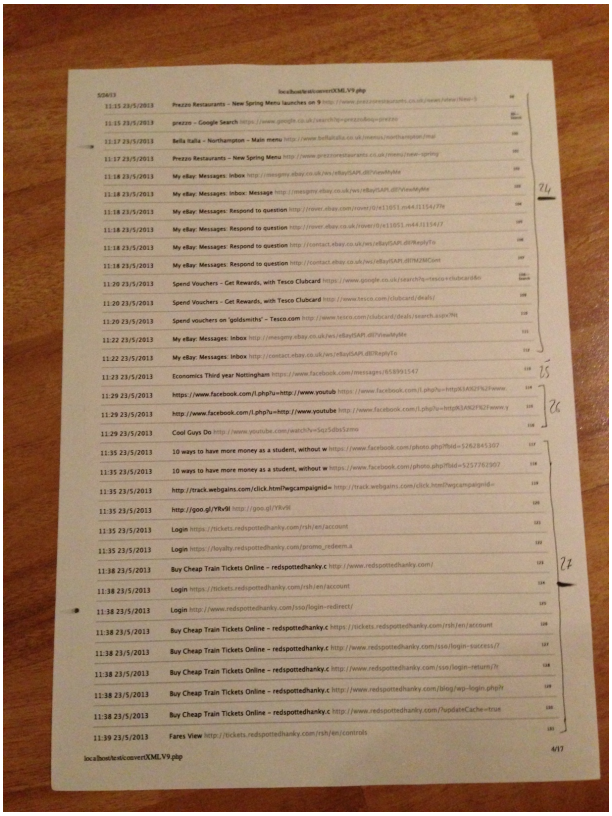
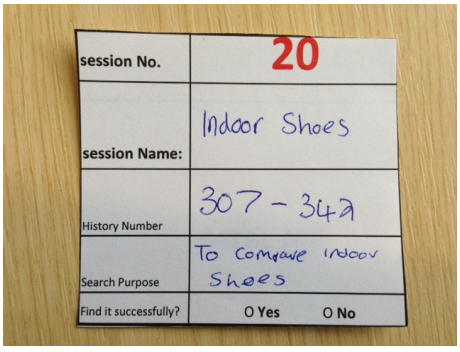


Fig. 4.3 Study 1: Card Sorting - Single Card

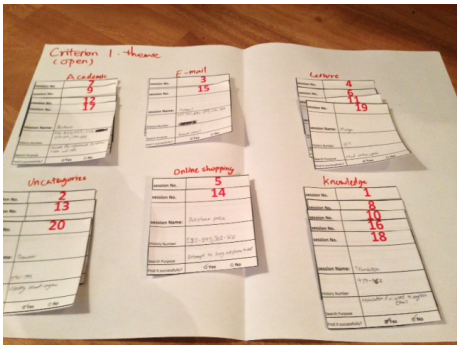


both non-technical and technical backgrounds. Nine were male and eleven were female. All of them were aged between 18 and 30. Eighteen out of twenty said they search online everyday and the other two said that they search online every three to five days. In the end, each participant was rewarded with a £15 Amazon voucher for the time heshe spent on the study.

Fig. 4.4 Study 1: Card Sorting - Clustering Cards



Fig. 4.5 Study 1: Card Sorting - Groups



4.3 Findings

4.3.1 Taxonomy

The thematic analysis [15] was adopted to analyse the transcript and also the generated sessions, and the whole process can be briefly summarised as 1) first, we captured all of the relevant informative words from the transcript (e.g., the reason) and also the self-identified sessions (i.e., session name), 2) then we iteratively collate these words list into some potential categories, 3) util we had a clear catgorisation.

Our qualitative analysis identified the following six key factors involved in determining different sessions: topic, task, phase, group of people involved in the activity, time gap, and multitasking. Table 4.1 summarises these key factors with details about when they cause a change in session, when there is an exception, and when they override other factors. The topic, task, and phase refer to the lexical clues about the activities grouped into sessions. The

Table 4.1 Taxonomy of Six Factors

	Factor	Creates a Session Boundary	Exception	Overrides another Factor
Content-Relevance	**Topic change**: Topic refers to the user's main intention, or higher-level work task, and it may be consist of one or more tasks or phases.	When Topic-change led to session change. Users may start a new session when the topic shifts. • P14: "*session 7 is about online shopping, and session 8 is related to my academic study, they are topical difference.*"	Topic-change may not lead to session change when the topic is too trivial or too broad. • P8: "*I grouped all of these [free-browse online shopping, social network] into one session because they are just free browsing, I don't have any particular purpose, just to relax.*"	Topic may override factors and join seemingly separate sessions. • When it overrides timeout and task-change: P14: "*all of information search in trip plan to europe before 1st of August should put into one session, including the accommodation, ticket , and places of interests searching.*"
	Task change: One topic may span several specific tasks, e.g. correspondes to distinct specific questions related to a big topic.	When task-change led to session change. In a big topic-based web acitivity, it may have mutliple specific tasks, which are relevant but different to each other. • P15: "*all of the specific problems searching related to the topic "Matlab" are put into separate sessions.*"	Task-change may not be able to lead to session change when the task is closely integrated or too small. • P17: "*these [topical-related] tasks are for different questions, but I want to group them together because some of them are just quick search and have only one query.*"	When the task or phase overrides other factors and bridge seemingly separate session together. • When it override time gap: P16: "*when I did some search yesterday and continue doing some more search on the same thing, I will put them into one session. Even if they have longer time gap.*" • When it override topic: the Matlab example from P15.
	Different phases: One topic may be made up of phases, they are more sequentially dependent when compared with the "Task".	When phase-shifts in one topic leads to session change. In a topic-based web activity, it may be identified as mutliple phases. • P4: "*In the flight ticket booking, Looking for information and final purchase are two different phases, because from checking price to purchase, I need a decision making and it takes time.*"	Phase-change may not lead to session change when phases may be too small to have a separate session. • P1: "*Searches on 'Burn a DVD' has two parts: 'How to burn it' and 'a software resource searching', and I put them into one session, because they are relevant and I didn't spread many sites.*"	
Social	**Different People** The group of people involved in the activity. e.g. the group of people user communicate with via their email or social network.	When group of people involved in the activity is changed, it leads to session change • P11: "*The gmail and uni emil should be put into different session, because I use the uni one to contact with my classmates and colleagues, and use the gmail for friends and family.*"	Some participants grouped all of adjacent activities across different social networks together. • P3: "*so while waiting, I will [...] either to check my personal emails, or because I use google chat a lot, or facebook, chat with my uni friends, and all of these should be put into one session, because they are just a break for me*"	When people override other factors and bridge seemingly separate sessions. • When it override topics, task, time: P6: "*I put all of the web activities from the same mailbox into one session, because the people I contacted with via the same mailbox are from same group.*"
Behaviour	**Time gap**: The time gap between web activity. It has been utilised as a main technique to divide session.	When the time gap is big enough to lead to session change, depending on other factors, such as task size and type of interruption. • P6: "*For the video, the acceptable time interval is less than 45mins. and for facebook, probably 1-2mins, and in academic search is less than 1h*" • P15: "*I put these two activities on one specific questions into one session, even they have more than 2 hours gap and it exceeds my acceptable time gap, but I knew it is interrupted by lunch.*"	When time gap is not considered as a factor in session division, esepcially for bigger, more Important activities. • P10: "*I don't mind the time, because they are for the same purpose, it is the same duty. So I put them together.*" • P14: "*because some of my information search may spread over days, for example, the information gathering on schengen visa takes me about three days, I will put all of them into one session even with days break.*"	When the time gap override other factors and bridge seemingly separate session together, such as the comments from P6 in the "Topic" factors.
	Multitasking: Sometimes, users may do multiple things concurrently.	Enough characteristically-diverse behaviour creates a session of "diverse activity", or a multi-tasking session. a session of un-connected web behaviour. • P6: "*I may feel borded when doing some task, so I probably stop and then go through my facebook, emails, or stream to have a break. I will put all of these during that period into one session - break session.*"	When the scale of interleaving activities are not trivial and they can be easily dividable. • P19 said "*my initial aim is to do academic info search then I switched to browse property info, and go back to academic again after a while. The property viewing in the middle should be put into a different session.*"	N/A

differences among them can be presented as follows: 1) topic refers to the broad goal of a series of activities that consists of one or more specific tasksphases; 2) tasks related to one topic are content-relevant but not sequential, such as the different *(task)* questions searched concerning the *(topic)* 'Java Programming'; 3) phases related to one topic are sequential (e.g., booking a flight ticket *(topic)* may first involve browsing cross sites *(phase 1 -> information gathering)* to gather the information before making the final decision and completing the transaction *(phase 2 -> transaction)*. These six factors are interrelated, and they represent common themes discussed by participants, rather than rules that can be directly applied. However, it is not a matter of *how often* each factor applies, but instead *how much* each factor applies at different times. Possibility for future work applying this taxonomy are discussed after the results.

Topic Change

The main topic was found to be one of the primary session delimiters in this study. Topic typically refers to the main idea of a user's intention, or hisher higher-level work task [17]. It may consist of one or multiple specific tasks or phases.

Creates a session boundary. Most participants discussed the topic of their work when marking boundaries in their web log. Participant 14 (P14) said: '*session 7 is about online shopping, and session 8 is related to my academic study, they are topical difference*' (see Table 4.1). P1 mentioned that he started a new session when shifting his activity from the online shopping on eBay to checking his email because he thought these activities were based on different topics. Most participants discussed the topic of their work when marking boundaries in their search logs, even during temporally consecutive periods of activity.

For example, some participants marked informal web activity prior to beginning work on a formal piece of work as different sessions because the topic was either to perform either some social 'housekeeping', or some news browsing before working on their project. One business student said '*For different news websites, I have quite different purpose for these. If I want to get some formal news, I mainly focus on the politics; I will think about finding some deeper things through these news, such as economics. If I just want to search some news for*

games, amusement, I don't think too much.'. Similarly, participants noted dividing the use of one online shopping resource into two sessions: '*If the shopping items are different, then they will be put into different sessions; the different shopping aims are the main reason.*'. In another example, a translator used the same website for different documents without a temporal gap between uses. In fact, the temporal gaps in their work were in the middle for breaks like lunch.

Exception. However, if the topic is too trivial to be identified as a single session, topic change may fail to cause session change and participants may just group trivial activities from different topics into one session instead. Partipant 8 (P8) said, *'I grouped all of these [free-browse online shopping, social network] into one session because they are just free browsing, I don't have any particular purpose, just to relax.'* P6 also stated: *'I may feel bored when doing some tasks, so I probably stop and then go through my facebook, emails, or stream to have a break. I will put all of these during that period into one session - break session.'*. In addition, if the topic is broad and its tasks are easily dividable, the tasks may be put into separate sessions rather than grouped together. This is explained further in task section.

Overrides another factor. Sometimes, topic may dominate and override other factors. P14 said:*'all of info search in trip plan to Europe before 1st August should be put into one session, including accommodation, ticket, and places of interests searching.'* In this case, the session extended across multiple days and the large time gap was not considered as important as the connectivity of the larger topic of trip planning. Furthermore, P14 bridged different specific tasks: accommodation, ticket booking, and researching places of interest into one session because of the one topic, trip planning, which overrode the tasks of accommodation and ticket booking.

Task Change

A topic may consist of several tasks. Participants often divided periods of activity into different tasks, where descriptions indicated that this was when these were more easily dividable, or larger in size.

Creates a session boundary. Specific tasks can be used to divide web activities into separated groups, even if they are related to a larger topic. Shifting tasks may create a session boundary. Within a large topic-based web activity, there may be multiple specific tasks that are topically relevant but different from each other (e.g., the Matlab example from P15 in Table 4.1. Tasks like questions searching such as 'what does error XXX mean in Matlab' and 'how to declare a variable in matlab' are both related to Matlab but were grouped into different sessions because P15 thought they were two different 'how-to' tasks, and separating them tidied the sessions up, which benefits their later revisitation or management.

Exception. When the scale of the tasks for one topic were relatively small, participants were less likely to make tasks sessions and would instead treat them as complementary or supporting 'missions' of the main topic. A comment from P17 suggests that several different but relevant tasks within a single search query should be put into one session, because he thought a session with a single item was meaningless (see the Europe trip planning example in the topic section).

Overrides another factor. Task is clearly related to the main topic in some forms, so projects that try to model common tasks in sessions help to determine thresholds and task detection. Task may override topic when they are easily dividable such as specific tasks in the aforementioned Matlab example, and it overrides the rule of 'putting topic-related activities into one session'. Task is also associated with specific collaborators and time impacts. In an example of when task overrides time, P16 said, 'when I did some search yesterday and continue doing some more search on the same thing, I will put them into one session. Even if they have longer time gap'. He/She decided to group continuous searching related to one technical problem across multiple days into one session despite it spanning overnight breaks, because the task was unchanged. Some tasks do have several sub-goals, but when they could be completed quickly, participants were less likely to call them sessions and more likely to consider them complementary or supporting search missions. In describing his/her preparation for an academic report, one participant described the process of interleaving different small tasks like finding references for different information but was not confident in defining finding references as separate from the session. This size concept and the ability to

clearly delineate one small goal from another small or main goal mean that when trying to model human search sessions, systems should consider thresholds' possible sub-goals before breaking them out as sessions.

Different Phases

Some types of activities have clear phases [96] for which progression can be predictable. There may be multiple phases related to one topic, for example. Compared with tasks, phases are more sequentially dependent on each other. Our participants also reported this behaviour, adding weight to the idea of whole-session relevance. These phases can be hierarchical or sequential, and participants noted that one phase might affect the activity in another one.

Creates a session boundary. One common example of this type was participants dividing periods of research and option comparison, as a separate session to then finding the best place to buy a product and then purchase it. P4 said: *'In the flight ticket booking, Looking for information and final purchase are two different phases, because from checking price to purchase, I need a decision making and it takes time.'* Another described two contiguous activities involving banking and bill paying as separate phases and thus separate sessions.

Exception. Not all phase-shifting led to session changes. Like with tasks, some participants said that although there were clear phases in the process, they were too small to be considered separate sessions. As P1 said, *'Searches on "Burn a DVD" has two parts: "How to burn it" and also a software resource searching, and I put them into one session, because I think they are relevant and I didn't spread many sites.'* These phases are easily identifiable: however, P1 thought the size of each phase was not big enough to warrant a separate session. P1 also said that if the downloading of software had involved learning and researching those would have become separable phases, highlighting the importance of size and delineatable goals for phases as well as tasks.

Overrides another factor. It is feasible that phases are simply a sequential instance of tasks, but this was not easy to determine from the qualitative data collected. The findings do have implications for projects examining supporting sessions with phases [96], which if grow to contain phases across separate sessions would have to adapt.

People

People are the special factors mentioned by most of the participants in the particular but frequent and popular web activities of email and social networks. People are somewhat related to the former three lexical factors (i.e., topic, task and phase), but they are a more specific factor related to the particular type of activity. I think extracting this particular type of factor can be helpful. The group of people involved in the activity was also a common theme in the interviews, although heavily related to others like task. This factor mainly applied in the online communication (e.g., the group of people a user communicates with via his/her email or social network). Related people could help identify a topic/task, but other contiguous periods of web activities were simply divided by the collaborators alone.

Creates a session boundary. 70% of participants preferred to put activities from different mailboxes and social networks into different sessions, because they utilise different platforms for contacting different groups of people (the domain-based implementation may work well for this). P11 described how contiguous use of email could be divided by the people involved: '*The gmail and uni emil should be put into different sessions, because I use the uni one to contact with my classmates and colleagues, and gmail for friends and family.*'

Exception. However, there are some exceptions to people change. A small group of participants grouped all of the activities across different social networks into one session when they were adjacent to each other. For example, P3 said: '*so while waiting, I will [...] either to check my personal emails, or because I use google chat a lot, or facebook, chat with my uni friends, and all of these should be put into one session, because they are just a break for me*'.

Overrides another Factor. Sometimes the group of people may override other factors, such as talking to specific people about multiple topics or tasks. P6 grouped all of the web activities from one mailbox even within a large time gap or activity differentiation into one session, saying, *'I put all of the web activities from the same mailbox into one session, because I think the people I contacted with via the same mailbox are from same group.'*

Time Gap

As with most research in this area, time gap has clearly been associated with methods used to divide sessions. Large time gaps (e.g., 30 minutes of inactivity) were repeatedly mentioned as separating sessions in our interviews, but the findings most notably highlight that they vary dramatically according to context. Time gap was mentioned by some of our participants, but none of them stick to 30 minutes. More than half of them said that they do not care about the timegap, and the remainder said that they consider the time gap, when it is a long night or lunch break, or that they have different time gap tolerance for different types of activities.

Creates a session boundary. Temporal gaps between activities were a common cause of separating sessions, whereas topics and tasks were frequently cited as anchoring sessions over a time gap. Large gaps, such as overnight breaks, usually divided sessions, P5 said, *'I did academic search about the "Learning environment" in different days, and I put them into separated sessions.'*. Furthermore, the acceptable time gap varies depending on types of web activity and the length of invested time, and some people even suggested that non-web activity gaps caused by real-life interruptions need to be as long as a few hours to divide a session. P6 said, *'For video, the acceptable time interval is less than 45 minutes. For facebook, probably 1-2 minutes, and in academic search is less than 1 hour'*.

Exception. In relation to the other factors discussed above, time did not divide temporally distant web activity when another factor became the overriding one. P14 highlighted how a topic can tie over a large period of activity together: '*[...] because some of my information search may spread over days, for example, the information gathering on schengen visa takes me about three days, I will put all of them into one session even with days break.*' Then P15 said, *"I put these two activities on one specific questions into one session, even they have more than 2-hours gap and exceeds my acceptable time gap tolerance, but I knew it is interrupted by lunch."*

Overrides another factor. In addition to clustering the activities with an 'excessive' time gap into different sessions, the time gap can also be applied to bridge some 'non-content relevant' activities with really small time gaps into one, such as the 'break period activities' from P6 above: he grouped all of the unrelated casual activities that happened in the break

period into one session because of the short time gap and trivial tasks. As a result, scaling the acceptable size of break in accordance with the size of the session determined so far might be a better way to model inactivity periods as a factor.

Multitasking

Because of the improved performance of web browsers that allows multiple tabs or windows to function at the same time, multitasking or interleaving is one of the most popular web behaviours. Participants frequently described activity in their logs as being caused by multitasking. Multitasking often accounted for divergent behaviour amongst larger sessions: however, participants also entered *states* of multitasking.

Creates a session boundary. Enough characteristically-diverse behaviour creates a session of diverse activity or a multitasking session, which is a session of non-content-relevant web behaviour. P6 said that he might also check his Facebook and email simultaneously during a serious working period to take a break. In this case, he preferred to put the break activity inside of the working session. The model of this situation is similar to 'one mainstream activity with some other trivial activities'.

Exception. P19 provided an example of when the scale of interleaving activities is not trivial and they are easily dividable: '*My initial aim is to do some academic information search, then I switched to browse property information, and go back to academic again after a while. The property viewing in the middle should be put into a different session.*' The model of this situation seems to be 'two or more mainstream activities interleave with each other', and the topical difference causes the session division.

Overrides another factor. To handle multitasking during other sessions (created by main goals), some approaches have simply ignored it and focused on things that match the current topical focus of the session. People may multitask in natural breaks, like between parallel tasks. These approaches to avoiding session changes seem relevant, but for sessions that are *identified* for multitasking it is important to learn not to model them to avoid unwanted incorrect support. The multitasking factor perhaps best highlights the risks for supporting sessions.

These comments and findings highlight the fact that these six interrelated factors contribute to the determination of users' session boundaries in different situations. The first three mainly focus on the content part of the online activities and make the decision based on the lexical relevance. However, telling the scale difference between them is still a significant challenge based on the log data analysis at this stage. The people factor is added mainly for activities involving other people such as online communication via email or social network. Because the relevant group of people is so closely related to different work activities, it can become a good indicator of online content. Time gap is a temporal technique applied in some prior research that still plays an important role in the session segmentation, but not in the way of a 'fixed time gap rule'. It varies dramatically according to individual preferences and the features of the activities themselves.

The exception and dominant between these factors is caused by 1) the individual difference on the scale preference and 2) people finding some tasks too trivial to be manually grouped, but it may be different with an automatic detection tool without any explicit input required from users.

4.3.2 Session Categorisation via Card Sorting

To understand what people thought about different types of sessions, card sorting was applied: which required participants to organise sessions into groups. There are mainly two types: open and closed. In the open card sorting, participants were asked to organise the sessions into groups that made senses to them and then name each group they created in a way that they felt accurately described the content. This helped us discover some unexpected findings. Then, in the closed card sorting, participants were asked to group these sessions using the pre-defined criteria. This enabled our research covering the dimensions we wanted to investigate.

Open Card Sorting

Table 4.2 lists the range of criteria participants chose to use in open card sorting. Nearly every participant began by using *purpose* to divide their sessions, creating groups like work,

entertainment, and social networking. The open card sorting provided some unexpected dimensions that differentiated sessions, such as *willingness* to do the activity, and whether or not sessions involved refinding via bookmarks. *Interestingness* was an unexpected but commonly used dimension. Although *interestingness* was defined differently than *willingness*, the separation of sessions was similar. Although difficult to utilise directly, these different dimensions may give us more inspiration to investigate other factors of session boundaries in the future.

Table 4.2 Dimensions of Open-Card Sorting

Purpose (19)	Willingness (1)
Language (2)	Academic or Not (2)
Text VS. Audio+Video (1)	Concentration (1)
Personal Use VS. Sharing (1)	Device (1)
Interestingness (5)	Region (Physical) (1)
Money related or not (1)	People involved (1)
Free VS. Paid (1)	Familiarity (1)
Long VS. Short Term (2)	Usefulness (1)
Google it or Not (1)	Successful or Not (2)
Search Engine Differ (1)	Via Bookmark or Not (1)

Closed Card Sorting

During closed card sorting, if not already used in the open process, we asked participants to group their sessions according to the following four criteria as shown in Table 4.3: importance, frequency, difficulty, and perceived duration in order to determine the relationship between people's perception on the session scale and those four dimensions. Furthermore, we noticed that the duration of some classified sessions were perceived as longer or shorter than objective measurements.

As a result, a *search session*[2] with more queries may lead to higher importance and difficutly but may happen less frequently. Meanwhile, a *browse-session*[3] with more pageviews may lead to higher importance, and the duration of them doesn't have obvious difference in the comparison.

Importance. *Search sessions* in the high importance group were longer and had notably more queries but fewer page views than other groups, which indicates that the query number

[2] if there are any web search queries submitted to one of the several major search engines (e.g. Google, Yahoo, Bing, Baidu) inside the session.

[3] if there are no search queries

Table 4.3 Analysis of Card Sorted Sessions, by Importance, Frequency, Difficulty, and Perceived Duration.

	Importance			Frequency			Difficulty			Perceived Duration			
	Low	**Med**	**High**	**Low**	**Med**	**High**	**Low**	**Med**	**High**	**FP-S**	**TP-S**	**TP-L**	**FP-L**
Number	67	31	48	75	25	46	71	35	40	8	52	6	45
Avg. Duration	15.7	18.6	17.3	15.8	24.6	14.2	9.7	22.3	24.8	32.3	4.2	73.3	18.9
Avg. Pageviews	19.2	15.5	16.9	17.7	20.7	15.9	12.0	22.5	23.4	15.0	8.6	56.5	23.7
Search Sessions													
Number	37	24	23	53	20	11	23	30	31	6	27	4	26
Avg. Duration	19.1	22.0	22.1	19.2	28.9	13.1	9.3	22.8	27.1	31.5	4.9	72.8	21.2
Avg. Queries	2.8	2.3	4.9	3.7	2.5	2.2	1.8	2.6	4.8	1.5	1.8	13.3	3.7
Avg. Pageviews	25.5	16.2	15.5	19.7	22.2	18.3	11.1	19.6	27.2	14.0	7.9	52.3	28.4
Browse Sessions													
Number	30	7	25	22	5	35	48	5	9	2	25	2	19
Avg. Duration	11.5	7.1	12.8	7.7	7.6	14.5	9.8	19.2	16.6	34.5	3.4	74.5	15.7
Avg. Pageviews	11.5	13.1	18.2	12.9	14.8	15.2	12.5	39.8	10.2	18.0	9.3	65.0	17.2

and the length of single pageviewing time may be an indicator of *search session* importance. Conversely, *browse sessions* in the high importance group had many more page views than other groups, indicating that the page views are related to the importance of a *browse session*.

Frequency. *Search sessions* in the low frequency group had more queries indicating that frequent search had fewer queries. Thirty-five out of forty-six sessions in the high frequency group were *browse sessions* with the highest average duration, perhaps because of some daily casual-leisure sessions like video streaming. Further investigation into the frequency of these sessions may reveal the pattern of daily session types.

Difficulty. The majority of high and medium-difficulty groups were *search sessions*, which implicitly indicates that query input may lead to higher difficulty. The *search session* in the high difficulty group lasted much longer and had more queries and page views. However, longer duration and more page views in *browse sessions* did not lead to higher difficulty, because the *browse sessions* in the medium group lasted the longest and had far more page views.

As a result, *search sessions* with more queries input may lead to higher importance and difficutly but happen less frequently. While *browse sessions* with more page views may be more importance, and their duration is not obviously difference.

Perceived Duration Besides the importance, frequency and difficulty, the session duration was also investigated. Duration includes two categories - actual duration and perceived duration - as shown in the confusion matrix in Table 4.4.

Table 4.4 Confusion Matrix of Actual and Perceived Session Duration

Versions	Perceived Long	Perceived Medium	Perceived Short
Actual Long	TP-L		FP-S
Actual Medium	FP-L	TP-M	FP-S
Actual Short	FP-L		TP-S

Rather than covering all three categories, we mainly focused on the long and short groups. Firstly, we mainly focused on the TP-L (both its perceived duration and actual duration are long) and TP-S (both its perceived duration and actual duration are short). Then there were another two groups that were perceived to be long or short when objectively not in those categorises: FP-L (its actual duration is not long but perceived as long) and FP-S (its actual duration is not short but perceived as short) as shown in Table 4.3. There are 45 FP-Ls (overestimation) and 8 FP-Ss (underestimation). Participants were more likely to overestimate rather than underestimate session duration.

FP-S and TP-S. The sessions in FP-S were percevied as short, but their actual durations were either medium or long. As a result, it was not surprising that their average duration and page views were much higher than those in the TP-S group. However, the query number in FP-S was lower than that in TP-S. This indicates that the query number in *search sessions* may have some effects on the perceived duration and lower query numbers may lead to underestimation.

FP-L and TP-L. The sessions in FP-L were percevied as long but their actual durations were either medium or short. The sessions in TP-L lasted much longer and had more page views, as well as many more queries if they were searches. There were no very significant findings concerning the reason causing the overestimation.

FP-S and FP-L. In the *search session* comparison between these two, the query number and pageviews in FP-L were more than twice those in FP-S, which indicates that query number and page views may have affect the duration perception, and more queries and page views may lead to overestimation.

4.3.3 Session Categories: Sellen's Six Categories and Elserwiler's Three Categories

Based on our preliminary investigation, some potentially interesting results relating to perceived duration, time of day, and use of queries were found. We consider each of these below in terms of two aspects: activity goal and activity context. For activity goal, we used Sellen et al.'s [103] six categories: finding, information gathering, browsing, transaction, communication, and housekeeping. This approach did not include email, so we added this as a seventh category. For activity context, we applied Elseweiler et al.'s [26] comparison between work and non-work (leisure) activities, including work, serious-leisure, project-leisure, and casual-leisure. At this early stage in the project we performed the classification based on corresponding examples given in the referenced work.

