NTNU

Norwegian University of
Science and Technology

Predicting the next click with Web log Process Mining

Suresh Kumar Mukhiya

Master in Information Systems
Submission date: June 2016
Supervisor: Jon Atle Gulla, IDI
Co-supervisor: Jon Espen Ingvaldsen, IDI

Norwegian University of Science and Technology
Department of Computer and Information Science

NTNU – Trondheim
Norwegian University of
Science and Technology

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE

Predicting The Next Click With Web Log Process Mining

Master Thesis

Submitted By:
Suresh Kumar Mukhiya

TDT4900 - COMPUTER SCIENCE, MASTER'S THESIS, SPRING 2016
SUBMISSION DATE: JUNE 2016
SUPERVISOR: INGVALDSEN, JON ESPEN; GULLA, JON ATLE

Abstract

This thesis proposes **a methodology for revealing deep content interaction models from real life web logs.** The methodology is applied on web log data from **Adresseavisen**, a regional news publisher in Norway. This thesis gives a brief literature overview of process mining, motivations of process mining, related works done with process mining and tools for process mining. In addition to this, the project compares process mining with data mining as well as web usage mining.
The dynamic nature of the Web and Information systems are becoming more and more intertwined with the operational processes. This gives possibility to record a multitude of event data and provides opportunity to use process mining to these data to extract process related information. Process mining is an active and innovative research area in recent years, where the goal is to extract process-related information from even logs by observing events recorded by some information system. Over last few decades, process mining has made its way as a new research field that focuses on analysis of processes using event log data.

The migration of today's news media business to personalized information delivery has created need for analyzing user behaviors on the news site and deliver personalized set of news item according to their preferences. There are several *recommendation algorithms* that are used for recommendation of news items. This thesis makes an endeavor to use event logs from news media site and use process mining on these data for process discovery. The discovered model from this process mining is used to predict the next click item for any anonymous reader. In addition to this, the thesis discusses about the value and implications of the extracted models and how these information can be consumed for business intelligence.

Keywords: *Process Mining, Business Process Management, Process Discovery, Petri nets, Workflow mining, Workflow management, process mining methodology*

Acknowledgments

This thesis is submitted to the **Norwegian University of Science and Technology (NTNU)** for the partial fulfillment of the requirements for a Masters degree. This work has been performed at the **Department of Computer and Information Science (IDI)**, NTNU, Trondheim in the spring of 2016.

This thesis would not have been possible without the support of many people. First of all, I would like to thank my supervisor Professor *Dr. Jon Atle Gulla* and co-supervisor *Jon Espen Ingvaldsen* at the Department of Computer and Information Science, Norwegian University of Science and Technology for their constructive feedback and helpful guidance to complete my thesis on time. I would also like to thank *Arne Dag Fidjestl* for providing resources to gather data from `http://www.adressa.no/`.

And finally, I would be falling in debt if I dońt thank my family back in *Nepal* who provided help and motivational support for the last two years that gave me the strength to work hard and finish my master. I would also like to express my gratitude towards all my friends who helped me with their constructive feedback and honest criticism.

Suresh Kumar Mukhiya
June, 2016

Table of Contents

List of Tables

List of Figures

Abbreviations

ACM	Adaptive Case Management
API	Application Programming Interface
BPM	Business Process Management
BI	Business Intelligence
BPI	Business Process Intelligence
BPI	Business Process Insight
CSV	Comma Separated Values
DM	Data Mining
EDM	Educational Data Mining
ETL	Extract, Transform and Load
IIS	Internet Information Server
JSON	JavaScript Object Notation
LA	Learning Analytics
PM	Process Mining
MSIE	Microsoft Internet Explorer
MXML	Mining XML
MOOCs	Massive Open Online Courses
OS	Operating System
SNA	Social Network Analysis
SMM	Social Media Marketing
WFM	Workflow Management
WUM	Web Usage Mining
XML	Extensible Markup Language
XES	eXtensible Event Stream

Chapter 1

Introduction

This master thesis examines how process mining can be applied on click logs data from Norwegian media sites to reveal contents relationships and readers behavioural characteristics. In addition to this, it studies state of art works in process mining. The project is carried out within the Department of Computer and Information Science at Norwegian University of Science and Technology (NTNU).

1.1 Objectives

The main objective of this thesis work is to examine how process mining can be applied on click logs from media sites and reveal contents relationships and readers behavioural characteristics. There has been spectacular growth of process mining in various business fields and workflow managements. This thesis aims to contribute in literature review of such related works to point out what is the state-of-the-art within web log process mining. Figure 1.1 shows the main objective of this thesis. As shown in the figure, the main objective is to use the log files of *adressa* stored in the database and use process mining approaches to discover processes.

The sub-goal of this work is to study tools available for process mining of web log data and and use tools like DISCO to discover process and use it for predicting the next click. The event logs from `http://www.adressa.no` is extracted using Cxense [1] API and are stored as JSON format. The stored data is used for process mining using DISCO. The models extracted from this analysis forms the basis for predicting next click logs for the readers.

1.1.1 Research Questions

To drive the study for research purpose, the project is shaped to answer following two research questions:

[1] https://www.cxense.com/

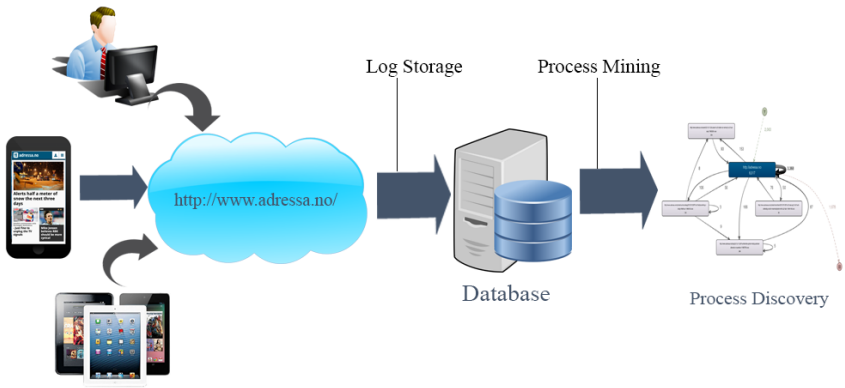


Figure 1.1: Illustration of the main objective of this thesis

1. What are the state of art works in web log process mining done so far?
2. What kind of news consumption models do different web sources drive?
   (a) What relationships do we see between web sources and news categories?
   (b) Do we observe different news consumption models on desktop and mobile traffic?

1.2 Motivations

During the last couple of decades, the digital media has gained a lot of popularity, that led to a drastic increase in the number of Internet news readers. The behavior of the readers has been changed distinctly from the traditional print news media to the digital news medias, where they can optimize the reading based upon their interests and behavior. With increase advancement of portable devices like smart-phones, and tablets, digital media has gained increased attention. At the start of 2015, 39 of the top 50 digital news websites have more traffic to their sites and associated applications coming from mobile devices than from desktop computers, according to **Pew Research Centers analysis of comScore** data [1]. In addition to this, increased application of *recommendation* and *personalization system* have addicted readers to spend more time on it leading to generate more data. In today's digital world, the volume of data being accumulated, accessed, reserved, and analyzed has exploded, in particular in relation to the user activities on the web and portable devices. One of the main challenges of todays organizations is to extract information and value from data stored in their information systems. In the past, large-scale data storage, processing, analysis, and modeling was only the domain of the largest organizations, but today many institutions are facing the challenge of handling and utilizing a massive amount of data [2].

As mentioned by the *The Four V's of Big Data* [3], huge **volume** of data are generated by Information System at very high **velocity**. There is tremendous variation of these type

of data referred as **variety**. Lastly, there is **veracity** in data which refers to uncertainty of extracted data to be completely correct. Data does not have to be **big** to be challenging, as data analytics questions are raised everywhere. Such questions can be answered by process mining. The objective of *process mining* is to use these *event data* to mine *process related* information [2].

In this thesis, the main motivation is to use event log files from Norwegian news website **adressa**, and use process mining tool to discover process readers follow on the site. Processing and analysis of these event data can reveal and gather knowledge about the relationships between content types and readers behavior. This process discovery approach can be used to predict the next click log and and define content recommendation strategies. In addition to this, there are several other motivations of process mining as described in section 2.2.

1.3 Approach

A big part of this thesis contributes on literature review of related work to point out what is the state-of-the-art within the field of *process mining*. Literature review consists of three parts - one part focuses on the theoretical backgrounds related to process mining, types of process mining and algorithms used in process mining, another part focuses on the related work in the field of process mining on web log data and the last part studies different types of process mining tools available till the date.

The implementation approach is to use the events log data files extracted from news site `adressa` and use process mining tool namely DISCO to discover processes. This result can be used to predict next click of the readers and for the recommendation purposes.

The analysis approach focus on inspecting the results obtained from the implementation phase and discuss on how this result can be used for business intelligence. Since, the process mining is not put into production in live site yet, empirical evaluation of the system is difficult. However, the evaluation strategy focuses on accuracy of process discovery and prediction of next click log.

1.4 Outcome

Approaches as mentioned in section 1.3 were used to carry out the entire project. Following are some of the major outcomes of the project:

1. Chapter 2 and chapter 3 is the result of literature study of *process mining* that discusses about theoretical aspects and related works of process mining.

2. Model mined using process mining tool DISCO by importing the event log files, show traffic originates form several sources like `www.facebook.com`, `www.startsiden.no`, `www.google.no` and others. Most of these traffic transits of news categories like *nyheter, 100sport, kultur and pluss*. In addition to this, news consumption on mobile or tablet devices are higher compared to that of desktop. Some of the other statistics and analytical model depending upon the frequency and performance metrics are discussed in chapter 5.

3. The results from the process mining shows that a set of most popular news categories mined are the best candidate to be recommended for a new visitor on the site. Further more chapter 6 and 7 describes some of the challenges and further works that can be done to improve the recommendation behavior of the system.

1.5 Outline

The thesis document is structured as follows:

- Chapter 1: **Introduction** - This chapter gives general overview of the thesis including problem domain, motivations, approaches used and brief overview of the results.
- Chapter 2: **Theory** - This chapter focus on underlying theories, standards and tools used to extract log files and implement the project.
- Chapter 3: **State of art** - This chapter focuses on the study of state-of-art researches done in process mining with web log events data.
- Chapter 4: **M ethodology** - This chapter describes how web log data has been extracted and transformed to make data ready for process mining tool like Disco.
- Chapter 5: **Analysis** - The chapter presents analysis results with examples of extracted models.
- Chapter 6: **Discussion** - This chapter discusses about the value and implications of the analysis and findings from this project.
- Chapter 7: **Conclusion** - This chapter presents final conclusions and suggestions for further work.

Chapter 2

Theory

This chapter gives a literature overview of the relevant theoretical background related to process mining and its offshoots. Section 2.1 presents brief overview of *process mining* and its scope. Section 2.2 discusses on motivations of processing mining in today's business. Section 2.3 discusses about various terms and technologies used about event logs data. There are different approaches of process mining and are summarized in section 2.4. Petri nets are important aspects of process mining. Brief overview of Petri nets and its relation with process mining is discussed in section 2.5. Difference between process mining and data mining is covered in section 2.6. In addition to this, brief overview of *Web Usage Mining* and its relationship with process mining is discussed in section 2.7. Further, sections 2.8 gives brief overview of different types of tools available for process mining.

2.1 Process Mining

As aforementioned, there is massive growth of digital world everyday that leaves huge amount of event data. The challenge is to exploit these event data in a meaningful way, for example discovery of system processes, identification of bottlenecks, problems anticipations and insight discovery. Process mining aims to *discover, monitor and improve real processes by extracting knowledge from event logs* readily available in today's information systems [2, 4].

Process mining complements existing approach *Business Process Management (BPM)*. BPM *combines knowledge from management sciences and applies this to operational business processes* [2, 9]. BPM on the other hand can be seen as an extension of Workflow Management (WFM). WFM focuses on the automation of business processes. Process mining is close to BPM life-cycle. Process mining sits between machine learning and data mining in one hand and process modeling and analysis on the other hand [2]. Figure 2.1 shows process mining lies in between process modeling and process analysis in one hand and between machine learning and data mining in another. Comparison between process mining and data mining is summarized in section 2.6. Figure 2.1 shows process mining

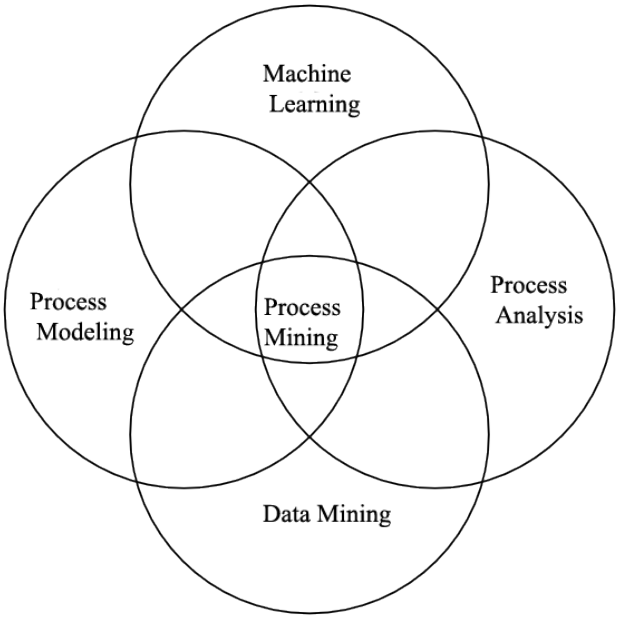


Figure 2.1: Comparison of process mining with other classical mining concepts

develops links between the actual processes and their data on the one hand and process model on the other hand [2].

2.1.1 Characteristics of process mining

Process mining focus on business process models and used to analyze the overall process. This section focuses on following important characteristics of process mining:

1. **Evidence-based BPM**: Process mining is based on observed behavior recorded in event logs where intelligent techniques are used to extract knowledge. So, Process mining is claimed as *evidence-based* BPM [2, 4].

2. **Fact-based**: Process mining is *fact-based* which means it is based on event data rather than opinions. The observed behaviors of the end users and machines are recorded in logs and are the main foundation of process mining [2, 4].

3. **Truly intelligent**: Process mining is *truly intelligent* meaning it learns from historic data and use the knowledge for model enhancement. It can be used to analyze the process in the system as well as can be used for conformance checking [2, 4].

4. **Process-centric**: Process mining is related to processes not only to data. It is not *data-centric* like other mining approaches [4].

2.2 Motivations of Process Mining

Data are important aspects of any organizations. Data can be processed to get information and information can be processed to get knowledge. Knowledge from these past data are driving forces for carrying out business in present and future. Different classical data mining techniques like *clustering, classification, regression, association and sequence mining* are used to analyze specific steps in overall process but it does not focus on *business process* [4]. Business process can be studied by using *process mining*. Some of the important motivations of process mining are summarized below:

- **Explosion of event data**: All the activities done by people, machines and software leaves trails called *event logs*. More and more events are being recorded thus providing detailed information about the history processes. As mentioned in section 1, high volume of data gives higher challenges of extracting useful information from it.
- **Discovering business process:** It is always interesting to discover if the users of the system follows some processes. Process mining techniques take event logs and discover process within it if exits.
- **Bottlenecks identifications in process:** Process mining can be used to identify bottlenecks in Information Systems by analyzing the event logs data. Using right processed data, one can find bottlenecks relating to missing steps, service interruptions or long process times. Hence, process mining can be used to identify and understand bottlenecks, inefficiencies, deviations, and risks [10].
- **Conformance checking with existing model:** For quality assurance, it is often required to check if reality as recorded in log confirms to the conceptual model. Process mining can be used for conformance checking provided conceptual model is available.
- **Recommendation of news items using process mining technique:** Process mining can be used to discover the process user reads news articles in general. These predicted process path can be recommended to new users visiting the site.

2.3 Event Logs

Process mining is concerned about events logs extracted from Information Systems. The definition of events and attributes is given in 2.1. Table 2.1 illustrates typical information present in an event log used for process mining. Event logs are the starting point of process mining. The main assumption about the event logs is, it contains data related to *single process.* Some of the important terminologies related to process mining are discussed below:

Case : A case is a specific instance of any process [8]. A *process* contains of several *cases*. Table 2.1 contains four cases. Each case has unique identifier referred as *case id.* Case 1 contains five *associated events*. The first event of case 1 is the execution of activity *Register the site* by Anne on December 30th, 2015.

Activity : Each event in a log refers to an *activity*. An activity forms one step in the process. It is well defined step in some process. Table 2.1 shows *Register the site, verify the site, login, search product and logout* as activities on the first case. A Process is composed of activities and the relation between activities. There could be following types of relations:

- Sequence- Activity X follows activity Y
- Concurrency- Activities X and Y happens mostly in the same time
- Loop- Activity X repeats a certain number of times
- Decision point - From activity X either activity Y or activity Z can be reached further

Event id : Each activity of the events can have unique identifier called *event id*. Table 2.1 also shows a unique id for this events *231456*. However, this is not commonly used for identification or mining purposes.

Timestamp : Timestamp[1] is a sequence of characters or encoded information identifying when a certain event occurred, usually giving date and time of day, sometimes accurate to a small fraction of a second. Table 2.1 shows a column with timestamp that is human readable. Timestamp can be in different formats according to configurations of servers where log files are saved.

Resources : In the table 2.1 each event is associated with a *resource*. Not all the log files contain resource information.

Data Attribute : Log files can contain one or more *data attributes* that provides additional information about the events. Table 2.1 contains cost as a data attribute. A log files may contain many other data attributes.

Traces : Events are linked to particular *trace*. An event log is a set of traces and the events within each trace are ordered in sequence. Each event in log is unique and can be linked to one trace [16].

Definition 2.1. (Event, attribute) Let ξ be the event universe, i.e., the set of all possible event identifiers. Events may be characterized by various attributes, e.g., an event may have a time-stamp, correspond to an activity, is executed by a particular person, has associated costs, etc. Let AN be a set of attribute names. For any event e $\in \xi$ and name n $\in$ AN: $\#_n(e)$ is the value of attribute n for event e. If event e does not have an attribute named n, then $\#_n(e) = \perp$ (null value) [2].

2.3.1 The Minimum Requirements for an Event Log

According to *Disco User Guide* [8], in order to use process mining tools, event logs should contain at least the following three elements:

[1] https://en.wikipedia.org/wiki/Timestamp

Case id	Event id	Time stamp	Activity	Resource	Cost
1	231456	30-12-2015:10:01	Register the site	Anne	111
	231457	30-12-2015:10:02	Verify the login	David	2
	231458	30-12-2015:10:03	Login	Ramesh	13
	231459	30-12-2015:10:04	Serch product	Suresh	3
	231460	30-12-2015:10:05	Logout	Anju	144
2	231461	30-12-2015:10:06	Login	Anne	10
	231462	30-12-2015:10:07	Add payment	Anju	188
	231463	30-12-2015:10:08	Buy Product	Ramesh	200
	231464	30-12-2015:10:09	Paid	Suresh	200
	231465	30-12-2015:10:10	Serch product	David	111
	231466	30-12-2015:10:11	Logout	Sita	12
3	231467	30-12-2015:10:12	Register the site	Anne	11
	231468	30-12-2015:10:13	Verify the login	David	22
	231469	30-12-2015:10:14	Login	Suresh	150
	231470	30-12-2015:10:15	Serch product	Ramesh	111
4	231471	30-12-2015:10:16	Register the site	Anju	100
	231472	30-12-2015:10:17	Verify the login	Sita	40
	231473	30-12-2015:10:18	Login	Anne	20
	231474	30-12-2015:10:19	Serch product	David	300
	231475	30-12-2015:10:20	Buy Product	Ramesh	300
	231476	30-12-2015:10:21	Paid	Suresh	300
	231477	30-12-2015:10:22	Edit Profile	Anju	2
	231478	30-12-2015:10:23	Logout	Sita	190
...	...	...	...	...	...

