



Igor Jakovljević, DI

Information Retrieval and Information Consumption in Multidimensional Information Space for Push Notification Systems

Doctoral Dissertation

to achieve the university degree of
Doctor of Technical Sciences
Doctoral degree programme: Computer Science

submitted to

Graz University of Technology

Supervisors

Assoc.Prof. Dr. Christian Gütl

Institute for Interactive Systems and Data Science, Graz University of Technology, Austria

Dipl.-Ing. Dr. Andreas Wagner

CERN, Switzerland

Second Reader

Prof. Dr. Michael Granitzer

Faculty of Computer Science and Mathematics, University of Passau, Germany

Institute for Interactive Systems and Data Science (ISDS)

Head: Kappe, Frank, Univ.-Prof. Dipl.-Ing. Dr.techn.

Graz, September 2023



Igor Jakovljević, DI

Informationsgewinnung und Informationskonsum im mehrdimensionalen Informationsraum für Push-Benachrichtigungssysteme.

Dissertation

zur Erlangung des akademischen Grades
Doctor of Technical Sciences
Doktoratsprogramm Computer Science

eingereicht an der

Graz University of Technology

Betreuer

Assoc.Prof. Dr. Christian Gütl

Institute for Interactive Systems and Data Science, Graz University of Technology, Austria

Dipl.-Ing. Dr. Andreas Wagner

CERN, Switzerland

Zweitgutachter

Prof. Dr. Michael Granitzer

Faculty of Computer Science and Mathematics, University of Passau, Germany

Institute for Interactive Systems and Data Science (ISDS)

Leitung: Kappe, Frank, Univ.-Prof. Dipl.-Ing. Dr.techn.

Graz, September 2023

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present dissertation.

Date

Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

Datum

Unterschrift

Abstract

The rise of the Internet has contributed significantly to the exponential growth of information, both within organisational boundaries and on the Web. This research is driven by the increasing difficulties that modern organisations, especially those that are large and interconnected like CERN, face in dealing with the massive amount of data they receive. Large organisations, employing thousands of individuals and engaging in complex and interdependent activities, are becoming hubs of data generation and communication. For instance, CERN, with its large workforce collaborating on extensive projects, illustrates this information-rich environment. However, this surge in data production has led to a new problem, named information overload. Users are exposed to vast amounts of data, making it difficult to determine the relevant information from the irrelevant. Traditional search engines, the primary tools for information retrieval, are no longer sufficient. The demand for personalised information retrieval tools has increased. To address this challenge, the CERN Notification Research Project was initiated. This project, which is currently being developed at CERN, offers user-specific notifications. Although it excels at distributing information based on user subscriptions, it lacks proactive recommendation capabilities. Unlike conventional recommendation systems, which use algorithms to propose personalised content, the CERN Notifications project follows a user-driven approach, respecting users' choices.

This doctoral dissertation aims to develop a privacy-preserving approach and recommendation system for information retrieval and recommendation within the context of a large organisation such as CERN. The main objectives of this research are to determine user consumption habits of users in large organisations and to prototype a Privacy-Preserving Information Retrieval and Recommendation System for large interconnected organisations. This research is based on a comprehensive literature survey of the current problems in privacy-preserving applications of machine learning and recommendation systems in large organisations.

The approach employed in this thesis encompasses a hybrid methodology, merging elements from Design Science (DS) research and the traditional Waterfall methodology to address the complexities of information retrieval and recommendation within large organisations like CERN. Design Science research, a design-orientated approach gaining traction in information systems research, offers a structured framework represented in cycles: relevance, design, and rigour. These cycles involve defining requirements based on the application domain, iterative development, and continuous refinement through knowledge gathering and sharing. Unlike the linear progression of the Wa-

terfall methodology, this approach integrates the flexibility of the DS design cycle, allowing for revisiting previous phases if issues arise. The five main phases of this hybrid methodology are sequentially linked, combining DS's iterative nature with Waterfall's structured approach. Phases include literature surveys, data analysis, system architecture design, development, and evaluation. These steps are meticulously interlinked, ensuring that the research objectives are met effectively.

The results of this research include three significant contributions. First, it introduces the incorporation of social media elements into the notification systems of large organisations, enhancing their information dissemination capabilities. Second, it emphasises the utilisation of anonymised data, ensuring both privacy and effective data utilisation. Third, it advances the field by employing a machine learning pipeline specifically designed for open data, fostering improved data-driven decision-making within large organisations.

The recommendation system developed for this thesis serves as a demonstration solution, showcasing the potential of privacy-preserving solutions in various systems. Our evaluation demonstrated that this system offers a viable alternative to traditional recommendation systems while ensuring data privacy. The findings and solutions presented provide valuable guidance for organisations such as CERN, aiming to harness the power of data to improve productivity and user satisfaction, while safeguarding privacy.

Kurzzusammenfassung

Der Aufstieg des Internets hat maßgeblich zum exponentiellen Wachstum von Informationen sowohl innerhalb organisatorischer Grenzen als auch im Web beigetragen. Diese Forschung wird durch die zunehmenden Schwierigkeiten getrieben, mit denen moderne Organisationen, insbesondere große und vernetzte wie CERN, bei der Bewältigung der riesigen Datenmengen konfrontiert sind, die sie erhalten. Große Organisationen mit Tausenden von Mitarbeitern und komplexen, voneinander abhängigen Aktivitäten werden zu Zentren der Datenproduktion und -kommunikation. Zum Beispiel illustriert CERN mit seiner großen Belegschaft, die an umfangreichen Projekten zusammenarbeitet, diese informationsreiche Umgebung. Dieser Anstieg der Datenproduktion hat jedoch zu einem neuen Problem geführt, dem sogenannten Informationsüberfluss. Nutzer sind einer riesigen Datenmenge ausgesetzt, was es schwierig macht, relevante von irrelevanten Informationen zu unterscheiden. Traditionelle Suchmaschinen, die Hauptwerkzeuge für die Informationsgewinnung, sind nicht mehr ausreichend. Die Nachfrage nach personalisierten Informationsgewinnungswerkzeugen ist gestiegen. Um dieser Herausforderung zu begegnen, wurde das CERN Notification Research Project ins Leben gerufen. Dieses Projekt, das derzeit bei CERN entwickelt wird, bietet benutzerspezifische Benachrichtigungen. Obwohl es sich hervorragend zur Verteilung von Informationen basierend auf Benutzerabonnements eignet, fehlen ihm proaktive Empfehlungsfähigkeiten. Im Gegensatz zu herkömmlichen Empfehlungssystemen, die Algorithmen verwenden, um personalisierte Inhalte vorzuschlagen, folgt das CERN Notifications-Projekt einem nutzerorientierten Ansatz und respektiert die Entscheidungen der Benutzer.

