Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a preprint version which may differ from the publisher's version.

For additional information about this publication click this link.
http://repository.ubn.ru.nl/handle/2066/127981

Please be advised that this information was generated on 2018-07-07 and may be subject to change.

Collecting a dataset of information behaviour in context

Maya Sappelli[1,2], Suzan Verberne[2], Saskia Koldijk[1,2] and Wessel Kraaij[1,2]
1. TNO and Radboud University Nijmegen
2. Radboud University Nijmegen
m.sappelli@cs.ru.nl, s.verberne@cs.ru.nl, saskia.koldijk@tno.nl, kraaijw@acm.org

ABSTRACT

We collected human–computer interaction data (keystrokes, active applications, typed text, etc.) from knowledge workers in the context of writing reports and preparing presentations. This has resulted in an interesting dataset that can be used for different types of information retrieval and information seeking research. The details of the dataset are presented in this paper.

1. INTRODUCTION

This research project is part of the project SWELL[1] (Smart Reasoning Systems for Well-being at home and at work). Our overall objective is to increase the physical and mental well-being of knowledge workers[2]. We monitor their behaviour and provide them with an unobtrusive digital assistant that advises them about fitness-improving and stress-reducing behaviour.

In light of this project, we have collected a dataset of human–computer interactions during typical knowledge worker's tasks [2]. In this data we find a large body of natural search behaviour data. Together with the detailed information of the user's computer activities, we think that this dataset is interesting for the Information Retrieval community because it describes information seeking behaviour in a work context. In this paper we describe how we collected and pre-processed the data and show statistics about the collected data. Additionally we will provide some examples of research that could be done with this dataset.

We will make the dataset available for research purposes.

2. METHOD

The main purpose of the data collection experiment that we carried out was to study stress among knowledge workers during a typical work day [2]. The setup of the experiment was aimed at collecting data to recognize user activities during their work at the computer. The subjects were asked to write reports on a total of 6 given topics and prepare presentations for three of the topics.

The experiment in which the data were collected captured three conditions: a) a neutral condition in which the participants were asked to work as they normally do; b) a condition in which they were time pressured and c) a condition in which they were interrupted with email messages. Each of the conditions lasted between 30 and 45 minutes. In the remainder of this section we will describe the tasks in more detail. For more information on the conditions and the stress related data collection we refer to Koldijk et al. [2].

2.1 Participants

We collected data from 25 participants with an average age of 25 (std 3.25). This number of participants is sufficient for within-subject measurements. There were 8 females and 17 males, and the participants were recruited among university students and interns at TNO[3]. 23 participants wrote their reports and presentations in English, two used Dutch. All participants received a standard subject fee for experiment participation. To motivate the students to do their best on the reports, they were told that the height of the fee was dependent on their performance.

2.2 Materials

The participants executed their tasks on a laptop computer equipped with Microsoft Office. The default browser was Internet Explorer with www.google.nl as start page. Also, uLog version 3.2.5[4] was installed. uLog is a key-logging tool that saves the active application, window title, url or file that is open, caption information, mouse movements, mouse clicks and keyboard activity with timestamps. Additionally the desktop was recorded with GeoVision's CCS[5] and the browser history was saved with IEHistoryView version 1.65[6]. Additional data types include: camera recordings for facial expressions, heart rate, skin conductance and 3D postures using Kinect. The participants had access to the general instructions of the experiment at all times. These can be found in Figure 1.

---

[1]http://www.swell-project.net

[2]A knowledge worker is a person whose job involves handling or using information. Nowadays, almost all office jobs are knowledge worker jobs.

---

[3]Dutch institute for applied scientific research

[4]http://www.noldus.com/human-behavior-research/products/ulog

[5]http://www.geovision.com.tw

[6]http://www.nirsoft.net/utils/iehv.html

"The aim of the experiment is to observe a knowledge worker during his/her tasks. To set a realistic scene you will be working on some typical knowledge workers task in an office setting, these include writing essays and preparing presentations. Additionally, you may receive e-mail messages during the experiment. You are allowed to read these, as they may contain useful information. In directory XXX you can find some material that may be helpful for your tasks, or you can use the internet to find information.
The experiment is made up of three blocks of activities. Before each block we will present you with an instruction for the activities in that block. After each block you will be asked to fill out a small questionnaire. The entire experiment will take about 3 hours, depending on your own speed. The amount of compensation you will receive depends on how many tasks you finished and the quality of your work. The minimal amount you will receive is 30 euros, the maximum is 40 euros. In each block you will be asked to prepare 3 presentations. At the end of the experiment we will choose one of these for you to present to us."

Figure 1: The general instructions that were given to the workers during the experiment.