Table 2.1: A section of example events log file

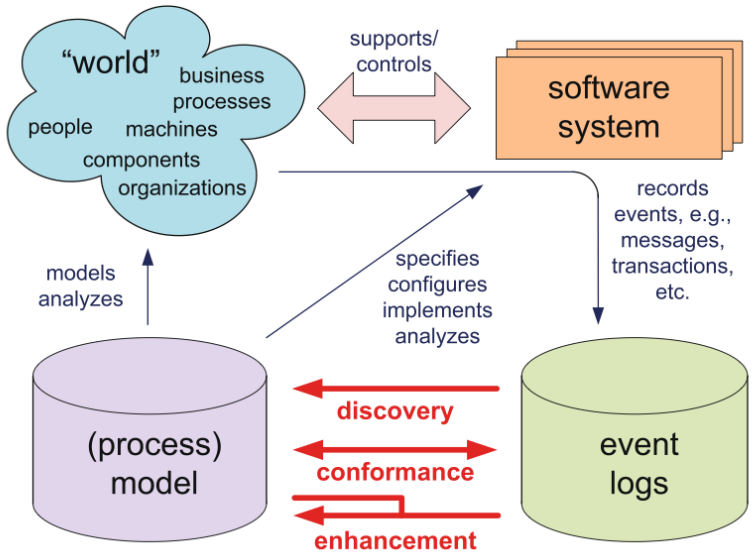


Figure 2.2: Three main types of process mining: *discovery, conformance and enhancement* [2]

1. Case Id
2. Activity
3. Timestamp

2.4 Types of Process Mining

Event logs can be used to perform three main types of process mining as shown in figure 2.2. The are - *process discovery, conformance and enhancements*. Section 2.4.1 discuss about process discovery. Section 2.4.2 discuss about what process conformance and lastly section 2.4.3 gives brief overview about process enhancement.

2.4.1 Process Discovery

A discovery technique takes an event log and produces a model without using any *a-priori information*. Process discovery problem is addressed by definition 2.2. One of such technique is used by α**-algorithm** [2, 11] which takes an event log and produces a Petri nets explaining the behavior recorded in the log. The algorithm is capable of constructing the Petri net automatically without using much additional knowledge [2].

Definition 2.2. General process discovery problem: Let L be an event log or as specified by the XES standard. A **process discovery algorithm** is a function that maps L onto a process model such that the model is representative for the behavior seen in the event log. The challenge is to find such an algorithm [2].

Applications of Process Discovery

As aforementioned, process mining can be used to discover process models. These models can be used for wide range of analysis as given below:

- For discussing problems among stakeholders.
- For generating process improvement ideas.
- For configuring a WFM/BPM.
- For analyzing bottlenecks in any systems or process models.
- For improving process flexibility or model enhancements.

Figure 2.3 shows the model discovered using log files from table 2.1. It shows the common process followed by most of the users. The model is generated using DISCO using only 4 cases as shown in the table above.

2.4.2 Conformance Checking

Conformance checking consists of use cases which have the intention of checking whether the process had the intended behavior in practice [2, 13]. Figure 2.4 shows the main idea of conformance checking. Process mining techniques automatically constructs a model without any a priori information. Conformance checking uses a model and event logs as input and the modeled behavior and the observed behavior is compared to find commonalities and discrepancies [2, 4].

Conformance checking can be used for following purposes:

- To find exceptions from the normal path: Conformance checking can be used to discover outliers of the process by inspecting at the exceptional behavior observed in practice [13].
- To find the degree in which the rules are obeyed: Conformance checking can be also used to check weather the rules and regulations are obeyed and to discover deviating cases [2, 4, 13].
- To find compliance to the explicit model: Conformance checking can help to compare the documented process model with the real process as observed in the event logs and identify what they have in common [13].
- As a starting point of model enhancement: Conformance checking as it helps to identify process fragments and deviation points, it can be seen as the starting point of model enhancement [4].

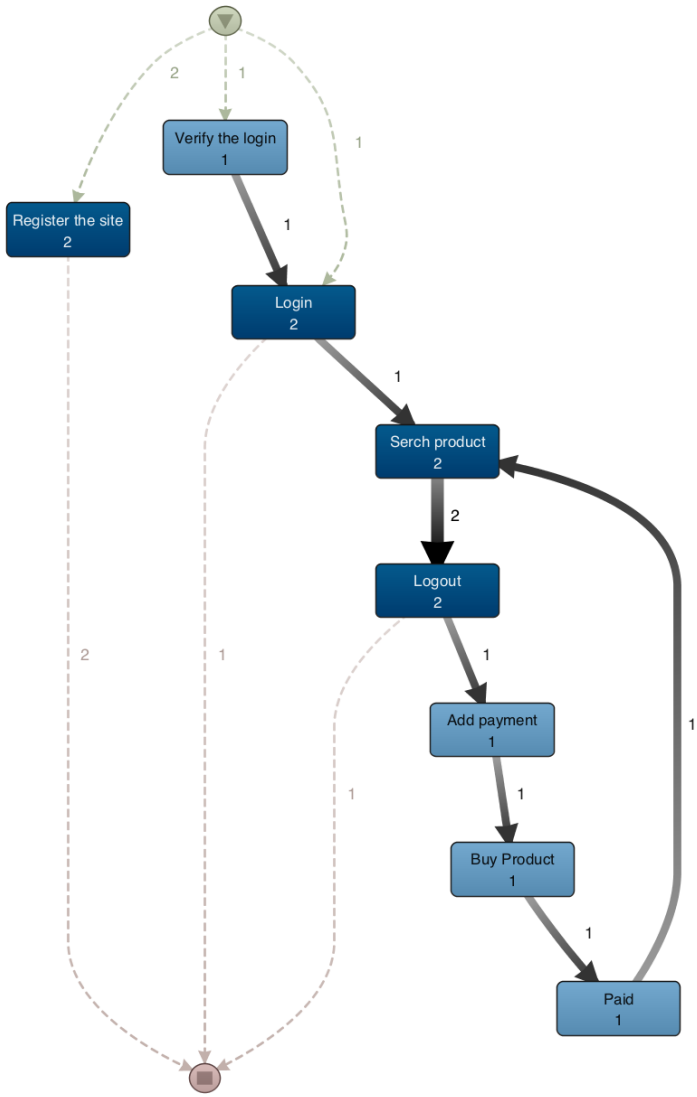


Figure 2.3: Discovered process model from event logs in table 2.1

Algorithms for Conformance Checking

According to Aalst [4], there are three approaches of conformance checking. They are discussed below:

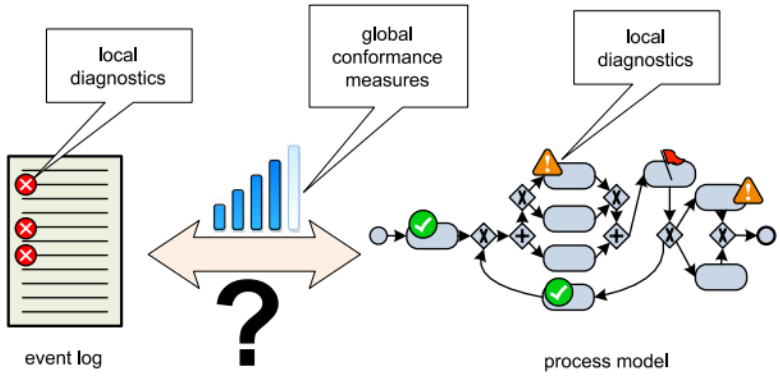


Figure 2.4: Conformance checking: comparing observed behavior with modeled behavior. [2]

1. **Comparing Footprints**: This approach creates the abstraction of the behavior allowed by the model. A *footprint* is a matrix showing causal dependencies between activities. For example, the footprint of an event log may show that x is followed by y but never the other way around. If the footprint of the corresponding model shows that x is never followed by y or that y is sometimes followed by x, then the footprints of event log and model disagree on the ordering relation of x and y [4].
2. **Replays**: This approach *replays* the event log on the model. One of the approaches is to replay the event logs on the model to check if it fits the model. A naive approach towards conformance checking would be to simply count the fraction of cases that can be parsed completely [2, 4, 13].
3. **Optimal Alignment**: This is once of the most advanced approach. This approach computes an optimal alignment between each trace in the log and the most similar behavior in the model [4].

2.4.3 Enhancement

As shown in figure 2.1, *enhancement* is the third type of process mining. The idea is to extend or improve an existing model using information about the actual process recorded in some event log [2, 4]. A type of enhancement is repairing the model to better reflect reality. For example, it is possible to analyze health quality performance by analyzing health record data to find degree of variation and use these discovery for enhancement of the health policies models [14]. It is also possible to construct social networks based on workflow of work and analyze resource performance [4].

2.5 Petri Nets

Petri nets are a graphical and mathematical modeling tool applicable to many systems that allows modeling of concurrency [2]. They are one of the oldest and best investigated pro-

Table 2.2: A computer program

A = 1
B = 2
C = 3
A = A+1
C = B+C
B = A+C

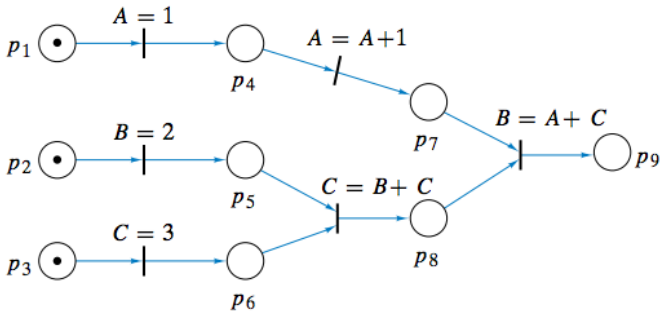


Figure 2.5: The program of Figure 2.5 as a Petri net. The tokens indicate that the conditions for executing A = 1, B = 2, and C = 3 are met. [36]

cess modeling language. Moreover, they are a promising tool for describing and studying information processing systems that are characterized as being concurrent, asynchronous, distributed, parallel, non deterministic, and/or stochastic [35]. As a graphical tool, Petri nets can be used as a visual-communication aid similar to flow charts, block diagrams, and networks.

Definition 2.3. Petri net: A Petri net [2] is a triplet N = (P,T,F) where *P* is a finite set of *places*, *T* is a finite set of *transitions* such that P $\cap$ T = $\emptyset$, and F $\subseteq$ (P $\times$ T) $\cup$ (T $\times$ P) is a set of directed arcs, called the flow relation. A marked Petri net is a pair (N,M), where N = (P,T,F) is a Petri net and where M $\in \mathbb{B}$(P) is a multi-set over P denoting the marking of the net. The set of all marked Petri nets is denoted N .

The mathematical definition of *Petri nets* is given by 2.3. A Petri net is made up of *places, transitions and tokens.* Transitions consume tokens from the input places and produce tokens in the output places. Less formally, a Petri net is a directed, bipartite graph where the two classes of vertices are called places and transitions. In general, parallel edges are allowed in Petri nets; however, for simplicity, we will not permit parallel edges.

Table 2.2 shows a typical compute program [26] and the corresponding *petri* net is depicted by figure 2.5. In the figure, the events (transitions) are the instructions, and the places represent the conditions under which an instruction can be executed.

Petri nets are often used in the context of process mining. There are several algorithms that employ Petri nets as internal representation used for process mining. α-algorithm and

.

Figure 2.6: Process Mining compared with data mining [15]

other region-based process discovery techniques use Petri nets as its internal representation during process mining [38].

2.6 Process Mining VS Data Mining

As aforementioned, process mining and data mining are related to each other. Figure 2.6 shows how process mining is connected to data mining. As mentioned in figure 2.1, process mining is built on two pillars- *process modeling and analysis* and *data mining* [2]. Process mining takes challenge to process large volumes of data like data mining.

However, process mining focuses on the process perspective. It includes the temporal aspects and looks at a single process execution as a sequence of activities that can be performed [15]. On the other hand, most of the data mining techniques analyze data from different perspective and summarize them into useful information. Most popular data mining techniques includes clustering, classification, regression, association rule mining, decision tree and sequence or episode mining. Data mining is used to analyze **specific step** in overall process not **business process**. Although both process mining and data mining start from data, data mining techniques are typically not *process-centric* and do *not* focus on event data. For data mining techniques the rows (instances) and columns (variables) can mean anything but for process mining techniques, we assume event data where events refer to process instances and activities. Moreover, the events are ordered and we are interested in *end-to-end* processes rather than local patterns. End-to-end process models and concurrency are essential for process mining. In addition to this, topics such as process discovery, conformance checking, and bottleneck analysis are not generally addressed by traditional data mining tools and methodologies [2, 34].

In data mining, data examples that do not match the general rules can be stripped away. Hence, generalization is very important to avoid what is called *overfitting the data*. Generalization is required in process mining to deal with the complex processes and understand the main process flows. With process mining, understanding the exceptions is important to discover inefficiencies and the points of enhancement [15].

Models generated from data mining are often trained to make predictions about the future related to similar instances in the same space. Data mining can give us insight into

causes of certain process behavior or predict the outcome of a running process. Process mining can tell us about how processes are carried out and how they can be discovered. Process mining focuses on end-to-end process and is possible because of growing availability of event data and new process discovery and conformance checking techniques [4].

There are many opportunities to leverage existing data mining techniques for process mining purpose. Case data attribute could also be analyzed in for association rule [16]. Event data attribute could be used to discover sequential pattern [17, 18] from a data perspective. In process mining, clustering techniques have been successfully used to group similar process instances for yielding more precise models [19, 20, 21], to mine higher-level activities [22, 23] and for tackling unstructured process such as hospital data [16]. Genetic mining algorithms were used for process discovery as described in [24].

2.7 Web Usage Mining

This section discuses on what web usage mining is and how it fits in the domain of process mining. Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [31]. The paper [31] provides an up-to-date survey of the rapidly growing area of **Web Usage mining**. With the growth of the web based applications, there is significant interest in analyzing web usage data to better comprehend web usage and apply the knowledge to better serve the users. This paper aims to address some of the challenges in web usage mining and hopes to be addressed by web research community.

The goal of web usage mining is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests [32]. Web usage mining contains three phases as follows [31, 32]:

1. **Preprocessing**: This stage deals with cleansing and partitioning of the click-stream data into a set of user transactions representing the activities of each user during different visits to the site. Preprocessing also deals with converting the usage, content and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

2. **Pattern Discovery**: In this stage, statistical analysis, database analysis, and machine learning operations are performed to obtain *hidden patterns* reflecting the typical *behavior of users*, as well as *summary statistics* on Web resources, sessions, and users. This stage draws upon methods and algorithms such as *statistical analysis, association rules, clustering, classification, sequential pattern mining, dependency modeling and other machine learning operations.*

3. **Pattern Analysis**: In this stage, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as *recommendation engines, visualization tools*, and *Web analytics* and *report generation tools.* The main motivation is to filter out uninteresting rules or patterns from the set discovered in the pattern discovery stage.

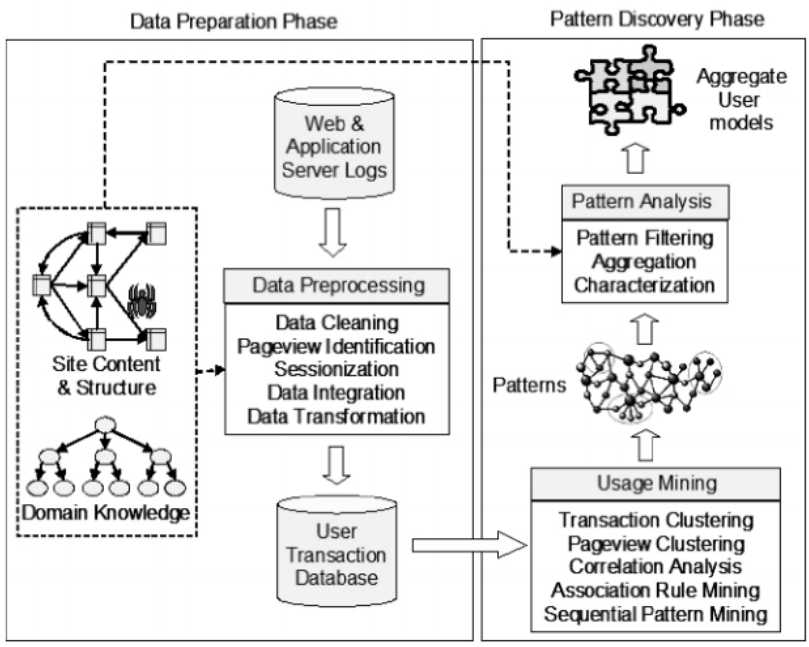


Figure 2.7: The Web usage mining process [31]

Figure 2.7 illustrates overall process of web usage mining. As mentioned in this section, the figure shows preprocessing consists of usage preprocessing, content preprocessing and structure preprocessing. Moreover, as depicted in the figure, the preprocessed data is used for pattern discovery and the discovered patterns are further studied in pattern analysis step.

Compare web usage mining and process mining

2.8 Tools for Process Mining

There are several tools that facilitates in gathering the logs files and performing process mining. ProM is one of the open-source framework as described in section 2.8.1. Section 2.8.2 describes about **Disco** that is used as the main tool in this thesis. There are other tools that helps to collect and convert the log files to make it appropriate for process mining as described in later sections.

Figure 2.8: ProM icon

2.8.1 ProM

The ProM[2] framework has been developed as a completely plug-able open source environment. It has been contribution from Netherlands, Germany, Italy, France, Australia, Austria, China, Spain, Portugal and Brazil [5]. ProM is an extensible and only comprehensive framework that supports a wide variety of process mining techniques in the form of plug-ins. It is platform independent as it is implemented in Java, and open source. Process discovery, conformance checking, social network analysis, organizational mining, decision mining, history based prediction and recommendation are all supported by ProM. It supports different process discovery algorithms. It requires process mining expertise to use tool and since it is open source, it adheres both merits and demerits of open source software [4, 6]. Figure 2.8 shows the icon of ProM software.

Features of ProM

- ProM 6 core is distributed as a downloadable package using the GNU Public License (GPL) open source license.
- It is platform independent as it is implemented in Java, and can be downloaded free of charge.
- There is increasing number of researchers and developers to contribute in the new plug-ins developments.
- Usage of the software is well documented including installation, case studies, publications and demonstration data.

2.8.2 Disco

Disco[3] is a commercial suite supported by the leading academic group in process mining at *Eindhoven Institute of Technology*. Disco has proved to be user friendly and very fast on big event logs while discovering high-level knowledge though it is not as rich in techniques as ProM [7]. Like ProM, Disco supports wide range of log export formats including *CSV files, MS Excel (XLS and XLSX) files, MXML and MXML.GZ files (ProM 5), XES and XES.GZ files (ProM 6), FXL Disco log files, and DSC Disco project files* [8].