Table 4.5 Perceived VS. Actual Duration by Activity Goal

	TP-L + FP-L	FP-L (Over-estimation)	TP-S + FP-S	FP-S (Under-estimation)
Finding	24	17 (70.8%)	36	3 (8.3%)
Info-gathering	35	15 (42.9%)	7	4 (57.1%)
Browsing	28	17 (60.7%)	5	0
Transaction	4	2 (50.0%)	5	2 (40.0%)
Communication	9	3 (33.3%)	5	0
Housekeeping	0	0	1	1 (100.0%)
Email	7	6 (85.7%)	7	0

First considering activity goal (see Table 4.5), the number of information-gathering sessions in (TP-L + FP-L) was five times that of (TP-S + FP-S), and the same is true of browsing. On the contrary, the number of finding sessions (TP-S + FP-S) was 1.5 times that of (TP-L + FP-L). Overall, users overestimated nearly 70% of (TP-L + FP-L) finding, 42% of information-gathering, 60.7% of browsing, 50% of transaction, and 85.5% of email were overestimated by users. Underestimation also occurred with finding, information-gathering, and housekeeping although overestimation was more frequent for finding, browsing, communication, and email sessions.

Table 4.6 shows that the number of casual leisure (TP-L + FP-L was three times that of those (TP-S + FP-S) and that 57.9% of defined long work, 68.2% of project leisure, and 53.8% of casual project sessions were overestimated by users, with lower levels of

Table 4.6 Perceived vs. Actual Duration by Activity Context

	TP-L + FP-L	FP-L (Over-estimation)	TP-S + FP-S	FP-S (Under-estimation)
Work	38	22 (57.9%)	31	2 (6.5%)
Serious-Leisure	8	2 (25%)	1	0
Project-Leisure	22	15 (68.2%)	23	5 (21.7%)
Casual-Leisure	39	21 (53.8%)	11	3 (27.2%)

under-estimation occurring. This encouraged a further study on the features of each kind of web activity to determine the main cause for incorrectly perceived duration.

4.3.4 Other Findings

The time of the day. Figure 4.6 shows that most of the information-gathering, finding and housekeeping sessions seemed to happen between 10:00 and 16:00, while more browsing, email, and communication activities were done between 22:00 and 0:00, which was labelled the 'before bed time' period. In addition, there was a peak around 14:00 in which more finding and information-gathering happened compared to other activities. Finally, at 23:00, general browsing was most prevalent.

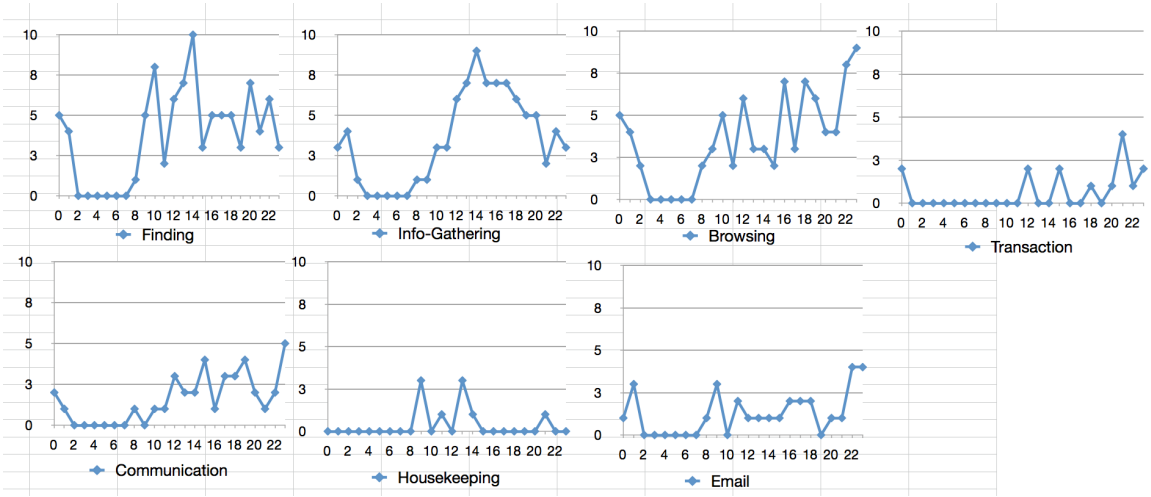


Fig. 4.6 Happening Time, by Activity Goal

Figure 4.7 shows that most of the serious leisure happened between 18:00 and 22:00. Most of the work activities happened between 11:00 and 18:00, which seems to fit within a

typical working day. In the 'before bedtime' period, the most frequent activity was casual leisure.

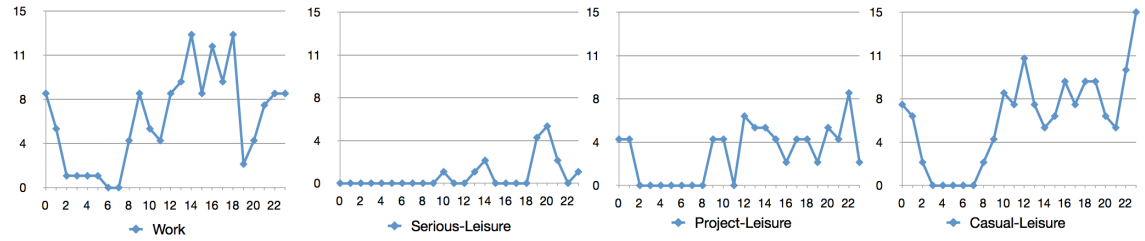


Fig. 4.7 Happening Time, by Activity Context

Combined with the two comparisons above, there seems to be some overlap between information gathering, finding, housekeeping, and work. There was also some overlap between browsing and casual leisure. Furthermore, these tend to suggest that there may be some patterns in users' online activity in their daily lives.

4.4 Summary

The taxonomy presents the factors in the session boundary detection, and the feature study of session delivers an idea about characteristic of session itself. The importance, difficulty, and frequency analysis of the session indicates that the query number, page views, and duration may have some effects on the judge of session scale. The overestimation and underestimation groups in the perceived duration indicate that we should estimate the perceived scale of a session by combining time and activity as indicators of importance and difficulty rather than just objective measurements. There are several insights that can be drawn from the taxonomy and session features. When thinking about the appropriate factors for different situations, the features of sessions play an important role. For example, the tolerance of a time gap when dividing sessions varies depending on the type of web activity, and it was much higher for larger and more important activities. In *search session*, we found a relationship between the number of queries and importance, with more queries potentially leading to higher importance and probably to overestimation of duration. In the *browse sessions*, more

page views led to higher importance. This seems to suggest that searches with more queries or browsing sessions with more pageviews should have higher time gap tolerance. Similarly, the number of queries was also a good indicator of frequent *search sessions*, where high frequency sessions had fewer queries and low frequency sessions had more queries. This may help to identify routine activities that can be more easily bounded as common sessions.

The results above provide three perspectives on how to determine the start and end of web sessions: 1) a taxonomy of factors that relate to the boundaries of sessions, along with notable exceptions and overrides; 2) insights into how those sessions manifested in web logs; and 3) how users perceived and categorised these sessions. Central to the contribution at this stage is that these factors have not been determined by researchers, but elicited from the users who created them. In relation to the four sub-questions set out at the beginning of this chapter, we discovered that the triggers that cause that start, break-point, dividing, and joining of sessions are dependent on six interrelated factors. The priority of each factor involved in different triggers requires further study, because it is significantly related to the scale of web activities themselves. Decision-making on the activity scale will be one of the most challenging asparts in the future. Although, for example, common triggers like large time gaps are typically considered to divide sessions, we saw sessions spanning overnight periods if the nature of the work task was large or important enough. Secion 4.3 also presented some insights into, for example, what makes a session important: *search sessions* were considered important if they included a large number of queries, and more page views were indicative of important *browse sessions*.

4.4.1 Applying the Taxonomy

Since the factors in our taxonomy were drawn from qualitative methods and are abstract themes that each relate to the boundaries discussed by participants, the subsequent challenge is to put the factors from the taxonomy into practice. It is not the case that simple rules can be derived from six factors, because each factor may have a certain amount of influence on a session ending. Consequently, the challenge in applying the taxonomy is to learn how to measure and track each of these factors, and then to discover how their thresholds, in

combination, create a boundary. This challenge is both the reason we cannot yet quantify the importance of each factor in our taxonomy and the primary motivator of our on-going and future work.

Chapter 5

Chrome Extension

The limited detail of each individual log has long been a bottleneck for many web research. Most of the prior research were based on the server query log (e.g., from AOL [11] or EXCITE [39]), which contains timestamp, URL, query, and clicked result. However, the other information listed below was missing. Consequently, a client-side log may be able to provide more insight into an individual's activity details.

- How much time did the user spend on each page (active dwell time)?
- How did the user get to this page (e.g., by clicking a hyperlink on a page)?
- When did the user open/close this page?
- How many times did the user shift his/her their attention off/onto this page, and when?
- If it is a search (e.g. in Google, Bing, Baidu, or Yahoo), what was the query?
- What did the user do on each page (e.g. click on some links)?

Compared to the server log, more interaction data on tabs/windows and the context switch between tabs can be recorded with the browser add-on. For example, Amazon may know when its users are on their Amazon website and which pages are viewed for how long, and eBay may also know its customers' data in the context of its website. While, the information of how and when users shift their attention between these different websites is not available in their server log. Knowning users' interactions between different websites (i.e., tabs/windows) can bring more insights. As a result, the browser extension becomes a good option for the

web activity research. According to the 2015 Global Desktop share report[1], more than half of desktop internet browsers are Google Chrome, and Google Search is also the dominant search engine[2]. In addition, there is a good community for Chrome extension development. Therefore, a chrome extension becomes the option for my research.

5.1 Design Detail

5.1.1 Single Log

The sample log captured from our Chrome extension is shown in Figure 5.1 and contains tab ID, log ID, window ID, URL, title, top 10 keywords, query (if it is a search), timestamp, and user interaction type. The top 10 keywords refers to the top 10 most frequent keywords in the webpage content. The user interaction type includes 11 categories as listed in Table 5.1, including tab opening, switching and closure, words typed in address bar, mouse clicks, and transition chain between pages/tabs.


```
"USER_ID": "formal_user_2",
"WIN_ID": 45,
"LOG_ID": 58,
"TAB_ID": 5,
"LOG_TIMESTAMP": 1441103533816,
"HB": "TULC_NonBlank",
"OPENER_TAB_ID": 0,
"TAB_STATUS": "UNFOCUSED",
"IS_TAB_INITIALISED": true,
"URL": "http://www.bbc.co.uk",
"TITLE": "BBC - Home",
"TOP10KWS": "bbc radio news watch listen audio video uk football match",
"QUERY": ""
```


Fig. 5.1 Sample Log

5.1.2 Data Storage

Local Machine: These logs were stored in local databases (IndexedDB)[3] in each participant's local machine. Compared to web storage (e.g., LocalStorage[4]), IndexedDB can store larger

[1] https://www.statista.com/statistics/544400/market-share-of-internet-browsers-desktop/

[2] Search Engine Market Share Report in 2015: http://gs.statcounter.com/search-engine-market-share/desktop/worldwide/2015

[3] https://developer.mozilla.org/en-US/docs/Web/API/IndexedDB_API

[4] https://developer.mozilla.org/en-US/docs/Web/API/Window/localStorage

amounts of structured data with a reasonable performance. It is a transactional database system that is similar to Sqlite to some extent. However, it is still a bit slower when compared to web storage like LocalStorage. Writing logs directly to IndexedDB can be a bit challenging when users interact with the browser in too quickly. Therefore, without requiring users to slow down their interactions, we designed the data storage flow shown in Figure 5.2: the logs were persistently stored in IndexedDB, which could take up space on the user's local machine. Because the writing and reading data from IndexedDB is much slower, all of the logs were firstly written to LocalStorage and then moved to IndexedDB when it was idle. Once the logs were moved to IndexedDB, they were wiped from LocalStorage. This enabled writing efficiency and eliminated the risk of data loss (e.g., from a sudden crash).

Remote Server: Participants could submit the local data to our remote server via Ajax[5]. It was then processed via PHP[6] script and stored in MySQL[7] in the server.

Regarding the special design delay between LocalStorage and IndexedDB, all of the log data was first put into LocalStorage because of its high efficiency in both reading and writing. The stored data was then delivered to IndexedDB in the damen program in the damon program in background. The goal was to 1) avoid data loss caused by a Chrome extension or browser crash and 2) avoid exceeding the size limitation of LocalStorage since there are only 5 MB for it in Chrome in MacOS.

5.1.3 Data Submission

In order to protect participants' privacy, a daily data submission and a manual operation were required to provide participants with full awareness about what had been submitted and when the data was sent. When participants tried to submit the data,the list of all logs that were going to be sent would be processed by a JavaScript[8] script and presented to the participants. They could then review or manage (as shown in Appendix B) their log data to remove any confidential information before submitting the logs to the server. The content fields (i.e., title,

[5]https://en.wikipedia.org/wiki/Ajax_(programming)
[6]http://www.php.net/
[7]https://www.mysql.com/
[8]https://developer.mozilla.org/bm/docs/Web/JavaScript

query, URL) of any log marked ‘private’ were replaced by the word ‘PRIVATE’, but the interaction data field would stay the same.

We were able to re-run these logs with different versions of algorithms to generate corresponding versions of sessions for evaluation. The details of these algorithms are described in Chapter 6. Since the logs were well formatted and stored in a database, they could easily be exported into different formats (e.g., JSON, CSV).

5.1.4 Pre-Processing Dummy Logs

Since one user interaction with the browser may create multiple logged actions and some of them also invoke the same log entry, the Chrome extension also processed the logs into a more usable format as shown in Figure 5.1. For example, opening a link in a new tab would create a series of log entries including a click and tab creation, as well as page content loading logs. Creating a new blank tab also includes the tab creation log, so pre-preprocessing some of the dummy system logs and extracting the actual user interaction data was important. In addition, the redirection on Google’s search page, Google Instant searches, and the browser cache each led to some unexpected dummy log entries that were not initiative log by the user. In order to better understand and exact interactions, therefore, the extension categorised groups of logs into the interactions shown in Table 5.1.

Table 5.1 Interaction Types

	Detail
CNT	Create new blank tab
CT	Close tab / window
SAB_BlankTab	Type query / URL in address bar in blank tab
SAB_NonBlankTab	Type query / URL in address bar in non-blank tab
TULC_BlankTab	Update tab via clicking hyperlink in blank tab
TULC_NonBlankTab	Update tab via clicking hyperlink in non-blank tab
OLNT	Create a tab / window via clicking hyperlink in tab
SW	Switch attention between tabs / windows
BOF	Click back or forward button
FEPNT	Create a tab via clicking hyperlink from an external application (e.g. MS Word)
LBIW	Switch attention off from the browser or go to Incognito window
SBFE	Switch your attention back from the external application

5.1.5 Test

Some unittest were applied for some functions and classes during the software development. Then, after the completion of beta version, the tool was firstly installed on my personal machine for one week, and then had been installed on three pilot participants' machine for one week. The data submission function was tested on every participant's machine in the induction meeting (see Chapter 6) with their allocated username and password.

5.2 Software Architecture

The whole project consisted of two parts: client (Chrome extension) and server (database + rerun algorithm + data visualisation for interview in experiment), as shown in Figure 5.2.

The Chrome extension has two aspects, namely background and ContentScript. The background performed several tasks: 1) capturing all of the logs, 2) storing this data into IndexedDB. The ContentScript is mainly in charge of obtaining page content information such as the top 10 keywords of the webpage. The programming language used here is JavaScript (jQuery) with the local database in IndexedDB. The server performed three tasks: 1) receiving the data from clients and storing it in a MySQL database, 2) re-running the algorithm on original log data to generate different versions of sessions, and 3) presenting the log data to participants in interviews and allowing them to annotate these logs. First, the data was presented to the user as shown in Figure 5.3: the detail of each log entry presented here was as shown in Figure 5.4. For each log, participants could tick the box when he/she thought it was a boundary, and the options for session selection would pop up as illustrated in Figure 5.5, allowing them to either create a new session or resume back to any existing session. After the user's self-annotation, there was a validation phase for users to double-check their own input, as shown in Figure 5.2. The programming language applied here was JavaScript(jQuery) and PHP with the database in MySQL.

Fig. 5.2 Chrome Extension Structure

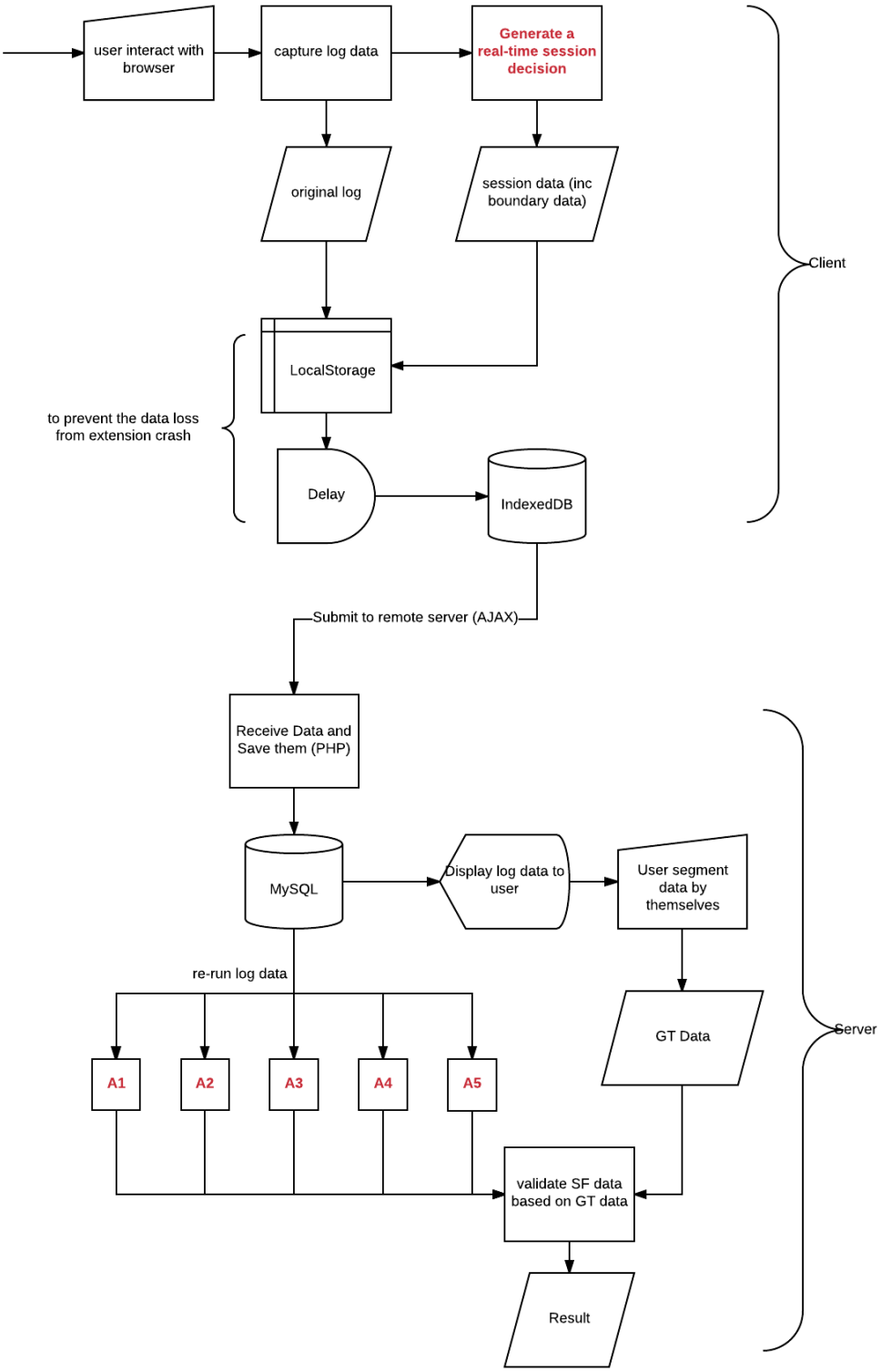


Fig. 5.3 Interview Interface for Data Input

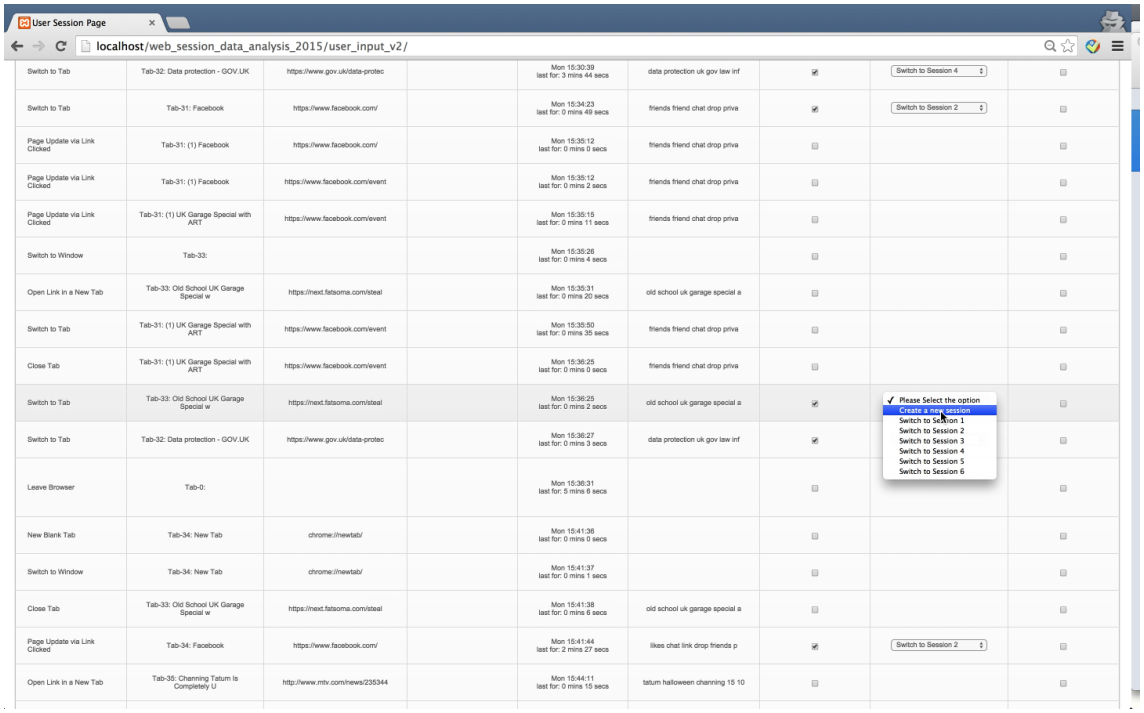


Fig. 5.4 Each Log Entry during the Interview

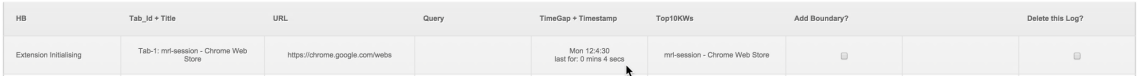


Fig. 5.5 Session Selection

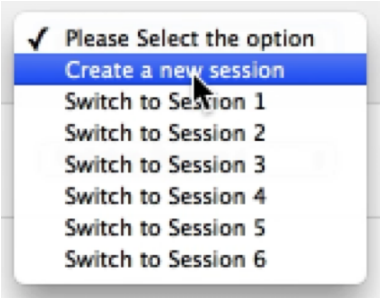


Fig. 5.6 Validation from User



Chapter 6

Study 2: Ground Truth Dataset

6.1 Goal

In accordance with the second research question - 'How are these self-identified sessions reflected in real-world logs of web activity, and what observable features represent their boundaries?', the goal of Study 2 was to determine what self-identified sessions look like in a more detailed web logs and generate a ground truth (GT) dataset for these sessions. We followed a similar experiment design to Study 1 in that we worked with participants to annotate and understand their own daily web data with no pre-defined tasks. The main difference from Study 1 is the log data. The log data applied in Study 1 was exported from the Firefox add-on History Export 0.4 with simple history logs in chronological order, and the recorded logs only refer to the logs containing timestamp, URL, and action (i.e., when the page/tab was closed or updated). These logs were much simpler than the reality during users' interactions with browsers, and some information (e.g., tab switching and navigation between tabs/pages) was missing. With the custom-built Chrome extension, more user interactions were captured. In addition, for each log entry more features were recorded, as shown in Figure 5.1 in Chapter 5. Moreover, all of these data was collected in a field study setting without any pre-defined tasks. Therefore, the ground truth corpus is a series of logs assigned to different sessions by the participants themselves rather than depending on automated techniques or manual labelling by researchers.

6.2 Participants and Protocol

Initially, twenty-one participants from both non-technical and technical backgrounds were recruited from across the university. However, one participant was removed based on failing to participant fully in the process. Of the remaining twenty participants, ten were male and ten were female, nighteen of them were between the ages of 18 and 30, while one was between the ages of 31 and 40. All particiapnts indicated that they search online everyday. Six were Mac OS users and fourteen were Windows users. The whole procedure was approved and pilot tested by the university's ethics committee, and every participant had been rewarded with a £15 Amazon voucher.

1) Setup phase: Participants began with a setup meeting; we met participants face-to-face to provide an introduction to the study and gain informed consent. Participants were shown what kind of data would be gathered, how to install the Chrome extension on their computers, how to submit data, and how to protect their privacy (i.e., they could remove the confidential data before data submission). The installation was then tested. We recommended that the particiapnts do daily data submissions with allocated username and password.

2) Logging phase: Participants were asked to search and browse as normal for approximately seven consecutive days. They all recorded more than enough data for the study.

3) Review phase: After the data collection, the participants were invited to engage in a 60-minute interview to review their own web data and annotate the sessions and boundaries. It was similar to Study 1, but with richer detail in each individual log. We actively avoided giving people a specific rule-based definition of 'session', and instead verbally provided an informal description of session as 'categories the logs serving the same information need or purpose into one session', which is closer to Jansen's [50] conceptual view of a session as a series of user interactions addressing a single information need. To explicitly identify sessions, all of the logs were presented to participants via a web interface (see Figure 5.3 in Chapter 5), and which allowed them to add a boundary and assign the type of the boundary: 'create a new session' or 'switch to an existing session'. In order to keep the task manageable within an hour, participants were asked to review the first 800 log entries from their web data.

6.3 Ground Truth Dataset: Session

6.3.1 Overview

In the 16,000 logs reviewed by 20 participants, a total of 227 user-determined sessions were identified in total. Several types of sessions were considered in the analysis. Based on whether or not they involved any search (query-input) activity, the sessions can be categorised as *search session* and *browse session*. Among these 227 sessions, 103 of them were *search sessions* and 124 were *browse sessions*. The results indicate that when modelling sessions, covering browsing is very important, because more than 50% of the sessions in our dataset are *browse sessions*. While most prior research has concentrated on *search session* rather than all web activity, Kong et al.'s work [61] on pre-search contexts suggests a strong relationship between search and browse in that many queries are triggered by different contexts in which browsing is one of the key factors. Our work below considered both *search session* and *browse session*, and it confirms many of Kong et al's conclusions about their relationship.

- **Search session**: If there are any web search queries submitted to one of several major search engines (e.g. Google, Yahoo, Bing, Baidu) inside the session
- **Browse session**: If there are no web search queries submitted to one of the several major search engines inside the session.

Aside from considering both *search session* and *browse session*, we also categorised the sessions into three groups according to their duration[1]: 103 *short session*, 67 *medium session*, and 56 *long session*. In order to eliminate the side effects of long breaks (like overnight gap beetween the adjacent logs) in *long sessions*, a one-hour threshold was used. This means that any inner session time gap of less than 60 minutes was considered part of the active duration, while any inner session timegap longer than 60 minutes was not considered part of the active duration.

- **Short session**: When session duration is less than five minutes.
- **Medium session**: When the session duration is equal to or more than five minutes and less than thirty minutes.

[1]The sum of active dwell time of all web pages inside one session

- **Long session**: When the session duration is equal to or more than thirty minutes.

6.3.2 User-Level

Based on each participants' data, the participants' sessions were very different from each other, as shown in Figure 6.3. The distribution of *search session* and *browse session* is presented in Figure 6.1 and the distribution of different session duration groups is presented in Figure 6.2. Figure 6.3 illustrates how our measures in these sessions varied between participants. The uneven shape of session duration and page duration indicates that there is a significant individual difference in time spent in sessions between users. For all participants, a large number of their sessions involved multiple tabs, domains, and pages. All of the users' average number of domains per session is more than 2. Some of them are even more than 10. Additionally, nearly half of participants had an average of more than 5 tabs in each session. For most participants, the average number of webpages visited in a session is over 5. All of these indicate that domain-based or tab-based segmentation may not be a successful method for clustering sessions, which is proved in more detail later on. Finally, and perhaps most interestingly, the average number of boundaries in a session is more than 10 for all users, which indicates that on average, a user switched to another session 9 times per session (since the first boundary is always to create the session). This indicates a high frequency of multitasking.