2.2.1 Tasks

In each of the conditions the participants randomly received two out of six tasks, one opinion task and one task that required more information seeking. The short descriptions of the opinion tasks were:

- "Your experience and opinion about stress at work."
- "Your experience and opinion about healthy living."
- "Your experience and opinion about privacy on the internet."

The informational tasks were:

- "5 Tourist attractions in Perth, West Australia."
- "A plan for a coast to coast roadtrip in the USA."
- "The life of Napoleon."

The longer descriptions of the tasks as presented to the participants will be available in the dataset.

2.2.2 E-mail messages

In the condition where participants were interrupted with email messages, these messages sometimes contained tasks or questions. This resulted in two additional topics in the data: Einstein and Information Overload. The exact content of the received email messages will be available in the dataset.

2.2.3 Questionnaire

We also collected responses to self-reporting questionnaires addressing Task Load, Mental Effort, Emotion and Perceived Stress. The participants were also asked about their interest in the topics and how complex the topic was to them. The outcome of this questionnaire will be discussed in a separate publication.

2.3 Procedure

At the beginning of the experiment the participants were asked to fill out some questionnaires about their health background. They were also given the general instructions. At the beginning of each condition the participants were asked to watch a relaxing movie for 10 minutes. This was necessary to get an adequate resting heart rate baseline for the stress research. After that, the participants could look at the given topics for the condition and start their work. They were told to give a signal when they were ready in the neutral condition, or to stop writing when the timer went off in the time pressure condition. After the experiment leader paused all sensors, the participant was given a next set of questionnaires related to stress and a questionnaire related to interest in the topics. The participant was allowed to take a break and walk around between the conditions.

3. RESULTING DATASET

The dataset we present contains all computer interaction data that was recorded during the experiment. Most importantly this dataset contains the data coming from the uLog keylogger as well as the data collected from the IEHistoryView. Figure 2 presents a small excerpt from a uLog datafile. This example shows the *event* 'Window "http://www.google.nl" – Windows Internet Explorer activated'.

```
</Event><Event>
  <EventType>Other</EventType>
  <EventAction>Window Activated</EventAction>
  <TimeStamp>2012-09-19T11:22:01.9745893Z</TimeStamp>
  <Date>2012-09-19</Date>
  <Time>11:22:01.9745893+02:00</Time>
  <Milliseconds>974</Milliseconds>
  <EventDescription>Window "http://www.google.nl/ -
  Windows Internet Explorer" activated.</EventDescription>
  <Control>
    <ControlType>window</ControlType>
    <ControlCaption>http://www.google.nl/ -
    Windows Internet Explorer</ControlCaption>
```

Figure 2: Example of uLog data

## 3.1 Event aggregation and manual labelling

Besides the raw keylog data, we also provide a variant of the dataset in which the collected history-files and uLog-files have been preprocessed. In this dataset, individual events are aggregated to meaningful *event blocks*. We defined the start of a new event block as either an application switch, or a change in window title. In total, our data collection consists of 9416 event blocks.

All the keys typed and all captions (mouse-over tooltips) that occurred within one event block are concatenated into strings and the number of mouse clicks per event block is counted. From the recorded Google URLs we extracted the queries that were entered using a regular expression.

For future training and testing purposes (e.g. for topic detection), we collected manual topic labels for the event blocks. For the labelling, we used the crowdsourcing platform Amazon Mechanical Turk because of the large number of event blocks. The event blocks were presented to the annotators in a desktop-like setting to mimic the desktop view of the user during the experiment (Figure 3). The annotators were asked to select 1 topic label and also indicate on a scale of 1-5 how certain they were of their decision. The event blocks were shown in random order, so they could not use any session information. The labels were the 8 topics, and an additional topic 'indeterminable' when the event block did not contain any identifiable topic, for example when just the website 'www.google.nl' was shown.

Table 1 shows an overview of the features collected per event block, with an example value for each feature. Table 2 shows the distribution of the labels in our data. 121 event blocks were not labeled because of problems with the annotating system (Amazon Mechanical Turk). Inter-annotator