[2] http://www.promtools.org/doku.php
[3] http://fluxicon.com/disco/

2.8.3 XESame

XESame[4] is an application that supports the extraction of an event log from non-event log data sources. The main objective of XESame is to extract event logs from data sources. The input format can be database tables, text files or even XML files. The output is an event log in the XES or MXML format [12].

2.8.4 OpenXES

OpenXES[5] is a reference implementation of the XES standard for storing and managing event log data. XES is an open standard for storing and managing event log data. The OpenXES library is a reference implementation of that standard in JAVA, which strives for strict XES compatibility, ease of development, and the best possible performance. OpenXES is released as open source under the terms of the GNU Lesser GPL (LGPL) license.

In addition to this, `http://www.processmining.org/tools/start` lists a set of other tools that are widely used with ProM including *ProMimport*[6]*, ProM CPN Library, MXMLib, and Process Mining Prom package.*

[4] http://www.processmining.org/xesame/start
[5] http://www.xes-standard.org/openxes/start
[6] http://www.promtools.org/promimport/

Chapter 3

State of Art

The concept of process mining is not new [2, 4, 5, 25, 26, 27]. It has been applied in many enterprise information systems for process discovery, conformance checking and enhancements. Several process-mining techniques are available and their value has been proven in various case studies [28]. Process mining techniques can be used to discover the real process, to detect deviations from some normative process, to analyze bottlenecks and waste, and to predict flow times [29, 2, 4].

There are several literature works[1] on process mining. This chapter discusses on some of the related work done in web usage mining. Section 3.1 discusses on how process mining is used in the field of E-commerce. Section 3.2 describes process mining work carried out in the education domain. Web usage mining is the application of data mining techniques to discover usage pattern from web data as described in section 2.7. Section 3.3 describes how process mining is used to analyze event data in *Coursera*[2]. Section 3.4 describes different pattern mining approaches form web usage point of view. Lastly, section 3.5 describes how the prediction and recommendation concepts in process mining can be applied to ACM.

3.1 Business Process Mining from E-commerce Web Logs

The paper *Business Process Mining from E-commerce Web Logs* [30] by *Nicolas Poggi* and et. al presents application of Business Process Management (BPM) methodologies in e-commerce website logs. In particular, the paper applies process mining techniques, basically Business Process Insight platform to analyze web user behavior. The authors experiment on custom click-stream logs from a large online travel and booking agency. They fist compare Web-clicks and BPM events, and then present a methodology to classify and transform URLs into events. The paper evaluates traditional and custom process mining algorithms to extract business models from web data. The models resulting from analysis,

[1] http://www.processmining.org/publications/start
[2] https://www.coursera.org/

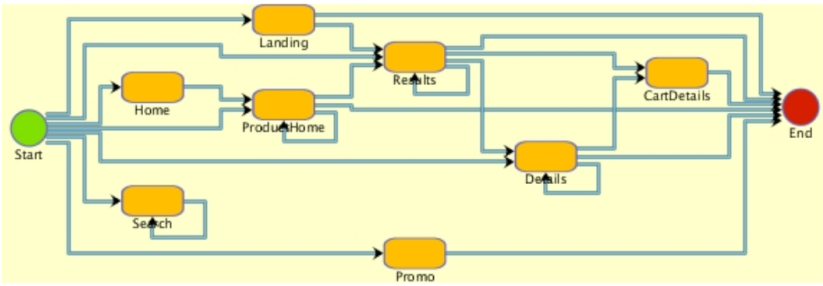
present an abstracted view of the relation between pages, existing points and critical path taken by customers.

The main motivation of the research, according to this paper is to use process mining technique to yield structured formal models of user behavior that can provide insights of potential improvement to the site. Online businesses reply on **web analytic** tools to inform their web marketing campaigns and strategic business decisions which are claimed by the paper to be not providing abstracted view of the customer's actual behavior on the site. Hence, Business Process Insight could provide simplified and correct understanding of their users' real interaction patterns on the site and their evolution. The paper claims to contribute in following three major areas:

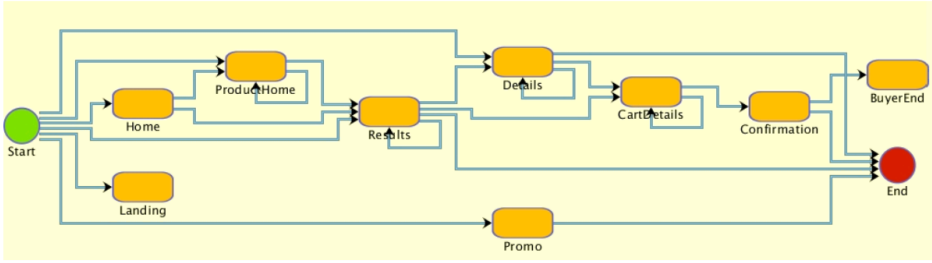
- Outlining how to transform web clicks into tasks suitable for analysis and modeling with BPM tools. The authors classify the URLs that correspond to web click logs into high level tasks that involves both manual and automatic *classification* techniques.
- Describing how to mine business processes that includes how regular web visitors and customer behave. Unlike most process mining algorithms that capture only the most common behavior in order to keep the resulting model simple enough, this paper addresses this issue with techniques such as *saturating* the dataset with low frequency behavior user tends to observer, *clustering* the process instances to extract pattern of behavior or using *knowledge-based* process mining algorithm.
- Evaluating the use of the knowledge-based mining algorithm under a variety of conditions and explaining it suitability to extract process models that abstract a complete over-view of user navigation from real, noisy data.

After applying Business Process Insight to analyze user behavior, the authors found that web navigation shares characteristics with traditional BPM activities such as *loops* and *parallel tasks*. It is seen that, sessions only span a few minutes on average and include no human intervention. In addition to this, the experiments results with discovery that, any analysis of web logs requires the classification of URLs to higher logical tasks, as the number of unique URLs become too big for human consumption and tradition mining algorithms. Finally, the paper also shows that clustering algorithms can automatically classify URLs, requiring only that each cluster be named. Figure 3.1 shows the resulting models by applying the knowledge based miner with default noise and window parameters to the *normal* (a) and *saturated* (b) data sets. The figure also shows the general workflow of the events with the main distinction being that the normal dataset does not contain the *Confirmation* and *BuyerEnd*. The figure also depicts loops in the both models where the loops are from the same originating event to itself, such as users iterative over the *Results* event.

Figure 3.2 illustrates the model generated by both *Heuristic* and *Knowledge-based miner* to a specific small cluster of customers and represents most common buying process. Moreover, the figure, demonstrates the critical path for buyers on the website thus giving information that most important pages should be optimized. It shows the most important pages consists of *CartDetails, Details and Confirmation* meaning that most buying sessions go strait to purchasing without much searching.



(a) Knowledge-based miner process model for the normal dataset



(b) Knowledge-based miner process model for the buyers saturated dataset

Figure 3.1: Knowledge-based miner process model for the buyers saturated dataset [30]

Figure 3.2: Process model of a customer for Heuristic and knowledge-based miners [30]

Like this paper, the thesis also deals with web click logs and usage of process mining techniques to analyze web user behavior. In order to make click logs suitable for process mining, log files are *pre-processed*. The JSON file is converted to CSV file using custom PHP scripts. The URLs like this paper is classified to higher level tasks, in particular to *category* in which it belongs. These categories are used as *activity* as referred in section 2.3. Unlike this paper which applies process mining in the field of e-commerce log data, this thesis applies *process mining* in the domain of news websites which face the challenge of lacking *concrete process model* as most of the sessions only span a few minutes on average. E-commerce websites undergoes a concrete process model like *product search, adding to card, checkout, payment and shipping* or a variants of this model. Unlike this, in news website no concrete process model is observed.

3.2 Process Mining in the Education Domain

The paper *Process Minining in the Education Domain* by *Awatef.hichaurcairns and et al.* applies process mining techniques in the educational domain. This work is done by **Altran Research and Altran Institute** in the context of project PERICLES[3]. The paper shows how social mining techniques can be used to examine and assess interactions between originators, training courses or pedagogical resources, involved in students training path. In addition to this, the paper proposes a two-step clustering approach to extract the best training paths depending on an employability indicator.

The paper focus on the fact that, education and training centers have started introducing more agility into their teaching curriculum in order to meet the fast changing needs of the job market and meet the time-to-skill requirements. The use of information and communication technologies in the educational domain generates large amount of data, which main contains insightful information about students, profiles the process they went through and their examination grades. These data are used for process mining input feed. The research done by the authors tends to develop generic methods, which could be applied to general education issues and more specific concerning professional training or e-learning fields. The research aims to develop generic methods which could be applied to general education issues and more specific ones concerning the professional training for following:

- The extraction of process-related knowledge from large education event logs, such as *process models* and *social networks* following key performance indicator or a set of curriculum pattern templates.
- Conformance checking of established curriculum constraints, educators' hypothesis and prerequisites with the educational processes.
- Enhancement of educational process models with performance indicators.
- The personalization of educational processes via the recommendation of the best course units or learning path to the students.

In addition to these, the paper also points out process mining issues in the *Education Domain* like - *handling voluminous data - large number of cases or events in event logs, handling heterogeneity and complexity and handling conceptual drifts.* The paper purposes clustering techniques for partitioning large logs into smaller parts that can be checked locally and more easily. To deal with heterogeneity and complexity, the paper suggests adoption of filtering, abstraction or clustering techniques that may help reducing the complexity of the discovered models. Finally, the paper develops a set of agenda for future in several directions that intend to combine the approaches proposed in the paper with other process mining techniques, which allow discovering interaction pattern from email dataset in order to discover interactions pattern between students in their collaborative learning tasks, communication actions and online discussions. Moreover, the proposed architecture will be implemented and deployed and tested on a distributed environment connected to several data sources and applications.

[3] httpu://e-pericles.org/

Table 3.1: Global statistics for the Coursera MOOC case study

Start Date	Nov 14, 2014
#Registered	43,218
# Visited course page	29,209
# Watched a lecture	20,868
# Browsed forums	5,845
# submitted an exercise	5,798
# Certificates (normal/distinction)	1,688
# Normal certificate	1,034
# Distinction Certificate	654
End date	Jan 8, 2015

Like this paper, the thesis also deals with web click logs from news website and usage of process mining techniques. The extracted from news website undergoes ETL processes, data pre-processing and then is used by process mining tool DISCO. This paper aims to improve both performance and readability of the mined students' behavior model in the context of e-learning where as this thesis tries to use web events logs for process discovery.

3.3 Learning Analytics on Coursera Event Data:A Process Mining Approach

The paper *Learning Analytics on Coursera Event Data: A Process Mining Approach* [37] by *Patrick Mukala, Joos Buijs, Maikel Leemans, and Wil van der Aalst* makes uses of *process mining* technique in order to trace and analyze students' learning habits based on MOOC data. The primary objective of this endeavor is to provide insights regarding students and their learning behavior as it relates to their performance. The analysis shows that successful students always watch videos in the recommended sequence and mostly watch in batch and the vice versa holds true. In addition to this, the research identifies a positive correlation between viewing behavior and final grades supported by *Pearson's Kendell's and Spearman's* correlation coefficients.

The paper uses **Coursera**[4] data that divides the data into three categories: general data, forums data and personal identification data. The data collected ran from November 11, 2004 to January 8, 2015. Table 3.1 shoes overall statistics of data used during this research. The paper limits the analysis to data about direct student behavior and considers three important dimensions in the analysis: *the general lecture videos viewing habit, the quiz submission behavior as well as a combination of both.*

Figure 3.3 shows dotted chart of a general viewing behavior for all the students having registered throughout the duration of MOOC focusing on when and how they watch the videos. The x -axis depicts the time expressed in weeks, while the y-axis represents students. Seven different colors represent different events at a given time as carried by students. The white dots show the timing when students viewed miscellaneous videos.

[4]http://wwww.coursera.org

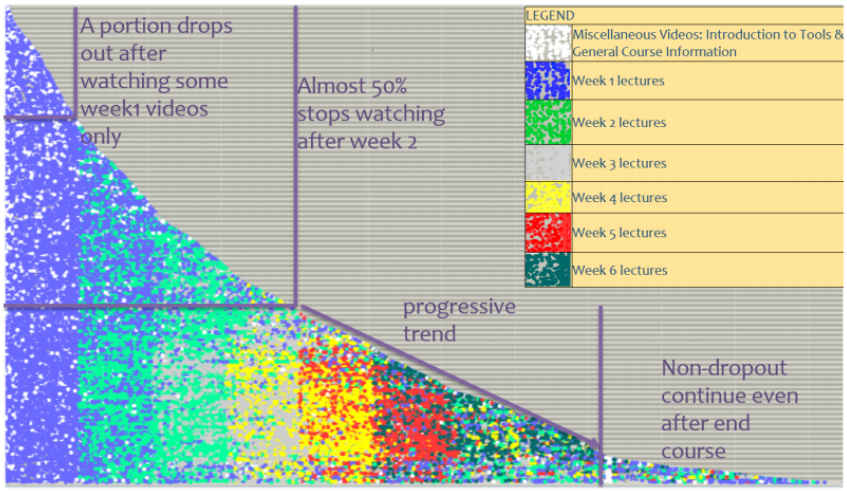


Figure 3.3: Dotted Chart depicting a general viewing behavior throughout the duration of the MOOC [37]

The analysis shows that:

- A significant number of student dropout throughout the duration of the course.
- Many stop watching after the first week but about 50% of students drop out after the second week of the course.
- Not all students watch the videos in sequence. Although, all of them watch Week 1 before watching Week 2. Some videos are watched repeatedly and a number of students progressively join the course later than the staring date.

Learning Analytics (LA) promises to provide insights from educational data. In addition to this, this paper proposes to use process mining in order to provide insightful analysis based on the actual student behaviors. After using process mining techniques on the MOOC data, the paper illustrates the way students watch videos as well as the interval between successive watched videos have a direct impact on their performance. The research also reveled that students who watched videos on regular basis and in batch are more likely to perform well than those who skip videos or procrastinate in watching videos.

Like this paper, this thesis also uses *process mining* approach to determine user behavior on news website. However, this paper uses process mining in education domain from event logs of Coursera.

3.4 Frequent Pattern Mining in Web Log Data

This paper *Frequent Pattern Mining in Web Log Data* [39] by *Renta Ivncsy, Istvn Vajk* investigates three different pattern mining approaches from *Web usage* point of view. The

different patterns in Web log mining are ***page sets, page sequences and page graphs.*** The paper deals with the problem of discovering hidden information from large amount of Web log data collected from servers. Introduction to the process of web log mining and demonstration of how frequent pattern discovery tasks can be applied on web log data in order to obtain useful information about the user's behavior are major contribution of this paper.

The paper starts by introducing different approaches of Web mining. As mentioned in section 2.7, this paper categorizes web mining into three categories namely - *Web content mining, web structure mining and web usage mining.* Web usage mining is the task of applying data mining techniques to discover usage pattern from Web data in order to understand and better serve the needs of user navigation on the web. Web usage mining consists of three different steps like *preprocessing, pattern discovery and pattern analysis.*

To support the research, the authors experiment with two web server logs files from `msnbc.com`[5] and ECML/PKDD 2005 DISCOVERY CHALLENGE[6]. The data in the log files were *preprocessed* which contained three different phases- cleansing of the data, identification of different sessions belonging to different users and conversion of data into the format supported by mining algorithms. The msnbc log data describes the page visits of the users who visited `msnbc.com` on September 28, 1999. Visits were recorded at the level of URL category and are recorded in time order. The data came from IIS logs for the site. For the sequence patterns, the row were converted such that they represent sequences. In the same way, **Click Stream** data was preprocessed to make it appropriate for mining.

After using frequent mining algorithms to the data set, it is noticed that for the mining process, beside the input data, the minimum support threshold value is required. It is one of the key issues to which value the threshold should be set. The *frequent item-set discovery* and the *association rule* mining was done using **ItemsetCode algorithm.** It is level wise "candidate generate and test" method based on **Apriori hypothesis** . Figure 3.4(a) shows the association rules generated from `msnbc.com` at a minimum support threshold of 0.1% and at a minimum confidence threshold of 85%. The results can be used to make advertising process more successful and change the structure of the portal. Figure 3.4(b) shows a part of the discovered sequences of the **SM-Tree algorithm** where percentage values shows the support for sequences. The main logic of SM-Tree algorithm is to examine the subsequence inclusion in a way that items of the input sequence are processed exactly once. In the figure, 2.07% are support of the sequence *misc→ocal* and so on. The frequent tree mining task is done using **PD-Tree algorithm**. PD-Tree algorithm has a new method for determining whether a tree is contained by another tree. This is done by using *push down automaton.* The results thus obtained by using these algorithm are useful information about the user's navigation behavior.

Like the approach used in this paper, this thesis utilizes the the same web usage mining approach including preprocessing, pattern discovery and pattern analysis. Unlike this paper which uses *frequent pattern* mining to analyze web log data, this thesis uses *process mining* for analyzing the web event log files from news website. The paper extracts categories to analyze the sequence mining and frequent mining using several algorithms as mentioned before, however, this thesis uses process mining approach to analyze the

[5]http://kdd.ics.uci.edu/databases/msnbc/msnbc.html
[6]http://lisp.vse.cz/challenge/CURRENT/

opinion & misc & travel	-->	on-air	90.26%
news & misc & business & bbs	-->	frontpage	90.24%
living & business & sports & bbs	-->	frontpage	90.00%
news & misc & business & sports	-->	frontpage	89.68%
news & tech & living & business & sports	-->	frontpage	89.00%
news & living & business & bbs	-->	frontpage	88.01%
frontpage & tech & living & business & sports	-->	news	87.87%
frontpage & opinion & living & sports	-->	news	87.81%
frontpage & tech & opinion & living	-->	news	87.60%
frontpage & tech & on-air & business & sports	-->	news	87.59%
news & misc & sports & bbs	-->	frontpage	87.55%
news & tech & on-air & business & sports	-->	frontpage	87.43%
news & living & business & sports	-->	frontpage	87.18%
news & business & sports & bbs	-->	frontpage	86.70%
misc & living & travel	-->	on-air	86.55%
tech & living & sports & bbs	-->	frontpage	86.52%
tech & business & sports & bbs	-->	frontpage	86.40%
news & misc & living & business	-->	frontpage	86.22%
on-air & business & sports & bbs	-->	frontpage	86.22%
news & tech & misc & bbs	-->	frontpage	86.18%
on-air & misc & business & sports	-->	frontpage	86.16%
tech & misc & travel	-->	on-air	86.09%
tech & living & business & sports	-->	frontpage	86.08%
news & living & sports & bbs	-->	frontpage	85.99%
misc & business & sports	-->	frontpage	85.79%
frontpage & tech & opinion & sports	-->	news	85.78%
news & opinion & living & sports	-->	frontpage	85.69%
misc & business & travel	-->	on-air	85.65%
news & tech & misc & business	-->	frontpage	85.63%
misc & business & bbs	-->	frontpage	85.57%
tech & living & sports & bbs	-->	news	85.49%
local & misc & business & sports	-->	frontpage	85.43%
news & opinion & business & bbs	-->	frontpage	85.32%
news & misc & living & sports	-->	frontpage	85.19%
news & on-air & business & sports	-->	frontpage	85.01%

(a)

misc → local	2.07%
frontpage → frontpage → sports	2.02%
local → frontpage	1.83%
on-air → misc → on-air	1.72%
on-air → frontpage	1.69%
on-air → news	1.51%
news → frontpage → news	1.49%
local → news	1.46%
frontpage → frontpage → business	1.35%
news → sports	1.33%
news → bbs	1.23%
health → local	1.16%
misc → frontpage → frontpage	1.16%
on-air → local	1.15%
misc → on-air → misc	1.15%
frontpage → frontpage → living	1.14%
local → frontpage → frontpage	1.13%
health → misc	1.12%
misc → on-air → on-air	1.10%
local → misc → local	1.09%
misc → news	1.06%
news → living	1.06%
on-air → misc → on-air → misc	1.00%

(b)

Figure 3.4: (a) Association rules and (b) Sequential rules based on the `msnbc` data [39]

category to see user reading behavior over several categories.

