Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Medieninformatik

Detecting Missions in Personal Web Archives

Bachelor's Thesis

Ludwig David Lorenz

1. Referee: Prof. Dr. Benno Stein
2. Referee: PD Dr. Andreas Jakoby

Submission date: January 28, 2024

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, January 28, 2024

.....
Ludwig David Lorenz

Abstract

"Missions" are a well-established idea in the segmentation of search query logs. This thesis aims to adapt this idea as a foundation for mission detection in personal web archives within the context of personal knowledge management. Therefore, a personal web archive is recorded and annotated. An annotation interface is developed and to address the problem of granularity in mission detection a hierachical mission annotation is proposed. On simple detection algorithms it is shown how features from query segementation can be used in personal web archives and how possible detection algorithms could be evaluated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The accelerated increase of available media and an economic shift towards information technology let historians coin a phrase for the recent decades: Information Age. [1] The stone, copper and iron age were named after the material the most influential tools was made of. What, then, is the transformative tool of *our* time, that is crafted from the essence of information?

The world wide web is one obvious answer. Started as a distributed storage for research, it became an interactive digital resource and economic platform. It is a great example of a tool that allows multiple parties, ranging from individuals to institutions, to collectively aggregate information and make it accessible. However, when it comes to a single individual, knowledge management is not a part of most's agenda.

Personal web archives can keep track of the information found on the world wide web. This is achieved by indexing and saving every website accessed by the user as demonstrated by Kiesel et al. [2018]. The information stored in a personal web archive gains in utility, if combined with other personal information. Therefore, before the possibilities of personal web archives and the detection of missions in them are explored, a brief look into the past is taken to understand the evolution of personal knowledge management and the concept of the personal knowledge base.

[1] Definition Information Age - Merriam Webster, visited on 12th January 2024, https://www.merriam-webster.com/dictionary/Information%20Age

Figure 1.1: Niklas Luhmann's personal Zettelkasten in the German city Bielefeld. *Source: Niklas Luhmann-Archiv, Fakultät für Soziologie, Universität Bielefeld, CC-BY-NC-SA 4.0*

1.1 Introducing Personal Knowledge Bases

A Personal Knowledge Base (PKB) represents a contemporary approach to managing and enhancing individual knowledge and information. In essence, a PKB is a, most commonly digital, repository that individuals use to collect, organize, and retrieve information relevant to their personal lives. This concept is not entirely new; rather, it has evolved in response to the changing dynamics of information consumption in the digital age. Historically, individuals have employed various methods to document and organize their knowledge. From commonplace notebooks and journals to more sophisticated personal libraries, people have sought ways to retain information for future reference. With the emergence of digital technology more radical approaches to personal knowledge management were introduced. The Memex, conceptualized in 1945 by Vannevar Bush, was a hypothetical personal knowledge device that allowed a user to link knowledge records similar to a hypertext system. From 1952 until 1997 the German sociologist Niklas Luhmann collected his knowledge written on little cards in drawers called a "Zettelkasten" that he manually linked with each other.

With over 90,000 cards it remains a mystery how Luhmann was able to recall the already existing cards when linking them to a new one. Luhmann [1981] explained that the "Zettelkasten" is not only a knowledge archive but instead an active tool of thought. It is worth to note that while Luhmann focused on the management of knowledge, others had a broader scope. The American system theorist R. Buckminster Fuller attempted to document his life as thorough as possible by collecting correspondence, bills, notes and sketches from 1917 to 1983. The collection, named "Dymaxion Chronofile" [Li et al., 2010], contains over 140,000 pieces of paper. Those extensive examples of Knowledge Management are passion projects of pioneers, while the emergence of digital technology has significantly transformed these practices, giving rise to the modern concept of Personal Knowledge Bases and making it more accessible and usable to everyone.

Methods for personal knowledge management evolved. Advanced note taking software like "Obsidian.md" allows users to manage a "Zettelkasten"-like Personal Knowledge Base by indexing, memorizing the existing notes and backlinking notes for them. With the exponential growth of online content, individuals find themselves overwhelmed with information from diverse sources. From the users perspective, Personal Knowledge Bases offer a central platform for the aggregation and organization of information scattered across different platforms, apps, and devices. As more and more information sources develop, individuals also face difficulties in recalling and connecting relevant pieces of knowledge. PKBs serve as a cognitive aid, helping individuals structure and categorize information according to their unique needs. While curating information that aligns with their specific needs and preferences, users may also be able to improve their own learning and recall ability and opening up the possibilities of life long learning.

1.2 Knowledge Bases as an Extended Mind

When your note helps you rediscover knowledge and make links to previous experience should your notebook receive credit for this cognitive process? The Extended Mind theory [Clark and Chalmers, 1998] challenges the traditional notion of the mind as being confined to the boundaries of the skull, suggesting that cognitive processes can extend beyond the brain and include external tools and artifacts. Personal Knowledge Bases form a coupled system with it's user and act as an external repository for information and knowledge offloading cognitive processes to external objects. The integration of digital tools, such as note-taking apps, calendars, and web archives, into the cognitive workflow allows individuals to access information beyond what is stored in their biological

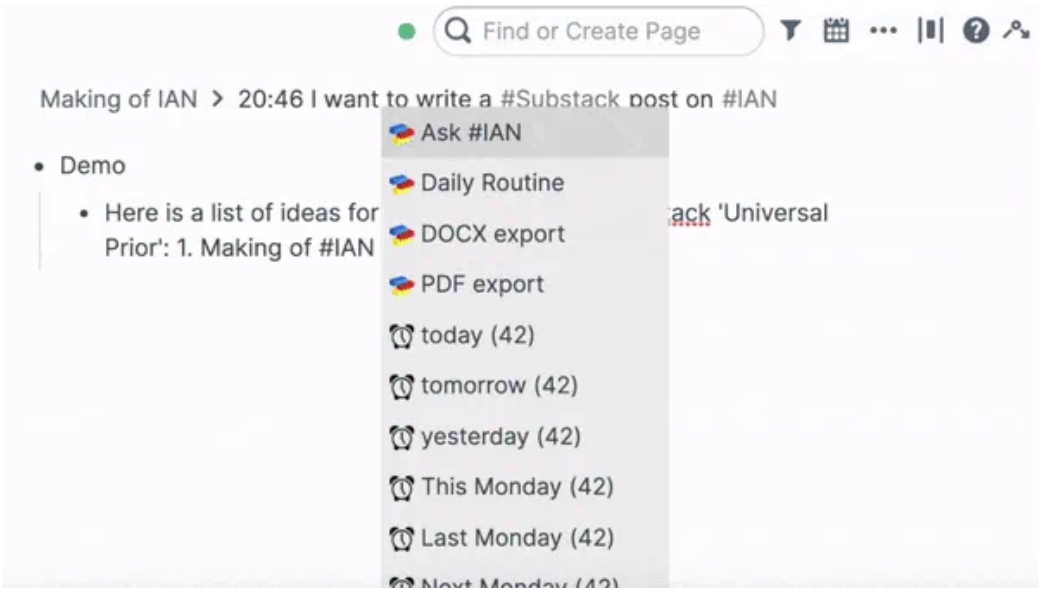

Figure 1.2: Demo of IAN, fine-tuned LLM on personal notes. *Source: Jan Hendrik Kirchner via Substack*

memory. This suggests that our cognitive abilities are depended on the tools we use, and the boundary between internal and external cognitive processes becomes blurred. Further, emerging technologies like machine learning on personal information open up the question if not only the access of information but even the creation of knowledge can be offloaded.

Experiments like #IAN, a fine-tuned large language model trained on a personal knowledge base show how such systems can be aware of the users experience and the context of the current mental process connected to the note. It was created by Jan Kirchner in August, 2021. [2] One year later, in November 2022, the commercial knowledge management platform Notion introduced their AI assistant which generates responses partially based on the content present in the knowledge base. Kirchner explained how he was inspired by Holden Karnodsky's idea to imagine yourself as a digital person. Instead of delegating responsibilities and tasks to others instead they are given to a deployed digital version of yourself in a computer.

By linking personal knowledge bases to the Extended Mind theory, one can view these external information systems as integral components of an individual's cognitive system. This also infers that a removal or constraint on access to a personal knowledge base damages one mental abilities. Those mental abil-

[2]Demo of IAN, visited on 12th Jan 2024, https://universalprior.substack.com/p/making-of-ian

ities are then also dependent on how accurately the knowledge base represents the digital self. In awareness of the before mentioned technologies it seems promising to combine Buckminster Fuller's take on personal information collection with Niklas Luhmann's approach of personal knowledge management. In explanation, the digital traces that we leave behind in our everyday interactions can form a mental frame to improve the recording and linking of ideas and knowledge.

1.3 Life Logging extends Personal Knowledge

Montoya et al. [2016] extends the PKB definition by implementing a framework to integrate recorded personal data to the knowledge base. It is reasoned, that not only information but also personal knowledge is spread over emails, messages, contact lists, calendars, location histories, and many other types of data. While commercial systems unfold as data traps, more efforts to self-recording personal data were adressed by the life logging movement. In that context, Li et al. [2010] proposed the notion of personal informatics systems as those that help people collect personally relevant information for the purpose of self-reflection and gaining self-knowledge. This describes a feedback loop between user and system. The system tracks the action of the user and thereby helps the user to reflect on short-term and long-term goals. The identification of goals from recorded data is a relevant computational aspect.

1.4 Personal Web Archives

A possible type of digital traces can be found in our browser history. For the purpose of this thesis, I recorded my own web activity for one month, forming a Personal Web Archive. To connect single web requests with ideas and knowledge present in a Personal Knowledge Base it is necessary to assess the intent behind each request. In the context of query logs for search engines, the problem of mission detection is already introduced as the problem to segment a query log to sets of queries which are connected with the same user's mission. The literature review in Chapter 2 dissects the current state of mission detection, drawing parallels between search query logs and the unique challenges posed by personal web archives. The varied definitions of information intent within existing literature are highlighted, emphasizing the need for a tailored approach when transitioning from query logs to personal web archives. Additionally, the categorization of web genres provides a reference, offering a comprehensive understanding of the diverse uses of websites. Chapter 3 introduces a conceptual framework which proposes a mapping of

mission detection within browser history, as well as outlining the adaptation of the cascading approach for mission detection in query logs. The methodology presented in Chapter 4 describes the technical details of recording the dataset, the annotation scheme and proposes an evaluation strategy for the introduced detection algorithms. Chapter 5 explains the experiment and the results obtained therefrom. In Chapter 6, it is discussed how useful the mission concept in personal web archive analysis is and which biases are present in the experiment setup.

Chapter 2

Related Work

Search engine providers store the interaction between their users and the search engine in a query log, where the submitted query, the time and a user identifier is stored. Through analysis of query logs, providers gain insights on user's web search behavior and further improve the search. The segmentation of query logs is a well-researched field. [Gayo-Avello, 2009]

Spink et al. [2006] identified a multitasking pattern in user web search sessions. This results in frequent topic changes in the query log. Jansen et al. [2007] addressed multitasking in query logs by defining sessions from a contextual viewpoint as a series of interactions that address a single information need. Jones and Klinkner [2008] identified a hierarchy in user search tasks and proposes an analysis in short-term goals and longer-term missions. A mission is introduced as an extended information need.

Lucchese et al. [2011] introduced the task-based session discovery problem. A task-based session represents a subset of queries in a physical session that relate to the same task of the user. In a similar way Hagen et al. [2011] focuses on identifying logical sessions in a physical session. In a logical session all queries address the same information need, but in contrast to task-based sessions all queries must be consecutive. Hagen et al. [2013] extends the idea of logical sessions to missions. A mission is a set of logical sessions that address the same extended information need. Table 2.1 compares definitions of physical sessions, logical sessions, missions and the mission detection problem, if given, among selected contributions. The concept of logical sessions is characterized in related work by diverse definitions. All express a shared idea of consecutive queries with the same reason, however, this reason is motivated differently with phrases ranging from task or goals to information needs, addressed in detail in section 2.1.