Figure 3: Example of presented event block to annotators. The typed text was shown in a box to the left of the file window
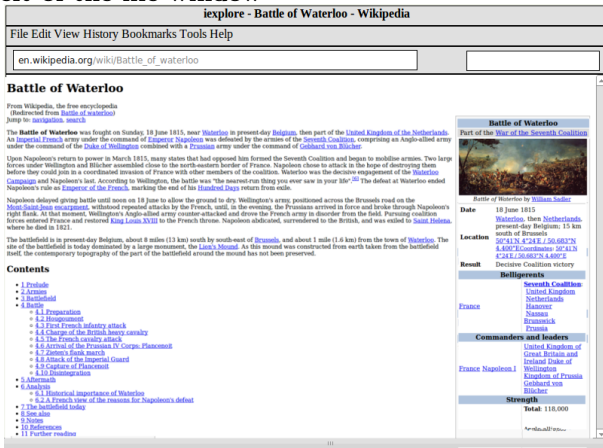
Table 1: Overview of features collected per event block, with example values

feature	example value
id	6
participant id	2
begin time	20120919T132206577
end time	20120919T132229650
duration (seconds)	24
# clicks	3
typed keys	we austra;i lia
application	iexplore
window title	Google - Windows Internet Explorer
caption	New Tab (Ctrl+T) New Tab (Ctrl+T)
url	http://www.google.nl/search?hl=nl& sclient=psy-ab&q=australia+&oq=australi
domain	www.google.nl
query	australia
Label	Perth

reliablity was measured on a stratified subset of the data with 10 random event blocks from each participant. Cohen's κ was 0.784, which indicates a substantial agreement.

3.2 Examples of analyses with the data

Several types of analyses are possible with the data. In previous work [2], we have studied the relations between stressors (time pressure, incoming e-mails) and sensor measurements (number of keystrokes, facial expression). We now discuss two types of analyses related to Information Retrieval: a query-central (system-oriented) and a behaviour-central (user-oriented) analysis. The multimodal aspects (camera and desktop recordings) of the data may be exploited in future research.

3.2.1 System-oriented analysis

As an example of a system-oriented analysis, we investigated the automatic estimation of relevance of URLs for each user query based on the user's interactions following the query. Per query, we extracted the URLs that were accessed in the query's event block and the event block after it. A common variable to estimate the relevance of a page is a dwell time of at least 30 seconds [1]. Since we have interaction data available, and we know that the user was

Table 2: Features in aggregated dataset and details on the labeled data

Total no. event blocks	9416
Average no. event blocks per participant	377
No. of 'Indeterminable' blocks	4347
No. of 'Einstein' blocks	117
No. of 'Information Overload' blocks	67
No. of 'Stress' blocks	612
No. of 'Healthy' blocks	637
No. of 'Privacy' blocks	269
No. of 'Perth' blocks	1248
No. of 'Roadtrip' blocks	1170
No. of 'Napoleon' blocks	828
No. of blocks failed to label	121

Table 3: Results of the query-central analysis

Total # of queries	980
of which followed by a click on a URL	732
of which followed by a switch to Word/Powerpoint	125
of which with control-c	15
with a dwell time of >= 30 seconds	44

collecting data for writing a report or a presentation, we not only calculated the dwell time on each web page but also registered whether the next active application was Word or Powerpoint, and if the user typed control-c in the browser before making this switch (copying text). Table 3 shows the results of this analysis. Note that the total number of queries in the first row includes duplicate queries that are recorded when the user clicks on a result and goes back in his browser. The table shows that if dwell time is used as only relevance criteria, only 44 pages would be judged as relevant. Taking into account the switches to Word and Powerpoint, this number is much higher. Table 4 shows an excerpt of the processed interaction data, focussing on a series of queries and the clicked web pages for those queries.

3.2.2 Behavioural analysis

Figure 4 shows an example of a behavioural analysis: a transition graph for the workers' information behaviour. It shows that when users are asked to write reports or prepare presentations on a relatively new topic, they spend more time on web pages than in the report they are writing, and they switch frequently between URLs and the report in order to gather the relevant information. The graph also shows the relatively frequent interruptions of e-mail, which is known to be very common for knowledge workers [3].

4. DISCUSSION

We encountered a few challenges in processing and analysing our dataset. First, the data that we collected is rich and comes from multiple sources. We found that combining the data from the keylogging software and the browser history software was not trivial, even with exactly matching timestamps. This was because the user could have multiple tabs active in the browser, with not all tab titles being separately recorded by the keylogging software. Second, users clicking one of Google's query suggestions sometimes led to incomplete queries and missing URLs. For example, we found that the query 'napol' lead to the URL http://nl.wikipedia.org/wiki/ Napoleon_Dynamite. We suspect that his happened because the user selected the suggested query 'napoleon dynamite' after the offset 'napol',

Table 4: Excerpt of the processed interaction data, focussing on a series of queries and the clicked web pages for those queries. Note the spelling errors in the queries that apparently not lead to information finding problems. The URL http://cm.g.doubleclick.net/push illustrates that in some cases, noise from advertisement was logged instead of the containing web page.