3.5 Next Step Recommendation and Prediction based on Process Mining in Adaptive Case Management

This paper *Next Step Recommendation and Prediction based on Process Mining in Adaptive Case Management* [41] by *Sebastian Huber, Marian Fietta , Sebastian Hof* investigates how *process mining* approaches for predictions and recommendations based on event logs can be integrated into existing ACM solutions.

ACM is a new paradigm that facilitates the coordination of knowledge work through case handling. Current ACM systems lack support for sophisticated user guidance and for next step recommendations and predictions about the case future. This paper investigates on process mining recommendation and prediction approaches and integrates them into an existing ACM solution. According to the paper, the goal was to come up with a prototype that gives next step recommendations and predictions based on process mining techniques. The models proposed, recommend actions that shorten the case running time, mitigate deadline transgressions, support case goals and have been used in former cases with similar properties. In addition to this, the authors makes a final evaluation to prove that the prototype is indeed capable of making proper recommendations and predictions.

The paper seems to contribute in three discrete ways as follows:

1. Literature review of publications from different research areas regarding predictions

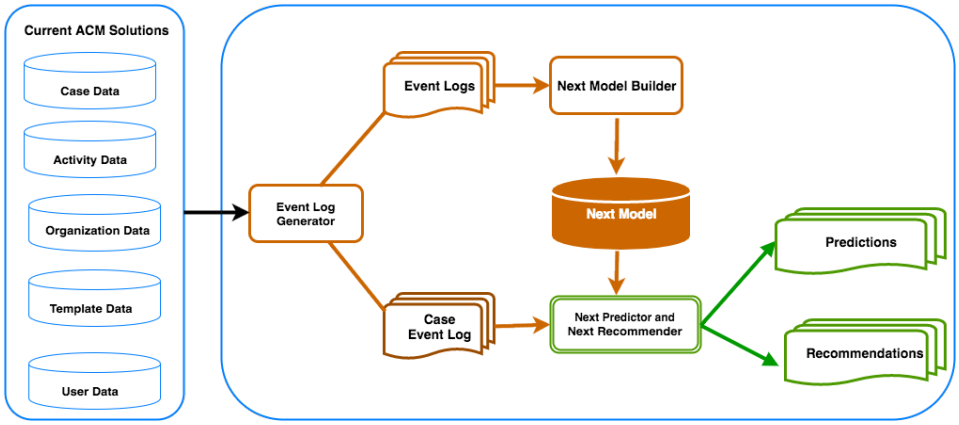


Figure 3.5: Architecture Overview of the prototype

and recommendations

2. Depiction of the conceptual approach for a software prototype based on these literature review

3. Evaluation of quality of prototype developed

As illustrated in figure 3.5, the prototype consists of four main parts - *the event log generator, the model builder, the generated model and predictor/recommender*.

- The **event log generator** produces log entries which represent the subjects activities in a standardized XES event log format. Those logs are based on historic events and operational data from the ACM system.

- The event logs are used by the **model builder** to construct underlying models based on different approaches. The model builder uses the stored activities to derive different models. Four models covering various aspects (time, deadline, decision, and goals) are created by applying different algorithms. Separate results are generated by every model, before they are combined into a composite model, called the next model.

- The **next model** is capable of predicting the future and recommending items based on various mining techniques. The used algorithms behind the next model are derived from data mining, statistics, and process mining.

- Generated models are used by the **predictor and recommender** to create predictions and recommendations for running cases. The next predictor and recommender is based on the next model and a partial trace of a currently running case. It takes all underlying models of the next models and applies the algorithms for each model.

The conceptual prototype developed is implemented as software by the authors. Collaborative Case Management is used as a platform and the prototype is applied as a plug-in

to extend the range of CoCaMas functionality [42]. The implemented model generates predictions and recommendations on the basis of different criteria. In addition to this, the prototype can predict remaining times, deadline violations, and goal support of running cases. A prototype evaluation based on fictive cases and activities is done that reveal the expected outcomes of tests.

Like this paper, the thesis analyzes the log files to generate process model that can be used for prediction of next click. However, this paper uses the event logs to describe how recommendation and prediction concept in PROCESS MINING can be applied to current ACM. The paper uses several factors and algorithms to make prediction and recommendation unlike this thesis which considers *popularity* as major factor to make prediction.

3.6 Summary

As listed in [40], there are several, works being carried out under the heading of *process mining*. Most of the works done with process mining deals in the field of medicine, ERP applications, E-commerce applications and others. Having countable number of works done in process mining with web logs data, some of them are summarized in this chapter.

Chapter 4

Methodology

This chapter focuses on how process mining was performed on event logs files including *data extraction, data preprocessing and process discovery.* Section 4.1 describes how events log data was extracted from Cxense. Section 4.2 describes pre-processing done the data. Finally, section 4.3 outlines different types of steps taken in DISCO for process discovery and other analysis as described in chapter 5.

4.1 Data Extraction

Figure 4.1 shows the general overview of getting data from Cxense API and using it for process mining. Data extraction is the first step in the process, which is followed by data preprocessing and then using it for process discovery. **Data extraction** is one of the important step which includes getting event logs from the media site. User behaviors on the news website are recorded anonymously by Cxense. Using its API, JSON event log files were extracted. The API provided several parameters during data extraction like *fields and length of data to be extracted.* The structure of JSON file is presented in Appendix A. The extracted JSON event log files contain fields as shown in table 4.1. As mentioned in section 2.3, *activity, timestamp, case and resource* are minimal requirements for *process mining.*

The table 4.1 shows different entries like *time, userCorrelationId, eventId, sessionStart, sessionStop, sessionBounce, browser, os, deviceType, url, refererUrl, referrerHost, referrerHostClass and referrerSocialNetwork.* The time is in UNIX format, which is converted into readable data format in data preprocessing stage. The *userCorrelationId* refers to the cross-site identifier used to differentiate devices/browser. This *userCorrelationId* is used as CASE in this work. Brief overview of terms is summarized in table 4.1.

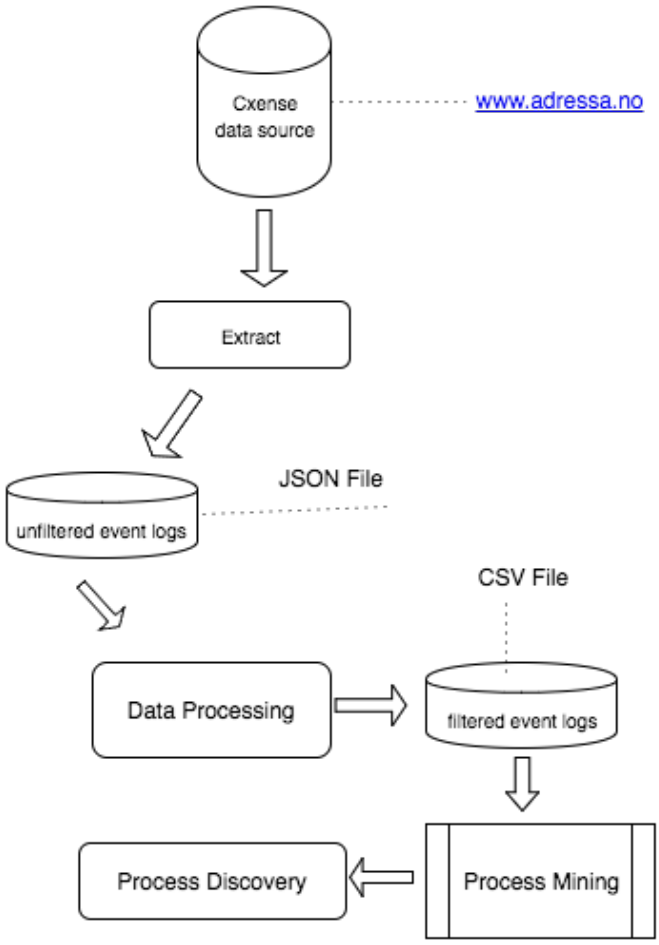


Figure 4.1: Overview showing the workflow of getting data and using it for process mining

4.2 Data Preprocessing

As aforementioned, preprocessing consists of converting the usage, content and structure information contained in the data sources into the data abstractions necessary for pattern discovery. This stage consists of converting the event logs into format suitable for process mining. As show in figure 4.1, *unfiltered event logs* i.e. JSON file is converted int CSV file, referred as *filtered event logs* in the figure.

4.2.1 Selecting case ID

From the available set of fields in table 4.1, *userCorrelationId* is chosen as the case ID, as it refers to cross-site identifier used to differentiate devices/browsers. The *eventId* on the other hand is used to differentiate distinct events from the same user, hence leaving

Table 4.1: JSON file entries brief definition

Name	Description
time	The time of event, measured in Unix time. Returned by default.
userCorrelationId	The cross-site identifier used to differentiate devices/browsers.
eventId	The identifier used to differentiate distinct events from the same user.
sessionStart	Indicates whether the event is considered as the first event in session.
sessionStop	Indicates whether the event is considered as the last event in session.
sessionBounce	Indicates whether the event is considered as the only event in session.
browser	The name of the browser.
os	The name of the operating system.
deviceType	The type of the device.
url	The URL of the visited page.
referrerUrl	The name of the site hosting the referrer page.
referrerHost	Classification of the referrer host.
referrerHostClass	The name of the referrer search engine.
referrerSocialNetwork	The name of the referrer social network.

userCorrelationId to be chosen as case Id.

4.2.2 Timestamp

Each event occurred at particular instance of time. The Cxense API recorded time in UNIX format which wan converted to readable data format using simple PYTHON function like `datetime.datetime.fromtimestamp(element["time"])`. Time in UNIX format like *1446753600* converts to *2015-11-05 19:59:58* in readable format giving required timestamp for the process mining tool DISCO.

4.2.3 Activity

Code snippet in Appendix B shows the logic behind extracting *activity*. Since *article's category* or *content type* is only recurring meta-information relating to articles, articles' category stood as best candidate for *activity*. The categories are important factors as they represent a way of identifying the readers interests. In addition to this, this project extracts referral hosts for each activities. As shown in table 4.1, each event is associated with particular *referrerUrl* like `http://google.com/search?q=cooking` or *referrerHost* like `google.com` or *referrerSocialNetwork* like `www.facebook.com`. To get the activity, the categories were extracted from *url*.

For example, for the event shown in listing 4.2, the activity extracted from URL is `m.facebook.com` and *kultur*. Each event in log file is duplicated to have *referrerHost* as activity, and *category* as second activity provided *referrerHost* is not `adressa.no` or *referrerHostClass* is not *internal*. *Timestamp* of the first activity is chosen as 2 second earlier than the later. In case of absence of *referrerHost*, *referrerSocialNetwork* is used as activity. Also for two or more *userCorrelationId* having same *referrerHost* is modified to be unique by appending unique *id* to it. This is achieved by code shown in Appendix C.

Listing 4.1: Pseudocode for generating artificial activity

```
Input: E = {e1, e2, e3, ..., en} (set of events to be
    processed in JSON Event File)
n = Total number of events
Output: C = {c1, c2, c3, ...., cm} (set of filtered events
    in SortedCSV file)

foreach entries in JSON Event File
{
  if(entry->referrerHost != 'adressa.no' ||
      refererHostClass !='internal')
  {
    - insert entries into sortedCSV
    - with activity = category to be extracted from URL
    - userCorelationId as entries just before it
  }
  else
  {
    - duplicate entries with
    - referrerHost as category with timestamp-2 second
    - second entries as done before
  }
}
```

Listing 4.1 represents pseudo code for generating artificial activity in the thesis. As mentioned before, the idea is to duplicate entries *referrerHost* as activity and timestamp-2 as timestamp. This way session is modified to generate new session id. If *referrerHost* is `adressa.no` or the *refererHostClass* is internal, then the entries is inserted with activity as category extracted from URL and userCorelationId same as entries before it.

Listing 4.2: An Example Event

```
{
      "time": 1446753602,
      "browser": "Chrome",
      "deviceType": "Mobile",
      "eventId": 302764119,
      "os": "Android",
      "referrerHost": "m.facebook.com",
      "referrerHostClass": "social",
      "referrerSocialNetwork": "Facebook",
      "sessionBounce": true,
      "sessionStart": true,
      "sessionStop": true,
      "url": "http://adressa.no/kultur/2015/11/05/%c2%abka-
          du-sei-ferr-n%c3%a5kka%c2%bb-11775939.ece",
      "userCorrelationId": "3aed656733365d25"
```

```
15      },
```

4.2.4 Resources

The event logs contained few candidates for *resources* including *browser, os and deviceType*. This thesis uses *deviceType* as the main resource and *os and browser* as *others* for process mining.

After applying scripts as mentioned in section 4.2, the final CSV hence achieved is ready for *process mining*. This file is imported into DISCO for further analysis.

4.3 Process Discovery

This step involves using processed event logs file for *process mining.* DISCO provides easy GUI interface to import, select the minimal required fields and start analysis. To use process mining on event logs, following steps are required:

- Download, install and register DISCO
- Import Data sets
- Start analysis

Downloading, installing and registering DISCO is straightforward as explained in **Fluxicon** website[1].

4.3.1 Importing Data sets

The event logs file is imported by clicking on import icon which is symbolized by the folder icon. The importer allows to choose the appropriate file from specified locations. This project uses CSV file as mentioned before. However, DISCO is pre-configured with various standard formats including *.mxml, .mxml.gz, .fxl, .dsc and .csv*.

4.3.2 Configuration Settings

Once CSV file is loaded in DISCO, it facilitates for configuration as shown in figure 4.2. DISCO is pre-configured to guess the columns as *case Id, activity, timestamp or resources*, but correct configuration can be set by the users before actual importing. DISCO allows to configure several other settings in addition to this like as follows:

- Configuration timestamp patterns
- Combining multiple case Id, activity or resource columns
- Swapping cases, activities and resources
- Importing pre-configured data sets

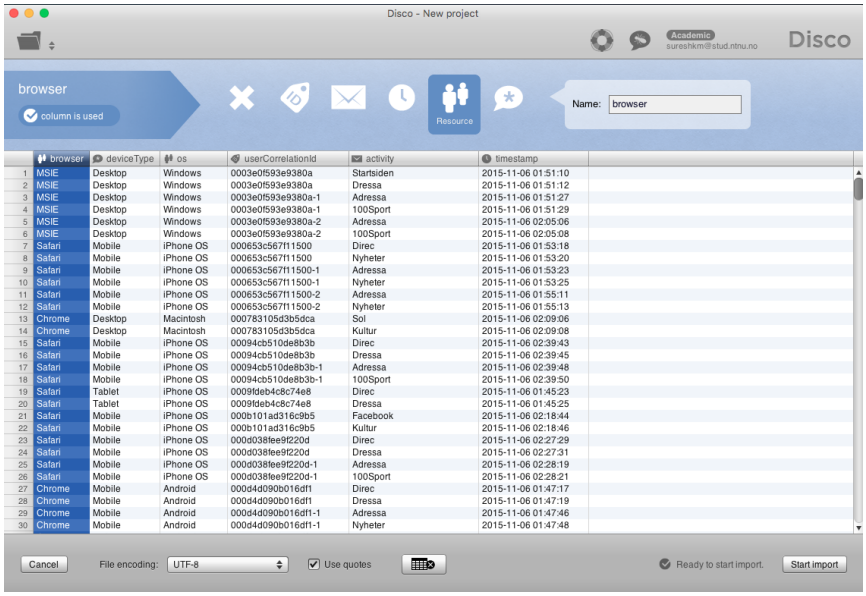


Figure 4.2: Import configuration screen in Disco

This thesis, process the JSON file to get filtered event logs with only six fields namely *activity, browser, device type, OS, userCorrelationId* and *timestamp*. Original JSON file contained 14 different fields as shown in table 4.1.

4.3.3 Analyzing Data Sets

After importing the data into DISCO, data analysis was done using three major analysis views namely *Map, Statistics and Cases*. Figure 4.3 shows three different views that was used to gather analytics from the web log data.

Map View

Map view shows process map that visualize actual flow of the process based on imported CSV data. Map view includes several elements like *canvas with process map, zoom slider, map detail controls, process map visualization options, filtering and animation.* Activities slider and path sliders are used to control the number of activities and the number of paths to be included in process map.

Statistics View

Statistics view is used to gather additional information and detailed performance metrics about the process. Global statistics about the log file, individual cases and variants was extracted from *global statistics* from **Statistics view**. In addition to this, statistics view

[1] http://fluxicon.com/disco/

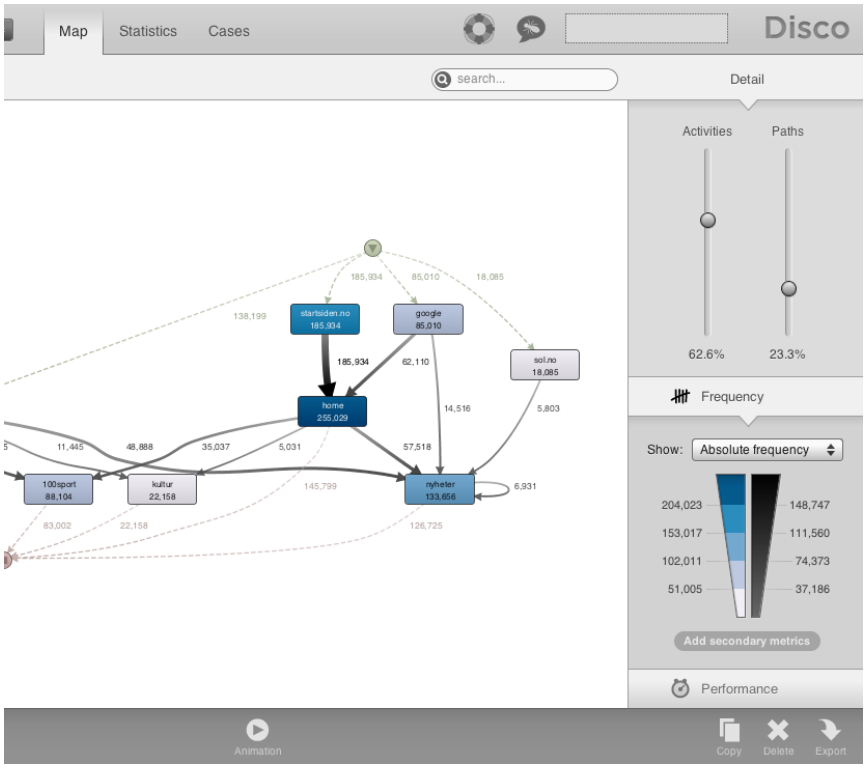


Figure 4.3: DISCO screen showing different controls and *map view, statistics view and case view*

was used to get other useful information like *activity statistics, resource statistics and attribute statistics.* Table 5.1 was constructed using Statistics view from the software. Global statistics provides chart to view view global metrics of the log files like *events over time, active cases over time, case variants, events per case, case duration and case utilization.* These metrics were can be used to analyze the log files at greater depth giving actual insight of the the process used in the system.