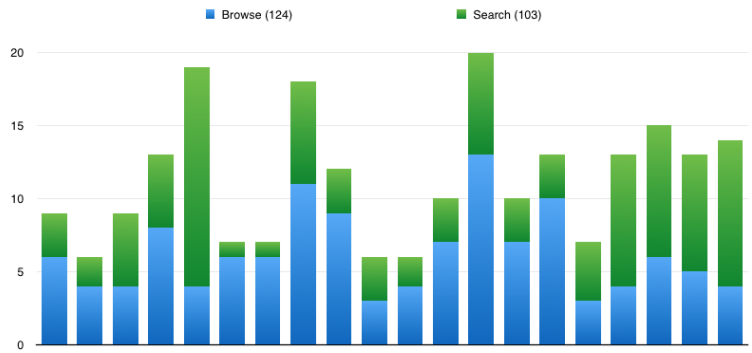


Fig. 6.1 Number of Search vs. Browse Sessions for Each Participant

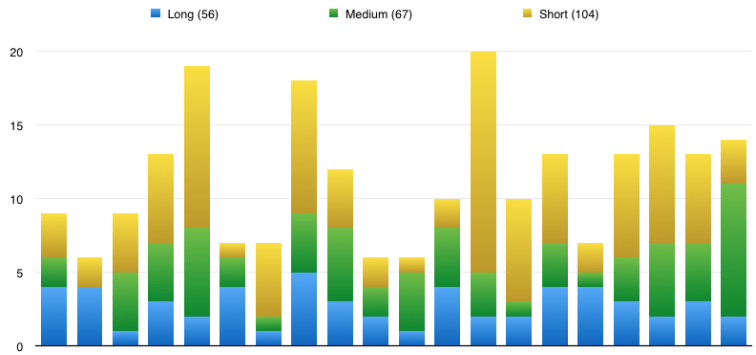


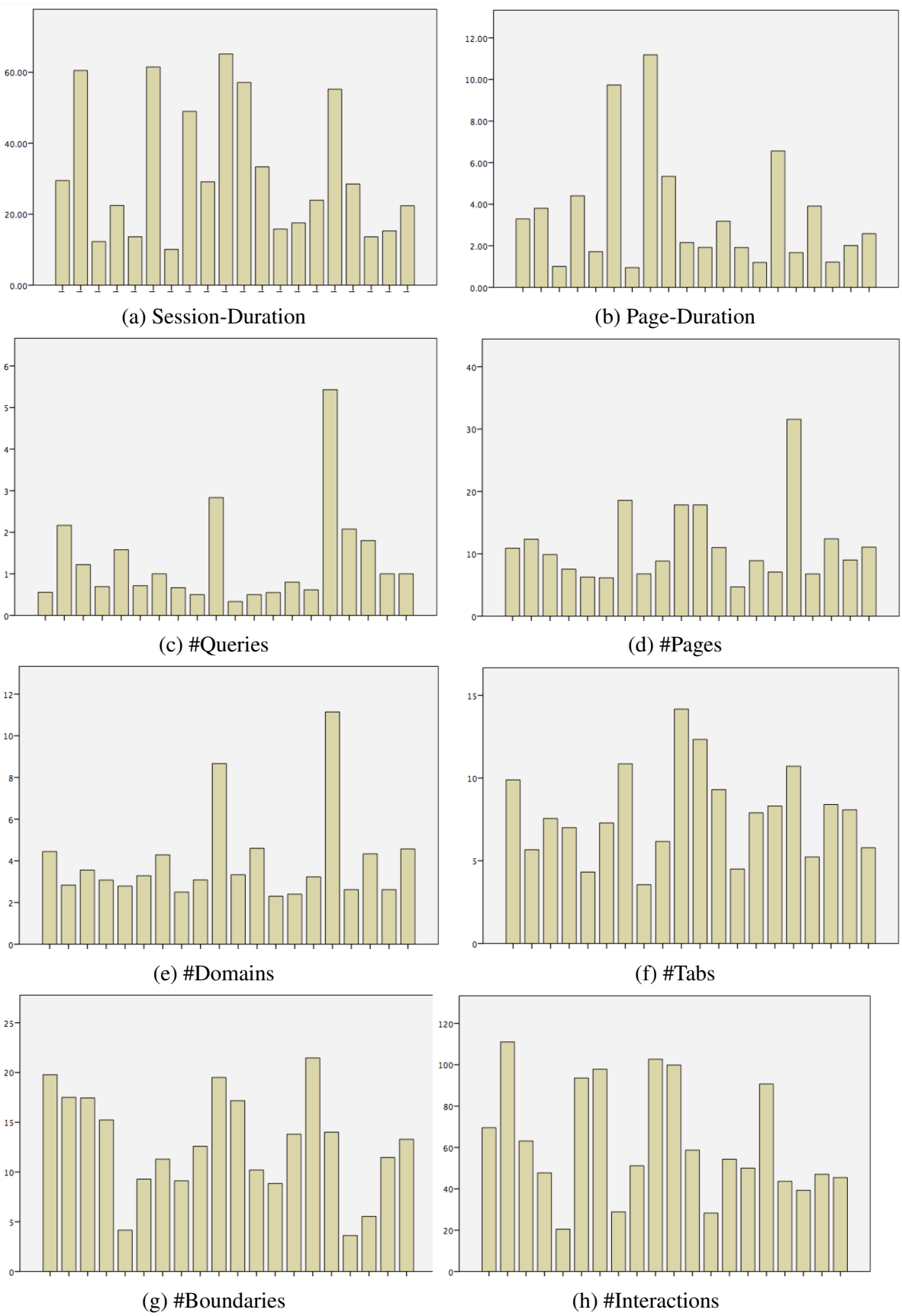
Fig. 6.2 Number of Long vs. Medium vs. Short Sessions for Each Participant

6.3.3 Session Level

Here we examine all the sessions as a corpus. As illustrated in Figure 6.4, the sessions were all very different from each other. The overview of these session characteristics is presented in Table 6.1 (i.e., Green: very strong correlation (> 0.8), light blue: strong correlation (0.6-0.8), orange: weak correlation). In order to characterise these sessions, we investigated the session details and their correlations with other logged actions. The correlation between characteristic may reveal some implicit findings to benefit further algorithm design.

Table 6.1 Session Characteristics

	Max	**Average**	**Std**
Session-Duration (mins)	282.92	28	51.3
Page-Duration (mins)	132.72	3.64	10.14
#Queries	24	1.18	2.46
#Pages	113	9.89	13.94
#Domains	28	3.62	3.95
#Tabs	61	7.11	9.24
#Boundaries	99	11.7	16.07
#Interactions	591	52.6	82.87
#SEI	12	0.21	0.95
#FEPNT	10	0.42	1.192
#OLNT_TAB	48	2.78	5.93
#SAB_Blank	7	0.87	1.23
#SAB_NonBlank	10	0.65	1.29
#TULC_Blank	7	0.53	1.09
#TULC_NonBlank	339	19.68	40.48
#R	32	0.68	2.70
#BOF	21	0.84	2.64
#SW	306	20.57	40.94
#CNT	11	1.21	2.07
#CT	35	4.16	5.89



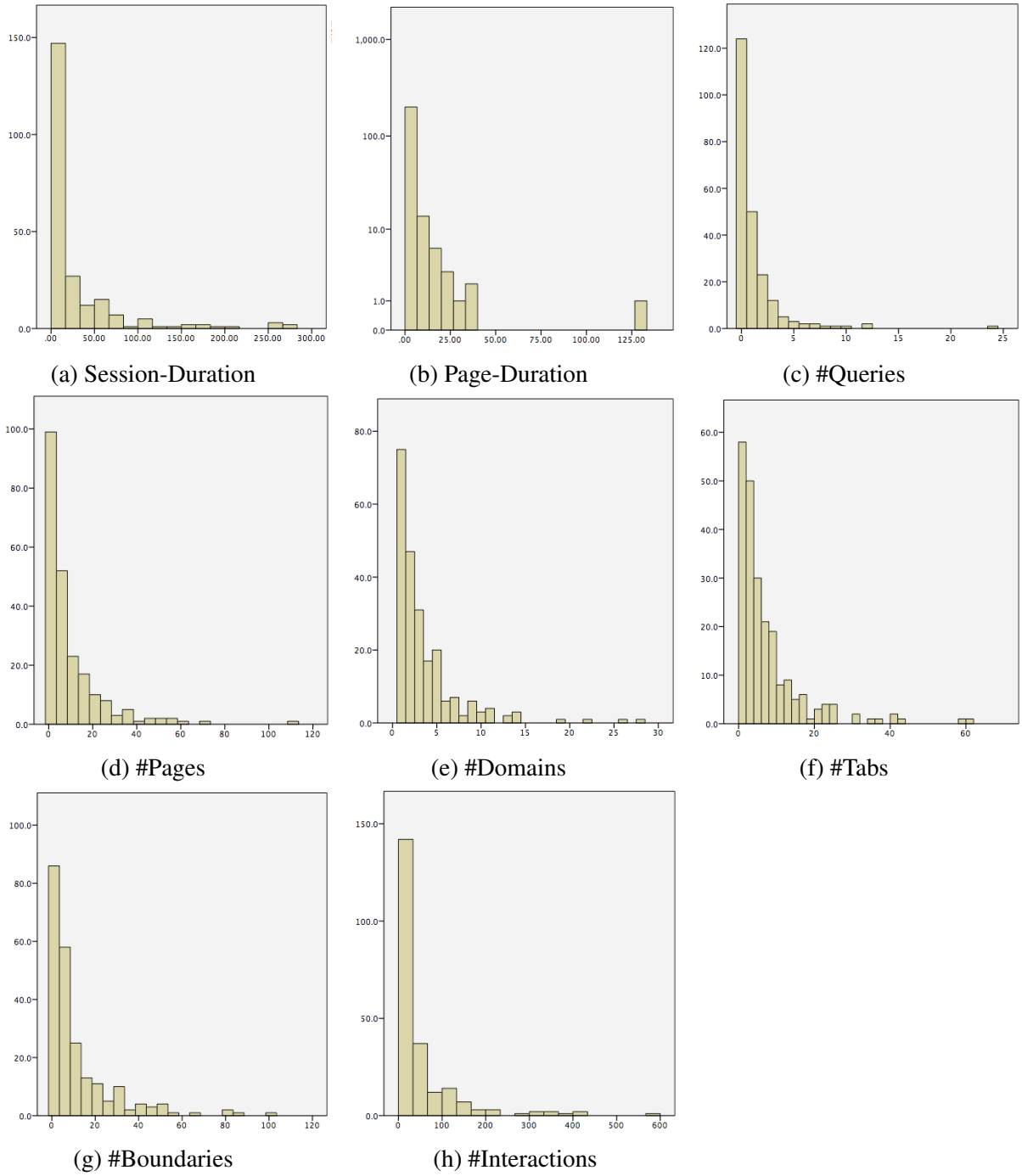
(a) Session-Duration (b) Page-Duration

(c) #Queries (d) #Pages

(e) #Domains (f) #Tabs

(g) #Boundaries (h) #Interactions

Fig. 6.3 Mean Session Measures for Each Participant



(a) Session-Duration (b) Page-Duration (c) #Queries

(d) #Pages (e) #Domains (f) #Tabs

(g) #Boundaries (h) #Interactions

Fig. 6.4 Distribution of Measures for All Sessions

	Session-Duration	Tabs	Boundaries	Pages	Domains	Interactions	SAB_NonBlank	OLNT	TULC_NonBlank	SW	BOF	CT	SAB_Blank	TULC_Blank
Session-Duration														
Tabs	0.774													
Boundaries	0.764	0.779												
Pages	0.733	0.737												
Domains	0.648	0.704		0.815										
Interactions	0.837	0.849	0.844	0.837	0.691									
SAB_NonBlank														
OLNT		0.757		0.624	0.638	0.678								
TULC_NonBlank	0.670		0.621	0.836	0.622	0.85								
SW	0.818	0.86	0.918	0.701	0.632	0.906		0.709	0.644					
BOF														
CT	0.619	0.811		0.706	0.672	0.738		0.792		0.704				
SAB_Blank														
TULC_Blank														

Fig. 6.5 Correlation Overview Table

Session Features

Since none of the datasets met the requirement of normality, Spearman's correlation was applied for the analysis. The correlation results are illustrated in Figure 6.5.

Session-Duration: Perhaps expectedly, **Session-Duration** had a strong positive monotonic correlation with **#Interactions** (0.837), **#Tabs** (0.774), **#Boundaries** (0.764), **#Pages** visited (0.733), and **#Domains** (0.648). Among these user interactions, **Session-Duration** was strongly correlated with **#SW** (0.818), **#TULC_NonBlank** (0.670), **#CNT** (0.641), and **#CT** (0.619). This means that longer sessions involved more interactions, especially SW (switching between tabs and windows) and TULC_NonBlank (following the hyperlinks). Longer sessions also involved more visits to web pages. There was a weaker correlation between **Session-Duration** and **#Queries** (r_s = 0.244, n = 227, p=0.007), which indicates that number of queries is not a good indicator of session duration (e.g., longer session may have fewer queries than shorter session). There is a further investigation of queries and related details in the *search sessions* section below.

#Pages: This factor varied dramatically and had a strong positive monotonic correlation with **#Interactions** (0.837), **#Domains** (0.815), **#Tabs** (0.737), and **Session-Duration** (0.733). Among these interactions, **#Pages** was strongly correlated with **#TULC_NonBlank** (0.836), **#CT** (0.706), **#SW** (0.701), and **#OLNT** (0.624). This means that a session with more webpages visited had more interactions, tabs, and domains. It also lasted a bit longer and was more likely to involve multitasking. Among these four strongly positively corre-

lated interactions, TULC_NonBlank and OLNT are two main indicators of exploring more webpages, especially TULC_NonBlank, which is very strongly correlated with **#Pages**. This illustrates that the change of web page visited via TULC_NonBlank and OLNT does not necessarily mean a change in session focus, and a big percentage of web pages visited within a session is triggered by clicking hyperlink. These large numbers of strong correlations, however, imply that **#Pages** visited cannot be easily be used as a predictor against other features.

#Domains: More than 70% of the sessions involved visits to multiple domains. This indicates that domain-based techniques for dividing sessions may not be very successful. **#Domains** had a strong positive monotonic correlation with **#Pages** (0.815), **#Tabs** (0.704), **#Interactions** (0.691), and **Session-Duration** as mentioned above. Among these interactions, **#Domains** was strongly correlated with **#CT** (0.672), **#OLNT** (0.638), **#SW** (0.632), and **#TULC_NonBlank** (0.622). As expected, this means that a session with more domains also involves more webpages visited and tabs created. In addition, there are also more interactions involved especially CT, OLNT, SW and TULC_NonBlank. Among these, OLNT and TULC_NonBlank have the potential to lead to new domain visits. This illustrates that visiting a new domain via TULC_NonBlank and OLNT seems to be more likely to continue a session rather than create a boundary. In addition, as illustrated in Table 5.1, there are several other kinds of interactions - namely SAB_NonBlank, SAB_Blank, TULC_Blank and FEPNT that can also lead to new domains, but **#Domains** was not positively correlated with these. This seems to imply that the new domain visited via SAB_NonBlank, SAB_Blank and TULC_Blank is more likely to lead to a session switch than TULC_NonBlank and OLNT.

#Tabs: More than 60% of sessions involved multiple tabs. As a result, tab-based methods for dividing sessions alone may not be very reliable. **#Tabs** had a strong positive monotonic correlation with **#Interactions** (0.849), **#Boundaries** (0.779), **Session-Duration** (0.774), **#Pages** (0.737), and **#Domains** (0.704). Among these interactions, **#Tabs** was strongly correlated with **#SW** (0.860) **#CT** (0.811), **#OLNT** (0.757), and **#CNT** (0.747). This means that a session with more tabs will involve more interactions, and more webpages visited, more domains, and more frequent multitasking (i.e., boundaries), and will have a longer

overall duration. Among these three positively correlated interactions, OLNT and CNT are used to create new tabs. There are several types of interactions (i.e., SAB_NonBlank, FEPNT, OLNT, and CNT) that can lead to new tabs, as seen in Table 5.1. As a result, the new tab created via OLNT is more likely to stay in the same session rather than break it. Because the tab after CNT is a blank tab (i.e., no page content), it is assigned to 'pending session' by default, which is a 'stay' in the current session. Additionally, interactions like SAB_Blank and TULC_Blank are not strongly correlated with #Tabs, so the decision on the tab after CNT still cannot be concluded as 'more likely to stay'. The variant version of tab-based session segmentation, which is to 1) make the new tab created stay in the same session and 2) make the new tab created via the other interactions (SAB_NonBlank, SAB_Blank, TULC_Blank, FEPNT) into a different session, may be useful.

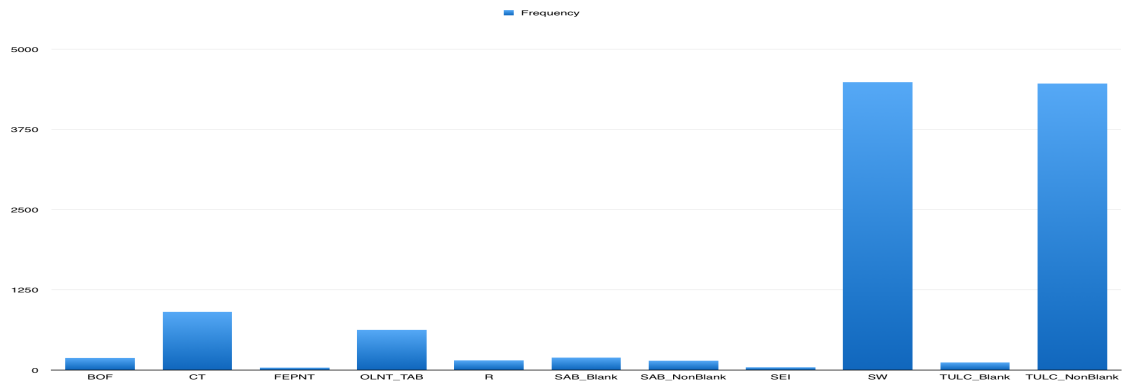
#Boundaries: More than 90% of sessions were interrupted by another session and later resumed, and therefore involved multitasking. **#Boundaries** had a strong positive monotonic correlation with **#Interactions** (0.844), **#Tabs** (0.779), **Session-Duration** (0.764), and **Page-Duration** (0.601). Among these interactions, **#Boundaries** was strongly correlated with **#SW** (0.918) and **#CNT** (0.655). This means that a session with more boundaries (more multitasking) had more interactions, tabs and webpages visited, and had a longer overall duration.

#Interactions: The number of interactions had a strong positive monotonic correlation with **#Tabs** (0.849), **#Boundaries** (0.844), **#Pages** (0.837), **Session-Duration** (0.837), and **#Domains** (0.691). Among these interactions, **#Interactions** had a strong positive monotonic correlation with **#SW** (0.906), **#TULC_NonBlank** (0.850), **#CT** (0.738), **#OLNT** (0.678), and **#CNT** (0.654). This means that a session with more interactions would primarily involve more webpages visited, more frequent multitasking, and more tabs created than other user interactions logged. Looking at the detail of interactions allows deeper insight into each particular type of log.

High-Level Interactions

One of the most unique details of our recorded log is the user interaction data as shown in Table 5.1. Figure 6.6 shows the overall frequency of each interaction across the whole dataset, and SW and TULC_NonBlank are the two most frequent interactions.

Fig. 6.6 Overall Frequency of Each Interaction



#SAB_NonBlank: This term refers to the number of searches from the address bar in non-blank tabs within each session. It was not strongly correlated with any other characteristics. Since it is one of the interactions with the potential to lead to new domains, or new pages, its weak correlation with **#Pages** and **#Domains** implies that SAB_NonBlank may mostly occur as a result of a shift of information need, leading to a session boundary.

#OLNT: This term refers to the interaction of clicking on a hyperlink or button to open it in a new tab. There was a strong positive monotonic correlation with **#Tabs** (0.757), **#Interactions** (0.678), **#Domains** (0.638), and **#Pages** (0.624). Among these interactions, **#OLNT** was also strongly correlated with **#CT** (0.792) and **#SW** (0.709). Generally, there are several interactions that can create a new tab (see Table 5.1. Since correlation between **#OLNT** and **#Tabs** is 0.821, compared to other interactions leading to the creation of a new tab, OLNT is more likely to be the main reason that more tabs are created in one session. As a result, the new tab created via OLNT seems to be more likely to continue a session rather than break it.

#TULC_NonBlank: This term refers to the interaction of clicking on a hyperlink or button to update the current tab webpage updated. It had a strong positive monotonic cor-

relation with **#Interactions** (0.850), **#Pages** (0.836), **Session-Duration** (0.670), **#Domains** (0.622), **#Boundaries** (0.621). Among these interactions, **#TULC_NonBlank** was also strongly correlated with **#SW** (0.644). As shown in Figure 6.6, it is one of the most frequent interactions during users' web activities, and combines the high correlation with **#Pages**. This indicates that more TULC_NonBlank will of course lead to more webpages and domains (as expected), and a large portion of new webpages visited occurs via TULC_NonBlank. Interestingly, however, **#TULC_NonBlank** was also positively correlated to **#Boundaries**, but analyses below indicate that it is not a good indicator of boundaries.

#SW: This term refers to the interaction of switching between windows and tabs. It is also one of the most frequent interactions seen in any one session. It had a strong positive monotonic correlation with **#Boundaries** (0.918), **#Interactions** (0.906), **#Tabs** (0.860), **Session-Duration** (0.818), **#Pages** (0.701), and **#Domains** (0.632). Among these interactions, **#SW** was also strongly correlated with **#OLNT** (0.709), **#CT** (0.704), **#CNT** (0.668), and **#TULC_NonBlank** (0.644). The very strong positive monotonic correlation between **#SW** and **#Boundaries** indicates that multitasking is closely related to tab switching, which supports the correlation between **#SW** and **#Tabs** noted above.

#BOF: This term refers to the interaction of clicking on the backward or forward buttons in a browser. It was not strongly correlated with any other details, but we examine it further when we categorise the types of sessions below.

#CT: This term refers to the interaction of closing a tab. It had a strong positive monotonic correlation with **#Tabs** (0.811), **#Interactions** (0.738), **#Pages** (0.706), **#Domains** (0.672), and **Session-Duration** (0.619). Among these interactions, **#CT** was also strongly correlated with **#OLNT** (0.792) and **#SW** (0.704). Naturally, this confirms that if there are more tabs have been created in sessions, then there will be more interactions about closing them (CT).

#SAB_Blank and **#TULC_Blank**: the first content update of a new blank tab (CNT) is always done via either searching in the address bar (SAB_Blank) or click on the link (e.g., a recently viewed page) or button (TULC_Blank). **#SAB_Blank** had a nearly strong positive monotonic correlation with **#Queries** (0.571). **#TULC_Blank** was not strongly correlated with any other characteristics.

Different Types of Sessions

As explained above, we objectively categorised sessions into key groups for analysis by determining whether they were *search sessions* or exclusively *browse sessions*, and into *short session*, *medium session* and *long sessions* (by **Session-Duration**). In all cases, Mann-Whitney U tests were applied because the data was not normally distributed.

Search session vs. browse session Based on the results of Mann-Whitney U tests in Table 6.2, there were several significant differences between *search session* and *browse session* in terms of **#Pages**, **#Domains**, **#Tabs**, **#Interactions**, **#SAB_NonBlank**, **#SAB_Blank**, **#TULC_NonBlank**, **#TULC_Blank**, **#CT**, and **#BOF**. *Search session* seems to have involved more webpages and domains visited with more tabs created and more interactions. Among these interactions, *search session* has more SAB_NonBlank, TULC_NonBlank, CT, and BOF than *browse session*. *Search session* is likely to involve visiting more webpages via TULC_NonBlank across domains, as well as more use of the backward and forward buttons (BOF). The **#Queries** in all *search sessions* ranged from 1 to 24 with a mean as 2.6. While, neither type of session was especially longer than the other; although the average page duration (**Page Duration**) is longer in the *browse sessions* in our sample, this difference was not significant. Similarly, switching windows and tabs (**#SW**) was more common in *search sessions*, but not significantly so. The similar number of boundaries between these two types of session indicates that either type is just as likely to be interleaved with other sessions.

Short session vs. medium session vs. long session Table 6.3 shows the differences between short, medium, and long sessions, but significant differences were not examined since they were inherently divided by size.

Perhaps as expected, in all session groups, the **#Pages** was positively correlated with link clicks (TULC_NonBlank - $r_{s\,(S)} = 0.765$ vs. $r_{s\,(M)} = 0.730$ vs. $r_{s\,(L)} = 0.774$), but not with OLNT($r_{s\,(S)} = 0.162$ VS. $r_{s\,(M)} = 0.267$ VS. $r_{s\,(L)} = 0.711$). The general frequency of TULC_NonBlank in the logs indicates that visiting new pages via link clicks is one of users' main ways to explore more pages, especially in *short session* and *medium session*. In the *long session* group,

Table 6.2 Characteristics of Search Session and Browse Session

	Search-Session (103)	Browse-Session (124)	Significant Diff
Session-Duration	31.3	25.3	U=5206, p=0.017
Page-Duration	2.4	4.7	
#Queries	2.6	NAN	
#Pages	14.3	6.3	U=3831.5, p=0.000
#Domains	5.2	2.4	U=3286.5, p=0.000
#Tabs	8.5	5.9	U=4946, p=0.003
#Boundaries	11.9	11.5	
#Interactions	64.1	43.1	U=5299.5, p=0.027
#SAB_NonBlank	1.02	0.35	U=3982.5, p=0.000
#SAB_Blank	1.45	0.40	U=2933.5.5, p=0.000
#FEPNT	0.35	0.48	
#OLNT	3.6	2.1	
#TULC_NonBlank	22.59	17.26	U=5007, p=0.005
#TULC_Blank	0.42	0.62	U=5289.5, p=0.006
#SW	25.37	16.58	
#CT	5.18	3.31	U=5215.5, p=0.015
#BOF	1.68	0.14	U=4176, p=0.000

Table 6.3 Characteristics of Short, Medium, and Long Session

	Short (104)	Medium (67)	Long (56)
Session Duration	1.6	14.2	93.7
Page Duration	0.7	2.6	10.4
#Queries	0.5	1.1	2.5
#Pages	2.8	10.7	22.1
#Domains	1.7	4.0	6.8
#Tabs	2.0	7.4	16.3
#Boundaries	3.0	11.4	28.3
#Interactions	10.2	47.1	138.0
#SAB_NonBlank	0.3	0.9	1.0
#SAB_Blank	0.5	1.0	1.4
#FEPNT	0.1	0.5	0.9
#OLNT	0.4	2.7	7.2
#TULC_NonBlank	4.3	17.1	51.3
#TULC_Blank	0.2	0.5	1.1
#SW	2.6	17.2	58
#CT	1.2	4.5	9.3
#BOF	0.3	0.6	2.2

some pages were also explored via OLNT. This implies that TULC_NonBlank is a good indicator of continuing sessions rather than boundaries (discussed below), as is OLNT in *long session*. Naturally, there are also some clear differences between these three comparisons. The **#Boundaries** in *long session* had a stronger positive correlation with **#Interactions** than that in *medium session* or *short session* ($r_{s\,(S)} = 0.565$ vs. $r_{s\,(M)} = 0.501$ vs. $r_{s\,(L)} = 0.705$). The **#Domains** in *long session* has a stronger positive correlation with OLNT ($r_{s\,(S)} = 0.138$ vs. $r_{s\,(M)} = 0.277$ vs. $r_{s\,(L)} = 0.776$). This means that in *long session*, new domains are more likely to be visited via OLNT than that in the medium and short Group. This seems to imply that the OLNT may be a stronger indicator for continuing (rather than breaking) a session, especially

for longer sessions. Another difference is that in *long session*, the **#Pages** were positively correlated to **#Tabs** ($r_{s\,(S)} = 0.373$ vs. $r_{s\,(M)} = 0.261$ vs. $r_{s\,(L)} = 0.732$). This seems to indicate that users may visit more pages in one tab during shorter sessions, as opposed to longer sessions. As a result, people tend to have more tabs in longer sessions and visit fewer pages per tab, which makes the tabs in longer session more coherent in content.