Table 2.1: Comparison of varying definitions for physical sessions, logical sessions and missions

Author [Year]	**Physical Session** *consecutive queries in a time window*	**Logical Session** *consecutive queries that address the same need*	**Mission** *set of queries that address the same need*	**Mission Detection Problem**
Silverstein et al. [1999]	two consecutive queries are part of the same **session** if they are issued within a five-minute time window	-	-	-
Spink et al. [2006]	-	**session**: are consecutive queries of a user that may have multiple goals or topics.	-	-
Jansen et al. [2007]	**web search episode**: temporal series of interactions among a searcher, a Web system, and the content provided by that system within a specific period.	**session**: series of interactions by the user toward addressing a single information need.	-	-
Jones and Klinkner [2008]	**search session**: is all user activity within a fixed time window	**search goal**: an atomic information need, resulting in one or more queries. Goals can still be interleaved.	**search mission**: a related set of information needs, resulting in one or more goals.	-
Gayo-Avello [2009]	**searching episode**: refers to the actions performed by a particular user within a search engine during, at most, one day	**topical session** or just **session**	-	-
Hagen et al. [2011]	-	**query session**: consecutive queries a user submits for the same information need	**search mission**: larger user search missions like planning the next vacation comprise of smaller goals like booking a flight, searching a nice hotel, etc.	-

Lucchese et al. [2011]	**Time-Gap Session**: maximum sequence of continuous queries with no more time in between than a given time threshold	-	**Task-based Session**: a subset of a (not necessarily consecutive) queries in a time-gap session or performing a given Web-mediated task	**Task-based Session Discovery Problem**: the Problem of finding all task-based sessions in one time-gap session
Lucchese et al. [2013]	**query session**, **user session**, **search session**: a specific set/sequence of queries issued by a user while interacting with a search engine	-	**user task**: is a set of possibly noncontiguous queries occurring within a search session which relates to the same need / where issued to achieve specific tasks, differentiated in intra-user and inter-user level (based on Lucchese et al. [2011])	-
Wang et al. [2013]	**session**: a segmentation of the query log by a fixed timeout threshold	**In-session Search tasks**: maximum subset of queries in one user session that correspond to a particular information need (based on Lucchese et al. [2011])	**Cross-session Search tasks**: maximum subset of queries in the query log that correspond to a particular information need	**Cross-Session Search Task Extraction**: partition the query log in cross-session search task consistent with the user's underlying information need
Hagen et al. [2013]	**physical session**: consecutive queries in the query log with a time gap in between below a certain threshold	**logical session**: consecutive queries with the same information need within the same physical session	**mission**: queries with the same information need within the query log	**Search Mission Detection**: identify queries a user submits for the same information need
Hienert and Kern [2019]	-	based on Gayo-Avello [2009]	-	-
Lugo et al. [2020]	-	**task-based session**: consecutive queries that relate to the same information need	-	-
Fischer et al. [2021]	**physical session**: consecutive queries with a time gap below a specific threshold (based on Hagen et al. [2013])	**logical session**: consecutive queries with the same information need in one physical session (based on Hagen et al. [2013])	-	-
Yu et al. [2022]	**physical search session** (based on Hagen et al. [2013])	**logical search session** (based on Hagen et al. [2013])	**search mission** (based on Hagen et al. [2013])	-

2.1 Search Intent and Information Need

Silverstein et al. [1999] defined sessions for query logs from a purely time-based perspective. The query log was segmented in sessions where each consecutive pair of queries in a session was not longer apart than a fixed time gap. In a time-based session a user might submit queries for many different reasons such as "planning a trip" or "checking the news". These queries can be intertwined, caused by multitasking behavior of the user. It is useful to define sessions from a reason-based perspective, many different definitions were proposed in following research.

Spink et al. [2006] addressed multitasking in sessions and redefined sessions as consecutive queries with the same topic and goal. Jansen et al. [2007] defines search sessions as serving a specific information need. Jones and Klinkner [2008] establishes the idea of a search intent in the context of mission detection as an extended information need. In a similar way, Hagen et al. [2013] introduced a mission as the set of queries a user submits for the same information need. Broder [2002] instead argues that through focusing on the information need behind a query the actual intent of the query is often overlooked. Intent is classified not only as informational, but also navigational or transactional. (1) Informational queries aim at static resources of knowledge on the web that don't force any interaction of the user except reading. (2) Navigational queries aim to reach a certain site that the user already has in mind (e.g. searching "Wikipedia"). And (3) transactional queries aim to reach a site to conduct a further interaction such as shopping.

Yu et al. [2022] took queries with additional signals and behavior patterns to classify missions with Broder [2002] proposed search intent categorization. It is stated that traditional information retrieval techniques focus on the information need of the user while the actual intent also impacts the relevance of a result. The search intent is assessed separately from the informational need. As one example the query for the German concert house "Elbphilharmonie" is submitted either with a transactional intent, i.e. to buy tickets for a venue, a navigational intent, i.e. to reach the specific website "`https://www.elbphilharmonie.de`", or an informational intent, i.e. to learn more about the "Elbphilharmonie". The difference between information need and informational intent is not stated clearly. Jansen et al. [2008] defines in a comprehensive literature review informational intent as "when a user addresses an information need, desire or curiosity". Following this notion, an information need implicates an informational intent.

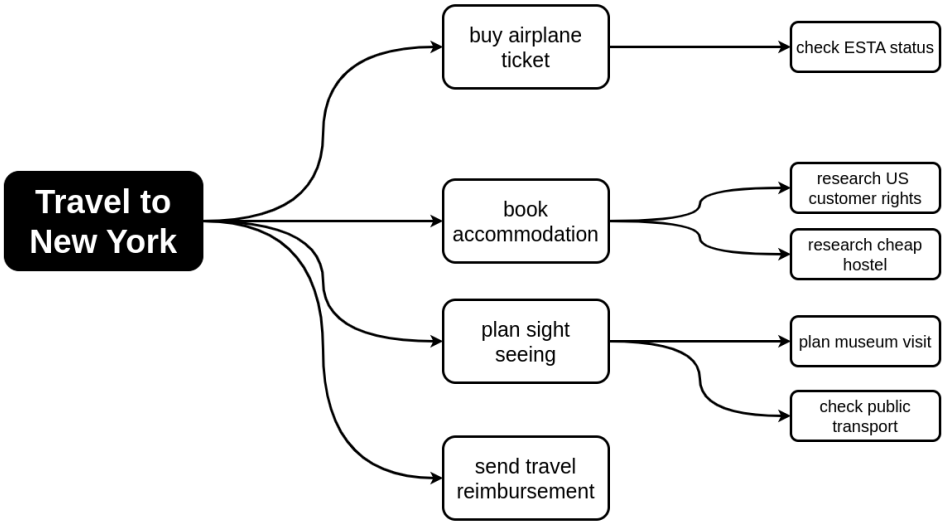

Figure 2.1: Tree Diagram to illustrate hierarchic tasks with the example of planning a visit to New York.

Another discussed aspect of search intent and information need is their granularity. As illustrated in figure 2.1 a task can be defined on different granularity levels and might contain subtasks. To address this in query log segmentation, two level of granularity have been defined so far: Logical Sessions and Missions. [Gayo-Avello, 2009]

2.2 Algorithmic Detection

2.2.1 Detecting Logical Sessions

Jansen et al. [2007] proposed an algorithm to detect a topic change in query logs. Between two consecutive queries, a topic change is detected if none of the following patterns apply: (1) Assistance, i.e. correction of previous query with search engine suggestion (2) Content Change, i.e. identical query on a different content collection like news or images, (3) Generalization, i.e. same topic as previous query but more general, (4) Reformulation, i.e. same topic as previous query and both queries contain common terms, and (5) Spezialisation, i.e. same topic as previous query but more specific.

Gayo-Avello [2009] compared the approach by Jansen et al. [2007] with a similar approach by He et al. [2002] and concluded on following patterns for logical session detection: (1) Repetition, i.e. both queries are identical, (2)

Specialization, i.e. terms have been added to the query, (3) Generalization, i.e. terms have been removed from the query, (4) Reformulation, i.e. terms have been added and removed from the query but the topic remains the same, and (5) New, i.e. the query is a different topic.

He et al. [2002] points out that a change of topic does not necessarily imply a change of the information need. With a combination of lexical and temporal features logical session breaks were detected. If the calculated value of these features exceeds a certain threshold, the record is marked as a session break by the algorithm. Özmutlu and Çavdur [2005] identifies, that this technique strongly depends on the variables and thresholds that were chosen for the algorithm. Variables that fit a certain query log are not necessarily best for another one.

Gayo-Avello [2009] proposed the geometric method that computes both, a temporal and a lexical distance. These values define a point in 2D space where a topic continuation area is defined. Hagen et al. [2011] proposed a cascading method that first applies cost-efficient features and if they are insufficient for a decision continues with more complex features: (1) The consecutive queries are analyzed for the query reformulation patterns described by Gayo-Avello [2009]. (2) The geometric method is applied. (3) The queries are compared through explicit semantic analysis. (4) The search results are compared.

The cascading method was updated by Hagen et al. [2013] so that the query log is (1) first divided in physical sessions with a time threshold of 90 minutes, (2) the query reformulation patterns are tested on continuous queries, (3) an enhanced test for lexical similarity and time inspired by the geometric method is applied, (4) the semantic similarity using explicit semantic analysis is computed, (5) the semantic similarity using Linked Open Data is computed, and finally (6) the search results are compared. Hagen et al. [2013] also point out that lexical similarity is not sufficient to detect a topic change as illustrated by following example. The two queries `istanbul archeology` and `constantinople` relate to the same information and share no lexical similarity.

2.2.2 Detecting Missions

Hagen et al. [2013] applied an adapted version of the cascade on logical sessions instead of queries to identify missions in a query log. The query log is first divided in physical sessions, which are then divided in logical sessions. The last record of each logical session is then compared with a sligthly modified cascading approach to the first query of every other logical session. If a continuation is detected both sessions are assigned to the same mission. It is also

adressed that not every query from every logical session are compared to each other. This is justified by the required cost effectiveness for an online scenario and conducted experiments suggest that the comparison of the last with the first query is often sufficient.

Lucchese et al. [2011] defined the Task-based Session Discovery Problem. Similar to a mission a task-based session is a set of not neccessarily contiguous queries. To evaluate the effectiveness of the approach the F_β score, Rand Index and Jaccard Index were calculated between the predicted task-based sessions and a ground truth.

Chapter 3

Theoretical Model

To adapt the problem of mission detection to personal web archives, following definitions are proposed. $\langle x_1, ..., x_n \rangle$ with $\forall i : 1 \leq i < n \longrightarrow x_i \leq x_{i+1}$ denotes an ordered set with n elements.

A minimal query log (Gayo-Avello [2009]) contains typically a unique identifier for the user or the session, the query string and a timestamp. Additionally, the result page number and the URLs clicked, if any, for each query are provided. Lucchese et al. [2011] denotes with $\mathcal{QL}$ a web search engine query log of the queries submitted by a set of users. In a similar way, a visit log for a single user is defined:

Definition 1 (Visit v & Visit Log $\mathcal{VL}$) *A website visit v is defined as a tuple $v = (t(v), URL(v), c(v))$ that contains the time $t(v_i)$ of access, the $URL(v_i)$ and the content $c(v_i)$ of the visited website.*
A visit log $\mathcal{VL} = \langle v_1, ..., v_n \rangle$ is an ordered set with respect to $t(v)$ for a single user.

Query logs ($\mathcal{QL}$) and Visit logs ($\mathcal{VL}$) both record user interactions in the web. Query logs focus on user requests to search engines and are created for understanding user search behavior, refining search algorithms, and assessing the efficacy of search results. (Gayo-Avello [2009]) Query logs are recorded on search engine provider's side for multiple users.

Visit logs contain user requests to any web page and record when which content is accessed. Visit logs are introduced to study the concept of missions beyond the scope of a search query, also recording any interaction that happens after the search results page. In our case, visit logs are recorded on the client's side and therefore only contain records that belong to one user. Since all search engine interactions are also recorded in the visit log, the user's query log can also be reconstructed from the visit log. On a single user level, the

query log is a subset of a visit log.

Similar to the time-gap session defined by Lucchese et al. [2011], a physical session is defined:

Definition 2 (Physical Session ϕ) *Let t_ϕ the maximum time gap threshold. The ordered set of consecutive queries $\phi_k = \langle v_s, v_{s+1}, ..., v_e \rangle \subseteq \mathcal{VL}$, with $s \leq e$, is said to be a physical session if it holds that:*

1. $\forall j : s \leq j < e \longrightarrow t(v_{j+1}) - t(v_j) \leq t_\phi$
2. $\nexists \phi_g \subseteq \mathcal{VL} : \phi_k \subset \phi_g$, *(there is no other physical session that contains ϕ_k and additional visits.)*

All physical sessions denoted as Φ form a disjoint partitioning of $\mathcal{VL}$.

The classification of intent by Broder [2002] also applies to a user visiting a website. Searching is one of many activities the web is used for. Rehm et al. [2008] divides web pages in different genres that hint on their intended use case. For example, genres like "Game", "Drama/Play" or "Pornographic" have entertaining purpose. Genres like "contact form", "discussion group" or "shop" aim at an interaction between the user and the web page. These provide an endpoint for the transactional intent in search queries. It must be noted, that the primary purpose of the website does not neccessarily define the user's visit intent. For example when the website was visited by accident e.g. searching "Google" on Bing. In the context of visit logs the term "information need" is not able to capture all intents for a visit. Therefore, in the proposed definition the terms "atomic task" or "short-term task" instead of "information need" and "task" or "goal" instead of "extended information need" are used. An atomic task is a simple task, that can not be divided in any subtasks.