Cases View

Cases view was used to inspect individual cases and see the raw data. This view was used to see individual logs entries or complete log entries, individual variants, list of cases, or searching particular cases. Any particular cases could be searched analyzed from this view.

4.3.4 Filtering

One of the important feature of DISCO is *filtering* that can be used to filter log data as the name implies. These filters can be used to narrow down the scope of analysis to be done. It helps to scale down the analysis in one hand and allows to inspect events as a finer level

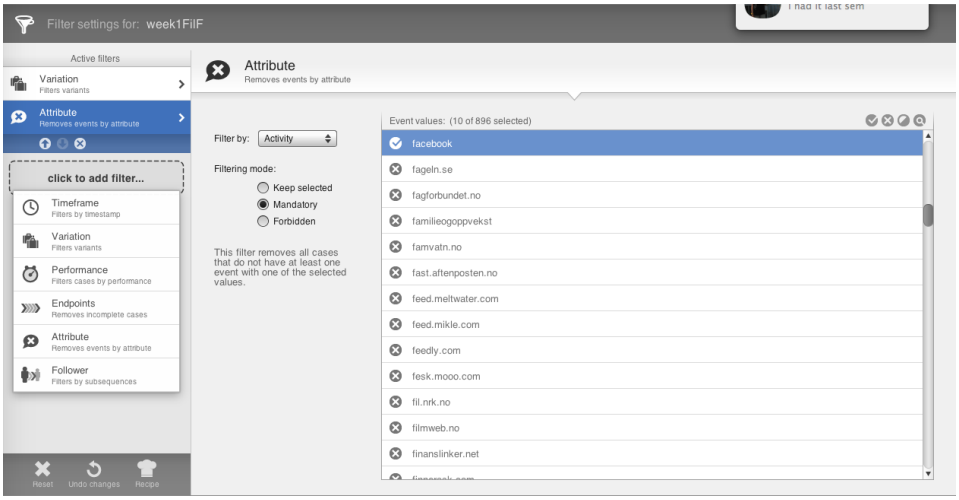


Figure 4.4: Using *Variation Filter* and *Attribute Filter* to filter event logs

on the other hand. In DISCO, When clicked on *filter* icon, the software presents user with a list of *recommended filter* suitable according to imported data. There are several types of filters available on the tool. The thesis make use of the following filters:

1. **Variation filter:** If DISCO finds that there is a high variation of case behavior, which generally leads to large and complex process models, it suggests to apply the Variation filter to help streamline some of these variation.

2. **Performance filter:** DISCO allows to know that cases in the log have highly varying throughput times and recommends the Performance filter to focus the analysis on cases that meets the specific performance requirements.

3. **Attributes filter:** The Attribute filter allows to filter events or cases based on arbitrary attributes in the data set. This filter allows to select *event column to be used* and *filtering mode*. For example to mine the model as shown in figure 5.10 and figure 5.11, attribute filters are used. Here for 5.10, *deviceType* is selected from *Filter by* drop down and *desktop* is highlighted from *event values*.

Two important filters used in this thesis are *variation filter* and *attributes filter*. As shown in the figure 4.4, there are different types of filters that can be used on the event log. *Variation filter* is the first filter used here. As DISCO detects high variation of case behavior, which typically leads to large and complex process models, variation filter is used to streamline some of these variation. Model mined in figure 5.4(a) and (b) are generated using variation filter containing 80% of the events, 92% of the cases and approximately 1% of the variants. Second filter is *Attribute filter* used to filter activities. This thesis tries to filter out the activities with traffic source coming from `www.facebook.com` and `www.google.com`. So, *attribute* filter is applied. Figure 5.9 is the model mined using both *variation filter* and *attribute filter*.

In addition to these features DISCO allows to import the mined model in different image format (JPEG, PNG) as well as allows to export the event logs, analysis etc in various formats(CSV, XML, XES). The extracted charts, figures and log files are used as basis for analysis in chapter 5 and chapter 6. Figure 4.3 shows these export icons that were used for exporting the results.

Chapter 5

Analysis

The chapter presents the results of the analysis with examples of extracted models from *process mining*. Section 5.1 shows the general descriptive statistics that provides some initial insights about the event logs, device types, operating systems and browsers. Section 5.2 describes general "spaghetti model" extracted from week one data. Moreover, it also describes the model with filtering applied on activities and transitions. General reading time of traffic at 80% of the events is analyzed in section 5.3 and discusses on mean and median reading duration of traffic. Section 5.4 discusses on referral model extracted from `Google` and compares with `Facebook`. Section 5.4.1 describes model produced from event logs with traffics originating from `Google` where as Section 5.4.2 analyzes the traffic coming from Facebook and Google. Section 5.5 analyzes traffic sources coming from Facebook and Google to inspect if there is different news consumption models.

5.1 General Statistics

The thesis analyzes the event logs of November 2016 in weekly basis. Table 5.1 illustrates some general statistics of the event logs file in weekly basis. As shown in the table approximately 14.5 million events are analyzed for the month. The *median case duration* for each weak is approximately 2 seconds. *Mean case duration* for each weeks are 17.7 minutes, 17.6 minutes, 16.6 minutes, 16.6 minutes and 11.2 minutes respectively. The entire month contained total of 60,894,043 distinct *cases*. 974 distinct activities were included in week one. Other weeks contained 916, 892, 912 and 514 distinct activities respectively. However, for analysis using *process mining*, only week 1 data is considered.

5.1.1 Browser Statistics

Table 5.2 shows statistics of browsers like *Safari, Chrome, MSIE, Firefox, Android, Opera, Unknown and Blackberry* used by traffic for reading news articles. As seen in figure 5.1,

Table 5.1: General statistics of log events according to weeks

Description	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**
Events	14,499,725	13,720,606	14,361,143	14,645,255	3,667,314
Cases	5,589,032	5,331,860	5,629,454	5,781,095	1,505,539
Activities	974	916	892	912	514
Median Case Duration	∼2 secs	∼2 secs	∼2 secs	∼ 2 secs	∼ 2 secs
Mean Case Duration	17.7 mins	17.6 mins	16.6 mins	16.2 mins	11.2 mins
Start	01.11.2015	08.11.2015	15.11.2015	22.11.2015	29.11.2015
End	08.11.2015	15.11.2015	22.11.2015	29.11.2015	01.12.2015

Table 5.2: Traffic using different types of browser

Browers/Weeks	Week 1	Week 2	Week 3	Week 4	Week 5
Safari	5,305,602	4,979,608	5,188,492	5,211,122	1,359,826
Chrome	5,173,591	4,921,125	5,016,990	5,166,649	1,312,062
MSIE	2,352,362	2,251,120	2,513,506	2,619,915	573,961
Firefox	1,138,396	1,077,261	1,127,878	1,137,346	294,218
Android	356,183	330,281	344,906	342,107	85,509
Opera	166,954	155,025	160,660	161,341	40,228
Unknown	5,743	5,405	7,730	5,806	1,219
BlackBerry	894	781	981	969	291

Safari[1] is one of the most preferred web browsers for reading articles. **Google Chrome**[2] on the other hand takes second position. *36.59%* of the users used *Safari* as web browser in Week 1 where as *Google Chrome* is used by *35.68%* of users. *16.22%* of traffic used MSIE and *7.85%* used *Firefox* in week one. The rank of other browsers used are shown in table 5.2 according to weeks.

5.1.2 Operating System Statistics

Figure 5.2 depicts different types of Operating Systems used by traffic when reading news articles on weekly basis. The actual statistics of the traffic and operating system used are shown in table 5.3. As shown in the table, **Windows** is most widely used Operating System. In week one, *38.3%* of users used *Windows*, *33.75%* used *iPhone OS*, *20.55%* used *Android OS* and *5.85%* of users used *Macintosh*. It is interesting to find that, the percentage of users using these operating system does not vary much according to week. In addition to this, only smaller fractions of users reading news on `http://www.adressa.no/` use *Linux* or *BSD* as Operating System.

[1] http://www.apple.com/safari/
[2] https://www.google.com.np/chrome/

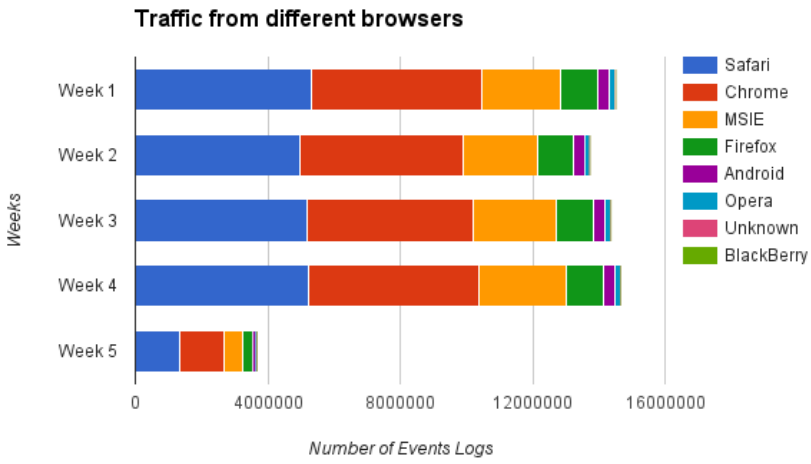


Figure 5.1: Traffic originating from different types of browsers

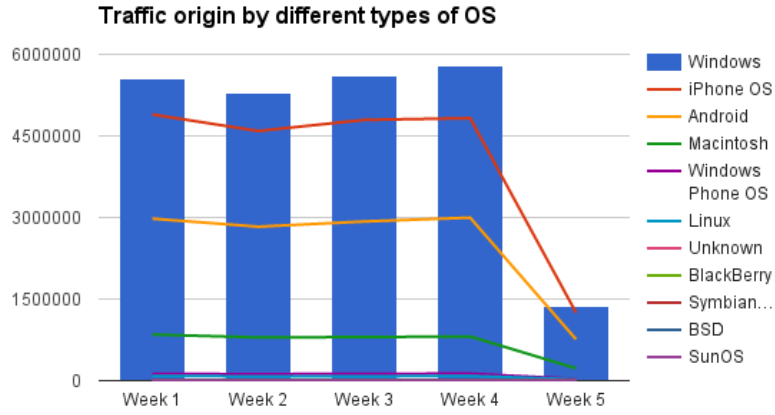


Figure 5.2: Traffic originating from different types of Operating System

5.1.3 Device Type Statistics

Table 5.4 contains the general statistics of users using different device types to read new articles. Figure 5.3 gives the graphical representation of the table. As seen in the table 5.4, *44.82%* of users used *Desktop* to read news articles in week 1 where as *37.85%* of users used *Mobile* and remaining *17.33%* used *Tablet*. As mentioned in section 1.2, it

Table 5.3: Traffic coming from different types of operating system

OS/Weeks	Week 1	Week 2	Week 3	Week 4	Week 5
Windows	5,553,642	5,288,353	5,616,818	5,776,406	1,372,715
iPhone OS	4,893,896	4,590,540	4,794,430	4,826,659	1,249,771
Android	2,979,226	2,830,724	2,926,902	2,997,181	759,129
Macintosh	848,283	795,256	801,587	812,487	226,207
Windows Phone OS	125,965	119,225	126,803	129,280	31,043
Linux	83,642	81,432	79,533	87,399	24,220
Unknown	13,360	13,473	13,061	13,860	3,734
BlackBerry	898	794	1,001	992	294
Symbian OS	498	417	452	447	90
BSD	167	267	436	396	84
SunOS	148	125	120	148	27

Table 5.4: General statistics of device types according to weeks

Device Type/Week	Week 1	Week 2	Week 3	Week 4	Week 5
Desktop	6,498,864	6,178,648	6,511,255	6,690,317	1,626,892
Mobile	5,488,695	5,193,762	5,362,413	5,479,596	1,392,069
Tablet	2,512,166	2,348,196	2,487,475	2,475,342	648,353

seems the number of traffic from mobile devices and tablets are more than that coming from *Desktop*. It is quite interesting to find the trend of more mobile traffic compared to desktop, is same over weeks of time.

5.2 General Model

This section describes the general model extracted from log files of week one data and gives answer to the research question "What kind of news consumption models do different web sources drive?" in section 1.1.1. In addition to descriptive statistics described in section 5.1, Week one data was used for process mining and the **Spaghetti Model** extracted from it is shown in figure 5.4(a). Each square in the image represents the name of the activity. As seen in the figure 5.4(a), default activities *starts* and *stops*, represented as circles. The arrows mark the transitions from one *activity* to another. The size and color of the activities illustrates insights about metrics. Thicker the arrow's line, higher is the transaction and darker the color, higher is the frequency. Figure 5.4(a) shows how the event logs from `adressa`, generates a complex *spaghetti* model when all the activities and transactions are visualized.

Figure 5.4(b) depicts a model extracted from the same event log as in 5.4(a), but with frequency threshold of 23.7% on activities and 3.4% on transition edges to show a comprehensible model. Some of the noticeable observations that can be made from figure

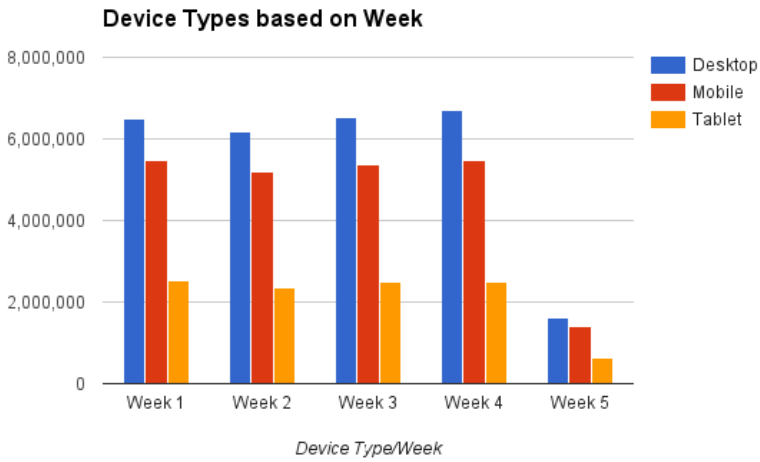


Figure 5.3: Traffic reading news on different device types

5.4(b) are enumerated below:

1. A large number of traffic originating from *facebook* goes to *news*(nyheter). Similarly, traffic from `startsiden.no`, `google.no` goes to home page.

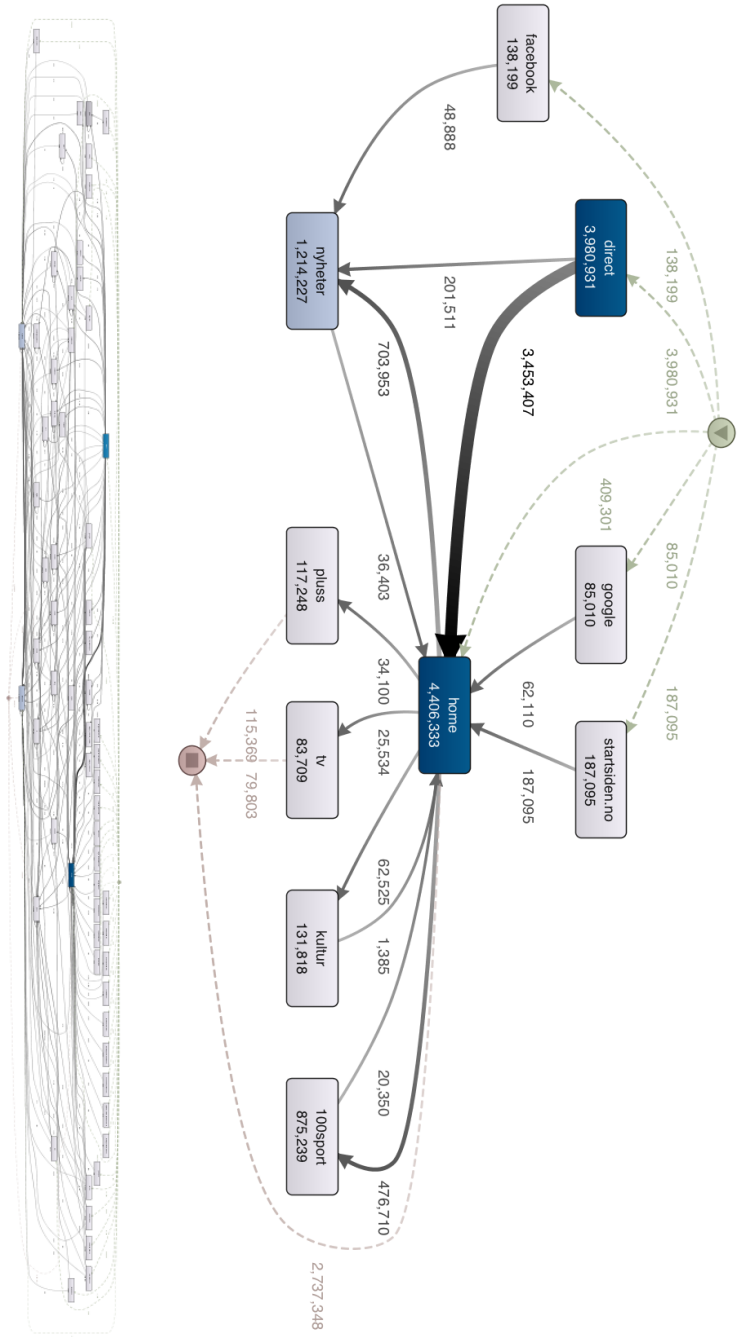
2. Most frequent **referrals** at 23.3% activities are `facebook.com`, `adressa.no`, `google.no` and `startsiden.no`. Traffic originating from `adressa.no` goes to either *100sport, home page or nyheter(news).*

3. At 23.7% activities, the most frequent news categories are *nyheter(news), kultur(culture), TV, pluss(Plus) and 100sport(Sports).* News is one of the most frequent items in the news categories which shows most of users coming to the site, starts the session by looking at the news items.

4. The figure also shows *category-activity* and transitions that are labeled with frequency values. The frequency values represent a measure of popularity. For example, news category *nyheter*(news) is accessed 1,213,119 times. Also, 48,888 users read news on this category originating from `facebook.com`.

5. Looping cycles are common for all the categories. This means that the next news article read is likely to be within the same category as the previous news article. For example, looping is seen in news category *home, news and 100sport* in this figure.

6. News categories has outgoing transitions to other categories-activities at the selected filtering levels. For example, in the figure, transition from home page to other categories like *pluss, tv and kultur* are seen. The figure also shows a large number of traffic transiting from home page to *news* category, reading several news articles and transiting back to home page.