Diese Doktorarbeit hat zum Ziel, einen datenschutzerhaltenden Ansatz und ein Empfehlungssystem für die Informationsgewinnung und -empfehlung im Kontext einer großen Organisation wie CERN zu entwickeln. Die Hauptziele dieser Forschung sind die Bestimmung der Konsumgewohnheiten von Benutzern in großen Organisationen und die Prototypisierung eines datenschutzfreundlichen Informationsgewinnungs- und Empfehlungssystems für große vernetzte Organisationen. Diese Forschung basiert auf einer umfassenden Literaturübersicht über die aktuellen Probleme bei datenschutzfreundlichen Anwendungen von maschinellem Lernen und Empfehlungssystemen in großen Organisationen.

Der in dieser Arbeit angewandte Ansatz umfasst eine hybride Methodik, die Elemente aus der Design Science (DS)-Forschung und der traditionellen Wasserfallmethode kombiniert, um den Herausforderungen der Informationsgewinnung und

-empfehlung in großen Organisationen wie CERN zu begegnen. Die Design Science-Forschung, ein in der Forschung zu Informationssystemen an Bedeutung gewinnender designorientierter Ansatz, bietet einen strukturierten Rahmen in Form von Zyklen: Relevanz, Design und Strenge. Diese Zyklen beinhalten die Definition von Anforderungen basierend auf der Anwendungsdomäne, iterative Entwicklung und kontinuierliche Verfeinerung durch Wissensgewinnung und -teilung. Im Gegensatz zum linearen Fortschreiten der Wasserfallmethode integriert dieser Ansatz die Flexibilität des DS-Designzyklus und ermöglicht es, zu früheren Phasen zurückzukehren, wenn Probleme auftreten. Die fünf Hauptphasen dieser hybriden Methodik sind sequenziell miteinander verbunden und kombinieren die iterative Natur von DS mit dem strukturierten Ansatz von Wasserfall. Die Phasen umfassen Literaturstudien, Datenanalyse, Systemarchitekturentwurf, Entwicklung und Evaluation. Diese Schritte sind sorgfältig miteinander verknüpft, um sicherzustellen, dass die Forschungsziele effektiv erreicht werden.

Die Ergebnisse dieser Forschung umfassen drei wesentliche Beiträge. Erstens führt sie die Integration von Elementen aus sozialen Medien in die Benachrichtigungssysteme großer Organisationen ein und verbessert dadurch ihre Informationsverbreitungsfähigkeiten. Zweitens betont sie die Verwendung anonymisierter Daten, um sowohl Datenschutz als auch effektive Datenverwendung sicherzustellen. Drittens trägt sie zur Weiterentwicklung des Feldes bei, indem sie eine speziell für offene Daten konzipierte maschinelle Lernpipeline einsetzt, die eine verbesserte datengetriebene Entscheidungsfindung in großen Organisationen fördert.

Das in dieser Arbeit entwickelte Empfehlungssystem dient als Demonstrationslösung und zeigt das Potenzial datenschutzfreundlicher Lösungen in verschiedenen Systemen auf. Unsere Evaluation hat gezeigt, dass dieses System eine praktikable Alternative zu herkömmlichen Empfehlungssystemen darstellt und gleichzeitig den Datenschutz gewährleistet. Die präsentierten Erkenntnisse und Lösungen bieten wertvolle Orientierung für Organisationen wie CERN, die die Kraft von Daten nutzen möchten, um Produktivität und Benutzerzufriedenheit zu verbessern und gleichzeitig die Privatsphäre zu schützen.

Acknowledgments

I would like to express my deepest gratitude to my parents, Zdenko and Dragica, for their unwavering love and support throughout my academic journey. Their encouragement, sacrifices, and belief in me have been the driving force behind my success. I am also grateful to my brother, Dario, for his constant support, encouragement, and guidance. His belief in me has been a source of inspiration and I am truly grateful for his unwavering support.

My heartfelt thanks go to my university supervisor, Christian Gütl, for his invaluable guidance and encouragement throughout my research. His insights and expertise have been crucial in shaping the direction of my work and I am grateful to him for his generous assistance.

I also thank my CERN supervisor, Andreas Wagner, for his support and guidance during my time at CERN. His expert knowledge and guidance have been invaluable in helping me navigate the complexities of the research environment at CERN.

I started my Ph.D. almost at the same time as my two colleagues Aleksandar Bobic and Alexander Steinmaurer. Thank you both for being part of my research journey, I enjoyed our discussions and rants about our research and personal things. Special thanks to Aleksandar Bobic for being a close friend and an exceptional colleague throughout the Ph.D. process, the COVID-19 pandemic, and general life in France and Switzerland.

Finally, I would like to thank my dear friends in Graz (specially Sanda Filipovic) and at CERN, for their unwavering support, encouragement, and motivation. Their belief in me and their unwavering support have been crucial in helping me overcome personal and professional challenges.

I am truly grateful to all those who have contributed to my academic journey and I am honoured to have had the opportunity to work with such a talented and supportive group of individuals.