Definition 3 (Logical Session λ) *The ordered set of consecutive queries $\lambda_{k,l} = \langle v_s, v_{s+1}, ..., v_e \rangle \subseteq \phi_k$ with $s \leq e$ is said to be a logical session if it holds that:*

1. $\forall i, j : (s \leq i, j) \wedge (i, j < e) \longrightarrow v_i$ *and* v_j *were issued for the same atomic task*
2. $\nexists \lambda_{k,g} \subseteq \phi_k : \lambda_{k,l} \subset \lambda_{k,g}$, *there is no other logical session that contains $\lambda_{k,l}$ and additional visits.*

All logical sessions in ϕ_k denoted as Λ_{ϕ_k} form a disjoint partitioning of ϕ_k. All logical sessions Λ_{ϕ_k} denoted as Λ form a disjoint partitioning of $\mathcal{VL}$.

A goal is defined by a related set of tasks that are executed to reach the goal. A task can belong to multiple goals.

Definition 4 (Mission μ) *Let $\lambda_i, \lambda_j \in \Lambda$ be any two logical sessions in the visit log. λ_i, λ_j are subsets of the same mission $\mu \subseteq \mathcal{VL}$ iff they were issued for the same task or goal. The set of all missions is denoted as M.*

Since $i = j$ is not excluded, each logical session belongs to at least one mission. The union of all missions results in the visit log. This definition differs from exisisting definitions in the choice that missions do not have to be mutually exclusive; in the same way that goals related to tasks are not mutually exclusive either. A task can serve multiple goals and a goal can be devided in subgoals.

Therefore, definition for the Mission Detection Problem has to be altered for compatibility with previous approaches on Mission Detection in query logs.

Definition 5 (Mission Detection Problem) *Find a partitioning of $\mathcal{VL}$ into $M' = \{\mu_1, ..., \mu_n\}$ missions.*

This constraints that (1) $\bigcup_{\mu \in M'} \mu = \mathcal{VL}$ i.e. the union of all missions results in the visit log and (2) $\forall \mu_i, \mu_j \in M' \wedge i \neq j : \mu_i \cap \mu_j = \emptyset$ i.e. each mission is mutually exclusive with the other missions. $M' \subseteq M$.

Chapter 4

Methodology

In the context of this thesis, the concept of logical sessions and missions in query logs were defined in the context of visit logs. In order to assess whether the proposed definitions are practicable, they will be used as a proof of concept in a small-scale trial study. The study is divided into four main steps:

1. recording a dataset
2. annotation of the dataset (ground truth)
3. performing automatic recognition using an algorithm
4. evaluation of the recognition with the ground truth

A dataset is required as the basis for this. Firstly, requirements for the dataset are defined and then a suitable procedure for recording the data is proposed. The dataset is annotated, an annotation scheme is proposed and an annotation interface is developed. Based on the annotated data, two simple recognition algorithms for logical sessions and missions are proposed. The results of the recognition algorithms are compared with the annotated sessions and missions using proposed evaluation measures.

4.1 Dataset Recording

For the recording of the dataset, requirements are first defined in comparison to query logs. The first section discusses the dataset specifics, focusing on a single-user visit log due to privacy and annotation concerns. It compares this log's 31-day duration to the AOL query log and highlights practical considerations.

The chosen time span aims to balance personal routines and the need for comprehensive data coverage. After that the recording method is explained.

4.1.1 Dataset Requirements

A visit log records all instances of a user's website requests within a designated timeframe. To accurately document a website visit, specific pieces of information are essential. This includes the timestamp indicating when the website was accessed, the URL of the website and the content of the web page in the form of an HTML document are fundamental requirements. The temporal span considered must be sufficiently extensive.

Query Logs that were explored in previous studies contain hundred of thousands of users. For this dataset the activity of only one person is recorded. This was decided due to an annotation-related and privacy-related consideration. In the context of annotation, it is questioned whether an annotator that is not responsible for the visits can actually assess the information intent behind the visit and label the visit with the correct logical session and mission. From a privacy viewpoint, a visit log contains sensitive information e.g. personal health information, banking transactions or personal account information. Therefore the annotation has to be conducted by the same person to avoid an invasion of the user's privacy. To ask multiple participants to record and annotate a visit log was not feasible in the context of this thesis.

The gold standard for query log analysis used by Gayo-Avello [2009], which was also used by Hagen et al. [2013], is the 2006 AOL query log containing 36.4 million interactions from about 650.000 users collected in 92 days. For the detection of missions the log has to include interactions of the same user from several days. The recording length for the visit log was set to 31 days, a third of the length of the AOL query log.

This time span appears practical in size to annotate. Notably, the visit log incorporates a user's query log, with all search queries made to search engines. Since my own browsing activity is recorded, personal observations also contributed to this decision. As I find myself returning to the same tasks after one day, a few days appear to be a good lower limit. At the same time, our everyday life consists of many routines and activities are repeated at fixed intervals, such as attending a lecture. The log therefore should also cover several weeks.

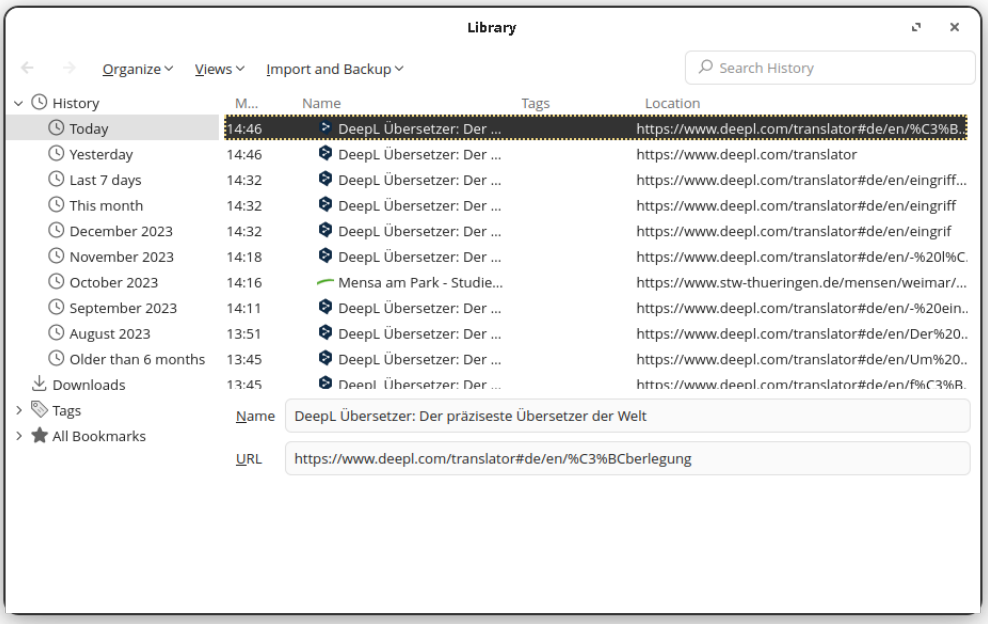

Figure 4.1: Screenshot of the browser history view in firefox.

4.1.2 Recording Method

The built-in browser history of many browsers, such as Firefox in figure 4.1, list all visited websites with URL and timestamp but do not store the content of the HTML page except the title. This is insufficient for a visit log. Fetching the content of the websites afterwards is no feasible approach, because the content of the website might be changed in between or no valid credentials to access the content can be provided anymore. The recording of the content therefore must happen directly when the user accesses the web page.

The proposed solution to this is the usage of a personal web archive proxy. In this scenario each request is forwarded by a proxy that also returns the response. The proxy server then handles the archiving of the requests and responses. A benefit of this method is that no additional software has to be installed on the users device. In theory the proxy can also handle multiple devices of the user in parallel. Such a proxy is part of the WASP prototype by Kiesel et al. [2018] for personal web archiving and search. WASP stores all requests and responses issued through the proxy in the standard Web archiving format WARC. From these WARC files, the content of the accessed web page is reconstructed.

In this study, the WASP application is installed on a remote server to record all HTTP requests and responses in the WARC format. All WARC records with the content type "HTTP responses" and with the HTTP content type "HTML" are extracted from the WARC archives. This drastically reduces the size of the dataset by removing any media like images and video streams. How useful information might be extracted from such media, e.g. through computer vision and transcription, should be adressed by future research. This study focuses on the HTML documents connected to the visited websites as the primary source of information for the detection algorithm.

4.2 Dataset Annotation

As mentioned in the previous section, the recorded user is also annotating the dataset. The annotation process is divided in two runs. In the first run, the logical sessions are annotated. In the second run, the missions are annotated. This approach is inspired by Hagen et al. [2013] who introduced a cascading detection based on first identifying logical sessions and then identify the missions out of the given set of logical sessions. Grouping the records first in logical sessions reduces the annotation expense for the missions. Instead of assigning individual records to missions, sessions can be assigned, that are already grouped by one information need.

In an example, the annotator annotates a wikipedia article about the "Spanish Inquisition" with the visit intent "to inform myself about the Spanish Inquisition". In the second round, the annotator groups multiple websites including the wikipedia article with the overarching mission "to write an essay on the development of moral institutions". The annotator relies on the website-specific visit intents from the first round which supports them to reflect on which other websites belong to the same mission.

For this purpose an annotation web interface was build with HTML and JavaScript in the front end and a Flask server in the back end to handle the annotation results. For the logical session annotation an overview page (see figure 4.2) presents all the physical sessions available.

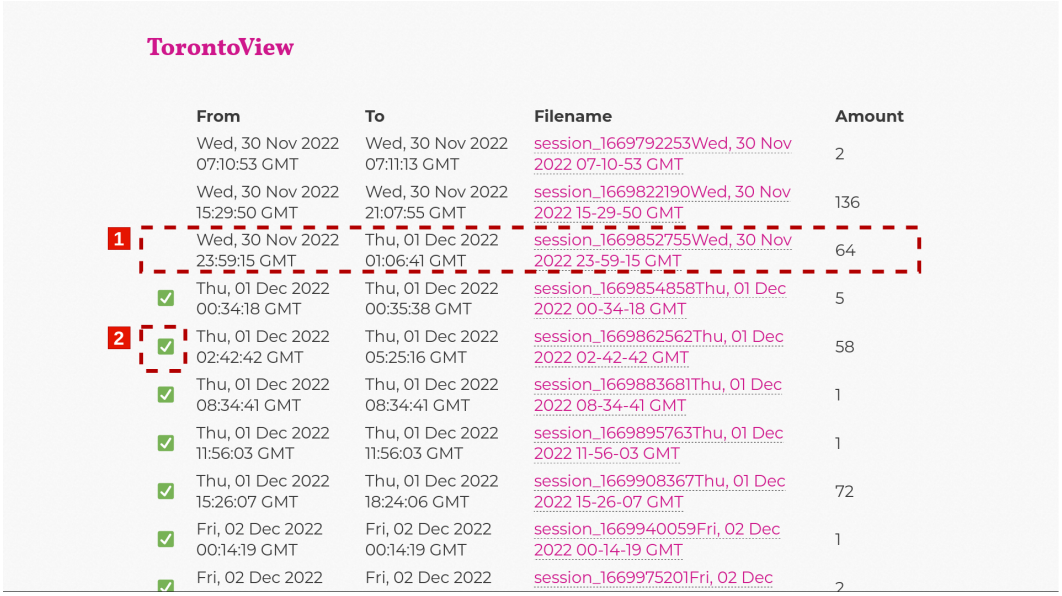

Figure 4.2: Screenshot of the physical session overview in the annotation interface. **(1)** Each physical session is displayed as a table row with start time, end time, filename with a link to the logical session annotation interface and the amount of visits contained in the session. **(2)** A little green check mark indicates wether the physical session was already annotated.