Combining length and type We chose to explicitly examine *search session* and *browse session* for each of the three duration groups; the colour-highlighted cells in Table 6.4 indicate a significant difference according to additional Mann-Whitney U tests.

Table 6.4 Comparison of Search vs. Browse sessions for each length group.

	Short ($< 5mins$)			**Medium** ($5-30mins$)			**Long** ($>= 30mins$)		
	Search (40)	**Browse (64)**	**Sig. Diff.**	**Search (35)**	**Browse (32)**	**Sig. Diff.**	**Search (28)**	**Browse (28)**	**Sig. Diff.**
Session Duration	2.0	1.4		15.6	12.6		92.9	94.5	
Page Duration	0.7	0.7		1.9	3.3	U=380, p=0.024	5.4	15.4	U=190, p=0.001
#Queries	1.4	0		2.1	0		5.0	0	
#Pages	3.5	2.4	U=903.5, p=0.009	13.7	7.4	U=282, p=0.000	30.4	13.9	U=203, p=0.002
#Domains	2.2	1.3	U=626, p=0.000	4.8	3.1	U=300, p=0.001	9.8	3.8	U=147.5, p=0.000
#Tabs	2.3	1.8	U=950, p=0.017	7.3	7.5		19.0	13.7	U=271, p=0.047
#Boundaries	2.2	3.4	U=980, p=0.037	10.0	12.9		28.0	28.5	
#Interactions	8.8	11.1		48.4	45.6		162.7	113.2	
#SAB_NonBlank	0.5	0.1	U=845.5, p=0.000	1.2	0.6	U=373.5, p=0.008	1.5	0.5	U=226, p=0.003
#SAB_Blank	0.9	0.3	U=628.5, p=0.000	1.5	0.5	U=265.5, p=0.000	2.2	0.6	U=173, p=0.000
#FEPNT	0	0.1	U=1140, p=0.031	0.2	0.9	U=438, p=0.031	1.1	0.7	
#OLNT	0.3	0.5		2.6	2.8		9.6	4.9	U=245.5, p=0.016
#TULC_NonBlank	2.5	5.4		19.3	14.6	U=387.5, p=0.030	55.4	47.3	
#TULC_Blank	0.0	0.3	U=891.5, p=0.000	0.4	0.7	U=408.5, p=0.022	1.0	1.1	
#SW	2.3	2.8		15.8	18.8		70.3	45.7	
#CT	1.3	1.1		4.3	4.7		11.8	6.7	U=249, p=0.019
#BOF	0.5	0.1	U=1002.0, p=0.002	1.1	0.1	U=365.5, p=0.002	4.1	0.3	U=188, p=0.000

According to Table 6.4, there are significantly more **#Pages**, **#Domains**, **#SAB_NonBlank**, **#SAB_Blank** and **#BOF** in *search session* than *browse session* in all duration groups. In Medium and Long Group, there are no significant difference in their **Session-Duration** but

significant differences were found in **Page-Duration**. *Browse session* tends to stay longer on a single page longer than *search session*: the difference seems to vary and become more obvious with the increment of session duration. In the medium group, *search session* tends to visit more pages than *browse session* but stay on single page for shorter lengths of time. In the long group, besides the significant difference in **Page-Duration**, there is also a significant difference in OLNT. *Search session* have more OLNT, which leads to more tabs created and more webpages and domains visited. Overall, as sessions became longer, *search session* and *browse session* became more distinctive in the logs, but differentiating between shorter sessions involves fewer clues. *Search session* tends to have more tabs created, more domains visited and also more pages via OLNT, but they stay on a single page for a shorter amount of time. *Search session* also seems to involve more interactions like SAB (searching directly or typing URL in the address bar), open more links and involve more clicks on back or forward button.

6.4 Ground Truth Dataset: Session Boundary

For each log, there are three primary possibilities: create new session (Create-Log), switch to existing session (Resume-Log), and continue current session (Continue-Log). Besides these three primary types, there are also other types listed (not included in the following analysis) in Table 6.5: Pending, Outside, and Private.

Table 6.5 Log Types

is Boundary?	Specific Type	Description
Yes	Create	Start a new session
	Resume	Switch back to one of the existing sessions
No	Continue	Stay in current session
Other	Pending	When there is no content on tab (Blank Tab Created)
	Outside	When user leave the browser and pay their attention to other programme (e.g. Type in MSWord)
	Private	When this log is marked by the participant as “Private” log

Therefore, our dataset does not explicitly identify the end of a session, because we could not be sure that a participant would never resume it; however sessions do have final entries. In this analysis, we analysed all of the occasions where participants either switched between sessions or created a new session. There were 227 Create-Logs and 2,434 Resume-Logs.

This sheer number of Resume-Log illustrates the frequency of multitasking in people's web activities. The vast majority of sessions were interrupted by other sessions and later resumed; many sessions were complexly interleaved. Below, we first examine the user interactions and logs around each boundary and then the nature of the periods of inactivity found around each boundary. We also examine whether those user interaction or time gaps primarily occurred around the boundaries or whether they occurred frequently at other times.

6.4.1 Interactions at Each Boundary

Table 6.6 shows the distribution of interactions around the Create-Logs. Searching in the address bar (SAB), including both SAB_Blank and SAB_NonBlank, is the key interaction associated with a user starting a new session; more than half of **Create-Log** transitions were instigated with an SAB (Search or Type URL in Address Bar). Opening a link within the same tab (TULC_Blank) was third, followed by opening a link in a new tab (OLNT). While searching in the address bar is quite distinctive, following a link clicked within a tab was one of the most common logged behaviours in the system (see Figure 6.6). Table 6.6 also demonstrates the distribution of interactions for all boundaries, including both create and resume. Searching in the address bar (SAB), either as SAB_Blank or SAB_NonBlank, is almost significant, highlighting its close relationship with boundary occurrence. The extremely low probability of TULC_NonBlank may also be a strong indicator of staying in the session. **All of these illustrated that users' interactions may be potential indicators for their changes in intention.**

Table 6.6 Probability of a Log Being a Create Boundary or a Boundary in General

	Create	Switch	Probability to be Create Boundary	Probability to be Create or Switch Boundary
BOF	0	11	0.0%	5.6%
SAB_Blank	90	58	44.8%	73.6%
TULC_Blank	38	61	30.2%	78.6%
FEPNT	3	17	2.0%	13.1%
SAB_NonBlank	33	50	22.3%	56.1%
TULC_NonBlank	7	200	0.2%	4.6%
OLNT	25	113	4.7%	25.7%
SW_TABS	7	1818	0.1%	38.1%

6.5 Summary

Our dataset shows notable variation in the number and size of sessions, both between users and within individual users. All of our participants engaged in multiple instances of both *browse session* and *search session*, as well as sessions that were short, medium, and long in length of time. However, there were also some more fundamental characteristics that most of the users had in common in terms of sessions and their boundaries. It was common, for example, for participants to have a large number of boundaries in most of their sessions. Combined with frequent tab switching, this confirms much prior research indicating that multitasking is one of the main behaviours in people's daily web activity [112]. Additionally, the multi-tab, multi-domain, multi-query nature of sessions across all participants also illustrates that in their minds, a real web session is far more complex than individual models of domain-based, tab-based, or query-based segmentation. These results all support the idea that session is a very complex concept.

From the characteristics investigation of all sessions, longer sessions were found to typically involve more webpages and domains visited as well as more tabs created with more interactions, especially tab switching (SW), link following (TULC_NonBlank and OLNT) and closing of tabs (CT). Conversely, no significant correlations were found between session duration and number of query, and it indicates that session duration has more to do with the way results are examined and less to do with querying behaviour. This is perhaps confirmed by the data indicating that link clicks (TULC_NonBlank) were the main way participants explored more pages with.

The comparison between *search session* and *browse session* shows that a *search session* is likely to be a session with more pages visited via TULC_NonBlank across domains, and increased use of backward and forward buttons (BOF) compared to a *Browse-Session*. With longer sessions, *search session* seems to create have more tabs and have more links clicked to open other pages via either TULC_NonBlank or OLNT to navigate further than *browse session*.

The comparison between different duration groups indicates that in *long session*, new domains are more likely to be visited via OLNT than that in *medium session* and *short*

session, and this implies that the OLNT may be a stronger indicator for continuation in longer duration session. Another finding is that people may create more tabs and visit fewer pages per tab in longer sessions.

Furthermore, TULC_NonBlank and OLNT are two main indicators of visiting more webpages inside a session, especially TULC_NonBlank. This indicates that the change of web page visited via TULC_NonBlank and OLNT is more likely to stay in current session. Additionally, sometimes, the new page could lead to a new domain, and the new domain visit via TULC_NonBlank and OLNT is also likely to stay in current session. Instead, a session switch, or starting a new session, was more closely tied to the address bar (SAB). These results indicate that SAB is a good indicator for having a session boundaries referring to the information need shift. The probability of SAB_Blank being a boundary is 73.6% (44.8% to be a create boundary) and the probability of SAB_NonBlank being a boundary is 56.1% (22.3% to be a create boundary). These indicate that there is a high possibility of leading to a boundary when typing into the address bar.

Chapter 7

Study 3: Five Versions of Algorithms and Fixed Timeout

After establishing a ground truth dataset and a deeper understanding of the boundary characteristics, five versions of rule-based algorithms were proposed. This problem was treated as a Data Stream Clustering [105]. In this case, the nature of data stream clustering could be described as follows:

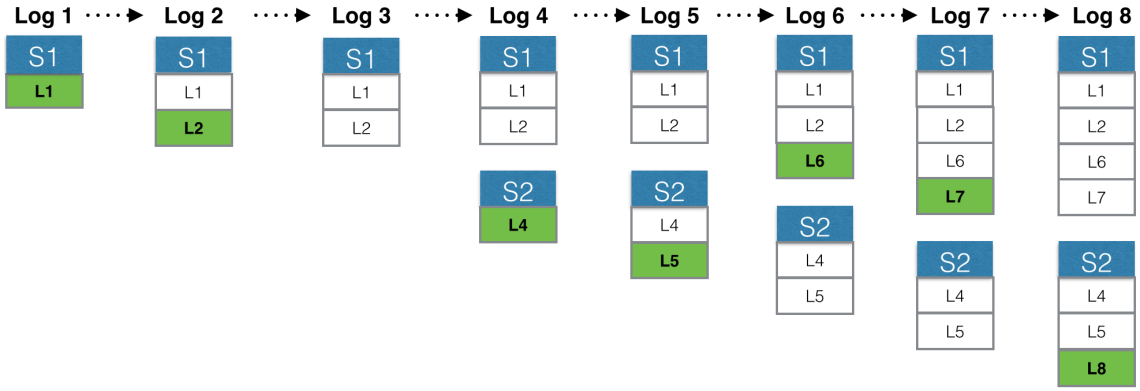
- Logs arrive continuously;
- There is no control over the order in which logs should be processed;
- The size of a stream is (potentially) unbounded;
- Logs are discarded after they have been processed. In practice, one can store part of the data for a given period of time and use a forgetting mechanism to discard them later; and
- The unknown data generation process is possibly non-stationary; that is, its probability distribution may change over time.

In principle, decision has to be made when a new log entry arrived, and be decided in what way it is related to existing sessions (clusters) in the data or whether it is sufficiently different to become a new session (cluster), as shown in the log sequence in Table 7.1 and its decision flow in Figure 7.1.

Table 7.1 Sample Log Sequence with Users' Self-Annotated Session Boundaries

Log id	URL	User Interaction	Title	Tab id	Session id	is Boundary?
Log_1	https://www.google....array+copy	SW	python array copy - Google	1	1	—
Log_2	http://stackoverflow.....-a-list	TULC_NonBlank	python - How to clone or copy a list-Stack Overflow	1	1	No
Log_3	chrome://newtab	CNT	New Tab	2	PENDING	PENDING
Log_4	https://www.facebook.....upcoming	SW	events	2	2	Yes
Log_5	https://www.facebook.....lists/	TULC_NonBlank	Facebook	2	2	No
Log_6	http://stackoverflow.....-a-list	SW	python - How to clone or copy a list-Stack Overflow	1	1	Yes
Log_7	http://stackoverflow.....-in-python	TULC_NonBlank	Mutability of lists in python - Stack Overflow	1	1	No
Log_8	https://www.facebook.....lists/	SW	Facebook	2	2	Yes

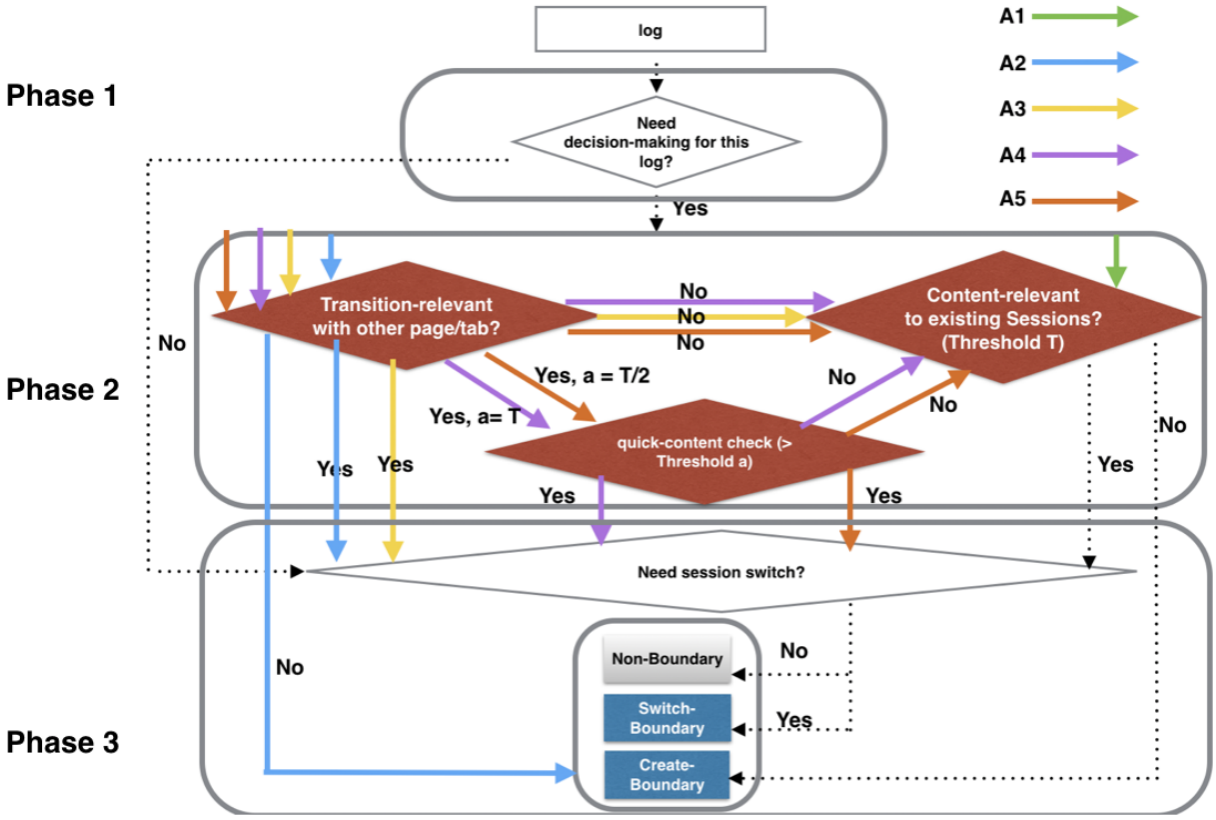
Fig. 7.1 Decision Flow of Sample Log Sequence



7.1 Whole Flow

The whole workflow starting from the arrival of new log to the final deicision of that log can be divided into these three phases as illustrated in Figure 7.2.

Fig. 7.2 Flow Chart of a Log Decision



Phase 1 it would determine whether it need to run through the decision-making in Phase 2. For example, for a log like 'Create a new blank tab' or 'Switch back to a tab whose session ID has already been determined', there is no need as there was no content change in a tab. Otherwise, it would move to Phase 2.

Phase 2 it is to find out whether this log is relevant to any existing sessions. In the relevance check, there are mainly three atomic checks across Content-Check and Interaction-Check.

Phase 1: Determine whether there is a session-decision needed for this log

```
1: function IS_PAGE_BLANK(log)
2:     if log.url == Blank_Page_URL then
3:         return True
4:     else
5:         return False
6:     end if
7: end function
8: function IS_TAB_CONTENT_UPDATED(log)
9:     tabId = log.tabId
10:    if tab[tabId].url == logId.url then
11:        return False
12:    else
13:        return True
14:    end if
15: end function
16: procedure IS_DECISION_NEEDED_FOR_LOG(log)
17:    if (then)Is_Tab_Content_Updatedlog
18:        if (then)Is_Page_Blanklog
19:            log.sessionId = PendingSesssionId
20:            return False
21:        else
22:            return True
23:        end if
24:    else
25:        ADD_LOG_TO_SESSION_WITHOUT_DETAIL_UPDATE(log, session[tab.sessionId])
26:        Return ralse
27:    end if
28: end procedure
```

They were further explained in the following subsection. Instead of running the comparison between a log and all of the other logs in the session, the comparison between the log and a session-central log was ran, and the session-central log will be updated in Phase 3 when there is a new log add to that session.

Phase 2.1: Determine the interaction relevance between adjacent logs

```
1: procedure IS_IT_TRANSITIONALLY_RELEVANT_WITH_PREVIOUS_LOG(log)
2:     if log.HB is in Transitional_Connected_HB_Groups then
3:         log.sessionId = preLog.sessionId
4:         Return True
5:     else
6:         Return False
7:     end if
8: end procedure
```

Phase 3 the log would be assigned to either existing sessions or new sssion based on the decision from the first two phases. When assigning the log to one session, the detail of that session will get updated at the same time. The way to get the session-centred log was inspired by the k-mean[1]: when adding the new node, the central node will be recalculated.

[1]https://en.wikipedia.org/wiki/K-means_clustering

Phase 2.2: Determine the content relevance between log and sessions

```
numOfComparisons = 4
function GET_TITLE_SIMILARITY_SCORE(log, session)
    if logId.title != null then
        simPercent = CALCULATESIMILARITYPERCENT(log.title, session.mainTitle)
        Return simPercent
    else
        numOfComparisons–
        simPercent = 0
        Return simPercent
    end if
end function
procedure CHECK_RELEVANCE_BETWEEN_LOG_AND_SESSION(log)
    matchedSessionArray = []
    for {q = 0; q < aliveSessions.length; q++} do
        sessionId = aliveSession[q]
        score = GET_SIMILARITY_SCORE_BETWEEN_LOG_AND_SESSION(log, session[sessionId])
        if score > curBestScore AND score >= THRESHOLD then
            curBestSessionId = session   ▷ Future: This list may be a good indicator for some divided sessions should belong to one
            matchedSessionArray.push(session)
            curBestScore = score
        end if
    end for
    for {q = 0; q < aliveSessions.length; q++} do
        sessionId = aliveSession[q]
        score = GET_SIMILARITY_SCORE_BETWEEN_LOG_AND_SESSION(log, session[sessionId])
        if score >= T1 then
            log.sessionId = sessionId
            ADD_LOG_TO_SESSION_WITH_DETAIL_UPDATE(log, session)
            Break
        elseIf score > curBestScore AND score >= THRESHOLD2
            curBestSessionId = session   ▷ Future: This list may be a good indicator for some divided sessions should belong to one
            matchedSessionArray.push(session)
            curBestScore = score
        end if
    end for
    if curBestScore >= THRESHOLD then
        log.sessionId = curBestSessionId
        ADD_LOG_TO_SESSION_WITH_DETAIL_UPDATE(log, session)
    else                                                    ▷ If not relevant with any existing sessions
        CREATE_NEW_SESSION(log)
    end if
end procedure
function GET_DOMAIN_SIMILARITY_SCORE(log, session)
    if log.domain != null then
        if log.domain in [major search engine domains list] then
            simPercent = GET_QUERY_SIMILARITY_SCORE(log, session)
            Return simPercent
        else if then log.domain == session.mainDomain
            simPercent = 1
            Return simPercent
        else if then log.domain in session.domainArray
            simPercent = log.domain's frequency / session.domainArray.length
            Return simPercent
        end if
    else
        numOfComparisons–
        simPercent = 0
        Return simPercent
    end if
end function
function GET_SIMILARITY_SCORE_BETWEEN_LOG_AND_SESSION(log, session)
    score = (GET_TITLE_SIMILARITY_SCORE(log,session)
             + GET_QUERY_SIMILARITY_SCORE(log,session)
             + GET_TOP10KWS_SIMILARITY_SCORE(log,session)
             + GET_DOMAIN_SIMILARITY_SCORE(log,session) )
             / numOfComparisons
    Return score
end function
```

Phase 3: Add log to session

```
1: function ADD_LOG_TO_SESSION_WITH_DETAIL_UPDATE(log, session)
2:     logId = log.logId
3:     session.logArr = session.logArr.push(logId)
4:     session.titleArray = session.titleArray.push(log.title)
5:     session.mainTitle = GET_MOST_FREQUENT_TERMS_FROM_THEM(session.titleArray)
6:     session.top10KWsArray = session.top10KWsArray.push(log.top10KWs)
7:     session.mainTop10KWs = GET_MOST_FREQUENT_TERMS_FROM_THEM(session.top10KWsArray)
8:     session.queryArray = session.queryArray.push(log.query)
9:     session.mainQuery = GET_MOST_FREQUENT_TERMS_FROM_THEM(session.queryArray)
10:    session.domainArray = session.domainArray.push(log.domain)
11:    session.mainDomain = GET_MOST_FREQUENT_DOMAINS_FROM_THEM(session.domainArray)
12: end function
13: function ADD_LOG_TO_SESSION_WITHOUT_DETAIL_UPDATE(log, session)
14:    logId = log.logId
15:    session.logArr = session.logArr.push(logId)
16: end function
```

7.1.1 Atomic Decision Making

Phase 2 above is concentrating on caluclating the relevance between the new log and existing sessions in terms of user interaction and content similarity. In general, there are three types of atomic checks described by Equation List (Three Types of Checks) as the three diamonds box in the Phase 2 of Figure 7.2: 1) Transition Check (TC), 2) Content Check (CC), and 3) Quick Content Check (QCC).

$$
\begin{aligned}
Transition-Check(TC) &= [\max_{i=1\cdot\cdot n}[sim^{action}(l,S_i)],\Theta_a] \\
Content-Check(CC) &= [\max_{i=1\cdot\cdot n}[sim^{content}(l,S_i)],\Theta_c] \\
Quick-Content-Check(QCC) &= [[sim^{content}(l,S_i)],\Theta_c^{'}])
\end{aligned}
$$

(Three Types of Checks)

Transition-Check (TC) is inspired by the relationship between boundary and user interaction explored in Chapter 6, as well as some prior research about the relationship between document relevance and user interactions [38, 23, 35], where a particular type of interaction seems to be correlated with the boundary occurrence. The purpose of TC is to check the transition connection between logs. For example, for a log like *"When in Tab_A, user click*

on one of the link in Search Result Page (SERP) and open it in Tab_B", there is a transition connection between Tab_A and Tab_B.

Table 7.2 Interaction in TC

	Detail	**Pass**
CNT	Create new blank tab	-
CT	Close tab / window	-
SAB_BlankTab	Type query / URL in address bar in blank tab	No
SAB_NonBlankTab	Type query / URL in address bar in non-blank tab	No
TULC_BlankTab	Update tab via clicking hyperlink in blank tab	Yes
TULC_NonBlankTab	Update tab via clicking hyperlink in non-blank tab	Yes
OLNT_BlankTab[2]	Create a tab / window via clicking hyperlink in blank tab	Yes
OLNT_NonBlankTab (OLNT)	Create a tab / window via clicking hyperlink in non-blank tab	Yes
SW	Switch attention between tabs / windows	-
BOF	Click back or forward button	Yes
FEPNT	Create a tab via clicking hyperlink from an external application (e.g. MS Word)	No
LBIW	Switch attention off from the browser or go to Incognito window	-
SBFE	Switch your attention back from the external application	-

Content-Check (CC) determines the content relevance between log and sessions, is used to decide whether a log belongs to any existing sessions or it has to create a new session. The CC consists of four comparisons: title, domain, query, and top 10 keywords (the top 10 most frequent words from the webpage content). The objective of this check is to determine whether a log is content relevant to any existing sessions. The logs that succeed in this check are those that have good content relevance with an existing session.

Quick Content-Check (QCC) is a special case of a CC that only compare the log to one particular session instead of all existing sessions. The reason for having a QCC is to run a content validation for transition-connected logs. The objective of this check is to determine whether there is strong enough content similarity between a log and a particular session. The logs that succeeded in this check were those that had transition connection and good content relevance with an existing session.

[2]No OLNT_BlankTab had been found in this study this time

7.2 Five Versions Algorithm

7.2.1 A_1 (Baseline): Solo CC

A_1 is a solo version of the CC. It is aimed at determining the boundary based on checking the content similarity between the log and sessions. The content comparison consists of four parts: 1) title comparison, 2) domain comparison, 3) top 10 most frequent words of webpage comparison, and 4) query comparison (if it is a search log).

7.2.2 A_2: Solo TC

A_2 is a solo version of the TC. According to our findings in Table 6.6 about the probability of each interaction being a boundary, the interactions like BOF (5.6%), TULC_NonBlank (4.6%), and OLNT (26.7%) illustrate that the transition connection between pages and tabs is less likely to break a session. Meanwhile, the interactions like SAB_Blank, TULC_Blank, and SAB_NonBlank are more likely to lead to a boundary. These indicate that the transition chain could be a good indicator for determining session boundaries. The basic rule of A_2 is that when the transition continues (i.e. the interaction passes the TC check as shown in Table 7.2), it will remain in the same session (*Continue-Log* or *Resume-Log*); otherwise, it will create a new session (*Create-Log*).

7.2.3 A_3: Hybrid - Trusts A_2

As the CC and TC utilise different principles to cluster logs into sessions, A_3 is a combination of them. As illustrated in Figure 7.2, if the interaction breaks the transition connection, then CC would be applied. Interactions with transition-connection do not involve CC. Therefore, the core idea is to skip the CC if the interaction has strong enough transition-connection. In practical implementation, this benefits the speed and memory.

Table 7.3 Be Clustered to One Existing Session

	Condition
A_1	Succeed in CC
A_2	Succeed in TC
A_3	Succeed in TC **OR** Succeed in CC
A_4	(Succeed in TC **AND** Succeed in QCC) **OR** (Succeed in CC)
A_5	(Succeed in TC **AND** Succeed in QCC) **OR** (Succeed in CC)

7.2.4 A_4: Hybrid - Checks A_2

While A_3 presumes that certain sets of user interactions are more likely to stay within a session, A_4 performs a QCC on these sets of user interactions. Using the same similarity threshold (threshold=0.3), A_3 checks the similarity of the new log, but only with the current session rather than all existing sessions. If the similarity is below 0.3, it reverts to a CC. Essentially, if the interaction is expected to keep the user in the same session but the content of the new page does not appear to be similar to the current session, it performs a CC against other sessions.

7.2.5 A_5: Hybrid - Half-Trusts A_2

As another variant of A_3, A_5 uses the same procedure as A_4 but uses a lower threshold (threshold = 0.15) when performinging QCC. Essentially, A_5 is similar to A_4, but has more confidence in the TC, and thus only performs a CC if the page looks very different from the current session.

$$A1 = CC$$

$$A2 = TC$$

$$A3 = TC \vee CC$$

$$A4 = (TC \wedge QCC) \vee CC$$

$$A5 = (TC \wedge QCC) \vee CC$$

(Five Versions)

As shown in the Equation List (Five Versions) above:

- A_1 is a pure CC that compares the log with all existing sessions to determine whether or not the maximal similarity exceeds the threshold.
- A_2 is a pure TC that checks whether or not the log is transition-related to any previous logs (i.e., the tab was opened via clicking a link in Tab_A).
- A_3 is a combination of a TC and a CC. The log is a *Create-Log* if it fails both the TC and the CC. Sequentially, the TC preccedes the CC.
- A_4 is a variant of A_3 with an extra validation step (QCC) for the logs that pass the TC. Therefore, the relevant condition is either 'it is transition-connected to $Session_1$ and also satisfied the quick content-validation' or 'it is content-relevant to $Session_1$'.
- A_5 is similar to A_4 with the only difference that lower threshold (half) was applied in QCC.