Contents

Abstract	ix
Acronyms	xxix
1 Introduction	3
1.1 Motivation and Background	3
1.2 Objectives	5
1.3 Methodology and Thesis Structure	6
1.4 Contributions	11
2 Background and Related Work	15
2.1 Information Generation and Consumption in Large Organisations . . .	16
2.2 User Profiling and Profiles	22
2.2.1 Users Information Collection	23
2.2.2 User Profiling Methods	26
2.2.3 Short-Term and Long-Term Interests	28
2.2.4 Filter Bubbles	28
2.2.5 Cognitive and Search Bias	29
2.2.6 Privacy Challenges in User Profiling	30
2.3 Information Retrieval and User Information Navigation	31
2.3.1 Brief History of Information Retrieval	31
2.3.2 Information Retrieval Methods	32
2.3.3 Information Retrieval Process	33
2.3.4 User Information Navigation	34
2.3.5 Subscriptions and Publishing Systems	35
2.4 Machine Learning	38
2.4.1 Brief History of Machine Learning	38
2.4.2 Main Challenges Of Machine Learning	39
2.4.3 Popular Machine Learning Approaches	40
2.4.4 Machine Learning Pipelines	43
2.5 Recommender Systems	45
2.5.1 Brief History of Recommender Systems	45
2.5.2 Collaborative Filtering	47
2.5.3 Content-Based Filtering	48

Contents

2.5.4	Hybrid Recommender Systems	49
2.5.5	Main Challenges Of Recommender Systems	49
2.5.6	Recommender Systems Evaluation Methodologies	52
2.6	Open Science Principles	56
2.6.1	Open Data and Open Information	58
2.6.2	Open Innovation	58
2.6.3	Importance of Privacy in Open Data	59
2.6.4	Open Data Initiatives	60
2.7	Related Work	60
2.7.1	Overview of Existing Approaches for Privacy-Aware Machine Learning in Large organisations	61
2.7.2	Privacy in Recommender Systems	64
2.7.3	Privacy in Community Detection	66
2.7.4	Importance of Open Science, Open Data, and Open Information	67
2.8	Summary	67
3	Analysis Of User Behavior	71
3.1	Contribution	71
3.2	CERN IT Department User Survey Analysis	72
3.2.1	Conclusion	76
3.3	CERN New Comers Analysis	76
3.3.1	Conclusion	78
3.4	CERN User Information Consumption Analysis	79
3.4.1	Setting and Instruments	79
3.4.2	Procedure	80
3.4.3	Results and Discussion	82
3.4.4	Comparing Work Related and Personal Information Consumption	85
3.5	Conclusion	94
3.6	Summary	94
4	Data Analysis and Exploration	99
4.1	Contribution	99
4.2	Large Organisations and Sensitive Data	100
4.3	Data Lift Framework	101
4.3.1	Define Purpose And Scope of Data	101
4.3.2	Data Collection and Classification	102
4.3.3	Risk Assessment	103
4.3.4	Data Transformation and Anonymization	106
4.3.5	Evaluation	107
4.3.6	Publishing	108

4.4	Application of the Data Lift Framework on CERN Mattermost Dataset .	109
4.4.1	Expert Interview	110
4.4.2	User Study	111
4.5	Summary	114
5	Applicability of Social Media Elements in Notification Systems	117
5.1	Contribution	117
5.2	Social Media Elements	118
5.3	Research Study	121
5.3.1	Study Design	121
5.3.2	Settings and Instruments	122
5.3.3	Procedure	125
5.3.4	Study Participants	126
5.4	Findings and Discussion	128
5.5	Summary	132
6	Requirements and System Design	135
6.1	Contribution	135
6.2	CERN Notifications System	136
6.2.1	Web Portal	137
6.2.2	Backend Service	137
6.2.3	Message Routing Service	138
6.2.4	Message Consumer Service	138
6.2.5	External Services	139
6.2.6	Notification System Main Activity	140
6.3	Requirements	141
6.4	Conceptual Architecture	143
6.4.1	Recommendation System	144
6.4.2	External Storage Service	145
6.5	Technology Decisions	145
6.5.1	Machine Learning Service	145
6.5.2	ML Model Storage	146
6.5.3	External Storage Service	146
6.5.4	Prediction Service	146
6.6	System Architecture	147
6.6.1	KubeFlow ML Service	148
6.6.2	KServe Prediction Service	148
6.6.3	Data Anonymisation and De-Anonymisation Service	149
6.6.4	MinIO Object Storage	150
6.7	Summary	150

Contents

7	Development	153
7.1	Contribution	153
7.2	Recommendation System Workflow	153
7.3	Kubeflow	155
7.3.1	Kubeflow Pipelines	157
7.4	KServe Custom Predictor	163
7.4.1	Custom KServe Predictor Implementation	164
7.5	CERN Notification System Integration	166
7.5.1	CERN Notification System Backend	166
7.5.2	Notification and Recommendation System Data Communication	166
7.5.3	CERN Notification System Recommendation Feedback Mechanism UI	167
7.6	Summary	168
8	Evaluation	171
8.1	Contribution	171
8.2	Data-based Evaluations	172
8.2.1	Data Sources	172
8.2.2	Data-based Evaluation of Implicit Recommendation Algorithms	174
8.2.3	Data-based Evaluation of Cluster Based Implicit Recommendation Algorithms	181
8.3	User Feedback-based Studies	189
8.3.1	User Feedback-based Study of Cluster Based Implicit Recommendation Algorithms	189
8.3.2	User Feedback-based Study of Implicit Recommendation Algorithms	194
8.4	Evaluation and Study Findings	199
8.5	Summary	201
9	Lessons Learned and Outcome	205
9.1	Retrospection	205
9.2	Outcome	206
10	Conclusion and Future Work	215
10.1	Conclusion	215
10.2	Limitations	216
10.3	Future Work	217
	References	221