4.2.1 Annotating Sessions

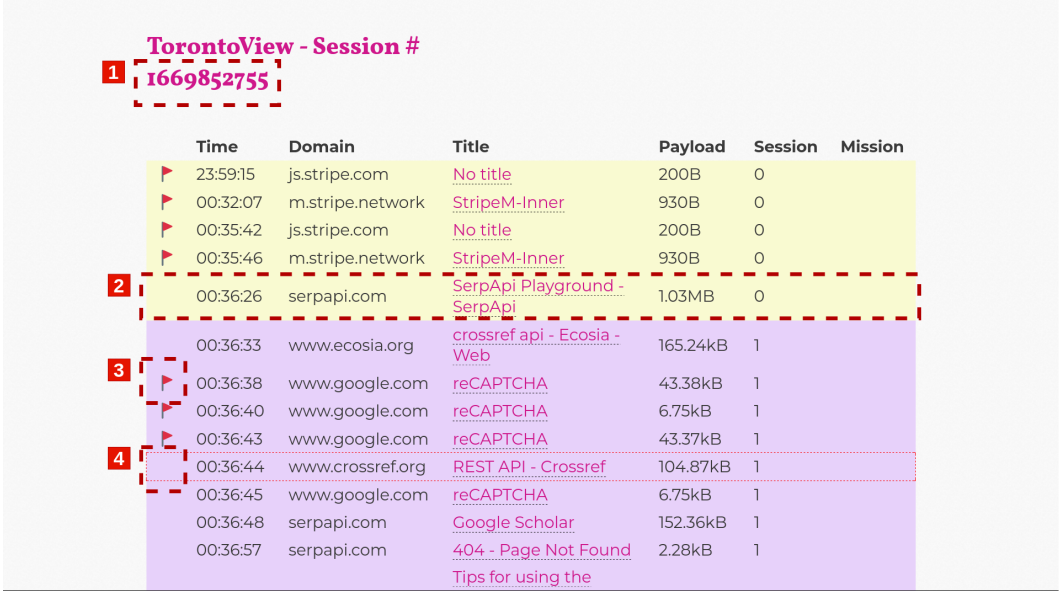

Figure 4.3: Screenshot of the logical session annotation interface. **(1)** The physical session ID is displayed at the top, **(2)** each visit record is displayed as a table row with time, domain, title, payload, associated logical session and associated mission, **(3)** a red flag indicates if a visit is marked as unintentional and **(4)** the current visit to annotate is highlighted with a red striped border.

The goal of the first run is to annotate logical sessions and remove unintentional visits from the dataset. Since hundreds, in rare cases even over thousand, of visit can be contained in one physical session, keyboard shortcuts have been implemented the speed up the annotation of a continuation with "c" and a break with "b". In addition "f" will annotate a continuation and flag the current visit as unintentional.

4.2.2 Annotating Missions

In contrast to previous work, the proposed definition for missions allows for the fact that a logical session can be contained by multiple missions. This decision was made to address the problem of task granularity in evaluation. By annotating hierarchical missions multiple degrees of granularity will be rewarded by the evaluation measure. The concept of a hierachical missun is illustrated in figure 4.4 where a mission can contain submissions and session. During annotation, multiple hierarchical mission trees will be created. This is displayed in the screenshot of the mission annotation interface in figure 4.5.

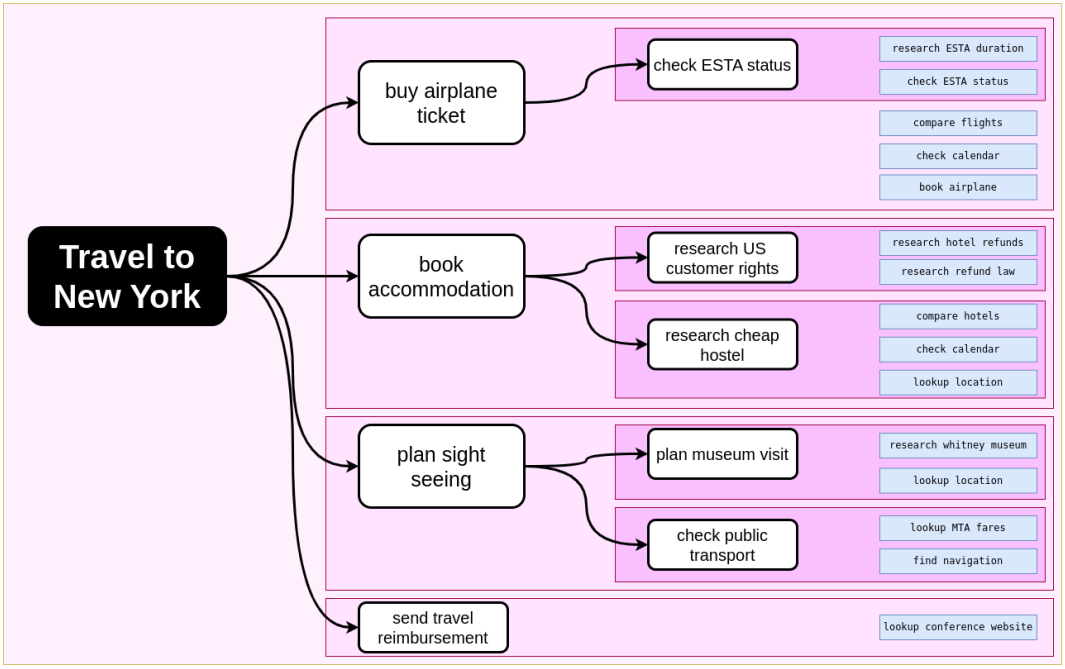

Figure 4.4: Example of planning a visit to New York with hierarchic Missions.

4.3 Algorithmic Detection

The study does not focus on complex heuristics for detection; rather, it aims to make reasonable initial choices to adapt the session and mission concept to visit logs. The decision to propose straightforward algorithms reflects a conscious approach. By keeping algorithms simple, the emphasis shifts towards a practical evaluation framework for future research. The primary focus is on providing guidance and ideas for potential enhancements and paths for future research.

Both, logical session detection and mission detection in query logs, use temporal, content-based and semantical features. In visit logs, the same temporal features can be utilized. [Lucchese et al., 2011] Content-based features are supposed to look at the lexigraphical level of queries. In visit logs a single visit is not associated with a query. However, methods are proposed to extract keywords from the website content associated with the visit that can be used for a content-based comparison. Semantical features are not proposed and remain an open task for future research.

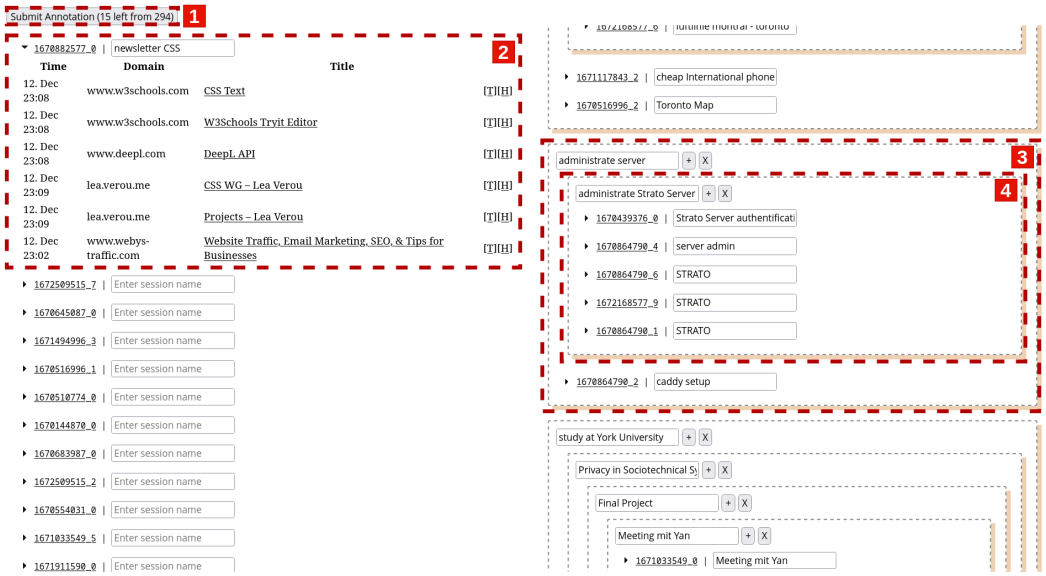

Figure 4.5: Screenshot of the logical session annotation interface. **(1)** A submit annotation button shows how many logical sessions are left to assign for a mission. These sessions appear as expandable draggable boxes on the left. **(2)** A box for a logical session start with the logical session id, which redirects the user to the related physical session annotation interface if clicked, and a little field to make notes during annotation. If expanded each intended visit of the logical session is displayed in a table with time, domain, title and two buttons that redirect to the result of the text extraction and the original HTML document. The box can be dragged to a mission on the right. On the right, multiple missions are stacked above each other. **(3)** A mission box contains a field for notes during annotation, a button to add a mission inside the current mission and a button to delete the mission if empty. **(4)** A mission box can also be inside another mission box and can be moved by dragging to each level of the hierarchy.

4.3.1 Detecting Logical Sessions

For the problem of logical session detection, time-based and content-based featured are considered. All introduced features are compared against each other in the experiment. The features represent basic building blocks used in previous work to define derived features or train classifiers [Jones and Klinkner, 2008]. The experiment shows how well conceptualy similar features work for the novel data class of visits compared to queries.

Time-based Feature

For the detection of logical sessions, multiple features are considered. Lucchese et al. [2011] defined the simple Timesplitting-t algorithm. Gayo-Avello [2009] used a simple time-based feature for the first splitting experiment. A purely time-based approach is often defined as a baseline. If these naive heuristics already produce acceptable results, more complex features are not cost-effective. [Hagen et al., 2011] Therefore a simple time-based approach is defined as the first feature.

Algorithm 1 Time-based detection with threshold t_λ

Require: $v_i, v_{i+1} \in \phi_k$, $\phi_k \subseteq \mathcal{VL}$
1: **if** $t(v_{i+1} - t(v_i) \leq t_\lambda)$ **then**
2: **return** 1
3: **else**
4: **return** 0
5: **end if**

Content-Based Features

In contrast to queries, lexical measures are not considered for the comparison of visits. Instead keywords are extracted from the visits out of two different sources: (1) Web page title and (2) the URL of the web page. For the web page title the title is split at every blank space to extract keywords. The URL is split at every "/", "-", "%20", "?", "q=" and "&" to extract the keywords.

For the comparison of keywords, two measures are proposed: (1) The existence of an intersection and (2) the Jaccard index. The existence of an intersection, i.e. $|A \cap B| > 0$, is the natural adaption of the reformulation pattern used by Jansen et al. [2007] for query detection. The Jaccard index is another measure used to compare terms of queries. ([Jones and Klinkner, 2008], [Lucchese et al., 2011])

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

For content-based features, the similarity of title-based keywords, URL-based keywords and the harmonic mean of both similarities is computed. In each case, existence of intersection and jaccard index is tested as a similarity function. This results in 6 features.

Link-Based Feature

A link-based approach results in a novel feature for the mission detection problem. Let $links(v)$ denote all links in the HTML markup of a web page. Two visits v_i, v_j are linked if:

$$|(links(v_i) \cap \{URL(v_j)\} \cup (links(v_j) \cap \{URL(v_i)\})| > 0$$

The feature is motivated by the self-observation, that when the user navigates on a web page, the task behind the web page visit does not change.

4.3.2 Detecting Missions

Similar to Hagen et al. [2013] it is asumed for the proposed mission detection algorithm, that all logical sessions have been correctly detected. A similarity function is defined for two logial sessions. In contrast to Hagen et al. [2013], the sessions are not merged by comparing only the first and the last item. While this has been shown to be efficient for query logs, as queries relate directly to an information intent, this may not be the case for visit logs covering the first and last visit in a logical session. Instead, an approach where each visit contributes to the similarity is favoured. Therefore, the content-based keyword similarity and linkage features are extended to the session level.

$$keywords(\lambda) = \bigcup_{v \in \lambda} keywords(v)$$

$$links(\lambda) = \bigcup_{v \in \lambda} links(v)$$

Given two logical sessions λ_i, λ_j, the keyword-based and link-based Jaccard index are computed. For the similarity function, both features are combined with a simple harmonic mean. Based on the similarity function a dissimilarity function is defined:

$$d_J = 1 - \frac{J(keywords(\lambda_i), keywords(\lambda_j)) + J(links(\lambda_i), links(\lambda_j))}{2}$$

The logical session together with the dissimilarity function, which is used as a distance function, forms a complete weighted graph. In this graph, related sessions are close to each other. To identify missions, these groups of sessions must be determined. This can also be seen as a clustering problem for which Jones and Klinkner [2008], Lucchese et al. [2011] or Wang et al. [2013] have proposed different unsupervised learning approaches. The more simple *QC-WCC* method by Lucchese et al. [2011], i.e. query clustering by weighted connected components, is selected in this experiment to merge logical sessions into missions. It should be noted that Lucchese et al. [2011] introduced the algorithm to compare queries and identify task-based sessions, which are similar to logical sessions as they contain queries. However, since the queries in a task-based session do not have to be consecutive and can be scattered in the physical session, the same concept can be used for clustering missions with an updated weight function.