7.2.6 Algorithmic Baselines and Alternatives

Establishing a baseline for this analysis was non-trivial. Beyond comparing the performance of these five algorithms to a range of fixed timeouts, A_1 was chosen as the baseline because of the tradition of content-similarity in task analysis history. The the comparison between A_1 and A_2 was run to test whether or not user interaction, compared to content-similarity, is a better solo indicator of session segmentation. Then the performance of combined versions (A_3, A_4, A_5) were evaluated. Since the objective of this study was to determine not only the most accurate but also the most efficient for practical real-time implementation, the efficiency was also considered. The speed of running each of these five algorithms to process the entire dataset was also shown in Table 7.4, and their accuracy was evaluated in the comparison to the performance of timeouts later in this chapter.

There were many alternatives to these five algorithms above, and the choice here was determined by a serie of factors including the speed-accuracy trade-off. As an alternative, TC could produce a probability estimate, and the algorithm could have always performed CC, even on low-probability interactions as shown in Table 6.6. The algorithm, however, runs in real-time in the browser and would be slowed down by running CC on low-probability pages. Furthermore, as described in the results section, the quicker and simpler check was

Table 7.4 Runtime in Milliseconds (ms): Efficiency

Threshold	A_1	A_2	A_3	A_4	A_5
0.1	2093	730	1499	1669	1693
0.2	3420	734	1141	1311	1321
0.3	5726	728	1268	1554	1500
0.4	9863	740	1304	2837	1693
0.5	15662	733	1362	4147	1844
0.6	17040	718	1445	4660	1852
0.7	20332	720	1466	6807	2792
0.8	27030	720	1458	14850	3269
0.9	36928	732	1549	28094	4270

found to dramatically improve the accuracy of the algorithm over the baseline (A_1). And the browser-level logged data had more detail and was more comprehensive than the log data applied in majority prior work in this are. Using a browser extension, many more interactions can be captured instead of just pages loaded (including submitted search query). Therefore, the time gaps in the logs were much more fine-grained (i.e., you can calculate how long have your attention been off from this tab with our log data). The result section below used the time gaps between page content changes (decision-needed logs) rather than between all logs in order to be more consistent with prior work. To match more recent work that has considered the similarities of queries [91, 95] and pages visited, A_1 was selected, which only uses a CC, as a reasonable baseline analog to prior work. The results below therefore evaluated the performance of the TC alone in A_2 (some prior work also considered behavioural aspects [38]), in combination with the CC (A_3, A_4, and A_5), and with content change-driven timeouts at different thresholds.

7.3 Design Concerns for Practical Implementation

7.3.1 Session Lifetime

One issue, that is different from other clustering problems, is session (cluster) lifetime. Determining whether or not a session is finished seems to be an impossible mission in this case, but keeping all of these sessions (clusters) alive is also far from realistic. While, in this stage, **in order to deliminate the effect from timegap in the results, the initial session timeout was set to be 1 week.**

7.3.2 Text Similarity Measurement

As shown in Figure 7.3, several types of measurements were considered to determine the text similarity, and the numbers represent the percentage of similarity between $String_1$ and $String_2$. In order to reduce the side effects from 1) user's mis-spelling in their queries (e.g., *hello* vs. *hallo*) and 2) random order of terms in a query or sentence (e.g., *javascript programming* vs. *programming javascript*), a variant of the longest common subsequence (LCS)[3] - iterative LCS rather than other popular similarity measures (e.g., cosine similarity (character-based) or Levenshtein distance / Jaccard distance (term-based)) was applied. Addionally, the stop word removal[4] was added before the similarity measure. After examining the performance of similarity thresholds between 0.1-0.9 shown in Table 7.5 and 0.3 was selected as the similarity-threshold.

Fig. 7.3 Pick the Right Similarity Measurement

Str1		hello	test	javascript	javascript programming	javascript hello world test programming	hello
Str2	hello world	hallo	tset	script	programming javascript	programming script	h
Levenshtein	9	80	50	60	9	15	17
NeedlemanWunch	50	90	75	80	52	53	50
Smith-Waterman	100	60	50	100	50	61	100
Smith-Waterman Gotoh	100	92	25	100	50	61	100
Smith-Waterman Gotoh Windowed Affine	100	92	25	100	50	61	100
Jaro	0	87	92	0	0	43	72
Jaro Winkler	0	88	93	0	0	43	75
QGrams Distance	0	57	33	60	71	43	18
Block Distance	0	0	0	0	100	29	0
Cosine Similarity	0	0	0	0	100	32	0
Euclidean Distance	0	0	0	0	100	15	0
Overlap Coefficient	0	0	0	0	100	50	0
original similar_text	0	80	75	75	50	38.6	33.3
variant_similar_text (In our extension)	0	80	100	75	95.5	59.6	0

[3] https://www.dyclassroom.com/dynamic-programming/longest-common-subsequence
[4] stop words library from Mirza Asif's github: https://github.com/mirzaasif/JS-StopWord

Text Similarity Measurement

```
1: procedure REMOVESTOPWORDS(str1, str2)
2:     str1 = str1.removeStopWords()
3:     str2 = str2.removeStopWords()
4:     Return str1, str2
5: end procedure
6: procedure FINDLONGESTCOMMONSTRING(str1, str2)
7:     if str1.length == 1 OR str2.length == 1 then
8:         threshold = 1
9:     else
10:        threshold = 0
11:    end if
12:    longestCommonStrLength = 0
13:    longestCommonStr = ''
14:    for {p = 0; p < str1.length; p++} do
15:        for {q = 0; q < str2.length; q++} do
16:            for {l = 0; l < Math.min(str1.length - p, str2.length - q); l++} do
17:                if str1[p+l] == str2[q+l] then
18:                    commonStrLength = commonStrLength + 1
19:                    if commonStrLength > longestCommonStrLength AND commonStrLength > threshold then
20:                        longestCommonStrLength = commonStrLength
21:                        longestCommonStr = str1.slice(p, p+l)
22:                    end if
23:                else
24:                    BREAK
25:                end if
26:            end for
27:        end for
28:    end for
29:    sum = longestCommonStrLength
30:    if sum > 0 then
31:        sub_str1 = str1.replace(longestCommonStr, '')
32:        sub_str2 = str2.replace(longestCommonStr, '')
33:        sum = sum + FINDLONGESTCOMMONSTRING(sub_str1, sub_str2)
34:    end if
35:    Return sum
36: end procedure
37: procedure CALCULATESIMILARITYPERCENT(str1, str2, sum)
38:    percent = 0
39:    str1_length = str1.length
40:    str2_length = str2.length
41:    if ( thenstr1_length + str2_length > 0)
42:        percent = sum * 2 / (str1_length + str2_length)
43:    end if
44:    Return percent
45: end procedure
```

Table 7.5 Accuracy of Using Similarity Thresholds 0.1 - 0.9 in Content Check

	Decision-need Logs F1-score	All Logs F1-score	Average
0.1	0.558	0.553	0.555
0.2	0.563	0.653	0.608
0.3	0.570	0.707	0.639
0.4	0.562	0.695	0.629
0.5	0.541	0.679	0.610
0.6	0.514	0.652	0.583
0.7	0.435	0.592	0.514
0.8	0.366	0.545	0.455
0.9	0.295	0.494	0.395

7.4 Algorithm Evaluation

7.4.1 Overview

Besides the five algorithms, different fixed timeout versions were also involved: 5 minutes (FT_5), 10 minutes (FT_{10}), 20 minutes (FT_{20}), 30 minutes (FT_{30}) and 60 minutes (FT_{60}). Table 7.6 characterised the sets of sessions generated by each algorithm. Among A_1-A_5, A_5 generated the closest number of sessions to the ground truth dataset: most of its measures, including **#Queries** per session, were the most similar to the GT dataset. Interestingly, FT_{10} also produced data that appeared to be similar to the GT dataset. These characteristics, however, do not tell us whether or not these produced sessions are similar to those in the GT dataset, and so Section 7.4.5 analysed their similarity in terms of whether or not they are coherently homogenous in comparison to the GT dataset. Table 7.6 also does not tell us whether the boundaries are in the right place, so Section 7.4.2 analysed the proportion of logs that are correctly classified as being boundaries (*Resume-Log* or *Create-Log*) or not (*Continue-Log*). However, this analysis included many logs those not go via Phase 2 (i.e., decision-making needed) in Figure 7.2, so Section 7.4.3 then focused on the decision-needed logs only to examine the proportion of decisions that were correctly made by each algorithm.

Table 7.6 Session Characteristic

	GT	A_1	A_2	A_3	A_4	A_5	FT_5	FT_{10}	FT_{20}	FT_{30}	FT_{60}
Session No.	227	278	553	137	271	194	362	229	160	127	94
#Queries	1.27	0.64	0.53	1.672	0.753	1.124	0.856	1.319	1.881	2.354	3.181
#Pages	10.4	6.478	4.586	16.08	7.347	11.026	9.727	13.983	18.831	22.937	29.66
#Domains	3.8	2.205	2.061	5.606	2.579	3.938	4.431	5.76	7.225	8.402	10.234
#Tabs	4.85	4.719	2.089	8.182	4.624	6.031	5.721	7.764	10.075	12.181	15.617
#Boundaries	11.7	14.061	6.056	16.891	12.096	13.804	1	1	1	1	1
#Interactions	50.13	40.36	20.32	81.912	41.41	57.845	36.227	57.266	81.963	103.26	139.511

The evaluation of performance of these five algorithms consisted of two parts: boundary (i.e., break-point of session) investigation and session content overlapping. In the boundary investigation, firstly, there was an overview of all logs. It then concentrated on the decision-needed logs. As shown in Figure 7.4, not all of the logs required the decision-making check run through it. As mentioned in the Phase 1 in Figure 7.2, the logs without any content update to the tab were categorised as the 'no decision needed' logs. For example, some of them

may just be caused by the tab-switching of some tabs whose session id had already been determined before: imagining that Tab_A was determined to be inside of $Session_1$ in Log_5, and then the user left this tab and switched to some other sessions before returning to Tab_A in Log_{20} without any content update in this tab. This would lead to a switch boundary (Resume-Log) without requiring any new decision-making via algorithm. In addition, for some non-boundaries (Continue-Log), there may be some decision needed where the algorithm decides to continue the session.

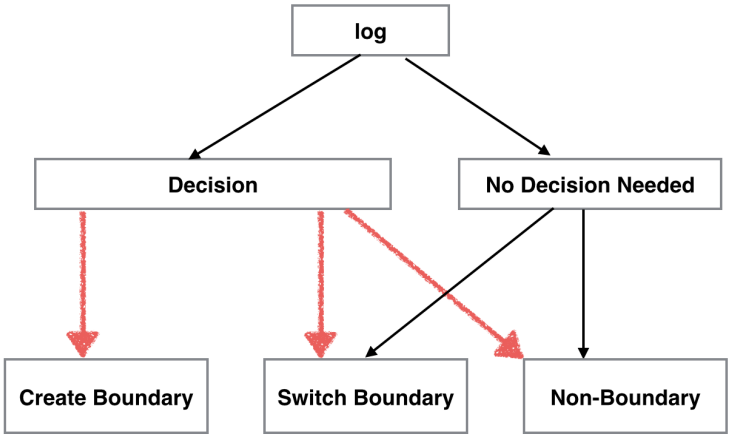


Fig. 7.4 The Relationship Between Logs and Decision-Needed Logs

7.4.2 Boundary Investigation: All Logs

This section mainly focused on the detection of boundaries for all logs including both decision-needed logs and no-decision-needed logs. Therefore, the boundary detection problem was converted into a binary classification problem featuring non-boundary (Continue-Logs) and boundary (including both Create-Log and Resume-Log). It was categorised as Confusion Matrix in Table 7.7 below. The predicted boundary and predicted non-boundary were the result from algorithm, and the actual boundary and actual non-boundary were from the GT dataset.

Precision, recall, and F1 measure were calculated to evaluate these algorithms' performance. Precision (TP / (TP + FP)) is a measure of a classifier's exactness, which describes

Table 7.7 Binary Classification Confusion Matrix - All Logs

	Predicted Boundary (in algorithm)	**Predicted Non-Boundary (in algorithm)**
Actual Boundary (in GT)	TP (True Positive)	FN (False Negative)
Actual Non-Boundary (in GT)	FP (False Positive)	TN (True Negative)

Table 7.8 Binary Classification Boundary Investigation Performance - All Logs

	Precision	**Recall**	**F1-Measure**
A_1	0.697	0.883	0.779
A_2	0.593	0.876	0.707
A_3	0.841	0.736	0.785
A_4	0.646	0.801	0.715
A_5	0.759	0.769	0.764
FT_5	0.5	0.052	0.094
FT_{10}	0.581	0.038	0.071
FT_{20}	0.656	0.03	0.057
FT_{30}	0.677	0.025	0.047
FT_{60}	0.766	0.021	0.04

'how many of the predicted boundaries are actual boundaries'. Higher precision indicates a higher percentage of TP (the logs that are actual boundaries and predicted boundaries) and a lower percentage of FP (the logs that are predicted boundaries but actual non-boundaries). Recall (TP / (TP + FN)) is a measure of a classifier's completeness, which describes 'how good the algorithm is at detecting the actual boundaries out'. Higher recall indicates a higher percentage of TP and a lower percent of FN (the logs that are predicted non-boundary but actual boundary). The F1 score (2 * Precision * Recall / (Precision + Recall)) conveys the balance between precision and recall.

Precision A_3 did the best in that more than 84% of its predicted boundaries were actual boundaries. A_5 also did well with more three quaters of its predicted boundaries being actual. In contrast, A_2 generated more unnecessary boundaries (i.e., higher False Positive) than the other four, and only 59.3% of its predicted boundaries were actual boundaries. Since A_2 was the solo transition check, this illustrates that nearly half of the transition-break logs are not boundaries, and also many sessions had transition break inside them. The 30% unnecessary boundaries (i.e., False Positive) in A_1 indicate that many sessions also have content shifts inside them. The best performance from A_3 illustrates that applying a CC on transition-break log filters out some unnecessary boundaries that are content-relevant but without a transition-connection with its last log. Even though A_5 did not perform as well

as A_3, it did better than A_4 and would probably work better with more trials with different thresholds tested. Nevertheless, these demonstrate that a combination (like A_3, A_4, and A_5) of content shifts and transition-break can be a better indicator than solo version.

Recall A_1 performed best and found 88.3% actual boundaries. This is not surprising, because A_1 is the most sensitive version among these five. The others also did well with A_2 finding 87.6% actual boundaries; A_4, 80.1%; A_5, 76.9%; and A_3, 73.6%. The progress from A_3 to A_4 or A_5 indicates that applying the QCC to transition-connected logs helps to find some boundaries that are transition-connected but without a good content-relevance to its last log. Nevertheless, the progress was not very significant. This illustrates that there were a group of boundaries those were transition-connected but with low content-relevance to its last log, and the percentage of them was not high. Overall, around 88.3% of the boundaries happened via content shift, around 87.6% of the boundaries happen via transition-break, and a small group of boundaries were transition-connected but content shift.

F1 Measure Taking both completeness and exactness into consideration, F1 measure was applied. A_3 performed the best (F1 measure = 0.785), as it did the best in precision with more than 84% of its predicted boundaries being actual boundaries and also did well in recall with only missing 26% of actual boundaries. A_5 also did well with higher recall than A_3 and acceptable precision.

Compared to Fixed Timeout The precisions (between 0.5 and 0.766) of the fixed timeout group (e.g., FT_5 = fixed timeout after five minutes of inactivity) illustrate that some boundaries had time gaps, but many time gaps also occurred within sessions. While, the extremely low recall of them illustrates that only a small portion of session boundaries have more than a five-minutes time gap. This further confirmes that fixed timeout is not appropriate for determining sessions.

7.4.3 Boundary Investigation: All Decision-Needed Logs

As shown in Figure 7.4, not all of the logs need the decision-making. In this section, only the decision-needed logs were concentrated on. The same strategy was followed as the previous analysis, and it was regarded as a binary classification problem of distinguishing between boundaries (i.e., *Create-Log* and *Resume-Log*) and non-boundaries (i.e., *Continue-Log*). The performance overview is presented in Table 7.9.

Table 7.9 Binary Classification Boundary Investigation Performance: Decision Needed Logs

Versions	TP	FP	FN	TN	Precision	Recall	F1-measure
A_1	594	58	837	3728	0.415	0.911	0.570
A_2	579	230	73	4335	0.716	0.888	0.793
A_3	503	110	149	4455	0.821	0.771	0.795
A_4	548	524	104	4041	0.511	0.840	0.636
A_5	533	250	119	4315	0.681	0.817	0.743

Then, instead of running the multi-groups comparisons directly, five pair-to-pair comparisons below were run to concantrate on some particular questions and the result for them were presented in Table 7.10:

- A_1 **vs.** A_2: *Which solo version is more useful: user interaction (A_2) or content similarity (A_1)?*
- A_1 **vs.** A_3: *The difference between A_1 and A_3 is the check run on the transition-connected logs (which is TC), so can this comparison be used to indicate that whether a transition-connected log is less likely to be a boundary?*
- A_2 **vs.** A_3: *Does the CC for transition-break logs benefit the accuracy or not?*
- A_4 **vs.** A_5: *Should a lower content relevance (QCC with lower threshold) validation be given to transition-connected logs?*
- **(A_4, A_5) vs.** A_3: *Does the QCC benefit the check on transition-connected logs or not?*

Table 7.10 Binary Classification Boundary Investigation Performance Summary - Decision-Needed Logs

Comparison	Precision	Recall	F1-Measure
A_1 **VS.** A_2	A_2	A_1	A_2
A_1 **VS.** A_3	A_3	A_1	A_3
A_2 **VS.** A_3	A_3	A_2	A_3
A_4 **VS.** A_5	A_5	A_4	A_5
(A_4, A_5) **VS.** A_3	A_3	A_4, A_5	A_3

A_1 **vs.** A_2 The precision and F1 measure in A_2 were much better than those in A_1. The recall of A_2 was also acceptable, even if it was not as high as that in A_1. In general, in the solo version comparison, user interaction (TC) is a better indicator compared to content check (CC).

A_1 **vs.** A_3 The precision and F1 measure in A_3 were much better than those in A_1. These indicates that transition-connected logs are more likely to belong to the same session. However, the lower recall of A_3 also illustrates that there is a small group of actual boundaries that have transition-connection with last log, which against the principle of TC, and it helps to explain the reason for having A_4 and A_5).

A_2 **vs.** A_3 The F1 measure of A_2 and A_3 was similar. The difference in precision ($A_3 > A_2$) and recall ($A_2 > A_3$) indicate that the content check for transition break logs can benefit the precision that helps to filter out some unnecessary boundaries, but it also causes some side effects such as missing some actual boundaries.

A_4 **vs.** A_5 The precision and F1 measure in A_5 were much better than those in A_4, but A_5 has a similar (or slightly lower) recall to A_4. This indicates that a lower content relevance requirement given to transition-connected logs can help to improve the performance, especially in terms of filtering out some unnecessary boundaries. However, this also led to missing some boundaries, because there were some boundaries that are transition-connected but also had acceptable content relevance with other logs even through they actually belong to two different sessions.

(A_4**,** A_5**) vs.** A_3 Based on the observation from Table 7.9, a QCC on transition-connected log did not bring significant benefit, but it helped to 'find out more actual boundaries', because the recall of (A_4, A_5) was higher than that of A_3. The weaker performance of QCC illustrates that there is a small group of boundaries missed in the transition-connected logs group. However, adding the extra content relevance validation did not significantly benefit

the performance, which indicates that for most transition-connected logs, the user interaction dominates the content relevance.

7.4.4 Boundary Investigation - All Decision-Needed Logs (Three-Classes Classification)

As shown in Figure 7.4, actual boundary detection is not as easy as the binary classification problem suggests by simply differentiating between boundary and non-boundary. Therefore, in the following analysis, this problem was treat as a multi-class (*Create-Log*, *Resume-Log*, and *Continue-Log*) classification problem, as described in Table 7.11 below.

Table 7.11 Three-Classes Confusion Matrix: Decision-Needed Logs

	Create in Algorithm	**Resume in Algorithm**	**Continue in Algorithm**
Create in GT	$TP_{-Create}$	E_3	E_5
Resume in GT	E_1	$TP_{-Resume}$	E_6
Continue in GT	E_2	E_4	$TP_{-Continue}$

Table 7.12 Three-Classes Detail

	TP	**FN**	**FP**	**TN**
Create	$TP_{-Create}$	$E_3 + E_5$	$E_1 + E_2$	$TP_{-Resume} + E_6 + E_4 + TP_{-Continue}$
Resume	$TP_{-Resume}$	$E_1 + E_6$	$E_3 + E_4$	$TP_{-Create} + E_5 + E_2 + TP_{-Continue}$
Continue	$TP_{-Continue}$	$E_2 + E_4$	$E_5 + E_6$	$TP_{-Create} + E_3 + E_1 + TP_{-Resume}$

Table 7.13 A_1: Three-Classes

	Create-Log in Algorithm	**Resume-Log in Algorithm**	**Continue-Log in Algorithm**
Create-Log in GT	87	94	27
Resume-Log in GT	48	365	31
Continue-Log in GT	143	694	3728

Table 7.14 A_2: Three-Classes

	Create-Log in Algorithm	**Resume-Log in Algorithm**	**Continue-Log in Algorithm**
Create-Log in GT	184	4	20
Resume-Log in GT	173	218	53
Continue-Log in GT	196	34	4335

In these Interaction Distribution Figures below, the left chart represents the *Create-Log*; the middle chart represents the *Resume-Log*; and the right chart represents the *Continue-Log*. Blue color means that when the decision between GT and A_X matched. The other two colors

Table 7.15 A_3: Three-Classes

	Create-Log in Algorithm	**Resume-Log in Algorithm**	**Continue-Log in Algorithm**
Create-Log in GT	73	83	52
Resume-Log in GT	28	319	97
Continue-Log in GT	36	74	4455

Table 7.16 A_4: Three-Classes

	Create-Log in Algorithm	**Resume-Log in Algorithm**	**Continue-Log in Algorithm**
Create-Log in GT	81	89	38
Resume-Log in GT	46	332	66
Continue-Log in GT	144	380	4041

Table 7.17 A_5: Three-Classes

	Create-Log in Algorithm	**Resume-Log in Algorithm**	**Continue-Log in Algorithm**
Create-Log in GT	79	88	41
Resume-Log in GT	38	328	78
Continue-Log in GT	77	173	4315

Table 7.18 Three-Classes: Create-Log

Versions	**Precision**	**Recall**	**F1-measure**
A_1	0.313	0.418	0.358
A_2	0.333	0.885	0.484
A_3	0.533	0.351	0.423
A_4	0.299	0.389	0.338
A_5	0.407	0.380	0.393

Table 7.19 Three-Classes: Resume-Log

Versions	**Precision**	**Recall**	**F1-measure**
A_1	0.317	0.822	0.457
A_2	0.852	0.491	0.623
A_3	0.670	0.718	0.693
A_4	0.414	0.748	0.533
A_5	0.557	0.739	0.635

Table 7.20 Three-Classes: Continue-Log

Versions	**Precision**	**Recall**	**F1-measure**
A_1	0.985	0.817	0.893
A_2	0.983	0.950	0.966
A_3	0.968	0.976	0.972
A_4	0.975	0.885	0.928
A_5	0.973	0.945	0.959

(i.e., green and orange), by contrast, shows the mismatch between them. It provides the performance detail of each user interaction for each log type.

Question: A_1 VS. A_2 Which solo version is more useful: user interaction (A_2) or content similarity (A_1)?

Fig. 7.5 A_1 vs. A_2: A_1 Interaction Distribution

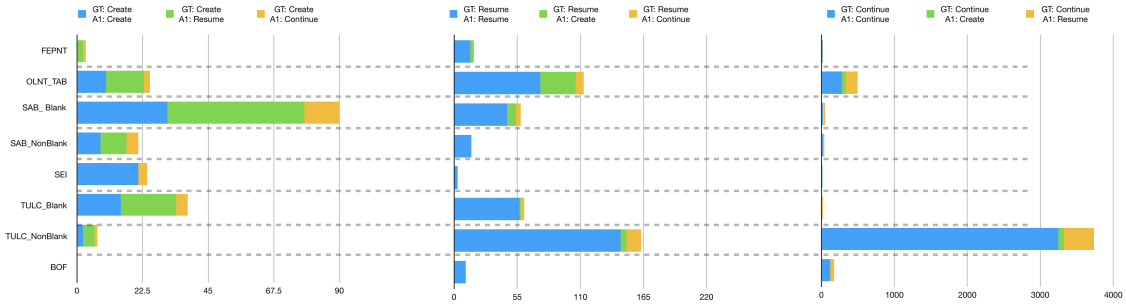
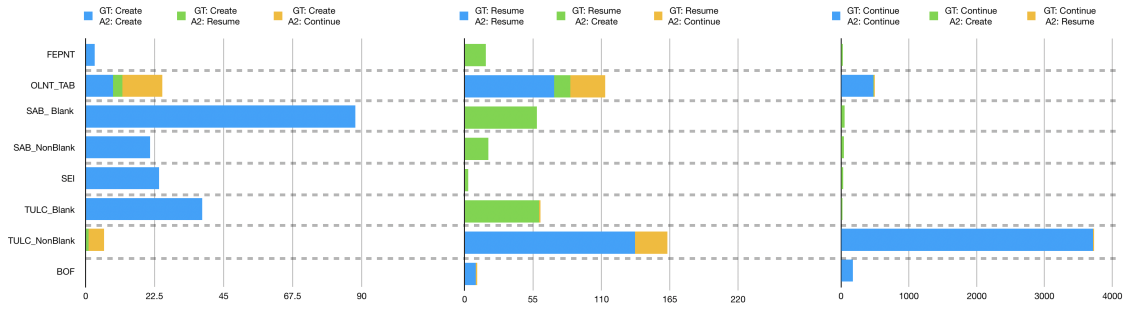


Fig. 7.6 A_1 vs. A_2: A_2 Interaction Distribution



The high recall of A_2 (0.885) in *Create-Log* detection illustrates the better performance of user interaction. The left chart on both Figure 7.5 and Figure 7.6 shows the performance detail in interaction level. A_2 succeed in these particular types of user interactions (e.g., SAB_NonBlank and SAB_Blank). Meanwhile, we find their performance difference on precision from the green bar in middle and right chart on both figures above. Even their precision looks similar via reading the number (A_1: 0.313 and A_2:0.333 in *Create-Log*), A_2 misrecognised more of SABs, TULC_Blank *Resume-Log* as *Create-Log* as shown in Figure 7.6. In contrast to the performance on *Create-Log*, A_1 with the longer blue bar achieved a higher recall (0.822 in *Resume-Log*) as shown in the middle chart. In addition, some particular types of user interactions (e.g., TULC_NonBlank and OLNT_Tab) in A_1 were more likely to be *Resume-Log*. Meanwhile, when determining the *Resume-Log*, A_1 and A_3 were found to performance differently on precision as shown in the green bar in left and right chart on both figures: A_2 had a much higher precision (A_2: 0.852; A_1: 0.317 in *Resume-Log*), and A_1 misrecognised a large percentage of SAB (i.e., SAB_NonBlank and SAB_Blank), TULC_Blank, and OLNT_TAB *Resume-Log* as *Create-Log* While, both of

them did very well in detecting *Continue-Log*, but the extremely good performance of A_2, especially in recall (see right chart in the two figures above), further confirms, the reliability of user interaction in this case.