List of Figures

1.1	Current Information and Communication Workflow at CERN (Ormancey et al., 2022)	4
1.2	Design Science Research Framework - graphic taken from Villanueva (2019)	7
1.3	Phases of the Waterfall Methodology based on Sherman (2015)	8
1.4	Flowchart of the Thesis Iterative Research Methodology	9
1.5	Structure of the Thesis	10
2.1	Popular Machine Learning Approaches	41
2.2	Open ML Pipeline Steps, image taken from (Jakovljevic, Gütl, & Wagner, 2022)	43
2.3	Classification Evaluation Representation taken from (Igor et al., 2023)	53
3.1	What is your age range?	73
3.2	Can we contact you for further information?	73
3.3	Preferred Device for Personal Usage (Female)	86
3.4	Preferred Device for Personal Usage (Male)	86
3.5	Preferred Device for Personal Usage (Preferred not to say)	87
3.6	Preferred Device for Work-Related Information Consumption (Female)	87
3.7	Preferred Device for Work-Related Information Consumption (Male)	87
3.8	Preferred Device for Work-Related Information Consumption (Preferred not to say)	88
3.9	Preferred Mediums for Personal Usage (Female)	89
3.10	Preferred Mediums for Personal Usage (Male)	89
3.11	Preferred Mediums for Personal Usage (Preferred not to say)	90
3.12	Preferred Mediums for Work-Related Information Consumption (Female)	90
3.13	Preferred Mediums for Work-Related Information Consumption (Male)	91
3.14	Preferred Mediums for Work-Related Information Consumption (Preferred not to say)	91
4.1	DataLift Framework Steps - Organisational Framework for Open Sourcing Data (taken from (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022))	101

List of Figures

4.2	Risk Dimension Evaluation Matrix (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022)	105
5.1	Codis Survey Tool User Interface	125
5.2	Simple Notification Information Display	126
5.3	Notification with Additional Information Display	127
5.4	System Usability Scale Detailed Results	129
5.5	Notification Element Ranking	131
6.1	Simplified CERN Notification System Architecture Overview based on (Ormancey et al., 2022)	136
6.2	Web Portal Main Page	137
6.3	Notification System Send Notification Activity	140
6.4	Conceptual Architecture for the Recommendation System	144
6.5	System Architecture for the Recommendation System	148
6.6	System Architecture for the Recommendation System	149
7.1	Workflow of the Recommendation System	154
7.2	KubeFlow Process	155
7.3	Figure 7.3 Recommended Item List UI for the CERN Notifications System	167
8.1	Simplified Diagram of the CERN Mattermost Dataset, taken from (Bobic et al., 2022)	173
8.2	Demonstration of a Simplified Graph Creation Process	182
8.3	Sample run showing similarities of users between found communities and Mattermost teams, taken from (Pobaschnig et al., 2023)	187
8.4	Similarities between discovered communities and Mattermost teams over iterations with threshold 52, taken from (Pobaschnig et al., 2023)	188
8.5	Procedure Steps of the Evaluation of Cluster Based Implicit Recommendations	190
8.6	Recommendation Email Sample with Anonymous Data	191
8.7	Procedure Steps of the Evaluation of Implicit Data Based Recommendations	195
8.8	Recommendation Explanation Notification	196

List of Tables

2.1	Methods for Collecting Active User-Provided Information	24
2.2	Methods for Collecting Passive Data	25
3.1	IT Department User Survey Report: Which methods do you use to communicate with colleagues? Taken from (Jones, 2017)	74
3.2	IT Department User Survey Report: How do you share documents with colleagues? Taken from (Jones, 2017)	75
3.3	Survey Sections and Question Details	80
3.4	Survey Phases and Participants Description	81
3.5	Demographic Characteristics of Study Participants	82
3.6	Participant Nature of Work at CERN	84
3.7	Utilization of Information Sources for Work-Related Information Retrieval and Consumption	85
3.8	Preferred Information Sources for Personal Usage	85
3.9	How often do you read news portals to get informed about topics related to personal interests	92
3.10	How often do you read news portals to get informed about work-related topics?	92
4.1	Privacy data classification (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022)	103
4.2	Data Anonymization Methods Matrix	107
4.3	Features for Dataset Evaluation	108
4.4	User Rating of Understandability per Framework Step	112
5.1	Summary of main social media elements taken from (Jakovljevic, Gütl, & Wagner, 2022)	119
5.2	General Questionnaire Questions	123
5.3	Article Feedback Questions	123
5.4	Usability of social media elements in Notification systems, taken from (Jakovljevic, Gütl, & Wagner, 2022)	124
5.5	Article Title and Validity	125
5.6	Questionnaire Results, taken from (Jakovljevic, Gütl, & Wagner, 2022)	128

List of Tables

5.7	Additional Information Ranking by Importance, taken from (Jakovljevic, Gütl, & Wagner, 2022)	129
5.8	Computer Emotion Scale Answer Emotion Distribution	130
5.9	Article Validity Evaluation Results	130
6.1	Functional Requirements for the Recommendation System.	142
6.2	Non-Functional Requirements for the Recommendation System.	143
6.3	Software packages for building reproducible machine learning applications. Taken From "Towards an Open Data based Privacy-Aware Reproducible Machine Learning Pipeline", by Jakovljevic, Wagner, & Gütl (2022)	146
6.4	Popular Machine Learning Model Serving Tools	147
7.1	Recommendations API Calls	166
8.1	Mattermost dataset entity properties were extracted for the purpose of calculating complex relevance factors, taken from (Bobic et al., 2022) . .	173
8.2	Factors for the Creation of Complex Measures	175
8.3	Recommender System Use-Cases and Definitions	178
8.4	The most performing configuration was identified by analyzing 2201 different attribute combinations. The values of the most performing combination are similar to the ones in literature (Hu et al., 2008; Renaud-Deputter et al., 2013)	179
8.5	Experimental Results of Best Performing CF Algorithms with Optimal Configurations	180
8.6	Recalculated five-number summary of members withing teams ignoring teams with one member.	182
8.7	Parameters and their description of the first table of each algorithm result.	183
8.8	Results including communities, modularity, and most important values of the five-number summary of similarities between Mattermost teams and found community with different algorithms at threshold 23. Values within columns represent mean and standard deviation over 25 iterations.	185
8.9	Nodes and edges of graphs with threshold 52 created with team and channel method.	188
8.10	Number of nodes, edges, and overall weight of the edges over different thresholds.	189
8.11	Participant Demographics of the Cluster Based Implicit Recommendation Evaluation	192
8.12	User Response Evaluation	193
8.13	Study Limitations	193
8.14	Participant Selection	197