Algorithm 2 Weighted connected components clustering with threshold c_λ

Require: $\Lambda \subseteq \mathcal{P}(\mathcal{VL})$
1: $V = \Lambda$
2: $E' = \Lambda \times \Lambda$
3: **for all** $\lambda_i \in V$ **do**
4: **for all** $\lambda_j \in V$ **do**
5: $d_J = 1 - \frac{J(keywords(\lambda_i),keywords(\lambda_j))+J(links(\lambda_i),links(\lambda_j))}{2}$
6: **if** $d_J > c_\mu$ **then**
7: $E' = E' \setminus \{(\lambda_i, \lambda_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: $M' = \text{Component-DFS}(V, E')$

Given all logical sessions Λ in the Query Log, the algorithm 2 starts by instantiating a graph with Λ for the vertices, $\Lambda \times \Lambda$ for the edges and $d_J : \Lambda \times \Lambda \longrightarrow [0, 1]$ as a weight function for the edges. An edge is removed from the set of edges, if the weight is above a given threshold c_μ. With the remaining edges, the components of the graph are extracted via a depth-first search algorithm 4. A component is a connected subgraph that is not part of any larger subgraph. The set of all components is returned as the set of predicted missions M'.

4.4 Evaluation

For the purpose of evaluation the problem of session detection and mission detection are considered independently. Similar in both cases is the comparing of a probe, the result of the detection algorithm, to a ground truth, the annotation by the user. But since both problems are modeled slightly different, different evaluation measures have to be applied to evaluate the difference between probe and result.

4.4.1 Evaluating Logical Sessions

Logical sessions are a segmentation of a physical session for both, visit logs and query logs. Therefore the same evaluation methods can be applied. Because no previous evaluation technique was tested on visit logs, multiple evaluation techniques from the related works have been selected for evaluation.

Precision and Recall

Precision and Recall are two properties of a binary classification system. Precision reflects the share of identified session breaks that are true session breaks. Recall represents the share of true session breaks identified by the model. F_β is a score that combines precision and recall where increasing β will increase the relative importance of recall over precision.

$$F_\beta = \frac{(1+\beta)^2 \cdot prec \cdot rec}{\beta^2 \cdot prec + rec}$$

In the session detection problem, it is aimed more for minimizing wrong session continuations (recall) than decreasing additional wrong session breaks (precision). This results in a higher β-value chosen for evaluation. In previous research $\beta = 1.5$ was chosen to emphasize wrong session continuations as the bigger problem compared to wrong breaks. ([He et al., 2002], [Gayo-Avello, 2009], [Hagen et al., 2011], [Hagen et al., 2013]) Both $\beta = 1.5$, and $\beta = 1$ as the standard measure will be evaluated in this study. The accuracy, as the fraction of correct classification divided by all classifications, will be evaluated as well.

Precision and Recall does not take into account how close a falsely detected break is to the true break. Therefore segmentations that nearly miss the true segmentation and segmentations that are far away from the true segmentation are penalized equally.

Windowdiff score

The P_k score introduced by Beeferman et al. [1999] is sensitive to near misses. It makes use of a sliding window of size k which moves record by record through the visit log. In each step for the two records at the end of the sliding window it is evaluated if the records of the probe segmentation (detected sessions) are in the same session or not. When the same evaluation comes to a different result for the two records in the reference segmentation (annotated sessions), an error counter increases.

Noteworthy weaknesses of the P_k score include a disproportionate penalty for False Negatives compared to False Positives, a sensitivity to variations in segment size, and an over-penalization of near-miss errors. WindowDiff, proposed by Pevzner and Hearst [2002], is based on the P_k score and improves the error metric by balancing out the penalties between false negatives and false positives.

4.4.2 Evaluating Missions

The effectiveness of the proposed mission detection algorithms depends on the degree of similarity between the predicted mission partition and the annotated missions of the ground truth. This asks to compare a set of sets that have no element in common, e.g. M', and a set of sets with intersections between them, e.g M. For this, the Jaccard Index, that was already used for mission evaluation [Lucchese et al., 2011], is altered to an evaluation function $E(M', M)$:

$$E(M', M) = \frac{1}{\sum_{\mu' \in M'} |\mu'|} \cdot |\mu'| \sum_{\mu' \in M'} (|\mu'| \cdot \max_{\mu \in M} J(\mu', \mu))$$

This evaluation function calculates the weighted contribution of each mission's best jaccard index with the ground truth. The same function can also be used to compare M and M' if M is also a partition of the visit log. In order to compare the evaluation measure, we extract a partition from M. Therefore each logical session λ is removed from a mission μ_i if there exists another mission μ_j with $\lambda \in \mu_j \wedge \mu_j \subset \mu_i$. The new set is denoted as M_L as in set of leaf missions. This is illustrated by the example mission in figure 4.6.

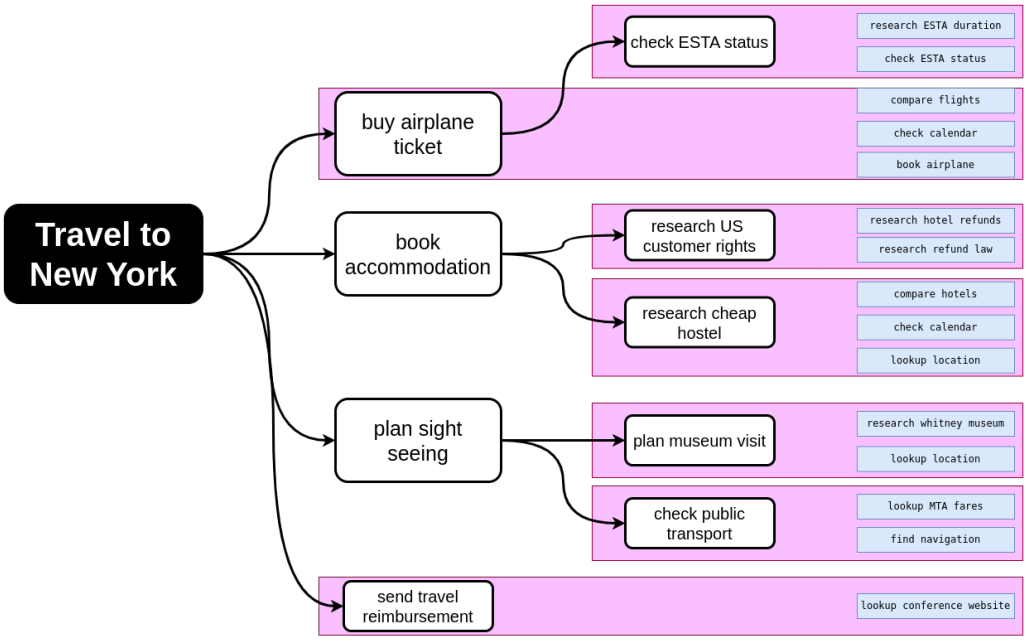

Figure 4.6: Example of planning a visit to New York with leaf missions.

Chapter 5

Experiment and Results

5.1 Dataset Recording

The dataset recording process involved the installation of WASP [Kiesel et al., 2018] as an archiving proxy on a personal virtual Linux server obtained from a hosting service. The server offered a network bandwidth of up to 100 MBit/s, which proved to be enough for the proxy to run with an unnoticed latency. The operating system in use was VPS Ubuntu 20.04 LTS 64bit, equipped with 4 CPU vCores, 8 GB of guaranteed RAM, and 300 GB of storage space, which was sufficient for the WARC records that summed up to be 15 GB in size after recording end. WASP was installed through Docker, and for secure HTTPS communication, it necessitated the trust of the certificate generated during the initial run of the personal WASP [1] instance. Notable difficulties were encountered when loading the extracted certificate onto mobile devices such as an Android smartphone and an iPad, while successful integration was achieved on a laptop. Therefore the idea of generating a visit log with multiple devices had to be postponed to future research. For the visit log in this study, one client was connected to the proxy with a self-signed certificate. The client was a Firefox Browser with version 106.0.1 at the start of the recording and version 109.0 at end of the recording running on Manjaro Linux x86_64.

The recording spanned from October 25, 2022, starting at 05:00:15 GMT, to January 24, 2023, concluding at 09:22:03 GMT. During the recording, I attended an exchange semester in Toronto, Canada. Therefore the local time-zone of the records in the visit log is Eastern Standard Time. The generated visit log is expected to resemble the ordinary life of an exchange student with activities ranging from studying to planning excursions and travel. It must be emphasized again, that not all web-mediated activities were recorded but only

[1] Further details and the codebase for WASP can be found on the GitHub repository webis-de/wasp.

those that were conducted through the Firefox browser on my laptop.

Table 5.1: Comparison of both datasets

	Recorded Archive	**Toronto Dataset**
First Record	Tue, 25 Oct 2022 05:00:15 GMT	Thu, 01 Dec 2022 00:34:18 GMT
Last Record	Tue, 24 Jan 2023 09:22:03 GMT	Sun, 01 Jan 2023 21:53:12 GMT
Recording Duration	91 days, 04:21 hours	31 days, 21:18 hours
Size in Records	504680	332806
Size in disk space	15 GB	12.3 GB
Physical Sessions	138	84
Logical Sesions	-	294

For the annotation and the experiment, from the raw recorded web archives the month of December was extracted. This month is used as the dataset from now on.

5.1.1 Filtering

During extraction, all HTTP requests that were not of content type `"text/html"` were removed. This reduced the dataset drastically in terms of file size and number of records as illustrated in figure 5.1. For the remaining WARC records that were HTTP requests, the most frequent domains were calculated to remove automatic and recurring Feed requests from the dataset. Since those were not intentionally issued by the user.

5.2 Dataset Annotation

Although the same person annotates the data that generated it, the annotation was a major challenge. Some pages, such as a "mushroom advent calendar", could still be remembered very well. Memories of other pages, especially those that were embedded in everyday routines such as study planning, quickly faded. Therefore, the annotation tool was constantly developed further during annotation, e.g. functions to display the extracted text or to render the HTML document were added when more features were needed for correct annotation. Even though in most cases, the title of the web page was sufficient.

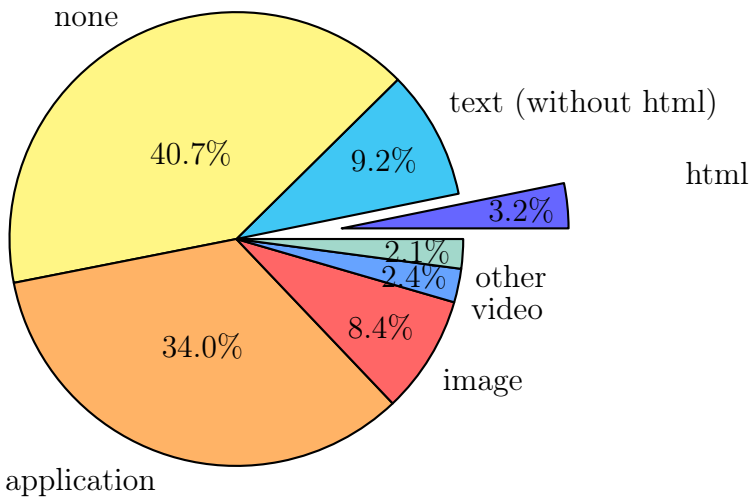

Figure 5.1: Pie Chart of `application/type` in HTTP responses by amount. The categories "binary", "font", "multipart", "Jpeg" and "audio" were combined in "other".

5.2.1 Annotating Sessions

Splitting into physical sessions and then annotating them individually was also advantageous from a practical point of view. This meant that the 5979 visits could be annotated in small portions. And it was more resource efficient for the program to load a small set of sessions rather than all at once. 84 physical sessions were further segmented in 294 logical sessions. 4561 visits were flagged as unintentional.

Filtering Unintentional Responses

The annotation interface displays all HTTP responses collected by the WASP instance during the dataset recording. This also includes HTTP responses that are no direct result of an intentional visit by the user. Such identified types of non-intentional responses are divided in foreground and background requests. Background requests are issued without a required interaction by the user such as RSS feed queries caused by the firefox plugin Livemarks that was installed during the recording (see physical session #1670275968). Not all feed queries were removed during the frequency based filtering before. Another type of background responses are requests caused by secondary resources loaded by a website. Non-intentional foreground responses are caused by pop-ups or forwardings that don't contribute to the user's information need. Such responses are advertisement that the user accidentally clicked on and log-in screens or

CAPTCHAs processed by the user in order to visit the actually intended site.

Responses lacking intentionality were identified and subsequently excluded from the annotation process for the experiment. Criteria for recognizing non-intentional responses that were identified during annotation include:

(1) CAPTCHA tests, such as those implemented by reverse proxy services like Cloudflare. (2) Response bodies of very small size, ranging between 0 and 3 KB. (3) Presence of "api" as a subdomain. And (4) HTML document titles displaying "No title," "Redirect," "Login," or "Loading" in the response.