Conclusion: the user interaction is a better solo indicator than content similarity. User interaction (A_2) did much better than content relevance (A_1) in *Create-Log* (90% of *Create-Logs* happen with some particular types of user interactions like SABs), were acceptable in *Resume-Log* (a large group of *Resume-Log* is more likely to be related to user interaction rather than content-relevance differences), and were also better in *Continue-Log*. However, the confusion between *Create-Log* and *Resume-Log* exists in both A_1 and A_2; therefore, these two solo versions are good for detecting the session break-point (boundary), but they are not perfectly ready for resuming back to the session via merging the discrete session-pieces. The individual differences in session scale (like the topic, task, and phase factors mentioned in the taxonomy) could also bring some barriers.

Question: A_1 vs. A_3 Is a transition-connected log less likely to be a boundary?

Fig. 7.7 A_1 vs. A_3: A_1 Interaction Distribution

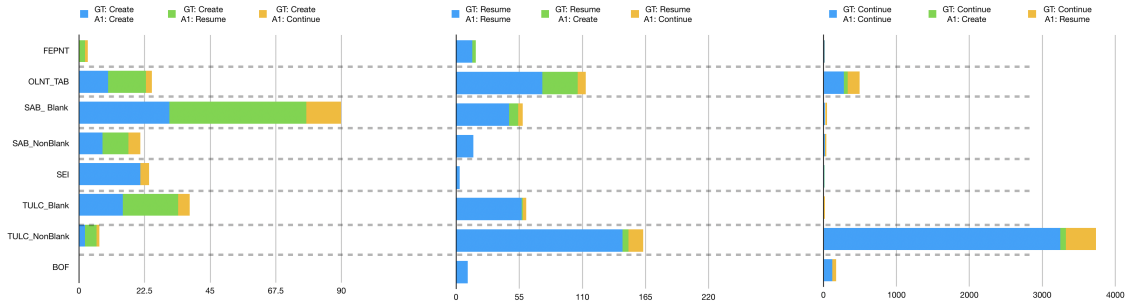
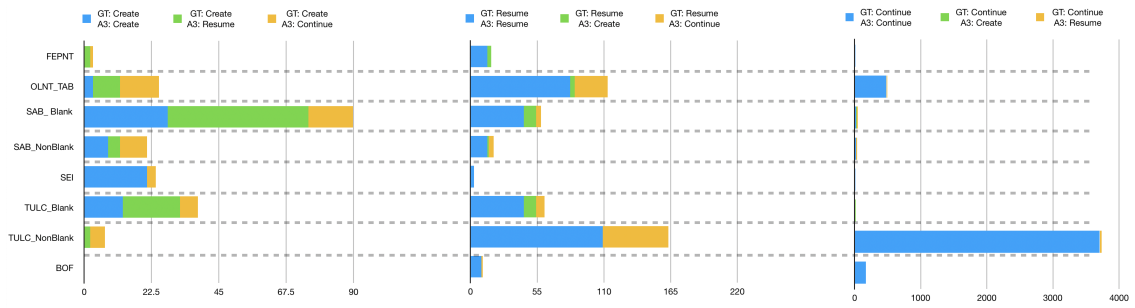


Fig. 7.8 A_1 vs. A_3: A_3 Interaction Distribution



Learnt from the left chart on both Figure 7.7 and Figure 7.8, A_1 and A_3's recall were similar (A_1: 0.418; A_3: 0.351), and they were both much lower than that of A_2 (0.885). This indicates that there is a group of *Create-Logs* that are content-relevant to other sessions but not transition-connected with them. As a result, it proves the possibility that the user interaction could dominate the content relevance sometimes. Meanwhile, there is not much difference between A_1 and A_3's precisions from the green bar in middle and right chart on both figures above, but the TULC_NonBlank. The better performance in determining *Continue-Log* (i.e., the x axis value scale in right chart is much larger) of A_3 led to a better precision (0.533) than that in A_1 (0.313). However, some improvement may be needed in recognising some TULC_NonBlank *Resume-Log*. Both A_1 and A_3 have some ambiguity in telling the difference between *Create-Log* and *Resume-Log*. The higher precision of A_3 when determining *Resume-Log* (A_3: 0.670, A_1: 0.317) indicates that the transition-check helped to filter out some FP for *Resume-Log* detection. Both A_1 and A_3 did well in the *Resume-Log* recall, especially A_1. The higher recall *Continue-Log* in A_3 (A_3: 0.976, A_1: 0.817) indicates that there is a group of *Continue-Log* that are transition-connected with their sessions but without obvious content-relevance. Then learnt from the right chart in Figrue 7.7 and Figure 7.8, A_1 performed a bit weaker than A_3 in OLNT_Tab and TULC_NonBlank when determining *Continue-Log*.

Conclusion: The high confidence on transition-connected logs is helpful. It can help to merge some small discrete session-pieces generated by content similarity checks. There is a large group of *Continue-Log* that do not have enough content relevance but do have transition connections with other logs like TULC_NonBlank. The high confidence of transition-connected logs benefits the performance but also has a problem with missing more *Resume-Logs*. This proves that content relevance is not a 'must' for connecting logs to one session and the users interactions can sometimes dominate. Meanwhile, both A_1 and A_3 have the same ambiguity as mentioned in the comparison above: distinguishing between *Create-Log* and *Resume-Log*.

Question: A_2 VS. A_3 Does the CC for transition break logs benefit the accuracy or not?

Fig. 7.9 A_2 vs. A_3: A_2 Interaction Distribution

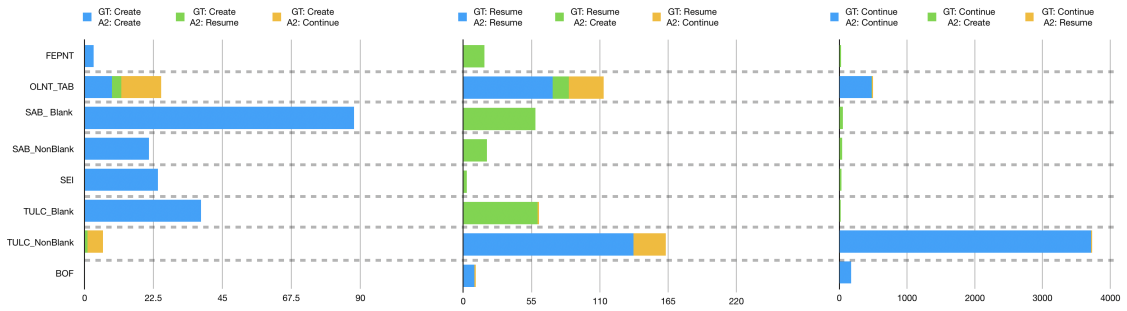
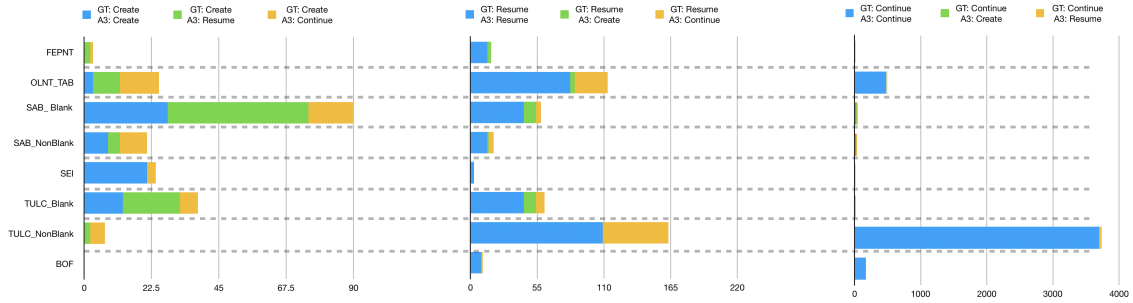


Fig. 7.10 A_2 vs. A_3: A_3 Interaction Distribution



This comparison had a similar result as the comparison between A_1 and A_2. For example, by observing the left chart on both Figure 7.10 and Figure 7.9, A_2 detected out more actual *Create-Logs* (i.e., recall - A_2: 0.885, A_3: 0.351), especially for these interactions: OLNT_TAB and SABs. While, A_3 found fewer *Create-Log*s than A_1 in general, but it generated much fewer unnecessary *Create-Logs* than A_2. The higher recall of A_3 (A_3: 0.718, A_2: 0.491) when determining the *Resume-Log* indicates that nearly three quarters of *Resume-Log*s comes with content shift, while the transition break rule can only recognise half of them out. The precision in A_2 (0.852) was higher than that in A_3 (0.670) when determining the *Resume-Log*.

Conclusion: The content check for transition break logs benefits the detection of Resume-Log, but also creates some side effects (i.e., telling the difference between Create-Log and Resume-Log) in Create-Log detection. There is a group of *Create-Log* that are content relevant but transition break with other sessions. Even A_3 found fewer *Create-Log*, but it also generated much fewer unnecessary *Create-Logs* than A_2 did. Therefore, user interactions can be a good indicator for *Create-Logs*, because most of the *Create-Log* come with some particular type of user interaction, but there should be further validation for these

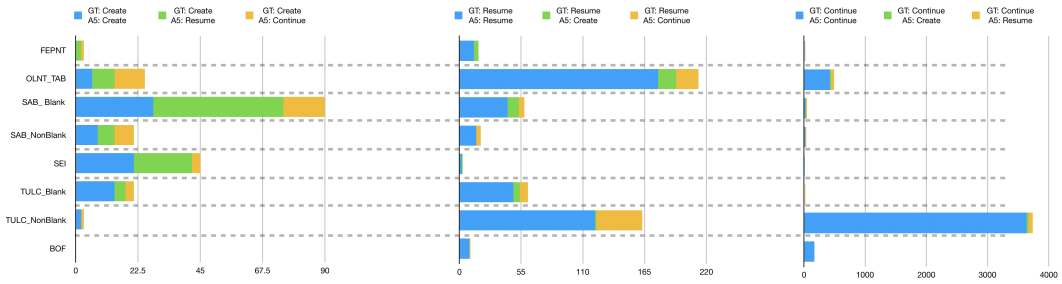
transition break logs before categorising them as *Create-Log*. However, applying the content-check here did not really benefit the detection of *Resume-Log*, especially when distinguishing between *Create-Log* and *Resume-Log*. For *Continue-Log*, this extra CC of transition break logs can help to reduce the mistake of misrecognising *Continue-Log* as *Create-Log* but can also lead to some mistakes of misrecognising *Continue-Log* as *Resume-Log*. However, the improvement from this extra check (i.e., the CC on transition break logs) seems to bring more benefit than the side effects.

Question: A_4 vs. A_5 Should a lower requirement of content relevance (QCC with lower threshold) validation be given to transition-connected logs?

Fig. 7.11 A_4 vs. A_5: A_4 Interaction Distribution



Fig. 7.12 A_4 vs. A_5: A_5 Interaction Distribution



The precision of A_5 (0.407) when determining *Create-Log* was higher than that of A_4 (0.299). This indicates that a lower content relevance requirement for transition-connected logs can benefit the precision of *Create-Log* detection. However, the similar weak performance of recall for both A_4 and A_5, when compared to A_2, further proves the previous finding: 'there is a group of *Create-Log* that are transition breaking but content relevant to other

sessions'. The precision of A_5 (0.557) when determining *Resume-Log* was higher than that of A_4 (0.414). Learnt from the middle chart in Figure 7.11 and Figure 7.12, both A_4 and A_5 achieved a good performance in recall, and A_5 performed better in OLNT_TAB. It indicates that lower content-relevance requirement for transition-connected logs in QCC could help to reduce the error of misrecognising *Resume-Log* as *Continue-Log*. When determination of *Continue-Log*, the precision of both A_4 (0.975) and A_5 (0.973) were strong. The big number in TP in A_5 (4315, see Table 7.17) than in A_4 (4041 see Table 7.16) and also the higher recall of A_5 indicates that lowering the content relevance requirement for transition-connected logs benefits the *Continue-Log* detection.

Conclusion: a lower requirement of content-relevance validation is helpful for transition-connected logs, especially in the detection of Create-Log and Resume-Log. Lower content relevance requirement for transition-connected logs in QCC can help to reduce the error of misrecognising *Resume-Log* to be *Continue-Log*. It also benefits *Continue-Log* detection.

Question: (A_4, A_5) vs. A_3 Does the QCC benefit the check on transition-connected logs or not?

When determining *Create-Log*, the precision of A_3 (0.533) was higher than that of both A_4 (0.299) and A_5 (0.407). This indicates that the QCC does not really provide obvious benefit, meanwhile, the TP numbers are 73 (A_3), 81 (A_4), and 79 (A_5) from Table 7.15, Table 7.16, and Table 7.17. This indicates that the QCC validation for transition-connected logs might help to detect out more real *Create-Log*. Learnt from the three Figures above, the recall of A_3 was a bit lower than that of A_4 and A_5 (especially A_5), and it indicates that QCC could benefit determining more *Resume-Log*. While, the recall of A_4 and A_5 was lower than A_3 when determining *Continue-Log*. These indicate that QCC could be beneficial but requires some further adjustment.

Conclusion: the QCC does not really provide obvious benefit for the detection, but it has created a slight improvement in detecting more Create-Log and Resume-Log. The non-obvious improvement indicates that the transition-connected characteristic may

Fig. 7.13 (A_4,A_5) vs. A_3: A_3 Interaction Distribution

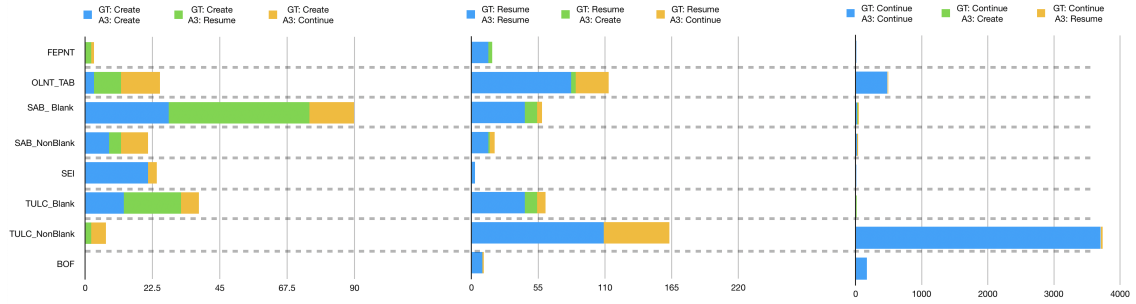


Fig. 7.14 (A_4,A_5) vs. A_3: A_4 Interaction Distribution

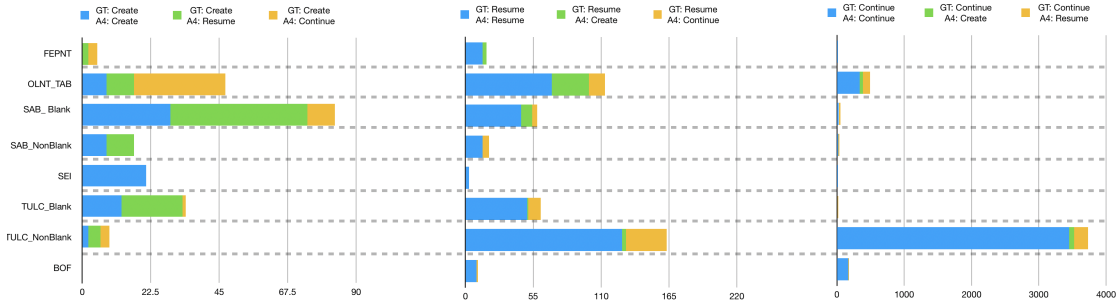
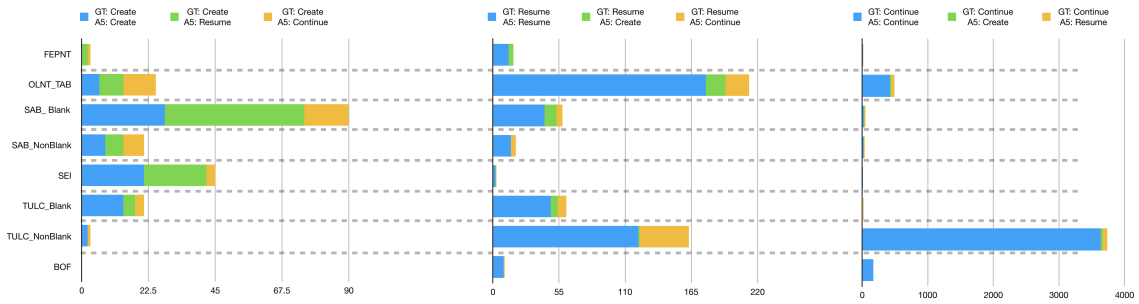


Fig. 7.15 (A_4,A_5) vs. A_3: A_5 Interaction Distribution



sometime dominate the content-relevance that was mentioned in previous comparisons, especially for *Continue-Log*. It could help to reduce the error rate of mis-recognising the *Create-Log* as *Continue-Log*, but it also increases the ambiguity between *Create-Log* and *Resume-Log*, as well as the ambiguity between *Resume-Log* and *Continue-Log*. While, since both *Create-Log* and *Resume-Log* are boundaries, QCC could benefit the boundary detection problem (binary classification) at some point. This does not mean that QCC is useless; actually, it has made some improvements.

7.4.5 Session Content Overlapping

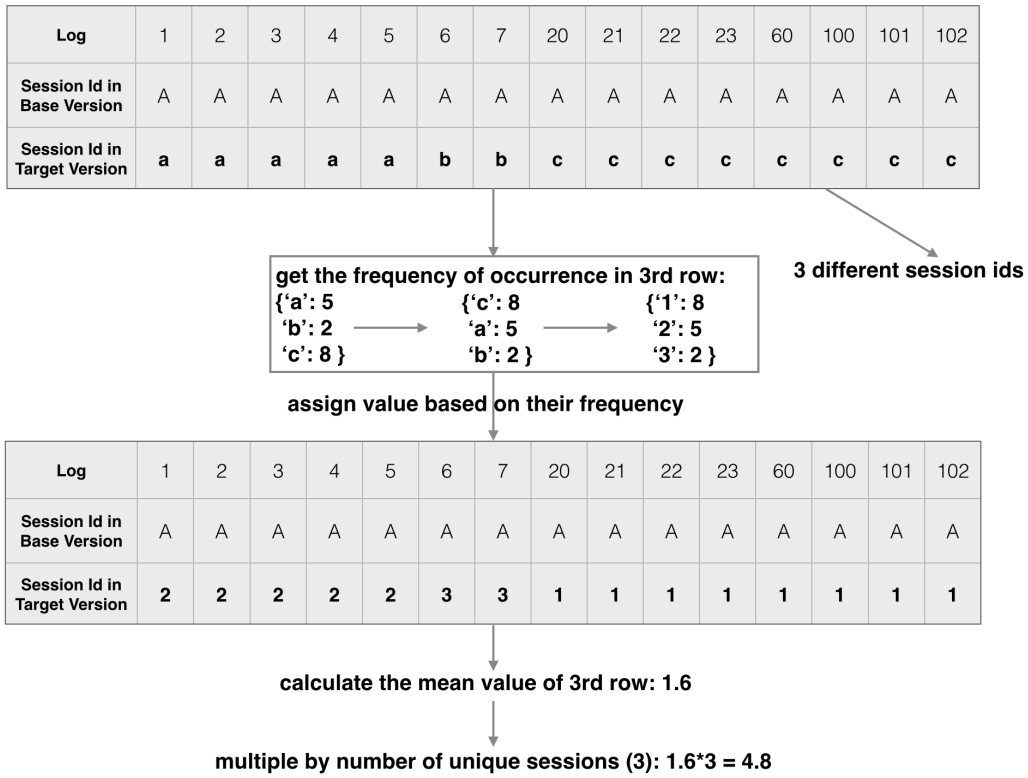


Fig. 7.16 Measure the Session Overlap Between Two Versions of Datasets

Besides the investigation of boundary above, the discrete session-pieces were also explored. Ideally, every log entry within a session should be from the same session in the GT dataset, and visa versa. Determining the correctness of session ID from different algorithms is non-trivial, so rather than exploring the correctness of session ID for each log, determining the log overlapping between two sessions can be a good indicator of session content similarity. Inspired by the Jaccard similarity coefficient [85], the session overlap was measured using the following steps shown in Figure 7.16:

- 1) Take all the logs from one session produced algorithmically (base version in Figure 7.16),
- 2) Identify the Session ID in the GT dataset (target version in Figure 7.16) for each of these logs (logs come from three sessions in Figure 7.16,
- 3) Count the number of logs from each of these target sessions,
- 4) Order these target sessions by decreasing frequency (e.g., c, a, b in Figure 7.16)

- 5) Assign an increasing integer to each Target session (‘1’ to each log from the most frequent Session c; ‘2’ to logs from the next most frequent Session a, etc)
- 6) Calculate the mean value of indexed scores, and
- 7) Multiply the mean score by the number of unique target session IDs.

The process in Figure 7.16 means that a session that is made up entirely of logs from one session in the GT dataset will get a score of 1; this algorithmically generated session would have high homogeneity. If an algorithmically generated session includes logs from more than one session (decreased homogeneity), the mean score increases from 1. Furthermore, we considered sessions made up from an increasing number of sessions to be worse than the same number of logs coming from fewer sessions. By multiplying the mean score by the number of sessions in Step 7, therefore, we guarenteed that mean index scores from a larger number of sessions received a higher final score than a mean index score produced by a lower number of sessions.

This process, however, tells that whether or not ‘different GT sessions have been merged into one algorithmic session’; the GT data was applied as the target (see Part 1 in Table 7.21. To understand whether ‘one GT session has been divided into multiple algorithmic sessions’, Part 2 reapplied the same process by swapping datasets, using GT data as the base version and the algorithmically generated sessions as the target version. As a result, as well as giving the mean score from each part of the process, Table 7.21 presents the combined score; 100% overlap would give a score of 1 in each stage, and thus a score of 2 in the sum column. Therefore, two different measurements were run here:

- Part 1 - GT data was used as the target version, and it helped to investigate whether some different actual sessions had been put into one session by these algorithms. A larger value indicates a worse performance and more missing boundaries.

- Part 2 - GT data was used as the base version, and it helped to investigate whether the actual sessions were divided into multiple pieces in these algorithms. A larger value indicates a worse performance and more unnecessary boundaries.

In Part 1, in the group of fixed timeout, the result value increases with the timeout increment, especially after 20, 30, and 60 minutes of inactivity. A higher value indicates a

Table 7.21 Session Content Overlapping Performance

	Part 1: GT as Target	**Part 2: GT as Base**	**Combined**
A_1	4.916	5.835	10.751
A_2	2.132	6.755	8.887
A_3	5.858	2.865	8.45
A_4	4.501	5.116	9.617
A_5	4.816	3.65	8.466
FT5	4.866	11.329	16.195
FT10	5.95	6.838	12.788
FT20	7.162	5.071	12.233
FT30	8.434	4.833	13.267
FT60	9.972	3.365	13.337

Table 7.22 Part 1: GT as Target

Comparison	**which one is better?**
A_1 **VS.** A_2	A_2
A_1 **VS.** A_3	A_1
A_2 **VS.** A_3	A_2
A_4 **VS.** A_5	A_4
(A_4, A_5) **VS.** A_3	A_4,A_5

worse performance and that more real boundaries were missed. However, Table 7.6 shows that these longer fixed timeout algorithms (i.e., FT_{20}, FT_{30}, and FT_{60}) produced a smaller number of sessions than the GT dataset contained, and so this result could be expected. The 5 pair to pair comparisons shown in Table 7.22 delivered some similar finding as boundary investigation in last section. For example, the better performance of A_2, when compared to A_1, double confirms that the user interaction is a better indicator than content relevance for determining more actual boundaries, and the comparison between (A_4, A_5) and A_3 indicates that having an extra QCC can help to pinpoint more actual boundaries.

Table 7.23 Part 2: GT as Base

Comparison	**which one is better?**
A_1 **VS.** A_2	A_1
A_1 **VS.** A_3	A_3
A_2 **VS.** A_3	A_3
A_4 **VS.** A_5	A_4
(A_4, A_5) **VS.** A_3	A_4,A_5

The result of Part 2 was a bit different from that of Part 1, and the higher value indicates more unnecessary boundaries. In the fixed timeout group, longer timeouts led to a better performance. This is because that 'longer timegap tolerance would lead to fewer session pieces'. For example, if the timegap is large enough that all of the logs from one participant get divided into one session, then no matter which session from the GT data is used as the

base, the target session will always have a full overlap with it. The better performance of A_1 in the comparison between A_1 and A_2 indicates that CC is more helpful in avoiding more unnecessary boundaries than user interaction. The comparison between $A_1$and A_3 indicates that there is a group of boundaries passed the TC, but the small difference between them illustrates that there are not many boundaries like this. The comparison between (A_4, A_5) and A_3 indicates that having a extra QCC can help to pinpoint more real boundaries.

In summary, the combination versions like A_3 and A_5 did better than the other algorithms. In the comparison between the solo versions, the TC did better than the CC in both Part 1 and 2. Its performance was nearly similar to those of A_3 and A_5 (i.e., the most successful algorithms). The differences between these three good performances (A_2, A_3, and A_5) indicates that they may be suitable for different occasions. The better performance in Part 1 indicates that more boundaries have been determined, while the better performance in Part 2 indicates that there are fewer unnecessary boundaries have been generated. A_2, which performed the best in Part 1, but not very well in Part 2, would be able to cover more boundaries but would also generate many unnecessary boundaries, making it as a more sensitive algorithm and leading to more session pieces. A_3 and A_5, which were better in Part 2 and as good as A_2 in Part 1, could reduce the rate of unnecessary boundaries, but with some cost in missing some real boundaries. Therefore, A_3 and A_5 may be able to generate a version of sessions that is a bit closer in scale to the GT data than A_2 could.

The trade-off values in the last column, which took both Part 1 and Part 2 into a consideration delivered a good evaluation of their performance in session overlapping. It shows that A_3 and A_5 scored the best with good performances in both Parts 1 and 2, marginally outperforming A_2. The worse performance from A_1 (the pure content-check) indicates that the CC alone is less effective - at least in our implementation - at determining boundaries. A_4, which double checks interactions that rarely lead to a session break with a CC, performed worse than A_3, which trusted the likelihood of the interactions. A_5 also double checked the interaction check, but at a lower threshold for content simiarity. The difference between A_3 and A_5 means that, according to this method of analysis, it is not clear whether the lower threshold check in A_5 improves A_3's performance. However, overall, a combination of

performing a CC on interactions with a higher likelihood of creating a boundary (A_3-A_5) improves the performance of a content-only check (A_1) and an interaction-only check (A_2).

7.5 Summary

The aim of this chapter was to evaluate the performance of these five algorithms (i.e., two solo versions and three combined versions). Firstly, the whole flow starting from the arrival of a new log and ending with the session decision to it was presented in Figure 7.2 with three Phases.

- 1) Determine whether there is a decision needed for this log
- 2) Determine whether there is relevance between this log and other existing sessions.
- 3) Add the log to the corresponding sessions or create a new session for it, and then update the session detail with the newly added log detail.

Then focusing on the Phase 2, three atomic decision making based on the content similarity and user interaction were introduced. After that, five algorithms were proposed with the different combinations of those three atomic checks, and the baseline was also picked. At the same time, some concern about the practical implementation were described.

Table 7.24 Performance Summary Based on F1 Measures

	(BI) All Logs	(BI) Decision-Needed Logs	(BI) Decision-Needed Logs (Three-Class)	Session Overlapping
Comparison$_1$: A_1 **VS.** A_2	A_1	A_2	A_2	A_2
Comparison$_2$: A_1 **VS.** A_3	A_3	A_3	A_3	A_3
Comparison$_3$: A_2 **VS.** A_3	A_3	A_2,A_3	A_2,A_3	A_3
Comparison$_4$: A_4 **VS.** A_5	A_5	A_5	A_5	A_5
Comparison$_5$: A_3 **VS.** (A_4, A_5)	A_3	A_3	A_3	A_3,A_5
A_1, A_2, A_3, A_4, A_5	A_3	A_2,A_3	A_3	A_3,A_5

In the end, the performance of these five algorithms including both boundary and session content were presented in Table 7.24, and A_3 did the best in general. In terms of *Comparison*$_1$ between A_1 and A_2, the better performance of A_2 at most of time indicates that user interaction is a better solo session boundary indicator than content relevance. In addition, user interaction (A_2) is also a 'cheaper' algorithm (see Table 7.4) with better runtime performance. The combined versions (A_3, A_4, A_5) of Content-Check and Transition-Check performed better

than the solo verisons (A_1, A_2) in general. The better performance of A_3 in most situations indicates that assigning the transition-connected log to the same session as the last log (i.e., high confidence on transition-connection) does turn out well, though it missed a few *Create-Log* and misrecognise a few *Create-Log* as *Resume-Log*. In $Comparison_3$ between A_3 and A_2, the better performance of A_3 indicates that applying the content-check on transition-break logs brings real benefit. The difference in $Comparison_4$ and the better performance of A_5 compared to A_4 implies that applying a lower threshold to the transition-connected logs on QCC (quick content validation) is also beneficial. Then in Chapter 8, more discussion about the algorithm would be presented along with some discussion on user interaction level and also future work.