List of Tables

8.15	Demographics of Participants	197
8.16	User Response Evaluation	198

Acronyms

CERN Conseil Européen pour la Recherche Nucleaire
CES Computer Emotion Scale
DS Design science
DSAM Decision Support Accuracy Metrics
EOU Ease of Use
GB Gigabytes
HCI Human–Computer Interaction
IR Information retrieval
IT Information Technology
LHC Large Hodron Collider
ML Machine Learning
NDCG Normalized Discounted Cumulative Gain
NLP Natural Language Processing
RS Recommender System
SAM Statistical Accuracy Metrics
SDI Selective dissemination of information
SME Social Media Element
SUS System Usability Scale
TB Terabytes
VSM Vector Space Model

1 Introduction

This chapter provides a general overview of the topics covered in this thesis. The first section (**Section 1.1**) explains the motivation and background of this work. **Section 1.2** outlines the main objectives of the thesis together with the research questions. Next, **Section 1.3** introduced the methodology used while working on this thesis and illustrates the general structure of the thesis. Finally, the main contributions are listed in **Section 1.4**, which include scientific publications and research results.

1.1 Motivation and Background

Since the creation of the Internet, the amount of data generated by humans has been increasing year by year because the world has become more data-driven. Every day people send emails, take photos, make videos, create documents, and use various data and information generation techniques (Forbes, 2018). This behaviour of rapid information generation also translates into the work space, especially within large and highly connected organisations (M.-C. Lee, 2016). As the amount of data produced by humans on the Web and within large organisations also rapidly increases, new challenges for navigation through information are formed. Today, navigating the web and information within an organisation normally requires using a search engine as the main entry point. This search engine can be one of the major search engines used to navigate the Web or/and an internal organisation search engine service. According to Jakovljevic et al. (2020), large organisations are organisations that employ more than 5000 individuals and involve large-scale corporate-controlled financial or business activities. For a large organisation to be interconnected, internal services and departments of the organisation must be dependent on each other and in constant communication via a plethora of channels.

An example of such an organisation is the European Organisation for Nuclear Research CERN, is an interconnected large international research organisation widely known for the largest particle accelerator called Large Hadron Collider (LHC). At CERN more than 20000 people who work together on various projects, communicate, discuss, and share large amounts of information. As seen in Figure 1.1, CERN has a variety of services that generate large amounts of information and share it with users through different devices.

As a result of a large number of services, the information generated daily has been

1 Introduction

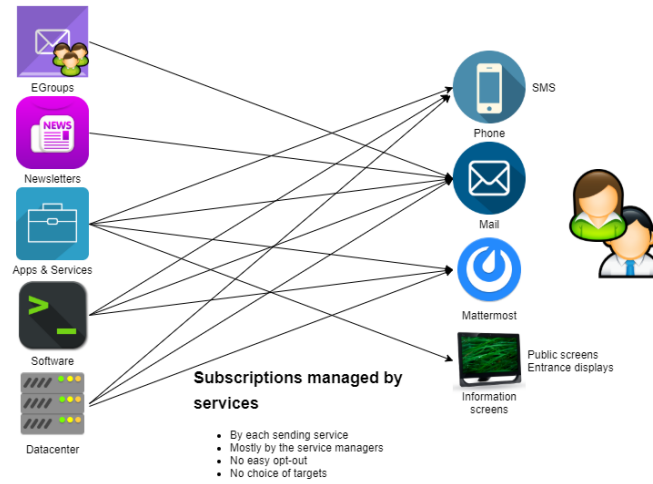


Figure 1.1: Current Information and Communication Workflow at CERN (Ormancey et al., 2022)

growing steadily. This increase in the amount of information available causes users to experience information overload. The difficulty of understanding an issue and effectively making decisions when provided with too much information about the issue is defined as information overload (Gunaratne et al., 2020).

One of the main issues with such large amounts of information in organisations is how to effectively retrieve information and re-find information and messages when needed. This requires not only a powerful search index within the organisation, but also access to a great variety of information outside the organisation. Many tools have been created for these use cases, and usually some of them are used simultaneously (M.-C. Lee, 2016). Searching for files in email conversations, searching for pieces of information in conversations, and the increase in information generation in organisations make it visible that there is a growing demand for tools that support information retrieval for large organisations. Today, navigating information within an organisation requires the use of a search engine as the main entry point. To reduce the effects of information overload, the demand for personalised approaches for search and information navigation has arisen. These approaches require the collection of personal user information, filtering unnecessary information and identifying valuable information, to create user profiles that can be used for personalization (Gauch et al., 2007).

One of such approaches that has been initiated at CERN is the CERN Notification Research Project. The CERN Notifications solution, currently being implemented, offers meaningful notifications in a heterogeneous environment of computing devices on both desktop and mobile devices. This new Notifications service provides opt-in and

opt-out notifications, which are delivered to devices of users of any flavour through live notifications or daily/weekly aggregations, depending on the user's preference, and provide access to a notification hub where all notifications are made available along with subscription options. The system aggregates various information from different services and sources and provides them to the end user according to his/her personal choice (Ormancey et al., 2022). The CERN Notifications project excels at collecting information from different sources but primarily acts as a distribution system rather than a recommendation system. It efficiently sends information to users based on their preferences and subscriptions but does not actively suggest or recommend what might interest individual users. In contrast to traditional recommendation systems that use algorithms to make personalised suggestions, the CERN Notifications project follows a more user-driven approach. It respects users' choices in selecting the notifications they receive.

To enhance the functionality of the Notifications project, the integration of a recommendation system is crucial. Such a system would empower the project to proactively suggest content, updates, and events that are highly relevant and potentially interesting to individual users. This proactive approach can boost user engagement by offering customised and timely suggestions. Users will not only receive information they explicitly subscribe to, but also discover content and events they might have otherwise overlooked. With the addition of a recommendation system, users can effortlessly access a curated selection of content aligned with their interests and responsibilities, streamlining the information consumption process and enhancing efficiency. In conclusion, expanding the CERN Notifications project with a recommendation system has significant potential to improve the user experience and optimise information consumption. By providing proactive suggestions, the project can ensure that users are not only well informed, but also actively engaged and satisfied with the service. This enhancement is in line with the evolving needs of large organisations like CERN, where efficient and personalised information delivery is of paramount importance.