Observing Session Entry Points

Throughout the annotation process, it became apparent that there are three primary entry points for initiating a logical session: (1) Search Engine Query or Browser Start Page, (2) Continuation from an Open Tab and (3) Forwarding from an application outside the Browser.

The second entry point neccessarily continues also the connected task in most cases, unless the opened page serves many different tasks as discussed in the next section. Unfortunately the recorded data does not include if the requested URL was from a tab opened. Also some tabs might still be cached in the browser. If so no request is issued and therefore no response recorded in the dataset. This is a flaw of the current method of recording that needs to be addressed in future research.

An interesting question for future research is how the proportions of entry points change with different device types. E.g. on a mobile device the third type might be more prominent due to different interaction paradigms like the frequent use of chat applications or the scanning of QR codes.

Handling Task Ambiguity

During annotation it was identified that several URLs reappear in a different information need. This proves that a simple mapping of URLs to information needs is no solutions to mission detection in visit logs. Examples for task-ambiguous websites are Ecosia, Google Maps, Deepl YorkU eClass, ChatGPT, the Wayback Machine. Deepl is a translation service. Translation tasks often reappear for different goals. (see physical session #1670439376). Ecosia is the primary used search engine that is often the start of a new logical session. Google Maps is used to look up locations during several different tasks. And YorkU eClass is the learning management of York university, the exchange university attended during the recording. The eClass portal, similar to moodle at Bauhaus universities, is used by all coursed taught at York University. All

tasks on the level of university courses can contain the YorkU eClass website. Another example are appointment surveys, that proved to be hard to correctly assign, because they often remain alone in their logical session. Appointment surveys fall in the category of the third entry point, they are send as a link and often opened from another external application.

Observing Information Need Digression

Table 5.2: Physical session #1670439376 excerpt

Time	Domain	Title	Content-Length
		⋮	
20:14:06	de.wikipedia.org	Tango (Standardtanz) – Wikipedia	70.31kB
20:49:46	www.ecosia.org	check einlösen - Ecosia - Web	142.13kB
20:49:59	www.ecosia.org	check einlösen wikihow - Ecosia - Web	114.47kB
20:50:06	de.wikihow.org	Liste der Genossenschaftsbanken in Deutschland – Wikipedia	251.83kB
20:51:55	www.ecosia.org	hypovereinsbank genossenschaftsbank - Ecosia - Web	186.73kB
20:52:14	de.wikipedia.org	Liste der Genossenschaftsbanken in Deutschland – Wikipedia	251.83kB
20:52:40	de.wikipedia.org	Bad Säckingen – Wikipedia	211.74kB
20:55:02	www.ecosia.org	fluglinie berechnen - Ecosia - Web	275.85kB
		⋮	

In comparison to query logs the users tendency to drift away from the original information need during browsing is apparent. This digression is caused by hyperlinks catering different aspects of the requested information. Physical session #1670439376 in table 5.2, as an example for this, contains a logical session initiated with searching how to redeem a check ("check einlösen" in German), the initial information need. The user then clicked on a WikiHow article that contains the technical term "cooperative bank" ("Genossenschaftsbank" in German). This term was unclear to the user and caused another search for the technical term. The user then clicked on a Wikipedia article with a list of all cooperative banks in Germany with the city of "Bad Säckingen". The user clicked on the link for the Wikipedia article of "Bad Säckingen" out of curiosity.

"Bad Säckingen" is a related aspect of the list of all cooperative banks in Germany. This list is a related aspect to "cooperative bank" which itself is related again to the question "how to redeem a check". Even though "Bad

Säckingen" and "how to redeem a check" do not seem related, they both belong to the same logical session. This aspect of digression is thought of as a potential challenge for mission detection algorithms that compare visit records instead of logical sessions to form missions.

Another example for digression in visit logs is apparent in physical session #1670725323 where the lookup of prices for an art museum initiated a search for the term "d/deaf" that was unknown to the user but mentioned on the prices section for reduced admission. In order to understand what the user has to pay at the museum, the initial information need, the user has to understand if they belong to the group "d/deaf" wich can be described as a conditional information need. Since conditional information needs are introduced by the initial information need, all responses related to these needs should belong to the same logical session.

5.2.2 Annotating Missions

Missions were annotated based on the previously annotated logical sessions. The 294 logical sessions were grouped in 75 missions. Only 22 of these 75 misions were root missions i.e. not contained by any other missions. A sunburst diagram of all annotated missions in figure A.1 is attached in the Appendix.

During mission annotation, it was observed that assigning logical sessions to topics is easier compared to assigning to a task. For example, many different tasks contained coding related logical sessions such as "look up delete tag in bs4 python". For the purpose of annotation, a two-phase approach turned out to be mentally easier. First, logical sessions were grouped by topic e.g. "Coding", after that, the topic was segmented according to the tasks e.g. "programming 2nd assignment foundations of digital media". It is important to note, that a general topic is not equal to a task. The same topic e.g. "Translating Services" can be found in different tasks e.g. "write an essay on personal information management".

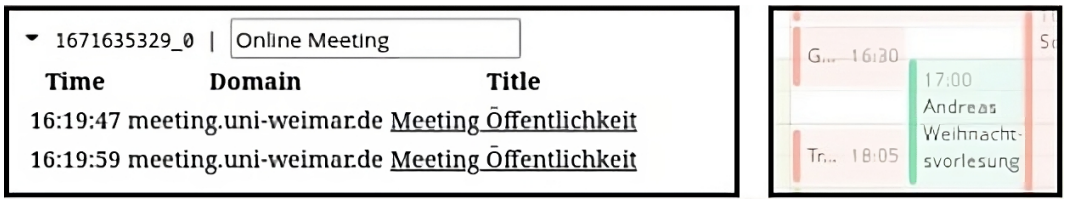

Figure 5.2: Left: Screenshot of mission annotation interface with the logical session "Online Meeting". **Right:** Screenshot uf user's calendar opened at the same time window with a green colored event named "Andreas weihnachtsvorlesung".

Also, many missions could not be created intuitively from the logical sessions alone. E.g. different topics that were researched online during "conversing with a friend" would not have been connected without the exclusive knowledge about the conversation by the annotator. This exclusive knowledge is a challenge for the detection algorithm. During annotation, the annotater made use of their calendar to assess the actual task that was connected to vague session descriptions as illustrated in Figure 5.2. Here, the atomic task was annotated as "Online Meeting" which is correct but does not leave any clues on why the user took part in the meeting. By looking at the user's calendar at the given timestamp the event with the name "Weihnachtsvorlesung" made clear that the logical session had to be assigned to the mission "study at bauhaus university".

5.3 Algorithmic Detection

A jupyter notebook was created for the implementation of the mission and session detection algorithms. [2] All visit that were flagged as unintentional have been removed from the dataset.

5.3.1 Detecting Logical Sessions

Time-Based Detection

The optimal time threshold t_λ was evalutated in the range from 1 to 500 seconds. In regards to the score, the optimal threshold is 109 seconds with $F_{1.5} = 0.636$. Figure 5.3 shows, that the precision rate, as well as the other evaluation measures, is quickly rising until 90 seconds. After that mark, the evaluation measures only improves slowly at the cost of recall. This might be an effect of Zipf's law, which can be observed in the dataset. Zipf's law states that in a decreasingly ordered list the n-th element is inversely proportional to n. Or in our case, in the first minute after a visit, a second visit occurs roughly twice as often compared to the second minute. Therefore, a low time threshold has a greater effect on the evaluation measure.

Table 5.3 shows the evaluation measures computed for (1) the time-based feature with $t_\lambda = 109s$, (2) if there is an intersection in domains, (3) if there is an intersection in URL-extracted keywords, (4) the jaccard index of URL-extracted keywords, (5) the intersection of title-extracted keywords, (6) the jaccard index of title-extracted keywords, (6) the intersection of both keywords

[2] The notebooks are linked in the GitLab repository of the thesis. `https://git.webis.de/code-teaching/theses/thesis-lorenz`

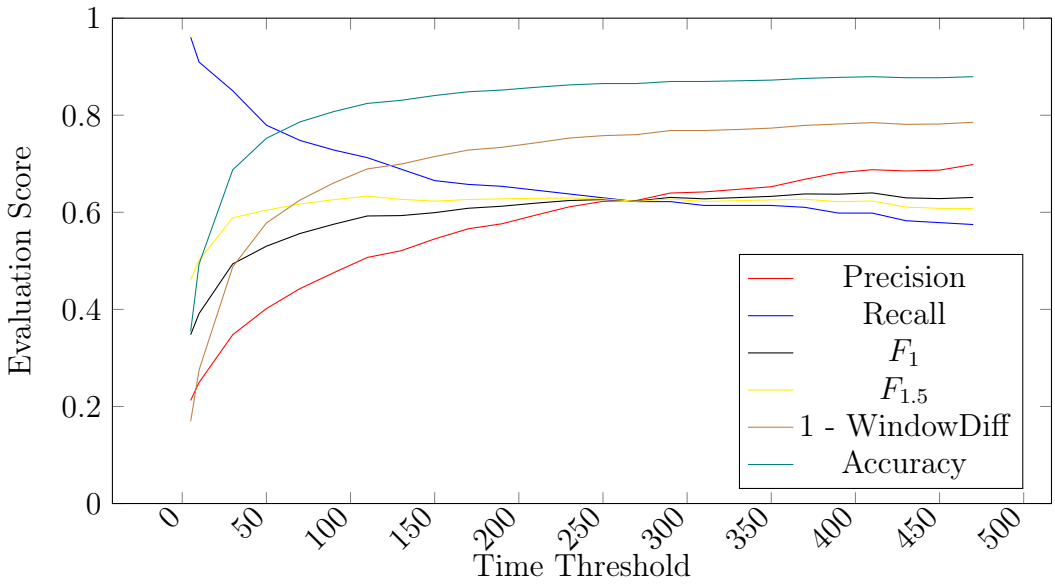

Figure 5.3: Line Chart with multiple evaluation measures in regards to the time threshold t_λ.

merged, (6) the jaccard index of both keywords merged and (7) if both visits are linked with a hyperreference. In all evaluation measures except recall, the simple time-based feature scores best. All content-based features have a higher recall rate, but a much lower precision rate.

Table 5.3: Computed evaluation measures for each feature. Best score per column is coloured in green. Worst score per column is coloured in red.

Feature	Precision	Recall	F_1	$F_{1.5}$	1 - Windowd-iff	Accuracy
Time	0.507	0.717	0.594	0.636	0.689	0.824
Domain	0.28	0.925	0.43	0.541	0.407	0.561
URL Keywords (Intersection)	0.281	0.972	0.436	0.554	0.375	0.55
URL Keywords (Jaccard)	0.215	1.0	0.354	0.471	0.179	0.347
Title Keywords (Intersection)	0.179	1.0	0.304	0.415	0.053	0.179
Title Keywords (Jaccard)	0.179	1.0	0.304	0.415	0.053	0.179
Joined Similarity (Intersection)	0.281	0.972	0.436	0.554	0.375	0.55
Joined Similarity (Jaccard)	0.215	1.0	0.354	0.471	0.179	0.347
Linkage	0.22	0.992	0.361	0.477	0.186	0.37

5.3.2 Detecting Missions

To detect missions, the logical sessions Λ from the annotated ground truth were loaded from the dataset. For each visit in each logical session the page title, title-based keywords, URL-based keywords and links in the HTML document were extracted. After implementing the distance function d_J, an adjacency matrix for all logical sessions was calculated. The algorithm 2 was implemented and tested with 10 different clustering thresholds c_λ evenly spaced in between 0.1 to 1.0. The resulting scores in the evaluation are plotted in figure 5.4. It was evaluated against the set of all annotated missions M and all leaf missions M_L. The evaluation scores are shown in 5.4.