(a) **(b)**

Figure 5.4: (a) Example of mined ”spaghetti” model before frequency filtering. (b) Example of mined model showing the 23.7% most frequent activities and 3.4% most frequent edges.

5.3 General Traffic Reading Time

DISCO allows to see different metrics in the **Process Map** including case *frequency and performance.* The performance metrics allow to know about the *time* spent in the different parts of the process extracted. Figure 5.5(a) depicts the general median reading time at 80% events extracted with 24.1% most frequent activities and 22.1% most frequent edges. Figure 5.6(b) shows mean reading time of 80% of events.

The figure 5.5 shows the traffic coming from external sites like `Facebook`, `Google`, `startsiden.no` and `adressa.no` to different news categories like *nyheter, 100sport, tv, kultur and pluss.* The median reading time of referral host to categories is *2 seconds* as these referral host categories are artificial activities created from events from log files as explained in implementation chapter section 4.2. DISCO allows different performance metrics like *total duration, mean duration, median duration, maximum duration and minimum duration.* For the model mined here, median duration is considered. As shown in the figure, the median reading time of users on the home page is *10.2* minutes and that of *nyheter* is 78 seconds. Figure 5.6 shows the performance metrics on the news category *nyheter*. Total reading duration on this category is almost *11.3* years. Median reading duration is *78* seconds and mean duration is about *61.3* minutes. DISCO has ability to show such performance metrics for each activity or path giving deeper insights of the activity. Maximum duration shows the largest execution times and delays that were measured during the process. Mean duration as in figure 5.6(b) shows the average execution times for each activity and the average idle times on each path. Mean reading time of *home page* is about *103.9* minutes. Reading loop is found in many categories including *nyheter, sports* and home page. Mean reading time of *nyheter* is about *61.3* minutes and that of *100sport* is about *41* minutes. These insights clearly indicate news reading time on the categories like *nyheter, sports, tv, kultur and pluss* are higher compared to other categories. This could be because of more engaging news content on these categories than others.

5.4 Referral Host Model

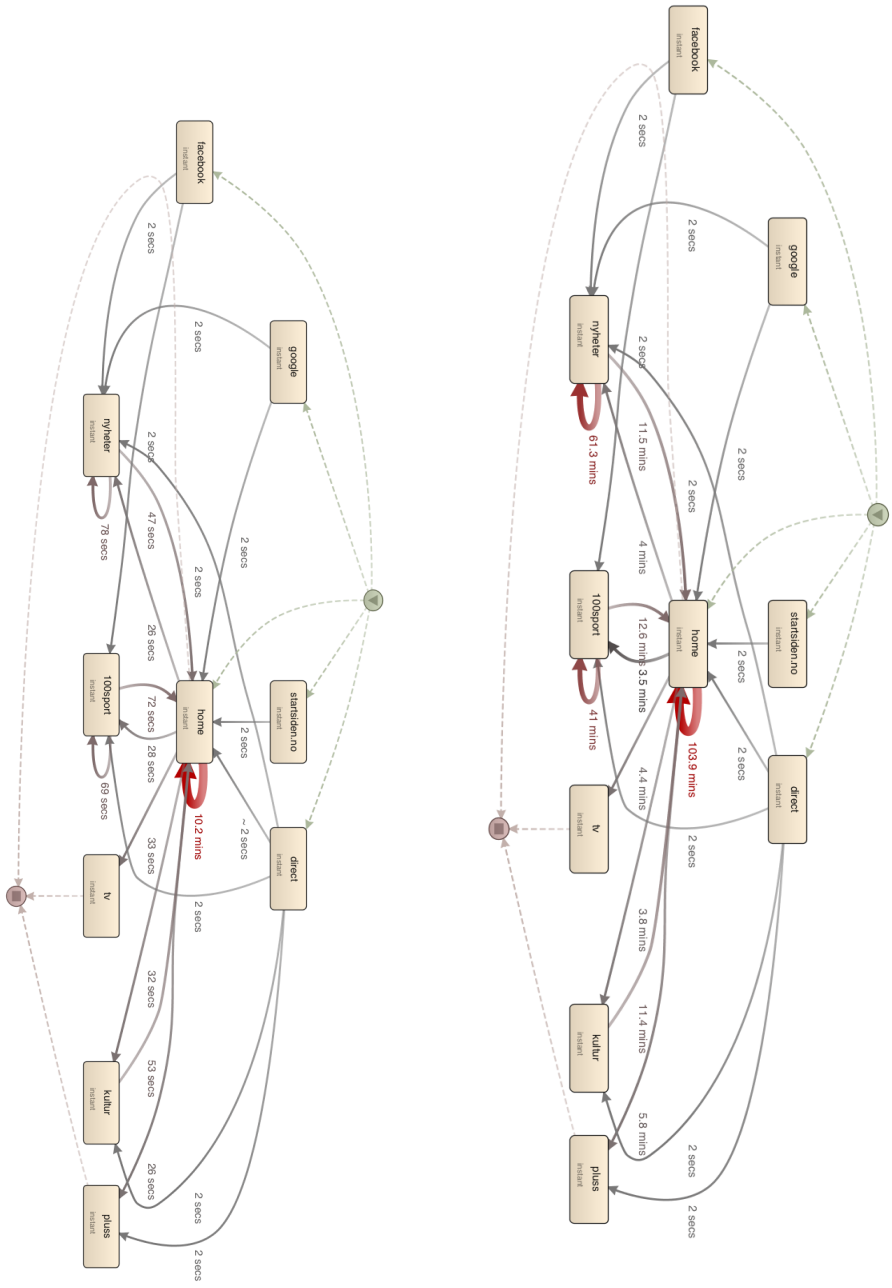
This section discusses on referral model extracted from `Google` and compares with `Facebook`. Section 5.4.1 describes model produced from event logs with traffics originating from `Google`. Section 5.4.2 compares traffic coming from Facebook and Google.

5.4.1 Google as Referral Host

DISCO has ability to produce model with various filters on activities and transitions paths. Figure 5.7 illustrates the mined model showing 21.2% frequent activities with 2.7% edges with referral host from `Google`.

Following are some of the important observations that could be made from the figure 5.7:

1. Most frequent news categories read by users coming from `Google` are *100sport, folk, nyheter,TV, kultur, search, pluss, bolig, forbruker and vaeret.* These news categories name are in Norwegian and its corresponding translations are given in Appendix in table C.1.



(a) **(b)**

Figure 5.5: (a) Median reading time. (b) Mean reading time at 24.1% most frequent activities and 22.1% most frequent edges

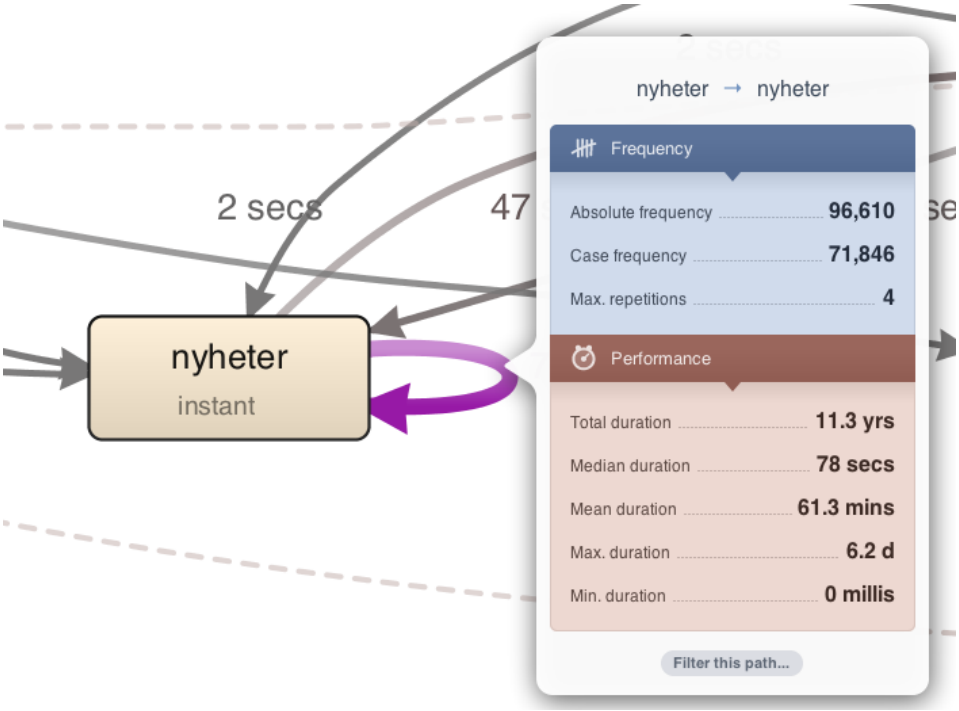


Figure 5.6: An overview badge of *nyheter* activity showing all metrics at a glance

2. In week one, *113,863* number of traffic originated from `Google` and landed on several news categories as shown in the figure. Darker color of `Google` indicates higher frequency of activities and thicker line from `Goolge` to home page indicates higher transactions.

3. Figure 5.8 depicts some of the most common news categories visited when traffic originated from `Google`. As seen in the figure, 45.53% of traffic coming from `Google` directly goes to home page of the site. This shows the users' searching behavior using browser. Instead of directly hitting the URL on the browser, users seem to hit `adressa` on the search bar and with `Google` being default search engine, lists the website on the top of search result page. Hence, most of the users land into the home page of the `adressa.no` from `Google`.

4. Second most popular category is *nyheter* which accounts 22.18% of total traffic in first week. Users from home page or categories like *vaeret, bolig, tv, pluss, search and kultur* seem to transit to *nyheter* categories. In addition to this, there is loop on the category indicating users reading different articles from the same categories.

5. *Sports and folk* are third and fourth most popular news categories with 12.55% and 6.27% traffic. Traffic to other news categories are shown in the figure 5.8.

6. As seen in the figure, users coming `Google` transit to *nyheter* or home page and

then to *nyheter*. This information can be used to recommend *nyheter* categories when users are on home page. The frequency values can predict what the user will click next most of the time, however these values does not say how much user preferred the contents.

7. One of the interesting observation that can be made from the model is the depth of interaction with the categories. Most of the users originating from `Facebook` transits to categories like *nyheter, kultur, pluss, 100sport, vaeret and tv*. Most of the traffic landing onto the *nyheter* categories continues reading news articles on the category and ends the session. Most of users tends to end session after reading from one category and then moving to home page. The reason could be people redirected by a certain news item shared on social media sites. They read the article and terminate the session.

Filtering activities is one of the advantages of process mining. It can be used to generate models enriched with various performance metrics like traffic originating from `Google` or `Faceook`. This analytics can give useful insights for business intelligence.

5.4.2 Comparing Google and Facebook as Referral Host

`Google` and `Facebook` have grown as a crucial source of incoming traffic, and have been vying with search as a source of new readers for some time. According to data from figure 5.9 `Facebook` surpassed `Google` as the top referrer for major news categories for the site. As seen in the figure, *149,698* users were referred from `Facebook` where as *113,863* users landed to the site using `Google`. As seen in the figure 5.9, `Facebook` drives traffic to news categories like *sports, vaeret, nyheter, tv, pluss and kultur.* Most of these traffic indicates news articles shared on `Facebook` by news website promotion group or the readers. Google on the other hand drives most of the traffic to news categories like *kultur, nyheter and home page.* It is interesting to to find that *nyheter* is one of the most read news categories on the site. 42,645 traffic were driven from `Facebook` to *nyheter* categories where as 9,355 number of traffic were driven from `Google`. In addition to this, news categories like *sports and nyheter* seems to engage users as seen with loops in the figure.

5.5 Referral Host with Device Types

This section compares traffic coming from various device types like desktop and mobile. Figure 5.10 shows model minded at 16.6% of most frequent activities and 2.5% of most frequent paths or edges. The model extracts traffic originating from `Facebook` and `Google` that are using desktop as main device type. On the other hand, figure 5.11 shows traffic originating from `Facebook` and `Google` using mobile and tablets as device type. To keep analysis consistent both the models were extracted at 16.6% of most frequent activities and 2.5% of most frequent paths. Mobiles and tablets are considered as similar entity for this comparison.

Comparison of above mined model extracted from week one log files shows following information:

Table 5.5: Traffic originating from Facebook and Google using different device type and transiting to different categories

Source/DeviceType	**Dektop**	**Mobile/Tablet**
Facebook	11.76 %	27.93 %
Google	22.59 %	14.31 %
nyheter	13.68 %	16.75 %
home	27.86 %	12.13 %
100sport	9.48 %	13.02 %
kultur	3.37 %	3.12 %
tv	3.3 %	2.16 %
pluss	2.99 %	2.17 %
folk	1.87 %	2.52 %
vaeret	1.2 %	0.72 %
incoming	0.84 %	0.47 %
forbruker	0.51 %	0.44 %
bolig	0.46 %	0.47 %
familieogoppvekst	0.22 %	0.27 %
meninger	0.21 %	0.47 %
reise	0.15 %	0.22 %
search	0.09 %	0.65 %
jobb	0.08 %	0.12 %
bil	0.07 %	0.1 %
digital	0.06 %	0.06 %

1. As shown in the table 5.5, *22.59%* of traffic coming `Google` used desktop to accessing the site, where as *11.76%* of users landed into site using `Facebook` and desktop. On the contrary, 27.93% of traffic coming from `Facebook` used mobiles or tablets as device type. *14.31%* of traffic was driven from Google and used mobile devices or tablets.

2. The table clearly shows that, traffic originating from mobile or tablets is always higher than that of desktop. On the contrary *27.86%* of traffic from desktop went to home page where as 12.13% of traffic from mobiles went to home page.

3. The number of visits per categories is higher from mobile devices than that of desktop. For example, *73,967* users from mobile visited *nyheter* with *7,697* looping around the same category. On the other hand, *30,670* visitors landed on *nyheter* category with *4,323* visitors looping around the category. This observation is valid for all other categories except *home page*.

To answer the research question mentioned above, the mined model shown in figure 5.10 and figure 5.11 and table 5.5 clearly show different news consumption models on desktop and mobile traffic. The number of traffic coming from mobile devices are higher than that of desktop. The rise in mobile phone traffic to online news sites is partly being fueled by the overall trend of using social media sites a major marketing channel. These

data in table 5.5 seems to show that desktops are being used to land directly on home page, while mobiles are used for more spontaneous, discovery-based landings on the news sites and its categories.

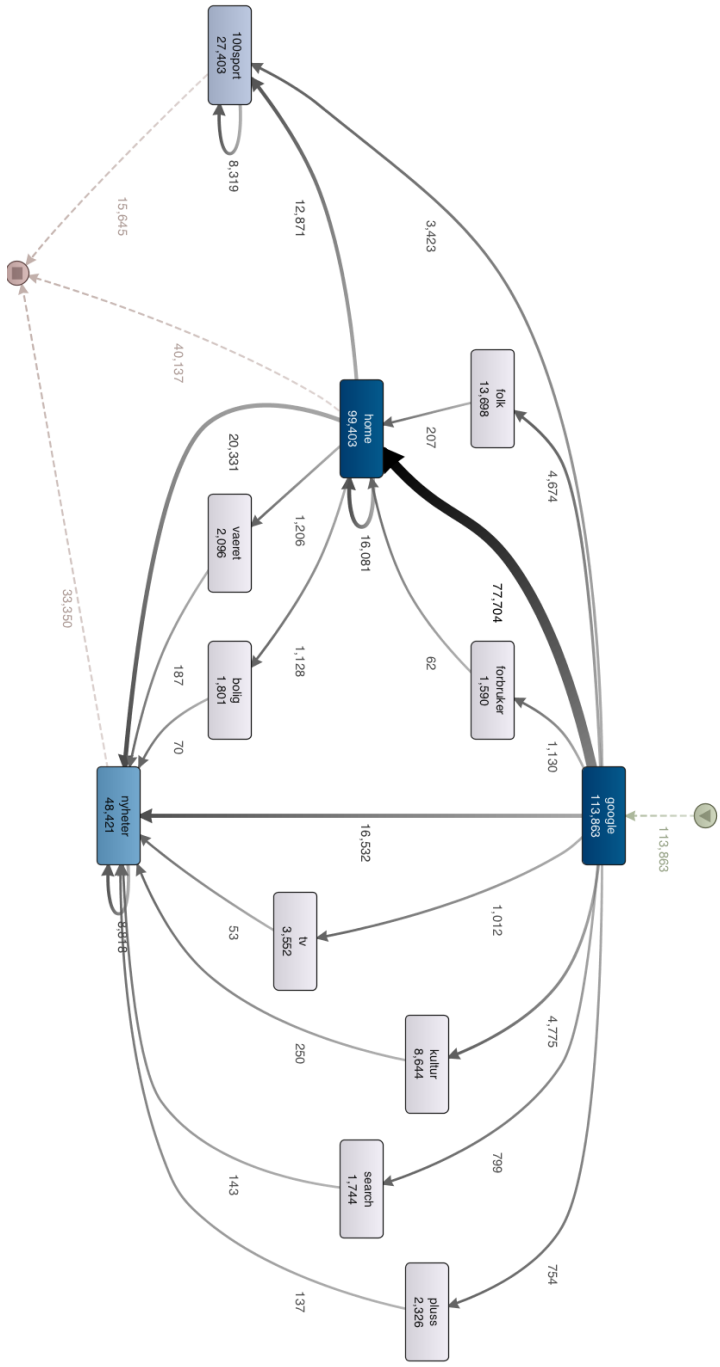


Figure 5.7: Example of mined model showing the 21.2% most frequent activities and 2.7% most frequent edges with referral host from `Google`

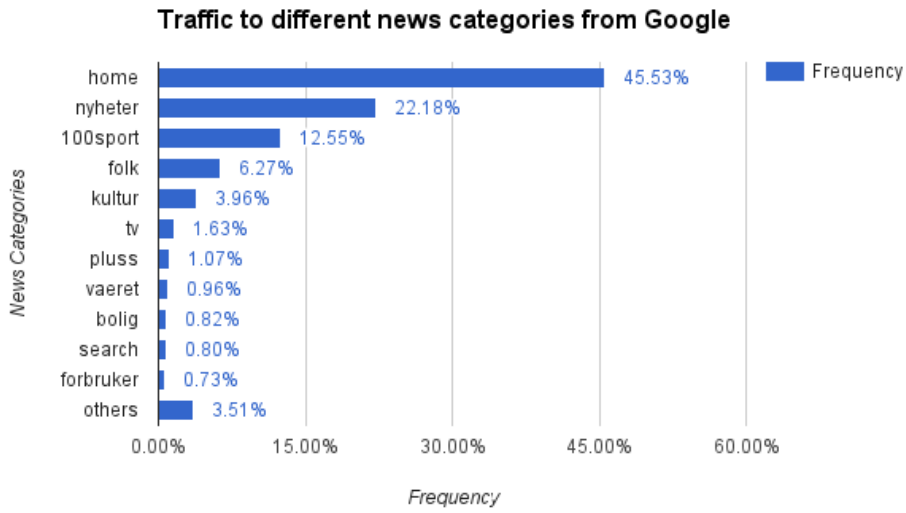


Figure 5.8: Traffic to different news categories coming from `Google`, extracted from figure 5.7