However, there is a strong demand to preserve privacy while providing the above-mentioned services. Unauthorised use and leaks of personal information, as well as deanonymization, have raised concerns about exploiting information about user behaviour and interests. Such incidents played a crucial role in highlighting the importance of user privacy and privacy in general (Bobic et al., 2022). Preserving privacy in a personalised system depends on ensuring that the user feels in control of their information and ensuring the integrity of that information Ghorab et al. (2013).

1.2 Objectives

The purpose of this research is to determine whether the CERN notification system can be extended to offer a wide range of information navigation and recommendation

1 Introduction

possibilities, taking into account access control at the item level (e.g., notifications, channels), user privacy and sensitive information, and user navigation history on each device used to consume the information. The technical implementation and research of the dissertation were carried out at CERN as part of the creation of the notification system.

The purpose of this doctoral dissertation is to investigate and prototype a privacy-preserving information retrieval and recommendation approach and system within the context of a large organisation, such as CERN. This prototype system was created and evaluated at CERN for their Notification System. Algorithmic methods and machine learning techniques for information retrieval and navigation were used to help the user navigate and retrieve information in the multidimensional information space of the Notification System CERN.

Taking into account the purpose of the research and the main objective of the thesis, the question that this work aims to answer is:

How can privacy-aware information retrieval, user navigation, and information visualisation methodologies be used to improve and complement the CERN Notification System in terms to make it proactive and to provide a highly flexible search and navigation functionality to the user?

To answer the main question of this research, four subquestions have been identified:

- **RQ1:** How do users behave and consume information in large, highly connected organisations?
- **RQ2:** How can sensitive organisational data be used for reproducible research and development?
- **RQ3:** How important are privacy and privacy preserving concepts for employees in large organisations?
- **RQ4:** How can sensitive information be used in a privacy-preserving way for the creation of a personalised recommendation and information retrieval system and what is the performance of such a system compared to traditional systems?

1.3 Methodology and Thesis Structure

Design science (DS) research is a relatively new design-orientated research approach that has received a lot of attention in the field of information systems research Villanueva (2019). Figure 1.2 displays the general phases in the DS process.

According to Figure 1.2, the DS research framework places the research in the design cycle between the environment and the knowledge base. Iterations in DS follow three cycles: the relevance cycle, the design cycle, and the rigour cycle. The relevance cycle is

1.3 Methodology and Thesis Structure

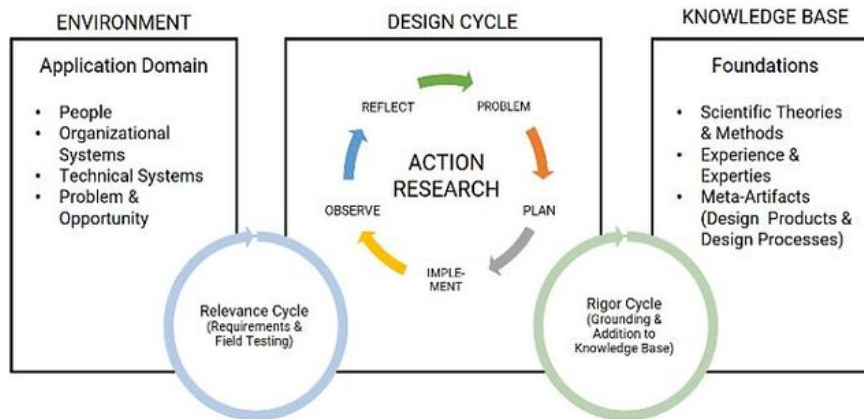


Figure 1.2: Design Science Research Framework - graphic taken from Villanueva (2019)

a step of the DS process that produces the requirements and defines the success criteria based on the environment, specifically the application domain (people, organisation, and others). Gathering and sharing knowledge about new scientific theories and methods, experiences, and expertise is the main task of the rigour cycle. The design cycle is the main step of the DS process that focusses on the development, evaluation, and use of the evaluation conclusions to refine the research. Each repetition of the design cycle takes input from both the relevance cycle and the rigour cycle (Simon, 1996; Horvath, 2007; Villanueva, 2019). Combining elements of engineering, computer science, information systems, and social sciences to create solutions that are practical and effective, the project focusses on making choices about what is possible and useful for the creation of possible futures, rather than what is currently existing. It focusses on the development and validation of prescriptive knowledge (Simon, 1996; Horvath, 2007; Villanueva, 2019).

When it comes to software development, the waterfall methodology is an example of a traditional approach that follows a linear and sequential process, where each phase of a project depends on the previous (Sherman, 2015). The project is deconstructed into a sequence of assignments, with the highest assignment level named phases. The waterfall approach requires phases that are completed in sequence and have exit criteria. A typical list of waterfall tasks is described in Figure 1.3.

The research methodology used in this work combines the steps of the two methodologies mentioned above (DS and Waterfall). It consists of five main phases, which are sequentially linked together as in the waterfall methodology. The phases are extracted from the DS methodology. Additionally, the phases are enhanced with knowledge of the waterfall methodology. Unlike the waterfall methodology, where it is not possible to return to a previous phase after completion, this methodology follows the DS concept