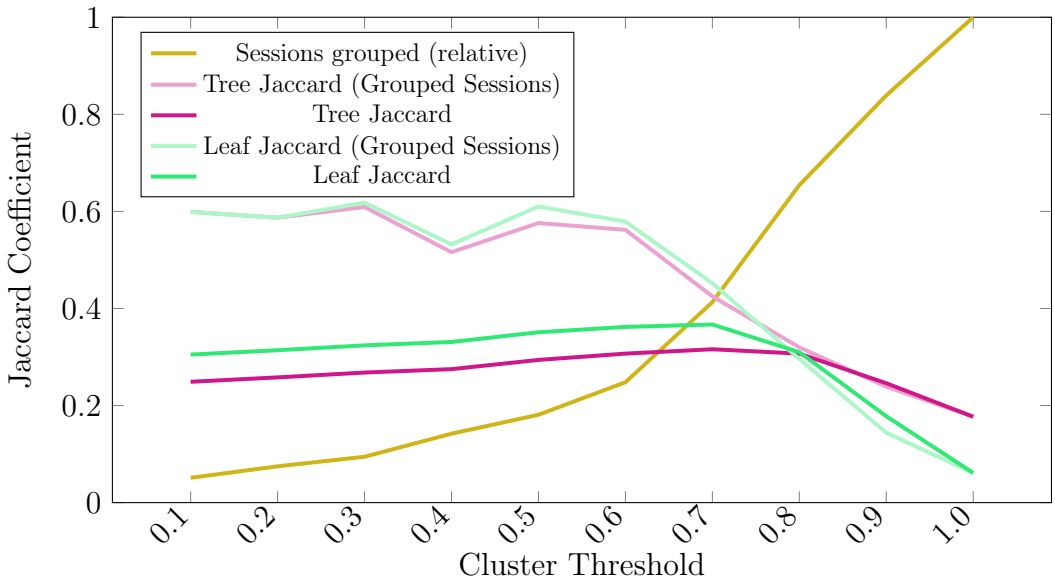

Figure 5.4: Line Chart with multiple evaluation measures in regards to the cluster threshold c_μ.

Table 5.4: Computed evaluation measures for different c_μ.

c_μ	$\lvert M' \rvert$	Session grouped (rel.)	J (Grouped Sessions)	J (Total Sessions)	J on M_L (Grouped Sessions)	J on M_L(Total Sessions)
0.1	5	0.0512	0.599	0.249	0.599	0.305
0.1	5	0.0512	0.599	0.249	0.599	0.305

0.2	7	0.0748	0.587	0.258	0.587	0.314
0.3	8	0.0945	0.609	0.268	0.618	0.324
0.4	14	0.142	0.516	0.275	0.532	0.331
0.5	17	0.181	0.576	0.294	0.610	0.351
0.6	22	0.248	0.562	0.307	0.579	0.362
0.7	30	0.413	0.425	0.316	0.452	0.367
0.8	27	0.654	0.320	0.307	0.296	0.310
0.9	9	0.839	0.239	0.246	0.144	0.178
1.0	1	1.0	0.177	0.177	0.061	0.0612

Chapter 6

Discussion

In this chapter the previous choices in adapting the mission detection problem to personal web archives are critically reflect. In the section on dataset recording the limitations of the current dataset is ilustrated in regards to the scope of recording and the sensitivity of the recorded data.

6.1 Dataset Recording

The dataset does not represent a complete personal web archive and can be improved. Other user devices can also be included in the data collection. This also results in other constraints for the definition of the visit log and the physical sessions. Furthermore, the current dataset does not record which pages the user has clicked on. This information could be important to improve the linking feature. Other interaction logs could also be collected, e.g. to provide an insight into what content on the website was viewed by the user.

The dataset contains sensitive personal information. In order to publish the dataset and annotation, it must be assessed how sensitive each website visit is to the privacy of the user as well as third parties related to the user. However, one advantage of the personal data is that gor the same timeframe, a calendar dataset can be extracted. The calendar dataset can be used together with the already annotated Toronto dataset to explore the problem of session and mission detection based on multiple personal data sources.

6.2 Dataset Annotation

During annotation, logical sessions were assigned to hierarchical missions. However, the mission definition introduced in section 3 also allows to assign

visits to multiple missions. This adds an additional layer for annotation and evaluation and was therefore disregarded in this experiment. The implications of this possibility should be further explored in future research. It would allow for cycles in the hierarchical mission annotation which would require to not only reimagine the annotation interface but also the evaluation measure.

The exclusive knowledge of the annotator, as explained in the example "conversing with a friend", is a natural boundary for the detection of sessions and missions. However, multiple data sources, like a calendar which provides the event "Meeting my friend" while the websites were visited, help to close the knowledge gap between annotator and detection algorithm.

6.3 Detection Algorithm

During the experiment, a simple session detection and mission detection algorithm have been evaluated. For both detection algorithms, the ground truth was only used for evaluation. The problem can also be redefined as a supervised learning task for future research. Therefore, the already annotated dataset can be split into a training set and a test set. To prevent overfitting, constraints need to be defined for the split. One reason for constraints: The tasks of the user depend on the time frame. On 21st of December the academic year ends and the tasks shift to a more travel-related focus. Splitting the dataset on a fixed date will result in overfitting.

As for now, the current detection algorithms score low in terms of f-measure and jaccard distance compared to related work in query log segmentation, where scores over 0.9 were achieved, ([Hagen et al., 2011], [Lucchese et al., 2011])

6.3.1 Detecting Logical Sessions

Based on the evaluation results for the selected features in table 5.3 it is noted, that all content-based features have a higher recall rate, i.e. more of the actual breaks have been identified, but are significantly worse in precision, i.e. identify more breaks than present. This suggests that in the dataset many continuations exist, where both visits do not share common keywords. This can be adressed, by proposing semantic features, assumed that both sessions share no lexical similarity but a semantic similarity. And it is also possible, that this is a result of the annotation process if logical sessions falsely combine two atomic tasks. This is hard to distinguish in annotation. For the next time, logical sessions should be annotated with a higher granularity. A promissing

approach for evaluating semantic similarity is the calculation of word embeddings for each set of keywords and the definition of a distance measure between two clusters in the vector space.

As expected, the jaccard index of the keywords performs worse in the F-score compared to testing if there is a keywords intersection. This is explained through the lower barrier for continuation in the second method, and therefore a higher precision through less breaks that are falsely identified. However, there is a small loss in recall compared to the jaccard index, possibly caused by task-ambiguous websites, e.g. a search engine, that share some content-based features but not the same task. This can be adressed by lowering the jaccard threshold c_J, currently set to 0.5, to improve precision or by improving the keyword extraction process. One possible approach for this is calculating the combined TF-IDF score for term frequency and inverse document frequency on the extracted keywords from the visit log or, more preferred if vailable, the annotated missions. This helps to identify keywords that are non-indicative for tasks and should be removed during keyword extraction. However, it should be noted that indication relates to the set of missions. E.g. the keyword "ecosia" is not indicative for all tasks, because it is the name of the primary used search enginge in the visit log. In contrast to that, the keyword "eclass" is indicative for the mission "study" but non-indicative for all submissions of this mission, becauses it is the name of the learning management system used by the university independently from the course.

The table 5.3 of evaluation measures also shows, that joining URL-extracted and title-extracted keywords does not improve the precision rate. On the selected dataset, there might be no knowledge gain in combining both features. Similarly, it should be asked if a combination of the other features can lead to a knowledge gain. Therefore, for the set of break indexes B_i, B_j, identified by each feature f_i, f_j, $|B_i \cap B_j| \div |B_i|$ should be computed as a measure for which features together identify different breaks. As a next step, all content-based features are to be combined in a content similarity function.

In previous research temporal and lexical features were combined as one of the most frequent measures for logical session detection. The geometric method proposed by Gayo-Avello [2009] with the updated time function by Hagen et al. [2013] could be chosen to combine content similarity of visits with their time difference. It should be ensured, that the content similarity high function scores significantly high for related visits. The similarity function defined in previous literature ([Jones and Klinkner, 2008], [Hagen et al., 2011], [Gayo-Avello, 2009]) relied on lexical similarity between query strings, which

is no appropriate feature for comparing keyword sets. The jaccard index used so far scores relatively low, due to the small intersection rate between keyword sets. Therefore, it is advised to conduct an investigation on the error frequency distribution, similar to Hagen et al. [2011], for the geometric method with the proposed content similarity function.

Algorithm 3 Combination of keyword similarity and time

Require: $v_i, v_{i+1} \in \phi_k$, $\phi_k \subseteq \mathcal{VL}$
1: $keySim = keywordsimilarity(v_i, v_{i+1})$
2: $f_{time} = 1 - \frac{t(v_{i+1}) - t(v_i)}{\phi_k}$
3: **if** $\sqrt{keySim^2 + f_{time}^2} < 1$ **then**
4: **return** 1
5: **else**
6: **return** 0
7: **end if**

The link-based feature faces some limitations. If two sites are linked, the user did not neccessarily visit the other site. This could be extracted with more effort from the visit log or by developing a custom browser plugin.

6.3.2 Detecting Missions

The distance function between two logical sessions is currently defined with the harmonic mean of keyword-based and link-based jaccard index. However, the proportion of both features is probably not ideal. Lucchese et al. [2011] combined the lexical-content $\mu_{content}$ and $\mu_{semantic}$ via a convex combination:

$$\mu_1 = \alpha \cdot \mu_{content} + (1 - \alpha) \cdot \mu_{semantic}$$

$\alpha = 0.5$ was chosen for simplicity in this experiment. In a second proposal, both characteristics were combined in a way that the second characteristic is only evaluated if the first characteristic has failed. In essence, a similar idea that was also used in the cascading approach by Hagen et al. [2011] and Hagen et al. [2013]. The distance fucntion can be adapted with this concept for a first basic improvement of the detection algoritm. A fundamental flaw of the algorithm itself is, that two highly connected subgraphs of vertices are detected as one component if only one edge remains after the edge pruning which results in two very distinct missions being merged because of one distant connection. This can be caused if two logical sessions contain intent-ambiguous websites like, translation services and search engines. For future research, other unsupervised learning algorithms should be considered.

The link-based feature faces some limitations. Since all the links are merged in one logical session, some links might falsely connect unrelated missions. Some websites are extensively linked like Google Ads, social media sites or the trusted shop logo. This effect can be prevented by merging URLs and links separately and comparing them against each other

6.4 Evaluation

6.4.1 Evaluating Logical Sessions

The windowdiff measure was proposed to address the issue of near break misses in session detection evaluation. However, the measure did not proved useful. According to the evaluation chart in figure 5.3, the measure scored almost equally to the $accuracy - c$ shifted by a constant c. It is assumed, that near misses are as probable as misses anywhere in the physical session. Since the windowdiff measure did not bring any knowledge gain in evaluation, it might be as well disregarded in future research.

6.4.2 Evaluating Missions

Two configurations of the ground truth were used for evaluation: (1) Annotated missions M, as illustrated in figure 4.4 and (2) leaf missions M_L, as illustrated in figure 4.6. The evaluation measures based on each dataset differ. Compared with leaf missions, evaluation with missions first results in slightly lower values for low c_μ. M_L contains more specific missions and lower c_μ will lead to more specific predicted missions. This is explained by how the leaf missions have been constructed. Missions that contain both, submissions and sessions, are split into their submissions and an additional mission with the direct sessions. This implicates $M_L \subsetneq M$ which is a counterintuitive result of this definition. To prevent this, additional missions can be annotated, such that all mission either have submissions. Also, compared with leaf missions, evaluation with missions results in higher values for higher c_μ. M contains more general missions and higher c_μ will lead to more general predicted missions. The evaluation based on hierarchical missions is useful in this case, because it can differentiate between a more general prediction and a simply worse prediction which was not possible with a mission partition of the log, like M_L , as ground truth. However, this can only be observed for high c_μ. For most c_μ the difference is not noticable.

Chapter 7

Conclusion

In conclusion, this thesis has comprehensively explored the concept of missions in web archives, aiming to lay the foundation to assess their utility within the framework of personal knowledge bases.

Chapter 1 introduced the idea of personal knowledge bases by giving Niklas Luhmann and R. Buckminster-Fuller as two examples for very different approaches. With covering contemporary approaches that make use of machine learning it becomes evident, that also personal data can contribute to personal knowledge management. In that context, the philosophical concept of the extended mind and the life logging movement are presented as two cultural aspects.

Chapter 2 reviewed the origin of mission and session detection. It gives an overview on how different notions of session and mission detection have been evolved. It also focuses on heuristics to detect sessions and missions. Chapter 3 formally defined the concept of sessions and missions for query logs. The proposed mission definition is less restricted compared to previous work to allow a higher degree in granularity. Chapter 4 explained why which steps are proposed to adapt the detection problem from query logs to visit logs. It also defines novel evaluation measures for both, session and mission detection. Chapter 5 illustrates the steps for a practical implementation. To evaluate the automatic approach for mission detection, a personal visit log was recorded and annotated. The dataset consists out of website records based of several WARC archives that were collected between 28th November 2022 and 20th January 2023. Each website record holds information about it's URL, content and the time when it was accessed by the user. After the end of recording, the provided information was used by the user to annotate each record with their website specific visit intent. In a second annotation round the user was

instructed to group the records with an extended visit intent. Chapter 5 also gives an overview about the several observations achieved during annotation and also state the evaluation results for the proposed algorithms.