Chapter 8

Discussion

8.1 Discussion of Algorithms

From the evaluation of both boundary investigation and session content overlapping, the performance of all these five algorithms were better than the fixed timeouts. As shown in Table 7.8, A_1 was able to detect out 88.3% of the actual boundaries; A_2, 87.6%; A_3, 73.6%; A_4, 80.1%; and A_5, 76.9%. All of these results were quite promising. Meanwhile, the fixed timeouts could only find less than 6% of boundaries mixed in with many unnecessary boundaries. A_1 also generated some unnecessary boundaries, which tended to split users' web activities into more small pieces; A_3 and A_5 did better than A_1. This problem could perhaps be overcome by integrating the session-merging afterwards. This could help to improve the precision, as suggested by Liao's design of a classification [66] of search queries.

However, distinguishing between *Create-Log* and *Resume-Log* still needs some improvement. As we know, a decision to a log consists of two parts: 1) correctly determining whether or not it is a boundary, and 2) if it is a boundary, making a decision about whether it would lead to a new session (*Create-Log*) or switch to other existing session (*Resume-Log*). These five algorithms did fairly well in Part 1, but making correct decision of Part 2 still requires further improvement. This decision might be biased by user's preference such as the scale of a session (e.g., whether the hotel booking in City A belongs to the session of hotel searching in City B, or whether the hotel booking in City A belongs to the session of

gathering transportation and place-of-interest information about City A). Such decisions are sometimes somewhat subjective, and bigger datasets are definitely required to improve the descion-making process.

8.2 Evaluation of Interactions

Learnt from the evaluation in Chapter 7, the combined versions performed better than solo versions in general. Actually, there could be further insight into the performance of interaction level performance of each algorithm as presented in Table 8.1 (F1 score), Table 8.2 (precision), and Table 8.3 (recall), so does their number of occurrence in the bracket.

Table 8.1 Performance Summary of Each Interaction (Based on F1 Measures)

		F1-Measure				
		A_1	A_2	A_3	A_4	A_5
Transition-Connected	**TULC_NonBlank (3905)**	0.384	0.789	0.716	0.460	0.645
	OLNT (632)	0.536	0.758	0.781	0.568	0.675
	BOF (186)	0.270	0.857	0.900	0.621	0.783
Transition-Break	**SAB_Blank (195)**	0.850	0.856	0.855	0.851	0.848
	TULC_Blank (120)	0.889	0.891	0.873	0.853	0.881
	SAB_NonBlank (78)	0.711	0.678	0.675	0.705	0.674
	SEI (47)	0.800	0.667	0.800	0.814	0.800
	FEPNT (42)	0.792	0.645	0.826	0.792	0.776

Table 8.2 Performance Summary of Each Interaction (Based on Precision)

		Precision				
		A_1	A_2	A_3	A_4	A_5
Transition-Connected	**TULC_NonBlank (3905)**	0.243	0.747	0.793	0.323	0.564
	OLNT (632)	0.376	0.822	0.867	0.427	0.598
	BOF (186)	0.156	0.818	0.900	0.474	0.692
Transition-Break	**SAB_Blank (195)**	0.813	0.749	0.841	0.840	0.834
	TULC_Blank (120)	0.852	0.810	0.878	0.857	0.864
	SAB_NonBlank (78)	0.640	0.513	0.675	0.646	0.630
	SEI (47)	0.727	0.500	0.727	0.750	0.727
	FEPNT (42)	0.679	0.476	0.731	0.679	0.655

TULC_NonBlank is the most frequent interaction with web page content update, and it is categorised as a transition-connected interaction. The high frequency illustrates how users explore further detail in their information seeking activities on browser. Based on the Table 8.1 about F1-measure, A_2 did the best with pure transition-check. It also achieved a good recall. A_3 and A_2 achieved a good balance with good performance in both completeness

Table 8.3 Performance Summary of Each Interaction (Based on Recall)

		Recall				
		A_1	A_2	A_3	A_4	A_5
Transition-Connected	**TULC_NonBlank (3905)**	0.918	0.835	0.653	0.800	0.753
	OLNT (632)	0.935	0.703	0.710	0.848	0.775
	BOF (186)	1.000	0.900	0.900	0.900	0.900
Transition-Break	**SAB_Blank (195)**	0.890	1	0.870	0.863	0.863
	TULC_Blank (120)	0.929	0.990	0.869	0.848	0.899
	SAB_NonBlank (78)	0.800	1.000	0.675	0.775	0.725
	SEI (47)	0.889	1.000	0.889	0.889	0.889
	FEPNT (42)	0.950	1.000	0.950	0.950	0.950

and correctness and efficiency in runtime (shown as Table 7.4), especially A_2. So having a high confidence about the transition-connection of TULC_NonBlank does bring benefit, the possible further improvement on it could be: adding some content-check with low requirement on the TULC_NonBlank interaction (e.g., try A_6 with a quarter of threshold used in A_4)

OLNT is the second most frequent interaction that leads to content update and new tab creation, and it is categorised as a transition-connected interaction. This indicates another popular way users prefer to do when seeking information on browser. As both OLNT and TULC_NonBlank visit the new web page via hyperlink or button clicking, there are some similarity between them, so does their performance over these five algorithms. The main difference between them is that A_3 performed better than A_2 especially in Recall. This illustrates that OLNT has higher possibility to lead to a information need shift compared to TULC_NonBlank. The similar further improvement from TULC_NonBlank could be applied here.

BOF refers to the interaction when the back or forward button is clicked, and it is categorised as a transition-connected interaction. A_2 and A_3 always preferred to put it into the same session as the previous tab/page in. All of the algorithms did well in the completeness (recall), but A_1 failed in precision. These indicates that there is a strong relevance between this log and last log when the back or forward button clicked. As a result, the page revisitation caused by BOF is more likely to stay in the same session. Learnt from current dataset, the Transition-Check is a strong enough rule for BOF in the session segmenetation.

SAB_Blank refers to the interaction when user is typing text (i.e., URL or query) into search address bar at the top of a browser when the current tab is blank[1]. The difference between SAB_Blank and SAB_NonBlank is that whether or not the tab the user is interacting with is blank. All algorithms did well in the completeness (recall) and their correctness (precision) performance was acceptable. Since A_2 directly makes the transition break interaction a boundary, its good performance in this interaction indicates that SAB_Blank is more likely to associate with information-need shift. As a result, the possible improvement on SAB_Blank could be that using A_3 as the base image, and try to apply content-check with various thresholds.

TULC_Blank refers to interactions like clicking on a hyperlink or button on a blank tab to get the tab updated (e.g., clicking on the 'Recent Visited Pages' on a default home page). Since this is very similar to SAB_Blank, the same possible improvement from SAB_Blank could be applied here.

SAB_NonBlank refers to the interaction when user is typing text (i.e., URL or query) into search address bar at the top of a browser when the current tab already has content. Users may suddenly interact with the browser address bar in a non-blank tab. The performance difference between SAB_Blank and SAB_NonBlank indicates that the interaction with SABs in different situations may indicate different user intentions. The high recall indicates that boundaries did occur with SAB_NonBlank; while some improvement was needed for the precision. Since A_1 did the best among these five, one possible improvement could be some changes applied to A_1 on with SAB_Blank, namely prioritising the current session more and giving a higher weight to the comparison between the log and the current session. This may help to reduce the number of unnecessary boundaries.

SEI is the interaction when a user begins to use his/her browser and some of the last-visited pages are recovered. The interaction is categorised as a transition break interaction, because its previous interaction information on these restored tabs (i.e., new Tab ID will be assigned

[1] chrome://newtab

to these stored tabs) are not available. Its recalls (see Table 8.3) were very high, and its previsions (see Table 8.2) were also acceptabe (i.e., A_2 was lower than others), this indicated that the content-check is necessary for SEI.

FEPNT is the most rare interaction, and it happens when a user clicks on the a link in another program like Word or a PDF file outside of the browser in order to visit a webpage. This is categorised as a transition break interaction because the original link is not from any of the browser tabs. The extremely low precision of A_2 and the mediocre value of A_1 indicate that a Content-Check is more appropriate than a Transition-Check here. Some further improvements are still needed for the Content-Check. The suggestion for SEI, SAB_NonBlank and SAB_Blank also applies here, and it could be beneficial to try to find it out whether or not it was relevant to the user's last remaining session before he/she left the browser.

8.3 Finding Summary

Both the Content (e.g., Query Similarity [50, 91, 73]) and Content-Ignorant (e.g., time gap [126, 11, 55] or user interaction [38]) were considered in this study. So was the combination of them, which was somewhat similar to some research that combined multi-factors. For example, the study by Lucchese et al's [72] combined the content-based (normalised levenshtein distance + Jaccard distance), semantic-based, and also timeout for pre-processing the break phase) but with the assumption of consecutive loggings without any interleaving with existing server logs. The data applied in this study was gathered in the field study without pre-defined task/topic outside lab environmentvery but with richer interaction detail.

8.3.1 Search VS. Browse

The work involved both search session and browse session. The big portion of browse session and the characteristic difference between them, for example, significant differences were found in Page-Duration between search and browse session in Medium and Long duration

group in Chapter 6, confirms Kong et al study [61] and Olston and Chi's study [89]. As mentioned in Kong et al' study, many interaction between user and browser could be query-free, and they found that there was a strong relationship between *searching* and *browsing* such that many queries were triggered by different types of context in which *browsing* was one of the key factors. In addition, Olston and Chi [89] stated that *browsing* and *searching* were the two predominant paradigms for finding information on web.

8.3.2 Fixed Timeout is not helpful

As shown in Table 7.8, A_1 was able to detect out 88.3% of the actual boundaries; A_2, 87.6%; A_3, 73.6%; A_4, 80.1%; and A_5, 76.9%. All of these results were quite promising. Meanwhile, the fixed timeouts could only find less than 6% of boundaries mixed in with many unnecessary boundaries. The fixed timeout is found to be useless after experimenting with 5/10/20/30/60 minutes, and it confirmes the finding from Montgomery and Faloutsos [80]'s study and also the study from Jones and Klinkners [55].

8.3.3 Domain-based or tab-based is not helpful

As there are more than 70% of the sessions involved visits to multiple domains and more than 60% of sessions involved multiple tabs, either domain-based or tab-based segmentation may not be the successful method for clustering sessions. Perhaps most interestingly, the average number of boundaries in a session is more than 10 for all users, which indicates that on average, a user switched to other sessions 9 times per session, and it confirms Ozmutlu et al.'s finding about high frequency multitasking [92].

8.3.4 User Interaction did better than Content Relevance

In terms of $Comparison_1$ between A_1 and A_2, the better performance of user interaction on session segmentation add more evidence to the findings from prior studies about the correlation between user interaction and other information as shown in the Table 6.6, and it proves the search behaviour study by Kim et al [59] and the behaviour-based approach

promised by Hassan et al [38]. In addition, user interaction (A_2) is also a 'cheaper' algorithm (see Table 7.4) with better runtime performance, and the advice of possible improvement to each interaction was presented in the following subsection.

8.3.5 Combined Versions perform better than Solo Versions

The combined versions did better than solo versions. The comparison between A_1 and A_3, differentiates between the two based on the different confidence level on transition-connected log. The better performance of A_3 in most situations indicates that high confidence on transition-connected log does turn out well, though it may miss a few Create-Log or misrecognise some Create-Log as Resume-Log. In $Comparison_3$ between A_3 and A_2, the better performance of A3 indicates that applying the content-check on transition-break logs brings real benefit. The better performance of A_5 in $Comparison_4$ with A_4 implies that applying a lower threshold to the transition-connected logs on QCC (quick content validation) could also be beneficial.

8.3.6 User Interaction

The association of each interaction type and session segmentation can be found in the TableTable 8.4 below:

Table 8.4 Possible Improvement for Each Interaction

	Well Performed Algorithm [2]	Advices for future improvement
TULC_NonBlank	$\mathbf{A_2}, A_3, A_4, A_5$	high confidence in Transition-Check, lower requirement on Content-Check validation (e.g., try A6 with a quarter of threshold that is used in A4)
OLNT	$A_2, \mathbf{A_3}, A_4, A_5$	high confidence in Transition-Check, lower requirement on Content-Check validation (e.g., try A6 with a quarter of threshold that is used in A4)
BOF	$A_2, \mathbf{A_3}$	high confidence in Transition-Check
SAB_Blank	$A_1, \mathbf{A_2}, A_3, A_4, A_5$	Content-Check with different level of thresholds, and prioritise the validation with current session
TULC_Blank	$A_1, \mathbf{A_2}, A_3, A_4, A_5$	Content-Check with different level of thresholds, and prioritise the validation with current session
SAB_NonBlank	$\mathbf{A_1}, A_4$	Content-Check with different level of thresholds, and prioritise the validation with current session
SEI	$A_1, A_3, \mathbf{A_4}, A_5$	pure Content-Check
FEPNT	$A_1, \mathbf{A_3}, A_4, A_5$	Content-Check with different level of thresholds, and prioritise the validation with last active session

8.4 Algorithm's Compatibility

Currently, the algorithms presented have only been run through the data gathered from the custom-built Chrome extension. However, they could potentially be run through other log datasets (e.g., AOL log). The structure of these algorithms is clearly explained in Chapter 7. Since all of the parameters of each element are adjustable, the absence of any criteria will not lead to a crash but rather to a corresponding score. To run it on a dataset, two issues have to be considered. The first is language: the CC of these algorithms was mainly designed for English. The similarity measurement can also be run on other languages, but the semantic measurement requires further attention. The interaction check (A_2) does not have this problem because it is content-ignorant. The second issue is fewer features: for example, the log entry from AOL web data consists of only these fields (e.g., timestamp, query, URL, title, and sometimes click data). The transition connection data (e.g., tab switching) is missing, and there are also some partial absences of content data (Top10KWs), Even that, a score with a comparison on query, title, and domain would still be available. Additionally, the thresholds and weights for each comparison requires some further testing and validation. As a result, these algorithms are compatible with a dataset including fewer features or even from a different language background. In addition, these algorithms have been implemented both locally and real-time with the user's interaction. I would allow the researcher to integrate his/her modified algorithm into the Chrome extension to run the real-time testing conveniently. With the growth of dataset scale and quality, a machine learning method would be the next step (some work has been done in one undergraduate student's final year project; the student used my GT dataset to integrate some machine learning methods for boundary detection and session segmentation).

8.5 The Potential Application

As explained in the existing tools section in Chapter 2, there are some demands from researchers, general users, and developers.

8.5.1 Researchers

Data Collection The Chrome extension can collect a comprehensive version of users' web data without any distraction during their interaction with the browser. In addition, it provides a privacy protection schema, and participants can 1) conduct confidential activities in an Incognito Window and 2) view a copy of the log data and filter out any confidential data. The installation of this tool is also qucik and easy with a direct link from the Chrome Web Store; no extra configuration is required. The extension was installed on the participants' laptops (both Mac and Windows) for more than one week, and some of them even kept it for one month, and there have not been any crash report. All of the log data is stored in a JSON format and can be converted to CSV directly or any other format conveniently. Compared to most of the existing online data collection tools, this tool has a custom-built system log to human interaction behaviour mapping, and more than 10 user actions can be extracted, making the logs more meaningful and easier to interpret.

Besides the logs, the sessions (grouping relevant logs into one) are also provided in order to provide a summary of participants' daily activity. There are five algorithms with different sensitivities, and different determining dimensions can be chosen. As a result, the researchers can get a copy of some more meaningfully clustered sessions, compared to the fixed timeout, and also do not need to spend too much time manually processing them.

Further Work on Session Segmentation The relevance judgement between visits is a long-term topic and there are many researchers are working on this area. Additionally, the combination of different factors applied in this implementation is not unique, since Lucchese et al's study [72] in 2011 and Liao et al.' study [66] in 2012 had also tried some combination of different factors or features or even different phases. However, the log data they tried with is bigger, but the features in each individual log is not as rich as what I gathered here. Tunning their ad-hoc algorithms on old datasets (e.g. AOL logs) or less-detail logs could limit their assessment of performance. This tool provides a convenient way to gather a dataset. The full details of log entries also enable a broader trial of different machine learning

models, because there are more features than other datasets, especially in terms of the users' interaction data and top 10 keywords from page content.

Further Step based on the Session Data After establishing a more accurate segmentation of logs into sessions, learning from the logs from the same session would provide a more direct influence on the returned search result rather than relying on the entire search history (i.e., Google's 'Preivous Query' Project).

In addition, a basic assistant version of web script for presenting these logs and gathering the GT data (see Chapter 5) can help to gather how is user's preference on these determines. Moreover, the performance evaluation methods presented in this thesis could also benefit their further evaluation design.

8.5.2 General Users

As suggsted in Chapter 2 by the examples of several existing Chrome extensions for tabs or pages management and also the background information about Chrome Extension in the beginning of Chapter 5, there is a market for these extensions because some general users would like to find a way to manage their web activity. This was reiterated during my study when most of the participants made a similar comment: *'I didn't realise that I spent so much time on XXX; I thought I only stayed there one or two minutes each time'*. The sessions determined by our extension could potentially be extended for a future use as 1) an automatic task-management tool to let users know how many tasks they have done today or how many pages they have visited, 2) an automatic time management tool to let users know how they spend their time on the browser and how their productivity is, and 3) some other add-ons like establishing the type of worker they are (i.e., whether they are more concentrated or more multitasking). Several posters[3] in the official Vivaldi browser forum had a positive discussion about requiring automatic grouping of relevant tabs to enhance their current tab stacking feature.

[3]https://forum.vivaldi.net/topic/1402/option-to-auto-group-tabs/9

8.5.3 Developers

Since the data gathered consists of both well formatted raw log data and also the raw session decision by our algorithms, they can be potentially exposed to other developer via the Restful API.

8.6 Future Work

Even though it followed the full research cycle and ended with a capable tool integrating the custom-designed algorithms integrated, some further work is still required to improve the performance. First, more data, more participants from more diverse backgrounds, and longer periods of data collection are always better. Since heavy manual processing was required to build the GT data of the sessions from participant themselves, more time is necessary and more subsequent studies have to be conducted to enhance the quality and scale of the GT dataset. The next possible step would be to make the GT-gathering web script available through a remote server to allow participants to use it remotely. This would enable us to recruit participants across occupations (not only university student), locations (not only the UK or other English-speaking countries), and cultures. We could also extend the period of study from one week to couple of weeks (e.g., one month). This would enable us to gather more insights, such as how different people's online behaviour on weekends vs. weekdays or during the daytime vs. the evening. All of the data gathered would be published to an online data repository formatted in either JSON or CSV, which would provide good data support for other researchers. Besides the tool improvement, the better data strategy should also be considered, and it includes both data storage, data processing, and data export.

Chapter 9

Conclusion

Through this study, I began with a user study focusing on theoretical taxonomy to understand the web session from user's perspective. Then a custom-built tool was developed in order to extend the essential log features. After that, a ground-truth dataset was worked out, then followed by several algorithms to assess the performance of different factors (i.e., user interaction vs. content-similarity). The following is a list of the interesting findings extracted during the entire PhD study:

1) Multitasking is the most popular web behavior in users' daily web activity, similar to Ozmutlu et al.'s finding in 2003 [92].

2) Participants do not always have a clear perception of their session duration: overestimation occurs more frequently than underestimation, and there is also an implicit possible regulation about the relationship between activity type and time of day.

3) In *search session*, people tends to visit more web pages and domains than during *browse session*, but they stay on a single webpage for a shorter amount of time, TULC_NonBlank is the most frequent interactions during their further page exploration.

4) Fixed timeout was not acceptable for most participants and perform very poorly in the session determination, similar to Jones and Klinkner's finding on this [55].

5) Domain-based, tab-based, or query-based segmentation is not successful enough in most situations.

6) Individual user preferences concerning session scale make the difference between *Create-Log* and *Resume-Log* is a bit ambiguous. This was also reflected in the taxonomy from Study 1, where the first three factors - topic, task, phase - refer to different scale of session. More data and more sample users are needed to make the further research on this.

7) In some cases, user interactions dominate traditional content relevance, and it definitely has some relationship with the user information need web. Some particular types of user interactions seem to be a strong indicator of session continuation and some of them indicate a high possibility of session shift.

8) The combination of user interactions and a content relevance check achieved a much better performance than the solo versions indicating that applying the CC to transition break logs creates real benefit.

9.0.1 Answering the Research Questions

For the first research question - 'How do people understand and describe their own web sessions, and what factors that create the divisions between them': the taxonomy in Chapter 4 summarised some of the factors people considered when segmenting their own logs, which has expanded the view of the factors involved in people's web activity clustering and als proved that fixed timeouts are not a meaningful enough delimiters. In addition, it also stated that there could be bias of the session scale in people's perception. Based on the taxonomy, the content relevance is always the primary factors considered by users, but this factor can sometimes be dominated by others in different situations. For the content relevance itself, there are also different levels of similarity (scale). These provide a good indication for the further steps. Besides these factors related to session segmentation, there are also other findings during it as listed in Points 1, 2, and 4 above.

For the second research question - 'How are these self-identified session reflected in real-world logs of web activity? And what observable features represent their boundaries?', knowing the list of theoretical factors related to the session segmentation is still far from the practical model, so the Chrome extension was built and a comprehensive log capturing have been provided. Then Implicit reflection on these practical logs, some of which are

even not noticed by the participants themselves, provided a more straightforward clue about further algorithm design. For example, the possibilities of each interaction type to being a boundary inspired the design of TC. It also proves that some previous existing segmentation: domain-based, tab-based, and query-based, are not successful enough. As the finding 3), 5) listed above.

For the third research question - 'How can we algorithmically identify web sessions from detailed web activity logs', based on the information extracted from the practical dataset in the second research question, the content relevance and transition relevance were found to play key roles. In order to evaluate these two indicators, five different versions of algorithm were proposed (see Chapter 7) and evaluated. The evaluation in Chapter 7 provides a through statement of the features of each algorithm. Points 6-8 above provide guidance for further updates to these algorithms. Even if it is not really possible to end with a perfect algorithm design at this stage, it has provided a complete way about these algorithm design and evaluation. Moreover, it has also considered the trade off between between accuracy and efficiency prepared for real-time implementation.

9.1 Contribution

9.1.1 Taxonomy

The taxonomy presentes the factors of the session determination and reflectes the complexity of session segmentation, which varies from one situation to the next. It also provided some regulations. It confirmes that content relevance is really important in the relevance judgement during user's web activity. At the same time, it indicates that content similarity is not everything; there are some other factors that require notice, as suggested by Point 7-8 above. In addition, the scale of sessions (i.e., the three factors in taxonomy: topic, task, phase) of content similarity requires further investigation. The whole taxonomy is not a collection of straight rules for session determination but provides a good theoretical explanation of the complexity of session determination. It could be extended by further user studies to

create more distinct rules about the session scales, such as telling whether the primary factors applied on particular occasion are correlated to the session type.

9.1.2 Tool Developed

The Chrome extension has provided comprehensive assistance for IR study: 1) it captures the user's full set of log data and maps these abstract system logs to meaningful user interactions, 2) it stores this log data for further investigation or study, and 3) it enables the data-streaming decision making algorithm, which was evaluated for its robustness for one week using 20 participants. I am going to release this tool to GitHub and the Chrome Web Store to allow other researchers in this area to use it for their studies (e.g., the data gathering part have already been used in the Search Literacy Project). It could provide a good support for field studies, allowing researchers to set up in web studies without any requirement of manual input during participants' interactions with the browser. Since it covers both browsing and searching, it could be used for contextual research on searching only, browsing-only, or combinations of the two. The well format of log data in JSON also enables further modification and research on the original dataset. Additionally, since most of the user interaction have already been categorised and indexed from system logs, the detail of each log is more user-friendly and easy to interpret. In addition, it has also involved the consideration of interleaving in terms of both active dwell time and user's attention focus status. It could be applied in either naturally diverse (disturbed or interleaving) or consecutive working patterns, as indicated in the modified version for the Search Literacy Project in Appendix A.

9.1.3 Dataset

Besides the Chrome extension, a GT session dataset was also generated. Even though it is not very large and the data is not perfect, it contains the users' self-identified sessions based on web activities from a natural setting. The detail of each log entry, especially concerning user interactions, is more comprehensive than in any other prior dataset. The richer detail of each

entry allows the researchers to try out different features and rules in order to involve more factors. Recently, this dataset was applied in one undergraduate student's final year project to integrate the machine learning methods for session segmentation, and the numerous features of each log entry provided good support for the whole study.

9.1.4 Algorithm Recommendations

This study did not develop a perfect algorithm for all situations, but it did create a clear workflow for already-built algorithsm, as well as improvement suggestions for each type of user interaction.

9.2 Final Remarks

Through this study, I achieved a good understanding of 1) users' web behaviour and 2) the gap between theoretical findings and practical implementation. This study has comprised a full-cycle research experience from the unknown -> some interesting findings -> the difference between practical implementation and theory -> implementing it -> evaluating it. At first, the detection of actual sessions sounded a bit subjective, but after an iterative study from theoretical findings to practical implementation, some regulations have been extracted and can potentially provide good support for further research. Additionally, because a tool will provide good software support for future field studies, not only in this project but also in other web-related studies. In addition, because it focuses on clustering the raw web data, there are some potential applications for utilising these data (e.g. automatic time management software on a browser) for general users. Even though it is still not really possible to find a way to perfectly matched all an user's session decisions, the features of these five algorithms have been evaluated, and the possible uses for each has also been discussed.

References

[1] Abrams, D., Baecker, R., and Chignell, M. (1998). Information archiving with bookmarks: Personal web space construction and organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, pages 41–48, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[2] Adar, E., Teevan, J., and Dumais, S. T. (2008). Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1197–1206, New York, NY, USA. ACM.

[3] Allan, J., Carterette, B., and Lewis, J. (2005). When will information retrieval be "good enough"? In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 433–440, New York, NY, USA. ACM.

[4] Alzahabi, R. and Becker, M. W. (2013). The association between media multitasking, task-switching, and dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, 39(5):1485.

[5] Aula, A., Jhaveri, N., and Käki, M. (2005). Information search and re-access strategies of experienced web users. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 583–592, New York, NY, USA. ACM.

[6] Baer, A., Saroiu, S., and Koutsky, L. A. (2002). Obtaining sensitive data through the web: An example of design and methods. *Epidemiology*, 13(6):640–645.

[7] Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[8] Bailey, P., Moffat, A., Scholer, F., and Thomas, P. (2015). User variability and ir system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 625–634, New York, NY, USA. ACM.

[9] Baillie, L. (2003). Future telecommunications: Exploring actual use. In *INTERACT*.

[10] Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424.

[11] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O., and Grossman, D. (2007). Temporal analysis of a very large topically categorized web query log. *J. Am. Soc. Inf. Sci. Technol.*, 58(2):166–178.

[12] Belkin, N. J. (1996). Intelligent information retrieval: whose intelligence? *ISI*, 96:25–31.

[13] Bowman, L. L., Levine, L. E., Waite, B. M., and Gendron, M. (2010). Can students really multitask? an experimental study of instant messaging while reading. *Comput. Educ.*, 54(4):927–931.

[14] Brasel, S. A. and Gips, J. (2011). Media multitasking behavior: Concurrent television and computer usage. *Cyberpsychology, Behavior, and Social Networking*, 14(9):527–534.

[15] Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

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

[17] Byström, K. and Järvelin, K. (1995). Task complexity affects information seeking and use. *Inf. Process. Manage.*, 31(2):191–213.

[18] Catledge, L. D. and Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide web. *Computer Networks and ISDN Systems*, 27(6):1065–1073.