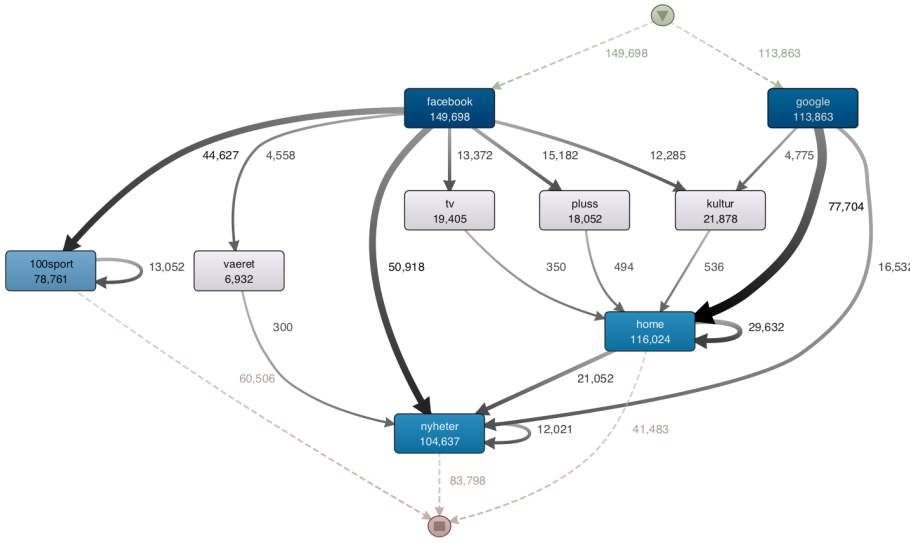


Figure 5.9: Comparing traffic coming from `Google` and `Facebook` at 15.3% most frequent activities and 2.5% most frequent edges

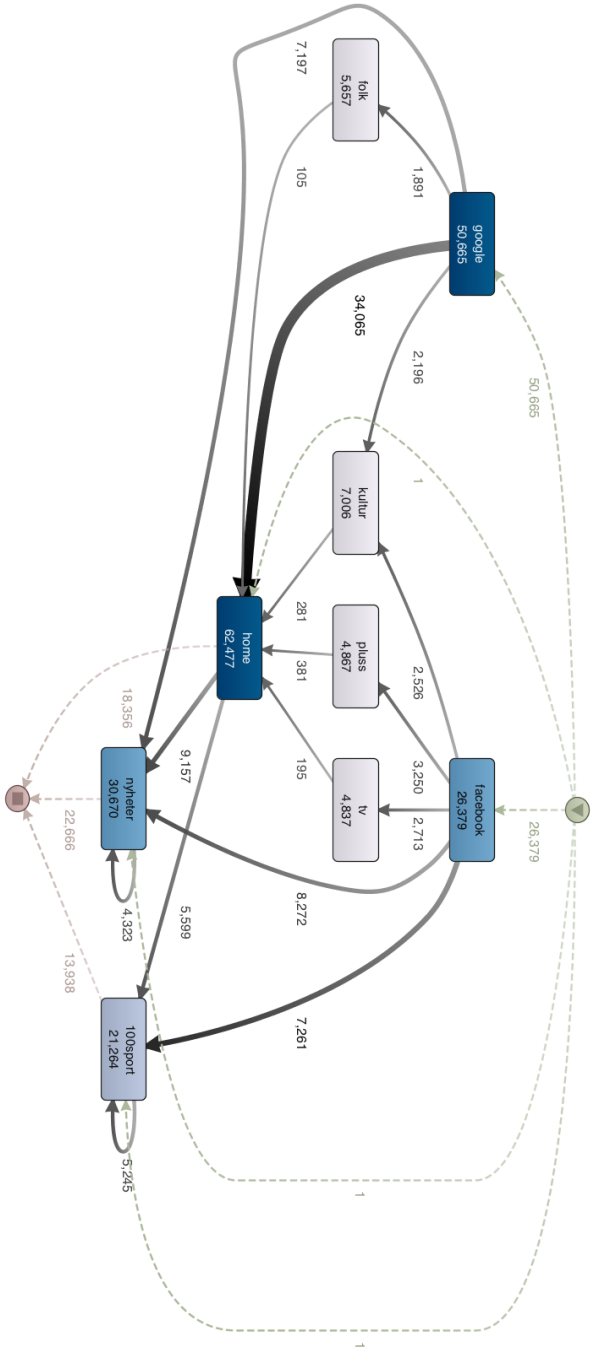


Figure 5.10: Google and Facebook Traffic originating from Desktop at 16.6% most frequent activities and 2.5% most frequent edges

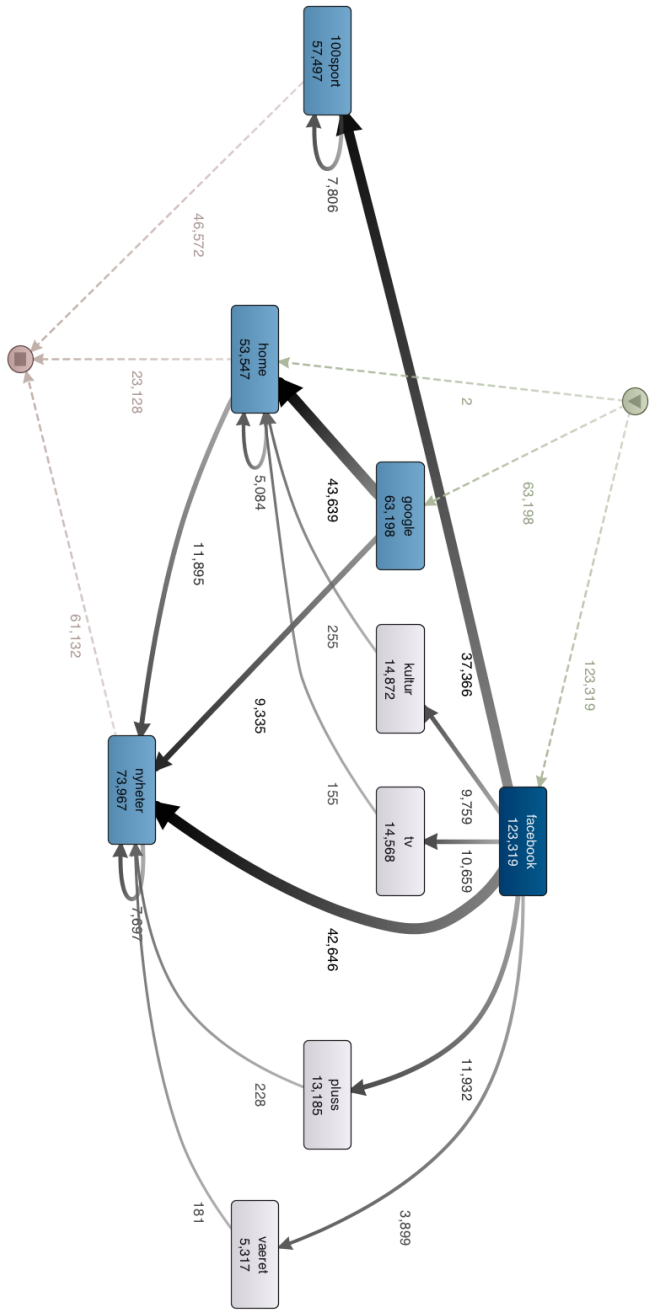


Figure 5.11: Google and Facebook Traffic originating from Mobile and Tablets at 16.6% most frequent activities and 2.5% most frequent edges

Chapter 6

Discussion

While the previous chapter 5 discussed about various analysis done with the example of mined model, this chapter discusses about the value and implications of the analysis and findings from this project. Section 6.1 answers the research question as mentioned in section 1.1.1. Section 6.2 presents how *process mining* can be used to predict next click for the news readers on the site. Finally, section 6.3 discusses some of the challenges faced during various approaches used in this thesis.

6.1 Answering Research Question

The thesis is shaped to answer some of the research questions as discussed in 1.1.1. This section answers those research questions.

1. **What are the state of art works in web log process mining done so far?**
   There are several works done in *web log mining*. Chapter 3 discusses some of the related works done in the field of web log mining. Like process mining used in *E-commerce web logs* as discussed in section 3.1, this thesis focuses on web click logs to analyze readers behavior. Similarly, section 3.2 discusses about process mining used in education domain which shows how social mining techniques can be used to examine and assess interactions between originators, training courses or pedagogical resources, involved in students training path. In addition to this, section 3.3 describes how process mining is used with web logs data extracted from MOOC site like `http://wwww.coursera.org`. This paper uses MOOC data to analyze students' learning behavior to provide insights regarding students and their learning behavior as it relates to their performance. The paper shows the result that successful students always watch videos in the recommended sequence and mostly watch in batch and the vice versa. Like this paper, this thesis uses process mining in *web logs*. However, the paper uses process mining in *education domain* or MOOC site where as this thesis uses process mining on data from news website. Finally, section 3.4 discusses three different pattern mining approaches from web

usage point of view. This paper also presents the problems for discovering hidden information from large amount of web log data collected from servers. Overview of the process of web log mining and demonstration of how frequent pattern discovery tasks can be applied on web log data in order to obtain useful information about the user's behavior are major contribution of this paper. The related works discussed in chapter 3 uses web logs data to mine useful information that are useful for business intelligence or provides useful insights about the process. Like these works, this thesis uses web log mining to extract some useful insight about the user reading behavior on the site. One of the key area of this thesis is **to reveal deep content interaction patterns connecting external sources and internal content types.** On the contrary, this thesis uses process mining to mine model that can be used for prediction of next click or as the basis for recommending news item from most popular news categories.

2. **What kind of news consumption models do different web sources drive?**
   News websites lack concrete process models like *E-commerce or appointment* websites where visitors follow definite process or path. In news website, users try to consume news items based on *personal preferences, recency and popularity*. One of the goal of this thesis was to discover what kind of news consumption models do different web source drive and how traffic follows between them. Section 5.2 discusses about the model mined from log files of week one data. Figure 5.4(b) is the model mined from these data that shows different web sources and associated news categories. The answer of following two questions accumulates the answer of second research question.

   (a) **What relationships do we see between web sources and news categories?**
   Figure 5.4(b), depicts different types of web sources and news categories that constitute most frequent activities at 23.7% and 3.4% most frequent activities. It is notable that most frequent web sources for the site are *startsiden.no, facebook, google, sol.no, adressa.alda.no, mobil.sol.no, bing.com, wazzup.polarismedia.no, onlineaviser.no, startsida.no, hemnestart.no, and others.* as listed in table C.2. The table C.2 shows most frequent web sources from week one data. On the other hand, most of the these traffic sources drive traffic to news categories to home page, *nyheter, 100sport, kultur, pluss, tv, vaeret, folk, bolig, incoming and others* as shown in the figure 5.4(b). Most of traffic coming from `startsiden.no` and `startsida.no` transits to home page. In addition to this traffic source, most of the other web sources drives traffic to home page as well. Traffic from `sol.no` converge to home page, news categories like *pluss, kultur, 100sports and nyheter*. Table C.3 shows different traffic sources and corresponding news categories they drive traffic to. Most of the traffic from `facebook.com` goes to categories like *forbruker, incoming, vaeret, tv, pluss, kultur, 100sport, nyheter and folk.*. Traffic from `Google` seems to land mostly on home page, and categories like *kultur, 100sport, nyheter, and folk.* As depicted in table C.3, `Facebook` drives most of the traffic to different categories. In the nutshell, most of web sources drives traffic to news categories like *nyheter, 100sport and kultur.* In

addition to this, Facebook seems to drive more traffic to the site compared to other web sources. This clearly indicates that **Adresseavisen** could clearly use Facebook and Google for promoting the news site and articles. Having a deep impact on searching and reading behavior of people with social media sites like Facebook and search engines like Google can be used by the site owner to enhance traffic to the site. Moreover, most popular categories shared on these web sources are likely to drive more traffic to the site. As SMM have proven to help in gaining a huge mass of traffic or attention through social media sites like Facebook and Google, making website social media friendly is one of the important aspect for **Adresseavisen**.

(b) **Do we observe different news consumption models on desktop and mobile traffic?**
Figure 5.10 and 5.11 are the mined model with traffic originating from Google and Facebook and from desktop and mobile respectively. Here, traffic from mobile and tablets are assumed to be coming from similar sources. Figure 5.10 shows model minded at 16.6% of most frequent activities and 2.5% of most frequent paths or edges. The model extracts traffic originating from Facebook and Google that are using desktop as main device type. On the other hand, figure 5.11 shows traffic originating from Facebook and Google using mobile and tablets as device type. To answer this research question, the table 5.5, clearly shows that, traffic originating from mobile or tablets is always higher than that of desktop. On the contrary *27.86%* of traffic from desktop went to home page where as 12.13% of traffic from mobiles went to home page. The mined model clearly follows the same path as discussed in [1] and gives information to make the site optimized for mobile device. There are several factors encouraging traffic from mobile compared to desktop. Having mobile devices evolved more and more every unit of time, mobile devices is released thats more capable and more powerful than the generation preceding them. In addition to that, current model mobiles are capable of handing almost all of their at-home and at-work tasks without need of external accessories. This capability of mobile or tablets make its wide use rather than its traditional approach for sending text messages or making call. The fact is clearly visible on the minded model 5.11. To answer the research question, the site http://www.adressa.no/ seems to have different news consumption models on desktop and laptop.

6.2 Prediction of next click

As the name of thesis implies, the main goal of this project is to predict next click based on process mining approach. There are several factors that determine the reading choices of the users in the news domain including *personal preferences, popularity and recency.* The main factor that this thesis considers is *popularity* which assume *"a new visitor is likely to click on most popular news item."*

For recommendation of news article, the most popular news categories is captured by using process mining approach. For example, in week one data, the most popular

categories as shown in mined model 5.4(b) are nyheter, 100sport, kultur and tv. For each new visitor on the site, the system recommends the news item form most popular news categories. In the nutshell, process mining approach can be used to mine most popular news categories or most popular news articles if articles are taken as activity and get the list of most popular news items which can be used for recommendation.

6.3 Challenges

An interesting project has come to an end. The insights obtained in the domain of process mining leaves with a better understanding of their importance. In addition to this, the project has not been without challenges. This section describes some of the challenges encountered during the project planning to project evaluation.

6.3.1 Big data aspects of process mining

The practical relevance of *process mining* is elevating as more and more event related data is available in one hand where as the ability of process mining tools to handle large of amount of data is very limited. For instance, a month of the data available from `www.adressa.no/` sized *13.48* gigabytes. To parse a week data of size *3.12 gigabytes* using the PYTHON script mentioned in Appendix A, it took 2.305 HOURS generating *907.8 megabytes* CSV file. To run the script, a single node machine *Macbook Pro* with *8 gigabytes* of RAM, *core i5* processor was used. Moreover, DISCO took more than 10 minutes to import over 14 million events with 5 million cases as shown in table 5.1. When using the log file for *process mining* in tools like DISCO with these data, the *filtering* tools on these took a lot of time to compute. In order to analyze the larger amount of data for entire month, it would require a lot of time and patience to generate the model. For news site like `adressa` that has millions of visitors per day and generates a huge amount of event logs per month, using tools like DISCO requires a lot of time. In order to use the analytics for business intelligence, from these data, it is required to mine the logs at month level rather than week level. With *process mining* tools available, it would be very time consuming to analyze a year data from event logs from news site. There does not seem to be support for *big data* in process mining. This was one of the biggest challenges when preprocessing the data, analyzing with tools and exporting the results. However, with support for different levels of *filtering approaches* on DISCO, it was possible to extract the models as discussed in chapter 5.

6.3.2 Data quality problem

Quality of data used for process mining is one of the important aspects as the quality of analysis depends on precision and completeness of data. This section describes some of the data quality problem faced during implementation. Following are the obstacles encountered during the implementation and analysis stage:

- **Incorrect structure of the data:** Minimum requirements for *process mining* is discussed in section 2.3. The log files extracted from the news site was in JSON

format which could not be imported into DISCO for process mining. Conversion of JSON file to CSV was not sufficient to match the standard for PM. The converted CSV files contained a large number of noise or which needed to be filtered out. It had to be preprocessed and re-structured to make the log files suitable for DISCO consumption.

- **Missing Data:** The log files did not contain all the data in uniform manner. Some of the data were missing like some event were logged without *referral host URL*. This might be because of incorrect logging or particular log fields value is unavailable for that event. In this thesis, if *referrerHost* were missing, the value for it was taken from *referrerHostClass* or from *referrerSocialNetwork* if *referrerHostClass* is missing as well. This was taken care at implementation stage.

- **Artificial activity timestamp:** As aforementioned, to study about new category referral source, this thesis uses artificial activities created using *referrerHost* and with timestamp of original activity minus 2 seconds. This step was taken to ensure that *referrerHost* always occurred before the main activity. **Timestamp** is one of the most important factor for *process mining* and using timestamp like this gives invalid process metrics on process map. For example 5.5(a)(b) shows 2 seconds mean and median reading time from traffic source to actual news categories. Also, the thesis does not consider session duration to filter activities. If the duration of session is greater than equal to a *threshold* value, breaking of activity to unique event would give better precision of process mining.

6.4 Pros and Cons of the methodology

- **Pros**
  Some of the advantages of implementing using this methodologies are listed below:

  - This methodology can be used to gather insight into deep content interaction patterns describing navigation patterns from external sources to internal content. This insight is useful for inferring the relationship between web sources and news categories.
  - In addition to this, the approach reveals some of the interesting statistics about the traffic sources, news categories popularity, browsers used and device type used by readers as described in chapter 5.

- **Cons**
  Some of the disadvantages of using this methodology are discussed here.

  - One of the disadvantage of this methodology is use of artificial timestamp that do not have meaningful timestamps on the external events. This is the reason of having shorter session time of approximately *2 seconds* between traffic source and categories. Knowing the meaningful timestamps would have been useful in extracting better performance insights.

- Another cons of this methodology is handling big data aspects of log file. It is extremely difficult to parse larger log files and import into available process mining tools.

Chapter 7

Conclusion

This chapter concludes this thesis by discussing about the limitations of the work done, and propose ideas for future work. First section 7.1 concludes the results and briefly answers the research questions. Secondly, section 7.2 discusses the limitations of the work done and proposes ideas for future work.

7.1 Conclusion

This thesis started with the motivation of exploring *process mining* on web log files from news website. While PM is trending over time, there is very few research done on using process mining on news log files. The thesis tends to cover various topics including the motivations for process mining, tools available for process mining, relationship between process mining and data mining and web usage mining, state of art works in web log process mining done so far, and methodology for revealing deep content interaction models from real life web logs.

As aforementioned, PM is an emerging discipline providing a sets of tools to provide fact-based insights and to support process enhancements. This new discipline builds on process *model-driven approaches* and *data mining*. Existing data mining techniques are too data-centric to provide understanding of the end-to-end processes in an organization. BI tools focus on simple dashboards and reporting rather than clear-cut business process insights. BPM suites heavily rely on experts modeling idealized to-be processes and do not help the stakeholders to understand the as-is processes [2]. Process mining techniques on the other hand facilitates organizations to uncover their actual business processes. *Process mining* is not only limited to process discovery. By tightly coupling event data and process models, it is possible to check conformance, detect deviations, predict delays, support decision making, and recommend process redesigns. With incredible growth of event data, there is wide range of opportunity for using *process mining* for *process discovery, conformance checking and model enhancements.* Some of the motivations of *process mining* are covered in section 2.2. There are wide range of applications of *process mining*. Some of the process mining works done with web logs data are discussed in chapter 3. This

chapter studies various works done with web log data using *process mining* and compares the works with the thesis.