Chapter 6 reviews the thesis contribution from a critical perspective. The conducted experiments cover diverse aspects, ranging from the adaptability of missions to the handling of granularity in tasks, to the practical details of experimental setup, dataset recording, annotation, and detection algorithms in the context of personal web archive. The consideration of privacy implications, the ideas for granularity respecting evaluation techniques, and the exploration of multiple personal data sources leave open ends for future research. Other aspects, like the proposed windowdiff evaluation measure, did not proved useful.

For future research, the identification of sessions and missions in personal web archives should be viewed as one component of a greater, more general, personal task and goal identification problem. In simple terms, figuring out sessions and missions in your personal web archives is like solving a bigger puzzle — it's just one part of understanding what tasks and goals you're aiming for in the digital and real world. By untangling this piece, we gebect a better picture of ouselves and what we want to achieve. For future research, the identification of sessions and missions in personal web archives should be viewed as one component of a greater, more general, personal task and goal identification problem. In simple terms, figuring out sessions and missions in your personal web archives is like solving a bigger puzzle — it's just one part of understanding what tasks and goals you're aiming for in the digital and real world. By untangling this piece, we become more aware of ourselves and what we want to achieve.

Acknowledgements

In regards to thesis I extend my heartfelt gratitude to **Johannes Kiesel** for his support and guidance throughout my thesis journey, which took a bit longer than expected. Because of that, I am so thankful for all of his patience and time, that he invested to help me find the right research topic and continue with my experiments.

I am also grateful for the inspiring interviews with **Magdalena Wolska**, **Jan Willmann**, **Franziska Klemmstein** and **Kevin Lang** that convinced me to pursue the research path for detecting missions in personal web archives.

Finally, I would also like to thank my parents and my friends **Olli**, **Linda**, **Agnes** and **Theo** who stayed by my side even during the dark days of academic life and always expressed their unwavering belief in me. Without them, I would probably study for another term.

Appendix A

Additional Material

Algorithm 4 Component splitting with depth first search

Require: $G = (V, E)$
1: **procedure** DFS(v, C)
2: $Visited = Visited \cup \{v\}$
3: $C = C \cup \{v\}$
4: **for all** $v_i \in V$ **do**
5: **if** $v_i \in Visited$ **then**
6: DFS(v_i, C)
7: **end if**
8: **end for**
9: **end procedure**
10: $\mathcal{C} = \{\}$
11: **for all** $v_i \in V$ **do**
12: **if** $v_i \in Visited$ **then**
13: $C = \{\}$
14: DFS(v_i, C)
15: $\mathcal{C} = \mathcal{C} \cup \{C\}$
16: **end if**
17: **end for**
18: **return** $\mathcal{C}$

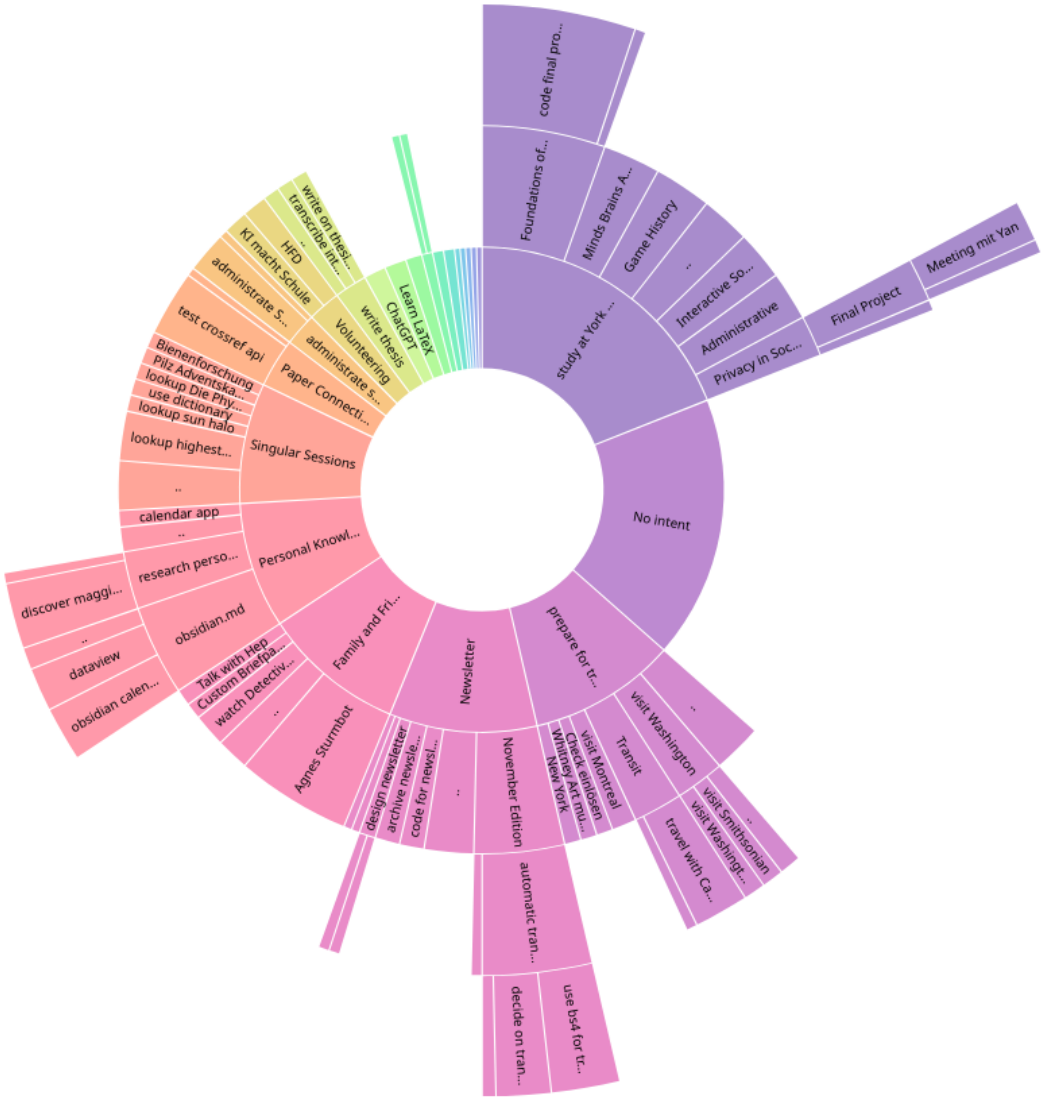

Figure A.1: Sunburst diagram of all annotated missions. The radial width of each mission in the sunburst relates to the number of logical sessions that are contained in this mission.

Bibliography

Doug Beeferman, Adam L. Berger, and John D. Lafferty. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, 1999. doi: 10.1023/A:1007506220214. URL https://doi.org/10.1023/A:1007506220214.

Andrei Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36, 9 2002. doi: 10.1145/792550.792552. URL http://dx.doi.org/10.1145/792550.792552.

Andy Clark and David Chalmers. The extended mind. *analysis*, 58(1):7–19, 1998.

Markus Fischer, Kristof Komlossy, Benno Stein, Martin Potthast, and Matthias Hagen. Identifying queries in instant search logs. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 7 2021. doi: 10.1145/3404835.3463025. URL http://dx.doi.org/10.1145/3404835.3463025.

Daniel Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179, 5 2009. doi: 10.1016/j.ins.2009.01.026. URL http://dx.doi.org/10.1016/j.ins.2009.01.026.

Matthias Hagen, Benno Stein, and Tino Rüb. Query session detection as a cascade. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 10 2011. doi: 10.1145/2063576.2063602. URL http://dx.doi.org/10.1145/2063576.2063602.

Matthias Hagen, Jakob Gomoll, Anna Beyer, and Benno Stein. From search session detection to search mission detection. In João Ferreira, João Magalhães, and Pável Calado, editors, *Open research Areas in Information Retrieval, OAIR 2013, Lisbon, Portugal, May 15-17, 2013*, pages 85–92. ACM, 2013. URL http://dl.acm.org/citation.cfm?id=2491769.

Daqing He, Ayse Göker, and David J. Harper. Combining evidence for automatic web session identification. *Inf. Process. Manag.*, 38(5):727–742, 2002. doi: 10.1016/S0306-4573(01)00060-7. URL `https://doi.org/10.1016/S0306-4573(01)00060-7`.

Daniel Hienert and Dagmar Kern. Recognizing topic change in search sessions of digital libraries based on thesaurus and classification system. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 6 2019. doi: 10.1109/jcdl.2019.00049. URL `http://dx.doi.org/10.1109/jcdl.2019.00049`.

Bernard J. Jansen, Amanda Spink, Chris Blakely, and Sherry Koshman. Defining a session on web search engines. *Journal of the American Society for Information Science and Technology*, 58, 2 2007. doi: 10.1002/asi.20564. URL `http://dx.doi.org/10.1002/asi.20564`.

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44, 5 2008. doi: 10.1016/j.ipm.2007.07.015. URL `http://dx.doi.org/10.1016/j.ipm.2007.07.015`.

Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout. *Proceedings of the 17th ACM conference on Information and knowledge management*, 10 2008. doi: 10.1145/1458082.1458176. URL `http://dx.doi.org/10.1145/1458082.1458176`.

Johannes Kiesel, Arjen P. de Vries, Matthias Hagen, Benno Stein, and Martin Potthast. WASP: Web Archiving and Search Personalized. In Omar Alonso and Gianmaria Silvello, editors, *1st International Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES 2018)*, volume 2167 of *CEUR Workshop Proceedings*, pages 16–21, August 2018. URL `http://ceur-ws.org/Vol-2167/`.

Ian Li, Anind K. Dey, and Jodi Forlizzi. A stage-based model of personal informatics systems. In Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden, editors, *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, pages 557–566. ACM, 2010. doi: 10.1145/1753326.1753409. URL `https://doi.org/10.1145/1753326.1753409`.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. *Proceedings of the fourth ACM international conference on Web search and data mining*, 2 2011. doi: 10.1145/1935826.1935875. URL `http://dx.doi.org/10.1145/1935826.1935875`.

Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Discovering tasks from search engine query logs. *ACM Transactions on Information Systems*, 31, 7 2013. doi: 10.1145/2493175.2493179. URL `http://dx.doi.org/10.1145/2493175.2493179`.

Luis Lugo, Jose G. Moreno, and Gilles Hubert. Segmenting search query logs by learning to detect search task boundaries. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 7 2020. doi: 10.1145/3397271.3401257. URL `http://dx.doi.org/10.1145/3397271.3401257`.

Niklas Luhmann. Kommunikation mit zettelkästen: Ein erfahrungsbericht. *Öffentliche Meinung und sozialer Wandel/Public Opinion and Social Change*, pages 222–228, 1981.

David Montoya, Thomas Pellissier Tanon, Serge Abiteboul, and Fabian M. Suchanek. Thymeflow, a personal knowledge base with spatio-temporal data. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 10 2016. doi: 10.1145/2983323.2983337. URL `http://dx.doi.org/10.1145/2983323.2983337`.

Huseyin Cenk Özmutlu and Fatih Çavdur. Application of automatic topic identification on excite web search engine data logs. *Inf. Process. Manag.*, 41(5):1243–1262, 2005. doi: 10.1016/J.IPM.2004.04.018. URL `https://doi.org/10.1016/j.ipm.2004.04.018`.

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguistics*, 28(1):19–36, 2002. doi: 10.1162/089120102317341756. URL `https://doi.org/10.1162/089120102317341756`.

Georg Rehm, Marina Santini, Alexander Mehler, Pavel Braslavski, Rüdiger Gleim, Andrea Stubbe, Svetlana Symonenko, Mirko Tavosanis, and Vedrana Vidulin. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association, 2008. URL `http://www.lrec-conf.org/proceedings/lrec2008/summaries/94.html`.

Craig Silverstein, Monika Rauch Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. doi: 10.1145/331403.331405. URL `https://doi.org/10.1145/331403.331405`.

Amanda Spink, Minsoo Park, Bernard J. Jansen, and Jan Pedersen. Multitasking during web search sessions. *Information Processing and Management*, 42, 1 2006. doi: 10.1016/j.ipm.2004.10.004. URL `http://dx.doi.org/10.1016/j.ipm.2004.10.004`.

Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W. White, and Wei Chu. Learning to extract cross-session search tasks. *Proceedings of the 22nd international conference on World Wide Web*, 5 2013. doi: 10.1145/2488388.2488507. URL `http://dx.doi.org/10.1145/2488388.2488507`.

Ran Yu, Limock, and Stefan Dietze. Still haven't found what you're looking for - detecting the intent of web search missions from user interaction features. *CoRR*, abs/2207.01256, 2022. doi: 10.48550/ARXIV.2207.01256. URL `https://doi.org/10.48550/arXiv.2207.01256`.