1 Introduction

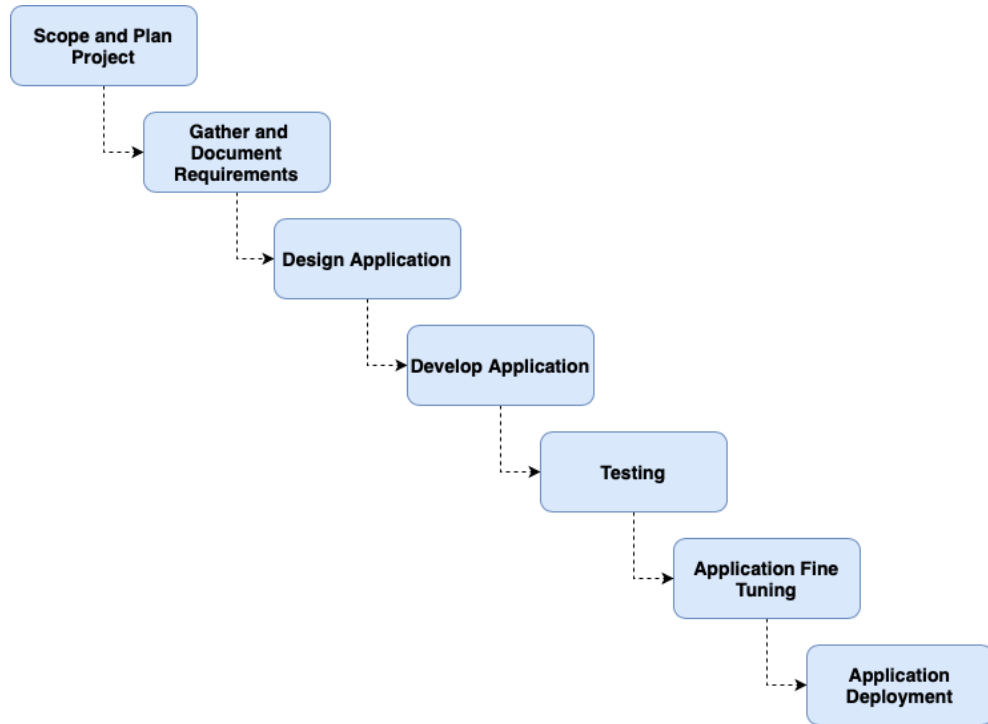


Figure 1.3: Phases of the Waterfall Methodology based on Sherman (2015)

of the design cycle. Thus, it is possible to return to a previous phase in the case that problems have been detected in the current phase. Figure 1.4 illustrates these five phases. This hybrid research methodology was used to meet the research objectives and challenges mentioned above.

Based on the selected methodology, the thesis structure was created to display the challenges and outcomes of each phase. Figure 1.5 shows the structure of the thesis and the relationship of the structure with the methodology used mentioned in Figure 1.4.

The initial phase consists of a literature survey and an investigation of related work. While **Chapter 1** provides motivation and determines the application domain, **Chapter 2** analyses and explores technologies, theoretical concepts, and methods related to the objectives of the thesis. Topics covered in this phase include determining how large organisations operate, focussing on information generation and user data processing and retention. Machine learning and recommendation systems and their link to privacy and privacy problems are also explored. The last area covered is open science principles and modern issues that arise from misusing organisational data (e.g. User Data, Sensitive Information, etc.). The goal of this phase is to link existing

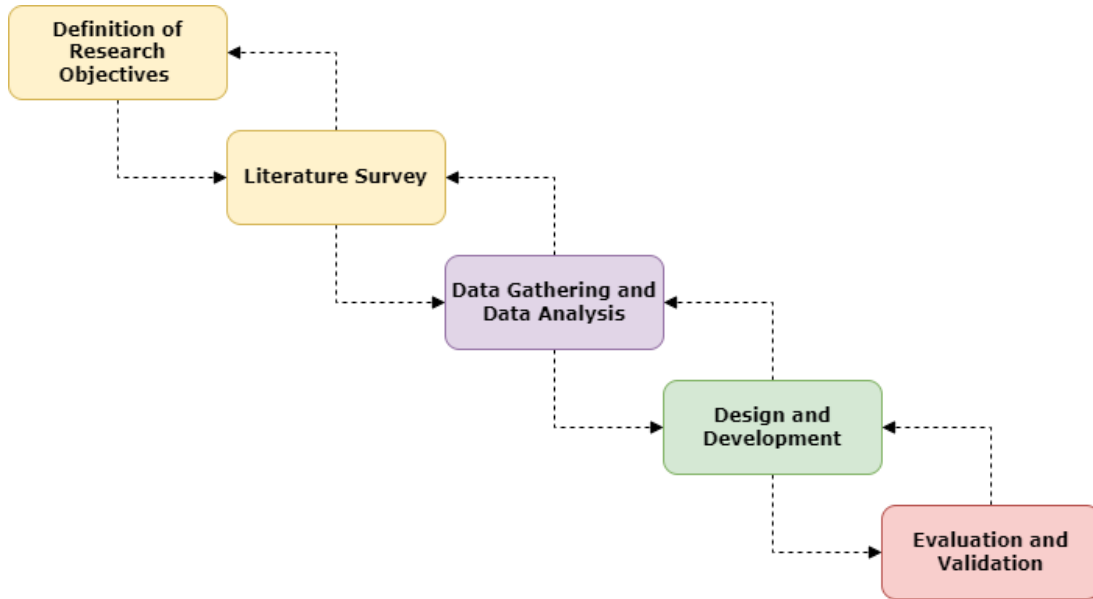


Figure 1.4: Flowchart of the Thesis Iterative Research Methodology

problems and issues, together with possible solutions to the research questions stated in the previous section.

The second phase begins with **Chapter 3** which inspects data generated from different user studies executed at CERN. The first study analysed is the CERN Newcomers Study, which focusses on determining new technological trends and information consumption behaviour among newcomers. The second study explored is the IT Department User Study. This study focused on determining users' daily work habits (e.g., tools used for communication, amount of meetings attended, ways of sharing information, etc.). The last study analysed was the CERN User Information Consumption Study. The main focus of this study was to determine the connections between user information consumption, daily habits related to information consumption, and user awareness of user privacy.

Chapter 4 describes a procedure for publishing open source organisational data. The need to enable data sharing for organisations without invading user privacy was the main motivator for this contribution. The created Data Lift Framework focused on the synthesis of easy-to-follow steps that would ensure privacy respecting safe data publishing. This framework represents the combination of multiple state-of-the-art data publishing frameworks. Furthermore, the application of this framework to the CERN Mattermost Dataset is described in this chapter.

Furthermore, **Chapter 4** utilises the data set resulting from the previous chapter.