The thesis uses the event log files extracted from the news website `www.adressa.no`, process the log files as described in chapter 4 and performs analysis as highlighted in chapter 5. The results show there are various traffic sources like *Facebook, Google*, `sol.no`, `startsiden.no` and `bing.com` that drives traffic to various news categories. The most popular news categories that most of the traffic are attracted to are *nyheter, kultur, 100sport and pluss*. Also, the news consumption by traffic from mobile or tablets are higher compared to that of desktop. That gives information for the site administrator to make the site more optimized for mobile devices or tablets as well. Today's website should be mobile friendly as it drives a tremendous amount of traffic compared to desktop devices.

When it comes to prediction of next click, one of the important features compared by this thesis is *popularity*. Most popular news categories are most likely to be clicked by a new visitor on the site. So, *process mining* can be used to get a list of most popular news categories or news items and when the system detects new visitors, they can be served with most popular news item. This is one of the aspects that shows how process mining be used for *recommendation* purposes.

7.2 Future work

Below are suggestions that were out of scope for this thesis, but could be interesting for further research:

- **Splitting the session based on time zone**: One of the important things that can be realized is the event log data contains events with longer sessions. The most obvious reason could be inactive users. Users that visit the page and close the device without closing the browser or session logs longer duration of session on log data. Having no other way to categorize such active and non-active users, longer session could be split to make a news one or two session based on some threshold. This would could give better performance metrics compared to what is done now.

- **Consider using process mining for big data:** As seen with this news website, how event log data for longer duration tends to be **big data**, *process mining* could handle such data to get better process insights. The sites with concrete processes as E-commerce sites like `www.ebay.com` or `www.amazon.com` produce large amount of data. These event data can undergo *process mining* to discover process insights, conformance checking as well as process enhancements. With tools available for *process mining* now, it is very time consuming to analyze larger set of data. With increasing use of process mining for process analysis, using it with big data can give better insight of the processes. In this thesis, using process mining with a month data could give real statistics for a month and the model extracted could be used for recommendation.

- **Consider other parameters when predicting next click log**: It is seen that recommender systems are becoming extremely common in recent years, and are applied

in a variety of applications. There are several ways how any news items are recommended to a new user. This thesis only considers one factor for recommendation which is *popularity*. When it comes to recommending, one of the important aspect is *personal preferences*. A new user when recommended with his/her preference item is most likely to be clicked. Not only that other factors like *recency* of the news is also important. One is likely to be interested in most recent news items compared to that of others. *Demographic* preferences can be one of the another factors for recommendation. Users prefer to know activities going around their own environment compared to others. All these parameters could have been considered to prepare a recommendation list that could be used to predict next click for a new user.

- **Evaluation of results by Adresseavisen and implementation of recommendation**: One of the way to increase the scope of the thesis is to use the extracted model as the basis for building recommendation system. In addition to this, evaluation of the models by **Adresseavisen** regarding how the results obtained are useful to them and how these can be used for business intelligence will help generate better model. These models can be used as basis for building recommendation system. Also, the RS system thus developed can be deployed on live site and measure actual performance.

Bibliography

[1] State of the News Media 2015, Pew Research Center, April, 2015 http://www.journalism.org/files/2015/04/FINAL-STATE-OF-THE-NEWS-MEDIA1.pdf, 01/12/2016

[2] *Wil M.P. van der Aalst* Process Mining - Discovery, Conformance and Enhancement of Business Process, 2011, SPRINGER

[3] *The Four V's of Big Data* http://www.ibmbigdatahub.com/infographic/four-vs-big-data, 02/02/2016

[4] *Wil M.P. van der Aalst*, Eindhoven University of Technology, Process Mining:Overview and Opportunities

[5] *Wil M.P. van der Aalst*, Eindhoven University of Technology, Challenges in Business Process Mining

[6] *Verbeek, H., Buijs, J., Dongen, B. VAN, and Aalst, W. van der 2010. ProM 6*,The Process Mining, Toolkit. In Proc. of BPM Demonstration Track 2010, M. L. Rosa, Ed. CEUR Workshop Proceedings Series, vol. 615. 3439.

[7] *B. van Dongen, A. de Medeiros, H. Verbeek, A. Weijters, and W. van der Aalst.* The prom framework: A new era in process mining tool support. In G. Ciardo and P. Darondeau, editors, *Applications and Theory of Petri Nets 2005*, volume 3536 of *Lecture Notes in Computer Science*, pages 444-454. Springer Berlin Heidelberg, 2005.

[8] Disco Userś Guide https://fluxicon.com/disco/, 01/22/2016, https://fluxicon.com/disco/files/Disco-User-Guide.pdf

[9] *M. Weske. Business Process Management:* Concepts, Languages, Architectures. Springer, Berlin, 2007.

[10] *Wil M.P. van der Aalst* No Knowledge Without Processes- Process Mining as a Tool to Find Out What People and Organizations Really Do.

[11] *W.M.P. van der Aalst, A.J.M.M. Weijters, and L. Maruster.* Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, 16(9):11281142, 2004.

[12] *ing. J.C.A.M. Buijs*, Mapping Data Sources to XES in a Generic Way, Master Thesis, Eindhoven, March 2010

[13] *Florian Daniel, Kamel Barkaoui, Schahram Dustdar (Eds.)* Business Process Management Workshops. BPM 2011 International Workshops Clermont-Ferrand, France, August 29, 2011 Revised Selected Papers, Part I

[14] *Matthew H. Loxton — Senior Analyst* , Process Discovery in Healthcare Quality Improvement, `http://wbbinc.com/resources/whitepapers/process-discovery-in-healthcare-quality-improvement`, 24/02/2016

[15] How Process Mining Compares to Data Mining, `https://fluxicon.com/blog/2011/02/how-process-mining-compares-to-data-mining/`, 03/03/2016

[16] *Rozinat, Anne*,Process Mining:Conformance and Extension,Technische Universiteit Eindhoven, 2010. - Proefschrift.

[17] *R. Agrawal and R. Srikant.* Mining sequential patterns, In Proceedings of the Eleventh International Conference on Data Engineering, pages 314, 1995.

[18] *H. Kum, J. Pei, W. Wang, and D. Duncan.*, ApproxMAP: Approximate Mining of Consensus Sequential Patterns. In Proceedings of SIAM Int. Conf. on Data Mining, 2003

[19] *R.P. Jagadeesh Chandra Bose and W.M.P. van der Aalst.*, Abstractions in Process Mining: A Taxonomy of Patterns. In U. Dayal, J. Eder, J. Koehler, and H.A. Reijers, editors, Business Process Management, 7th International Conference, BPM 2009, Ulm, Germany, September 2009. Proceedings, volume 5701 of Lecture Notes in Computer Science, pages 159175. Springer, 2009.

[20] *R.P. Jagadeesh Chandra Bose and W.M.P. van der Aalst.* ,Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models. In Proceedings of the 5th International Workshop on Business Process Intelligence (BPI), 2009.

[21] *G. Greco, A. Guzzo, L. Pontieri, and D. Sacca.*, Mining Expressive Process Models by Clustering Workflow Traces. In Proc of Advances in Kowledge Discovery and Data Mining, 8th Pacific-Asia Conference (PAKDD 2004), pages 5262, 2004

[22] *C.W. Gunther, A. Rozinat, and W.M.P. van der Aalst*, Activity Mining by Global Trace Segmentation. In Proceedings of the 5th International Workshop on Business Process Intelligence (BPI), 2009.

[23] *A.K. Alves de Medeiros, A. Guzzo, G. Greco, W.M.P. van der Aalst, A. Weijters, B.F. van Dongen, and D. Sacca.*, Process Mining Based on Clustering: A Quest for Precision. In A. ter Hofstede, B. Benatallah, and H.-Y. Paik, editors, BPM 2007 Workshops, volume 4928 of LNCS, pages 1729. Springer, 2008.

[24] *A.K. Alves de Medeiros*, Genetic Process Mining. PhD thesis, Eindhoven University of Technology, Eindhoven, 2006.

[25] *W.M.P. van der Aalst and A.J.M.M. Weijters*, Process Mining: A Research Agenda, Department of Technology Management, Eindhoven University of Technology, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.

[26] *R. Agrawal, D. Gunopulos, and F. Leymann.* , Mining Process Models from Workflow Logs. In Sixth International Conference on Extending Database Technology, pages 469483, 1998

[27] *J.E. Cook and A.L. Wolf.* , Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology,7(3):215-249, 1998

[28] *IEEE Task Force on Process Mining. Process mining case studies.*, `http://tinyurl.com/ovedwx4`, 2013

[29] *Diego Calvanese, Marco Montali, Alifah Syamsiyah, Wil M.P. van der Aalst*Ontology-Driven Extraction of Event Logs from Relational Databases, Free University of Bozen-Bolzano, Italy.

[30] *Nicolas Poggi, Vinod Muthusamy, David Carrera, and Rania Khalaf*Business Process Mining from E-commerce Web Logs

[31] *Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan* Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data

[32] *Bing Liu*, Web Data Mining, 2016/03/115, Page 449-479

[33] *Awatef HICHEUR CAIRNS, Billel GUENI, Mehdi FHIMA, Andrew CAIRNS and Stphane DAVID1 Nasser KHELIFA*, Process Mining in the Education Domain, International Journal on Advances in Intelligent Systems, vol 8 no 1 & 2, year 2015, `http://www.iariajournals.org/intelligent_systems`

[34] *Wil M.P. van der Aalst*, No Knowledge Without Processes Process Mining as a Tool to Find Out What People and Organizations Really Do

[35] *TADAO MURATA, FELLOW, IEEE* , Petri Nets: Properties, Analysis and Applications, *Invited Paper*

[36] *R. Johnsonbaugh*, Discrete Mathematics, `http://condor.depaul.edu/rjohnson/dm7th/petri.pdf`, 2016/03/18

[37] *Patrick Mukala, Joos Buijs, Maikel Leemans, and Wil van der Aalst*, Learning Analytics on Coursera Event Data: A Process Mining Approach , `http://ceur-ws.org/Vol-1527/paper2.pdf`, 2016/04/17

[38] *Wil van der Aalst*, Decomposing Petri Nets for Process Mining A Generic Approach

[39] *Renta Ivncsy, Istvn Vajk* , Frequent Pattern Mining in Web Log Data, Department of Automation and Applied Informatics, and HAS-BUTE Control Research Group Budapest University of Technology and Economics

[40] *Process Mining* , `http://www.processmining.org/publications/start`, 05/31/2016

[41] *Sebastian Huber , Marian Fietta, Sebastian Hof* , Next Step Recommendation and Prediction based on Process Mining in Adaptive Case Management

[42] *Huber, S., Lederer, M., Bodendorf, F. (2014)*: IT-enabled Collaborative Case Management: Principles and Tools. In: Waleed W. Smari, Geoffrey C. Fox, Mads NygaŁrd (Eds.): Proceedings of the 2014 International Conference on Collaboration Technologies and Systems, Minneapolis, Minnesota, USA, IEEE by The Printing House, Inc., Stoughton, p. 259 - 266.

Appendix A

JSON file Structure

```
{
  "start": 1446753600,
  "stop": 1446757200,
  "events": [
    {
      "time": 1446753600,
      "browser": "Safari",
      "deviceType": "Mobile",
      "eventId": 866111861,
      "os": "iPhone OS",
      "referrerHostClass": "direct",
      "sessionBounce": false,
      "sessionStart": false,
      "sessionStop": true,
      "url": "http://www.adressa.no/100Sport/handball/
         article631540.snd",
      "userCorrelationId": "84124c03debba5a8"
    },
    {
      "time": 1446753600,
      "activeTime": 31,
      "browser": "Chrome",
      "deviceType": "Desktop",
      "eventId": 1604174112,
      "os": "Windows",
      "referrerHost": "adressa.no",
      "referrerHostClass": "internal",
      "sessionBounce": false,
      "sessionStart": false,
```

```
29          "sessionStop": false,
30          "url": "http://adressa.no/incoming/2015/11/05/udi-har
                -f%c3%a5tt-innfridd-knappe-halvparten-av-%c3%b8
                nsket-sitt-11773223.ece",
31          "userCorrelationId": "0eb8a9aa4243bc54"
32      },
33      ...
34      ]
35  }
```

Appendix B

Converting JSON file into CSV file

```
1  import csv
2  import json
3  import datetime
4  import sys
5
6
7  BROWSER = 0
8  DEVICE_TYPE = 1
9  OS = 2
10 USER_CORRELATION_ID = 3
11 ACTIVITY_1 = 4
12 ACTIVITY_2 = 5
13 TIME_1 = 6
14 TIME_0 = 7
15 TIME_SUBTRACTION = datetime.timedelta(seconds=2)
16
17 def loadJson():
18     inputFile = sys.argv[1]
19     with open(inputFile) as data_file:
20         print "Input file: " + inputFile
21         data = json.load(data_file)
22     return data["events"]
23
24 def processToCsv(data):
25     firstOutput = sys.argv[2]
26     with open(firstOutput, 'wb' ) as csvfile:
27         writer = csv.writer(csvfile, delimiter = ',',
               quotechar=',', quoting=csv.QUOTE_MINIMAL)
28         for element in data:
```

```
29              time0 = datetime.datetime.fromtimestamp(element
                    ["time"])
30              time1 = time0 - TIME_SUBTRACTION
31              try:
32                  activity1 = element["referrerHost"]
33              except:
34                  activity1 = ""
35              try:
36                  activity2 = element["referrerHostClass"]
37              except:
38                  activity1 = ""
39              if activity1.lower() == "adressa.no" or
                    activity2.lower() == "internal":
40                  activity1 = "ok"
41              if activity1 != "ok":
42                  try:
43                      activity1 = element["referrerHost"]
44                  except:
45                      try:
46                          activity1 = element["
                                referrerSocialNetwork"]
47                      except:
48                          try:
49                              activity1 = element["
                                    referrerHostClass"]
50                          except:
51                              activity1 = "Others"
52              cleanUrl = element["url"][7:]
53              try:
54                  activity2 = cleanUrl[cleanUrl.index("/") +
                        1:]
55                  try:
56                      activity2 = activity2[:activity2.index
                            ("/")]
57                  except:
58                      activity2 = activity2
59              except:
60                  activity2 = "Home"
61              try:
62                  activity2 = activity2[:activity2.index("?")
                        ]
63              except:
64                  pass
65              writer.writerow([element["browser"], element["
                    deviceType"], element["os"], element["
```

```
                userCorrelationId"], activity1, activity2,
                time1, time0])
    print "First output: " + firstOutput
def getKey(item):
    return item[USER_CORRELATION_ID]
def getKey2(item):
    return item[TIME_1]
```

Appendix C

Filtering the CSV file

```
def sortDataToCsv():
    dict = None
    userCorrelationId = ""
    fisrtOutput = sys.argv[2]
    lastOutput = sys.argv[3]
    with open(fisrtOutput, 'rb') as f:
        fieldnames = ['browser', 'deviceType', 'os', '
           userCorrelationId', 'activity','timestamp']
        with open(lastOutput, 'wb', ) as csvfile:
            writer = csv.writer(csvfile, delimiter = ',',
               quotechar=',', quoting=csv.QUOTE_MINIMAL)
            writer.writerow(fieldnames)
            for row in sorted(csv.reader(f), key=lambda x:
               (getKey(x), getKey2(x))):
                if "facebook" in row[ACTIVITY_2].lower():
                    row[ACTIVITY_2] = "facebook"
                if "facebook" in row[ACTIVITY_1].lower():
                    row[ACTIVITY_1] = "facebook"
                if "google" in row[ACTIVITY_2].lower():
                    row[ACTIVITY_2] = "google"
                if dict != None and row[USER_CORRELATION_ID
                   ] in dict and row[ACTIVITY_1].lower() !=
                    "ok":
                    dict[row[USER_CORRELATION_ID]] = str(
                       int(dict[row[USER_CORRELATION_ID]])
                       + 1)
                    userCorrelationId = row[
                       USER_CORRELATION_ID] + "-" +str(dict
                       [row[USER_CORRELATION_ID]])
```

```
21                      writer.writerow([row[BROWSER], row[
                            DEVICE_TYPE], row[OS],
                            userCorrelationId, row[ACTIVITY_1].
                            lower(), row[TIME_1]])
22                      writer.writerow([row[BROWSER], row[
                            DEVICE_TYPE], row[OS],
                            userCorrelationId, row[ACTIVITY_2].
                            lower(), row[TIME_0]])
23                  else:
24                      if dict == None or row[
                            USER_CORRELATION_ID] not in dict:
25                          dict = {row[USER_CORRELATION_ID]:0}
26                          userCorrelationId = row[
                                USER_CORRELATION_ID]
27                      writer.writerow([row[BROWSER], row[
                            DEVICE_TYPE], row[OS],
                            userCorrelationId, row[ACTIVITY_2].
                            lower(), row[TIME_0]])
28          print "Final output: " + lastOutput
29  def main(argv):
30      startTime = datetime.datetime.now()
31      print "Start Time: " + str(datetime.datetime.time(
            startTime))
32      processToCsv(loadJson())
33      sortDataToCsv()
34      endTime = datetime.datetime.now()
35      print "End Time: "+ str(datetime.datetime.time(endTime)
            )
36      print "Elapsed Time: "+ str(endTime - startTime)
37
38  if __name__ == '__main__':
39      if len(sys.argv) < 4:
40          print "You must set three arguments!!!"
41          sys.exit()
42      main(sys.argv[0])
```

C.1 Miscellaneous

Table C.1: Translations of categories name from Norwegian to English

Norwegian Name	**English Name**
100sport	sport
home	home
bolig	residential
folk	folk
forbruker	consumer
vaeret	weather
nyheter	news
tv	tv
kultur	culture
search	search
pluss	plus
reise	travel
incoming	incoming
tema	theme
bil	car
digital	digital
folk	folk
familieogoppvekst	family and upbringing
video	video

Table C.2: Referral sources extracted at 80% of events data with 100% most frequent activities and 33.1% most frequent edges

Activity	Frequency	Relative frequency
startsiden.no	185,934	1.6 %
facebook	138,199	1.19 %
google	85,010	0.73 %
sol.no	18,085	0.16 %
adressa.alda.no	9,890	0.09 %
mobil.sol.no	4,695	0.04 %
bing.com	3,194	0.03 %
wazzup.polarismedia.no	2,465	0.02 %
onlineaviser.no	1,834	0.02 %
startsida.no	1,308	0.01 %
hemnestart.no	1,296	0.01 %
t.co	1,219	0.01 %
dinstartside.no	1,199	0.01 %

Table C.3: Major web sources driving traffic to corresponding news categories

Activity	**News categories**
startsiden.no	home
facebook	forbruker, incoming, vaeret tv, pluss, kultur, 100sport, nyheter, home
google	home, kultur, 100sport, nyheter, folk
sol.no	home, pluss, kultur, 100sport, nyheter