Query	Clicked URL	Time on page	Key	next application
the life of napole	http://en.wikipedia.org/wiki/Napoleon	19 seconds		iexplore
the life of napoleon bonaparte	http://en.wikipedia.org/wiki/File:Napoleon_in_His_Study.jpg	8 seconds	CTRL+C	WINWORD
how to write a biography	http://homeworktips.about.com/od/biography/a/bio.htm	31 seconds		WINWORD
facts about napoleion	http://www.sheppardsoftware.com/Europeweb/factfile/Unique-facts-Europe10.htm	11 seconds		iexplore
facts about napoleon bonaparte	http://cm.g.doubleclick.net/push	7 seconds		WINWORD
family tree napoleon bonaparte	http://www.genery.com/sites/all/themes/gen2/images/screen/bonaparte.png	3 seconds		iexplore

4: Transition graph for the information behaviour of the workers. The size of the states represents their relative frequency. The state 'query' represents events where the active application is the web browser, in which a Google query and its results are shown. The state 'Google' represents events where a Google page is active without query. The state 'OtherURL' represents events where the active application is the web browser, with another URL than Google. The transition probability between states S_1 and S_2 was calculated as $count(S_1 \rightarrow S_2)/count(S_2)$. Only transitions with a probability > 0.1 are shown.
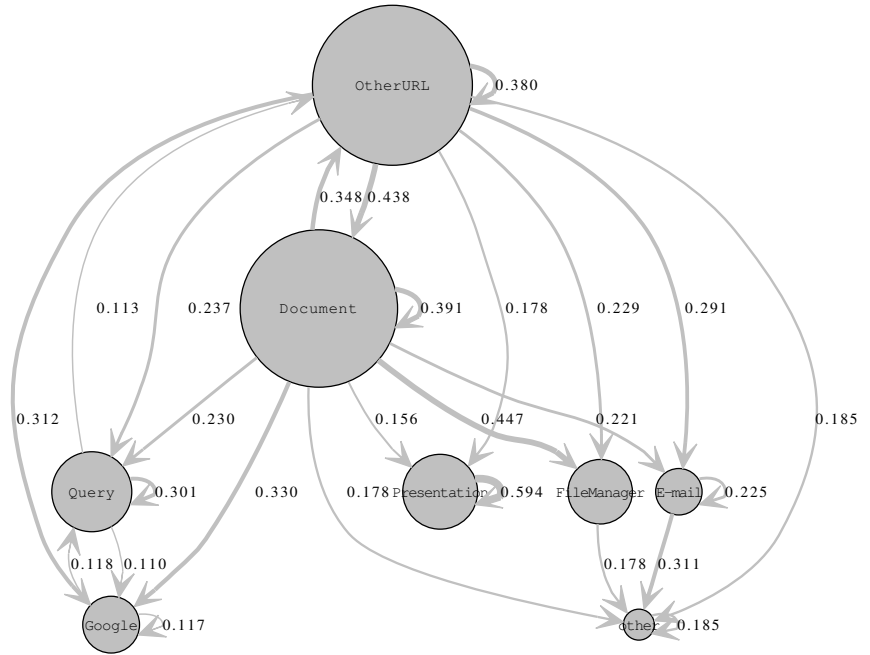
and then clicked the Wikipedia URL. Third, in some cases the window title of the browser did not change when a user clicked on a result (especially when the click was a result from Google Images), which caused the clicked URL to be included in the same event block as the query, and dwell time was missing for this particular URL. Fourth, the browser logging resulted in a lot of noise. We had to filter out a large amount of on-page social media plugins, advertisements and icons. In addition, browsing the Google domain leads to many additional URLs. An extreme example was 25 occurrences of the Google Maps URL http://maps.google.nl/maps?hl=nl &q=usa&bav=on.2 in one event block.

5. CONCLUSIONS AND FUTURE WORK

We collected and preprocessed a dataset of information behaviour of knowledge workers in the context of realistic work tasks. The data set is relatively small in terms of the number of participants, but large in terms of types of information collected. The contributions of this dataset are: 1) it includes different types of data, including keylogging data, desktop video recordings and browser history; 2) the information seeking behaviour is completely natural because it results from the recording of user behaviour during report writing, e-mail reading and presentation preparation; and 3) the search activities have been recorded together with the context of these other tasks, which allows for future research in context-aware information retrieval.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the 21st international conference on World Wide Web*, pages 569–578. ACM, 2012.

[2] S. Koldijk, M. Sappelli, M. Neerincx, and W. Kraaij. Unobtrusive monitoring of knowledge workers for stress self-regulation. In *Proceedings of the 21th International Conference on User Modeling, Adaptation and Personalization*, 2013.

[3] S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, pages 276–283. ACM, 1996.