[Chen and Cooper] Chen, H.-M. and Cooper, M. D. Using clustering techniques to detect usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*, 52(11):888–904.

[20] Cockburn, A., Greenberg, S., Jones, S., McKenzie, B., and Moyle, M. (2003). Improving web page revisitation: analysis, design, and evaluation.

[21] Cockburn, A., Greenberg, S., McKenzie, B., Jasonsmith, M., and Kaasten, S. (1999). Webview: A graphical aid for revisiting web pages. In *Proceedings of the OZCHI*, volume 99, pages 15–22.

[22] Cockburn, A. and McKenzie, B. (2001). What do web users do? an empirical analysis of web use. *Int. J. Hum.-Comput. Stud.*, 54(6):903–922.

[23] Diaz, F., White, R., Buscher, G., and Liebling, D. (2013). Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1451–1460, New York, NY, USA. ACM.

[24] Donato, D., Bonchi, F., Chi, T., and Maarek, Y. (2010). Do you want to take notes?: Identifying research missions in yahoo! search pad. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 321–330, New York, NY, USA. ACM.

[25] Downey, D., Dumais, S., and Horvitz, E. (2007). Models of searching and browsing: Languages, studies, and applications. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2740–2747, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[26] Elsweiler, D., Wilson, M. L., and Kirkegaard Lunn, B. (2011). understanding casual-leisure information behaviour. In *New directions in information behaviour*, pages 211–241. Emerald Group Publishing Limited.

[27] Feild, H. (2013). Crowdlogger home. https://crowdlogger.cs.umass.edu/index.html. Accessed: 2017-05-08.

[28] Feild, H. A. and Allan, J. (2013). Using crowdlogger for in situ information retrieval system evaluation. In *Proceedings of the 2013 Workshop on Living Labs for Information Retrieval Evaluation*, LivingLab '13, pages 15–18, New York, NY, USA. ACM.

[29] Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168.

[30] Gauch, S. and Smith, J. B. (1993). An expert system for automatic query reformation. *Journal of the American Society for Information Science*, 44(3):124.

[31] Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822 – 1843.

[32] Goecks, J. and Shavlik, J. (2000). Learning users' interests by unobtrusively observing their normal behavior. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, IUI '00, pages 129–132, New York, NY, USA. ACM.

[33] Greenberg, S. and Cockburn, A. (1999). Getting back to back: Alternate behaviors for a web browser's back button. In *Proceedings of the 5th Annual Human Factors and the Web Conference*.

[34] Greifeneder, E. (2016). The effects of distraction on task completion scores in a natural environment test setting. *J. Assoc. Inf. Sci. Technol.*, 67(12):2858–2870.

[35] Guo, Q. and Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 3601–3606, New York, NY, USA. ACM.

[36] Gwizdka, J. and Spence, I. (2006). What can searching behavior tell us about the difficulty of information tasks? a study of web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22.

[37] Hagen, M., Gomoll, J., Beyer, A., and Stein, B. (2013). From search session detection to search mission detection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 85–92, Paris, France, France.

[38] Hassan, A., Jones, R., and Klinkner, K. L. (2010). Beyond dcg: User behavior as a predictor of a successful search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 221–230, New York, NY, USA. ACM.

[39] He, D., Göker, A., and Harper, D. J. (2002). Combining evidence for automatic web session identification. *Inf. Process. Manage.*, 38(5):727–742.

[40] He, J. and Yilmaz, E. (2017). User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 67–76, New York, NY, USA. ACM.

[41] Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition.

[42] Hembrooke, H. A., Granka, L. A., Gay, G. K., and Liddy, E. D. (2005). The effects of expertise and feedback on search term selection and subsequent learning: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(8):861–871.

[43] Herder, E. (2005). Characterizations of user web revisit behavior.

[44] Hirsh, S. G. (1997). How do children find information on different types of tasks? children's use of the science library catalog.

[45] Hölscher, C. and Strube, G. (2000). Web search behavior of internet experts and newbies. *Computer networks*, 33(1):337–346.

[46] Huang, J. and Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86. ACM.

[47] Huang, J. and White, R. W. (2010). Parallel browsing behavior on the web. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 13–18, New York, NY, USA. ACM.

[48] Huffman, S. B. and Hochster, M. (2007). How well does result relevance predict session satisfaction? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 567–574, New York, NY, USA. ACM.

[49] Jansen, B. J., Booth, D. L., and Spink, A. (2009). Patterns of query reformulation during web searching. *J. Am. Soc. Inf. Sci. Technol.*, 60(7):1358–1371.

[50] Jansen, B. J., Spink, A., Blakely, C., and Koshman, S. (2007a). Defining a session on Web search engines. *JASIST*, 58(6):862–871.

[51] Jansen, B. J., Spink, A., and Pedersen, J. (2005). A temporal comparison of altavista web searching: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(6):559–570.

[52] Jansen, B. J., Zhang, M., and Spink, A. (2007b). Patterns and transitions of query reformulation during web searching. *International Journal of Web Information Systems*, 3(4):328–340.

[53] Jesper, S., Clough, P., and Hall, M. (2013). Regional effects on query reformulation patterns. In *International Conference on Theory and Practice of Digital Libraries*, pages 382–385. Springer.

[54] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2).

[55] Jones, R. and Klinkner, K. L. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 699–708, New York, NY, USA. ACM.

[56] Jung, S., Herlocker, J. L., and Webster, J. (2007). Click data as implicit relevance feedback in web search. *Inf. Process. Manage.*, 43(3):791–807.

[57] Kaasten, S. and Greenberg, S. (2001). Integrating back, history and bookmarks in web browsers. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 379–380. ACM.

[58] Kellar, M., Watters, C., and Shepherd, M. (2007). A field study characterizing web-based information-seeking tasks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):999–1018.

[59] Kim, Y., Hassan, A., White, R. W., and Zitouni, I. (2014). Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 193–202, New York, NY, USA. ACM.

[60] Kjeldskov, J., Skov, M. B., Als, B. S., and Høegh, R. T. (2004). *Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field*, pages 61–73. Springer Berlin Heidelberg, Berlin, Heidelberg.

[61] Kong, W., Li, R., Luo, J., Zhang, A., Chang, Y., and Allan, J. (2015). Predicting search intent based on pre-search context. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 503–512, New York, NY, USA. ACM.

[62] Lau, T. and Horvitz, E. (1999). Patterns of search: analyzing and modeling web query refinement. In *UM99 User Modeling*, pages 119–128. Springer.

[63] Law, A. S., Logie, R. H., and Pearson, D. G. (2006). The impact of secondary tasks on multitasking in a virtual environment. *Acta Psychologica*, 122(1):27 – 44.

[64] Lazonder, A. (2000). Exploring novice users' training needs in searching information on the www. *Journal of Computer Assisted Learning*, 16(4):326–335.

[65] Lee, F. J. and Taatgen, N. A. (2002). Multitasking as skill acquisition. In *Proceedings of the twenty-fourth annual conference of the cognitive science society*, pages 572–577.

[66] Liao, Z., Song, Y., He, L.-w., and Huang, Y. (2012). Evaluating the effectiveness of search task trails. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 489–498, New York, NY, USA. ACM.

[67] Liu, C., Gwizdka, J., Liu, J., Xu, T., and Belkin, N. J. (2010a). Analysis and evaluation of query reformulations in different task types. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 17:1–17:10, Silver Springs, MD, USA. American Society for Information Science.

[68] Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J., and Zhang, X. (2010b). Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 69–78, New York, NY, USA. ACM.

[69] Liu, J., Liu, C., Cole, M., Belkin, N. J., and Zhang, X. (2012). Exploring and predicting search task difficulty. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1313–1322, New York, NY, USA. ACM.

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

[71] Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. (2011a). Identifying task-based sessions in search engine query logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 277–286, New York, NY, USA. ACM.

[72] Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. (2011b). Identifying task-based sessions in search engine query logs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 277–286, New York, NY, USA. ACM.

[73] Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. (2013). Modeling and predicting the task-by-task behavior of search engine users. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 77–84, Paris, France, France.

[74] Lui, K. F. H. and Wong, A. C.-N. (2012). Does media multitasking always hurt? a positive correlation between multitasking and multisensory integration. *Psychonomic Bulletin & Review*, 19(4):647–653.

[75] Marchionini, G. and White, R. (2007). Find what you need, understand what you find. pages 205–237.

[76] May, K. (2012). Google introduces flight tracking system in search results, using your booking emails. https://www.tnooz.com/article/google-introduces-flight-tracking-system-in-search-results-using-your-booking-emails/. Accessed: 2017-02-14.

[77] Mehrotra, R., Bhattacharya, P., and Yilmaz, E. (2016). Characterizing users' multi-tasking behavior in web search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 297–300, New York, NY, USA. ACM.

[78] Milic-Frayling, N., Jones, R., Rodden, K., Smyth, G., Blackwell, A., and Sommerer, R. (2004). Smartback: Supporting users in back navigation. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 63–71, New York, NY, USA. ACM.

[79] Miwa, M. (2001). User situations and multiple levels of user goals in information problem solving processes of askeric users. In *Proceedings of the ASIST Annual Meeting*, volume 38, pages 355–71. ERIC.

[80] Montgomery, A. L. and Faloutsos, C. (2001). Identifying web browsing trends and patterns. *Computer*, 34(7):94–95.

[81] Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 272–281, New York, NY, USA. Springer-Verlag New York, Inc.

[82] Morris, D., Morris, M. R., and Venolia, G. (2008). Searchbar: A search-centric web history for task resumption and information re-finding. pages 1207–1216. ACM.

[83] Needleman, R. (2009). Yahoo search pad: An online notebook that watches you. https://www.cnet.com/news/yahoo-search-pad-an-online-notebook-that-watches-you/. Accessed: 2017-05-08.

[84] Nield, D. (2016). 5 reasons to use to vivaldi instead of chrome or firefox. http://fieldguide.gizmodo.com/5-reasons-to-use-to-vivaldi-instead-of-chrome-or-firefo-1771284757. Accessed: 2017-02-14.

[85] Niwattanakul, S., Singthongchai, J., Naenudorn, E., and Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, page 6.

[86] Norman, D. (1988). *The Psychology Of Everyday Things*. The Psychology of Everyday Things. Basic Books.

[87] Norman, D. A. and Draper, S. W. (1986). *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[88] Ohly, S., Sonnentag, S., Niessen, C., and Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology*, 9(2):79–93.

[89] Olston, C. and Chi, E. H. (2003). Scenttrails: Integrating browsing and searching on the web. *ACM Trans. Comput.-Hum. Interact.*, 10(3):177–197.

[90] Ozmutlu, H. C. and Çavdur, F. (2005). Application of automatic topic identification on excite web search engine data logs. *Inf. Process. Manage.*, 41(5):1243–1262.

[91] Ozmutlu, S. (2006). Automatic new topic identification using multiple linear regression. *Inf. Process. Manage.*, 42(4):934–950.

[92] Ozmutlu, S., Ozmutlu, H. C., and Spink, A. (2003a). Are people asking questions of general web search engines? *Online Information Review*, 27(6):396–406.

[93] Ozmutlu, S., Ozmutlu, H. C., and Spink, A. (2003b). Multitasking web searching and implications for design. *Proceedings of the American Society for Information Science and Technology*, 40(1):416–421.

[94] Palen, L. and Salzman, M. (2002). Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 87–95, New York, NY, USA. ACM.

[95] Piwowarski, B., Dupret, G., and Jones, R. (2009). Mining user web search activity with layered bayesian networks or how to capture a click in its context. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 162–171, New York, NY, USA. ACM.

[96] Raman, K., Bennett, P. N., and Collins-Thompson, K. (2013). Toward whole-session relevance: exploring intrinsic diversity in web search. In *Proc. SIGIR 2013*, pages 463–472, New York, NY, USA. ACM.

[97] Rha, E. Y., Mitsui, M., Belkin, N. J., and Shah, C. (2016). Exploring the relationships between search intentions and query reformulations. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*, page 48. American Society for Information Science.

[98] Rieh, S. Y. and Xie, H. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Inf. Process. Manage.*, 42(3):751–768.

[99] Robertson, S. E. and Hancock-Beaulieu, M. M. (1992). On the evaluation of ir systems. *Information Processing & Management*, 28(4):457–466.

[100] Rose, D. E. and Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA. ACM.

[101] Rugg, G. and McGeorge, P. (1997). The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems*, 14(2):80–93.

[102] Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[103] Sellen, A. J., Murphy, R., and Shaw, K. L. (2002). How knowledge workers use the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 227–234, New York, NY, USA. ACM.

[104] Shneiderman, B., Byrd, D., and Croft, W. B. (1997). Clarifying search: A user-interface framework for text searches. Technical report.

[105] Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., Carvalho, A. C. P. L. F. d., and Gama, J. a. (2013). Data stream clustering: A survey. *ACM Comput. Surv.*, 46(1):13:1–13:31.

[106] Spink, A. (2002). A user-centered approach to evaluating human interaction with web search engines: An exploratory study. *Inf. Process. Manage.*, 38(3):401–426.

[107] Spink, A., Bateman, J., and Jansen, B. J. (1999a). Searching the web: A survey of excite users. *Internet research*, 9(2):117–128.

[108] Spink, A., Griesdorf, H., and Bateman, J. (1999b). A study of mediated successive searching during information seeking. *Journal of Information Science*, 25(6):477–487.

[109] Spink, A., Jansen, B., and Ozmultu, H. (2000). Use of query reformulation and relevance feedback by excite users. *Internet Research*, 10(4):317–328.

[110] Spink, A., Ozmutlu, H. C., and Ozmutlu, S. (2002). Multitasking information seeking and searching processes. *J. Am. Soc. Inf. Sci. Technol.*, 53(8):639–652.

[111] Spink, A. and Park, M. (2005). Information and non-information multitasking interplay. *Journal of Documentation*, 61(4):548–554.

[112] Spink, A., Park, M., Jansen, B. J., and Pedersen, J. (2006). Multitasking during web search sessions. *Inf. Process. Manage.*, 42(1):264–275.

[113] Stalgaitis, C. and Glick, S. N. (2014). The use of web-based diaries in sexual risk behaviour research: a systematic review. *Sexually Transmitted Infections*, 90(5):374–381.

[114] Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Inf. Process. Manage.*, 28(4):503–516.

[115] Sullivan, D. (2008a). Google now notifies of "search customization" & gives searchers control. http://searchengineland.com/google-now-notifies-of-search-customization-gives-searchers-control-14485. Accessed: 2017-02-14.

[116] Sullivan, D. (2008b). "previous query" refinement coming to hit google results. http://searchengineland.com/previous-query-refinement-coming-to-hit-google-results-13743. Accessed: 2017-02-14.

[117] Sullivan, D. (2013). Google's impressive "conversational search" goes live on chrome. http://searchengineland.com/googles-impressive-conversational-search-goes-live-on-chrome-160445. Accessed: 2017-02-14.

[118] Summit, R. K. (1967). Dialog: An operational on-line reference retrieval system. In *ACM 1967*, pages 51–56. ACM.

[119] Sun, X. and May, A. (2013). A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices. *Advances in Human-Computer Interaction*, 2013:1–9.

[120] Sutcliffe, A. and Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10(3):321 – 351.

[121] TAUSCHER, L. and GREENBERG, S. (1997). How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47(1):97 – 137.

[122] Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 415–422, New York, NY, USA. ACM.

[123] Thomas, J. C. and Kellogg, W. A. (1989). Minimizing ecological gaps in interface design. *IEEE Software*, 6(1):78–86.

[124] Vakkari, P., Pennanen, M., and Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39(3):445 – 463.

[125] Venkatesh, V., Brown, S., and Sullivan, Y. (2016). Guidelines for conducting mixed-methods research: An extension and illustration. *Journal of the Association of Information Systems*, 17(7):435–495.

[126] Wakeling, S. and Clough, P. (2016). *Determining the Optimal Session Interval for Transaction Log Analysis of an Online Library Catalogue*, pages 703–708. Springer International Publishing, Cham.

[127] Waller, M. J. (1997). Keeping the pins in the air: How work groups juggle multiple tasks. *Advances in interdisciplinary studies of work teams*, 4:217–247.

[128] Wang, Q. and Chang, H. (2010). Multitasking bar: prototype and evaluation of introducing the task concept into a browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 103–112. ACM.

[129] Waxman, S. (2006). At an industry media lab, close views of multitasking. *New York Times*, 15.

[130] White, R. W. and Drucker, S. M. (2007). Investigating behavioral variability in web search. In *Proc. WWW2007*, pages 21–30. ACM.

[131] White, R. W., Dumais, S. T., and Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 132–141, New York, NY, USA. ACM.

[132] White, R. W., Ruthven, I., and Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 35–42, New York, NY, USA. ACM.

[133] WILSON, T. (1981). On user studies and information needs. *Journal of Documentation*, 37(1):3–15.

[134] Wolfram, D., Wang, P., and Zhang, J. (2009). Identifying web search session patterns using cluster analysis: A comparison of three search environments. *J. Am. Soc. Inf. Sci. Technol.*, 60(5):896–910.

[135] Xu, Y. and Mease, D. (2009). Evaluating web search using task completion time. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 676–677, New York, NY, USA. ACM.

[136] Yeykelis, L., Cummings, J. J., and Reeves, B. (2014). Multitasking on a single device: Arousal and the frequency, anticipation, and prediction of switching between media content on a computer. *Journal of Communication*, 64(1):167–192.

[137] Zohrabi, M. (2013). Mixed method research: Instruments, validity, reliability and reporting findings.

Appendix A

Search Literacy Project

Table A.1 Interaction Types in Search Literacy Project

	Detail	**SERP only?**
HOVER_ENTER, HOVER_LEAVE	The mouse movement over result / other elements in SERP	**Yes**
UPDATE_T_via_CLICK_QUERY_SUGGESTION	Click on the “Query Suggestion" Component in SERP	**Yes**
UPDATE_T_via_CLICK_NAVI_MENU	Click on the “Navigation Menu" Component in SERP	**Yes**
UPDATE_T_via_CLICK_PAGE_NUMBER	Click on the “Page Number" Index Component in SERP	**Yes**
UPDATE_T_via_SEARCH_RELATED_TO	Click on the “Search Related To" Component in SERP	**Yes**
UPDATE_T	Click on any other click or button to update the tab	**No**
OPEN_LINK_T, OPEN_LINK_W	Open link in new Tab / Window	**No**
SCROLL	The scroll-bar movement	**No**
LEFT_CLICK, RIGHT_CLICK, WHEEL_CLICK	Three different types of mouse click	**No**
BOF	Back or Forward Button Click (Next to Address Bar)	**No**
OPEN_FIND_BAR	When user open the “Search Bar" (Short-cut: Ctrl+F or Cmd + F)	**No**
Search_AB	Type query into Address Bar	**No**
Type_URL_AB	Type url into Address Bar	**No**
Search_SF	Type query into Search Field	**No**
Create_T, Create_W	Create new blank Tab / Window	**No**
Close_T, Close_W	Close Tab Window	**No**
SWITCH_T, SWITCH_W	Switch focus between Tabs / Windows	**No**
LEAVE_BROWSER	Leave Browser or Go to Incognito Window	**No**
OPEN_FROM_EXTERNAL	Open Click From External Programme in New Tab	**No**

Table A.2 SERP Component

Position	Name	Is Result?
Head	QUERY_INPUT_FIELD	**No**
	NAVI_MENU	**No**
	QUERY_SUGGESTION	**No**
Body	NORMAL_RESULT	**Yes**
	NORMAL_RESULT_WITH_IMAGE_VIDEO	**Yes**
	ADS	**Yes**
	TOP_CONTAINER	**Yes**
	IMAGES_FOR	**Yes**
	RIGHT_SIDE_BAR	**Yes**
	SOCIAL_MEDIA	**Yes**
	PEOPLE_ALSO_ASK	**No**
	SEARCH_RELATED_TO	**No**
Bottom	PAGE_NUMBER	**No**

Appendix B

Instruction for using Chrome Extension

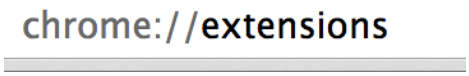
What is this study about:

The main aim of this study is to look into people's web data to work out a more intelligent model for determining user web session. Before the testing, you will be asked to have Google Chrome Web Browser and a Chrome Extension (developed by our researcher) installed on your personal computer to track your web data for 7 days. Then after that, you will be invited to an 80-mins interview (Audio recorder will be applied). It will be a discussion with one of our researchers about your own web data. You will receive £15 Amazon voucher after this study.

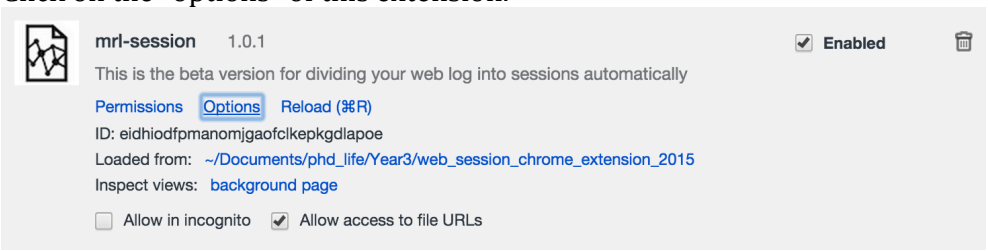
You will receive from one of our researchers including the private link for the chrome extension and also your userId and password. Please do not share these information with any others.

How to submit your data to remote server

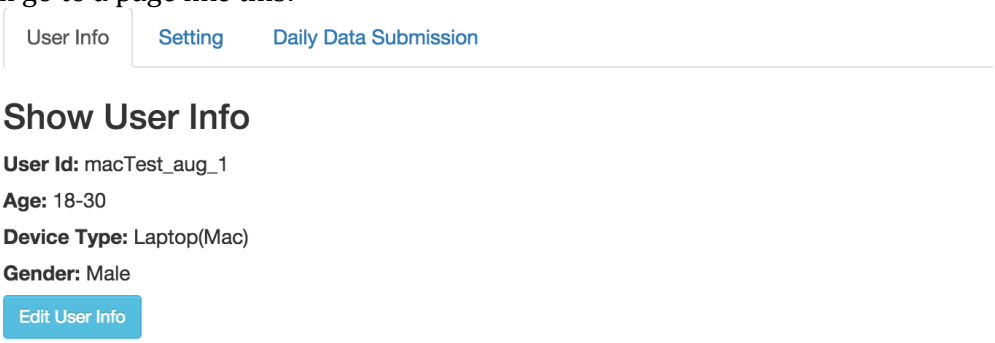
I. Go to your chrome extension page: chrome://extensions



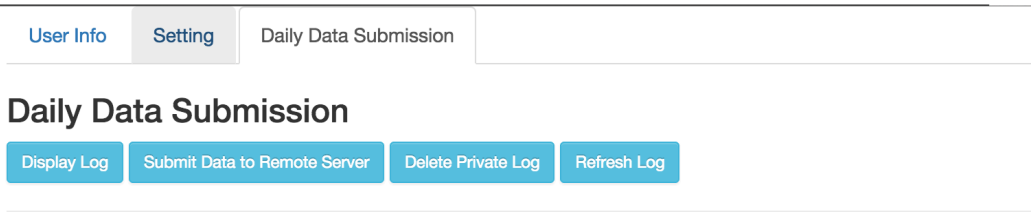
II. Click on the "options" of this extension.



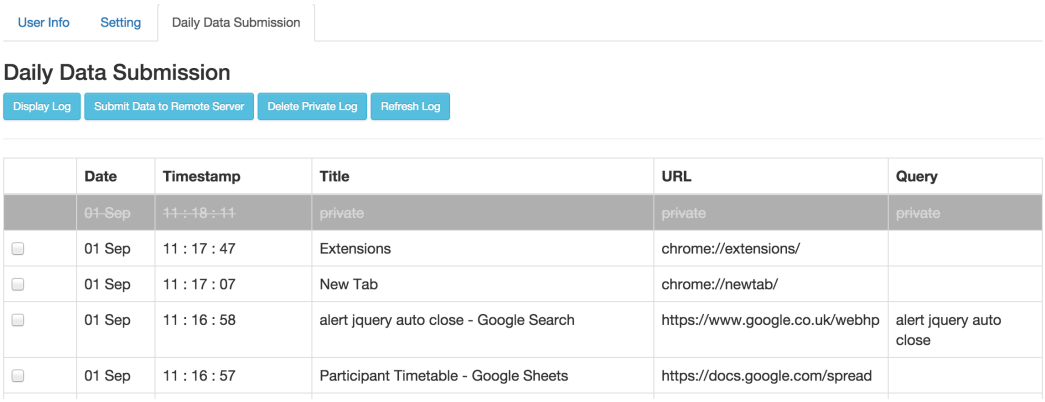
III. It will go to a page like this:



IV. Go to the Tab "Daily Data Submission"



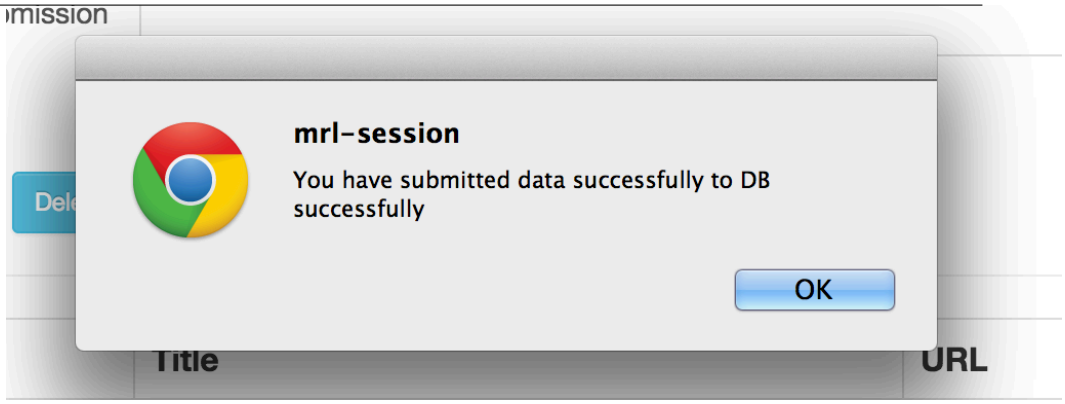
V. If you want to protect your own privacy, please read this part carefully. Otherwise, jump to point
There are 2 ways to protect your privacy
- Use the "Incognito" window for some activities, such as logging to some online banking. There is no data tracking in this window.
- Even you forget to use the "Incognito" window for some private data, you can still delete them before synchronsing it to my server. Here is how to do it:
    - Click on the "Display Log" button, it will return a list of the logs waiting to be submitted, like this



	Date	Timestamp	Title	URL	Query
	~~01 Sep~~	~~11 : 18 : 11~~	private	private	private
☐	01 Sep	11 : 17 : 47	Extensions	chrome://extensions/	
☐	01 Sep	11 : 17 : 07	New Tab	chrome://newtab/	
☐	01 Sep	11 : 16 : 58	alert jquery auto close - Google Search	https://www.google.co.uk/webhp	alert jquery auto close
☐	01 Sep	11 : 16 : 57	Participant Timetable - Google Sheets	https://docs.google.com/spread	

    - If you want to delete any particular log, you can tick the checkbox in front of it. And then press the button "Delete Private Data". Then, you can press the "Refresh Log Data" to see the updated log

VI. After the privacy protection, you can submit your data to remote server by clicking on the button "Submit data to remote server". As your data size may be big. After clicking on that, can you please wait for several seconds until you can a window alerted "Successfully!!!" like this:



If it returned an error, please contact me: psxcy1@nottingham.ac.uk

VII. After this data submission, you can just close this option page and continue your interaction with browser. Please try to submit your web data once per day (otherwise, the log file may be too big)

Important Things Need to Keep in Mind

Because this extension is just a beta version, so please follow the instruction, otherwise, it may not work.

1. Please use the provided "userId" and "password", you just need to type them in your first data submission. The later submission won't need any more authorisations.
2. When submitting data, if you didn't receive the alert window with "Successfully!!!", please contact me via email ASAP.
3. Please don't share the chrome extension link with other people.

Interview

After the data collection, there will be an 80 mins.

If you have any questions, please feel free to contact Chaoyu Ye: psxcy1@nottingham.ac.uk