1 Introduction

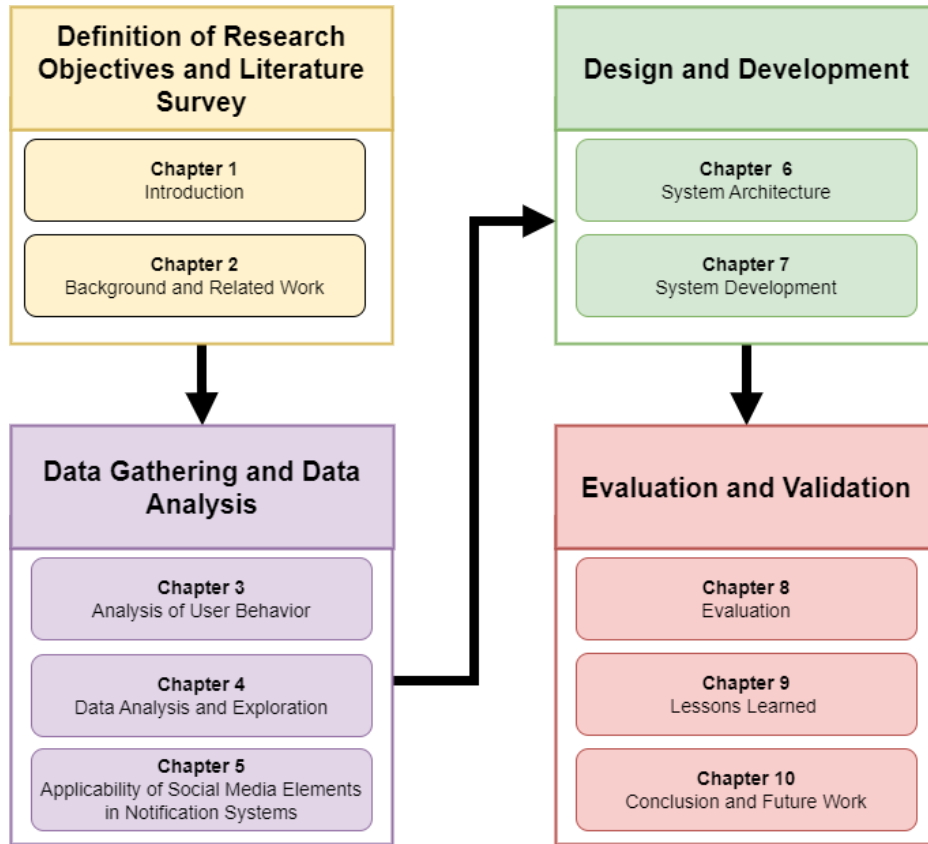


Figure 1.5: Structure of the Thesis

The application of statistical analysis, data exploration, and network analysis combined with community detection is discussed. State-of-the-art recommendation system algorithms are utilised to process the outcomes of data exploration and analysis. These algorithms undergo rigorous evaluation to quantify their effectiveness and identify optimal performers based on established performance metrics and criteria.

Lastly, **Chapter 5** investigates the stagnating design of the notifications displayed to users. Social media elements are explored to integrate them into the notification design. As part of this investigation, a user study was conducted to determine the user's opinion about novel notification display designs. The preferability of the use of various elements of social media were analysed.

Phase three consists of the design and development of the prototype system together with the architectural design of the system. **Chapter 6**, the first chapter of this phase, is based on the previous chapters and defines the Notification System, the requirements

set by the organisation to extend it, and the conceptual architecture of the extension. Technological decisions are discussed in this chapter and aligned with the general goals of the thesis. In **Chapter 7** details related to the technical implementation of the prototype system are described. These details also include methods of integration with different existing systems.

The last phase is dedicated to the evaluation of the prototype system and discussion of the lessons learnt, with additional future ideas for the expansion of research. **Chapter 8** describes the evaluation methods used, such as statistical evaluations of Machine Learning models, system usability scales, user satisfaction questionnaires, etc. The limitations, challenges, and insights gained during the work are discussed in **Chapter 9**. While **Chapter 10** reflects on and concludes the work carried out during this thesis. In addition, various future research opportunities are introduced.

1.4 Contributions

This section provides an overview of the scientific publications and research contributions made during the work on this thesis. Research contributions range from theoretical novelties and improvements to existing frameworks and methods to the enhancement of a novel notification system. The main contributions of this thesis are summarised below:

1. **In Depth Review of Privacy Preserving Methods for Recommender Systems** - The literature survey phase of this thesis has identified current problems in privacy preserving applications of recommendation algorithms in large organisations. Additionally, this phase has also summarised current work in the area of privacy-aware machine learning and its link to open science principles.
2. **Development of a Framework for Open Sourcing Organisational Data** - An additional contribution of this thesis is the creation of an organisational framework for open-sourcing data. This framework, named Data Lift, represents the aggregation of different existing frameworks from various regions around the world operating in different jurisdictions. It combines best-practises and simplifies the steps necessary to generate open data from sensitive organisational data.
3. **Evaluation of State-of-the-Art Recommendation Algorithms** - Various state-of-the-art machine learning algorithms for recommender systems have been evaluated as part of this work. The focus of this evaluation was on comparing the performance of such algorithms with implicit (anonymised organisational data) and explicit data (data containing sensitive information).

The following section contains a list of scientific publications in conference proceedings, book chapters, and journal articles that are directly related to the thesis. They covered the main aspects and concepts crucial to the thesis.

1 Introduction

- **Jakovljevic, I., Gütl, C., & Wagner, A. (2023).** Privacy-Preserving Collaborative Filtering: Evaluating a Machine Learning Recommender System in a Large Interconnected Organization. In 5th International Open Search Symposium
- **Jakovljevic, I., Gütl, C., & Wagner, A. (2023).** Privacy-Preserving User Clustering: The Application of Anonymized Data to Community Detection in Large Organizations. IARIA JOURNALS
- **Jakovljevic, I., Gütl, C., & Wagner, A. (2023).** Exploring Information Consumption Patterns Among Users in Large Organisations: A Survey Analysis of CERN Users. MDPI - Multidisciplinary Digital Publishing Institute, [In Review]
- **Jakovljevic, I., Gütl, C., & Wagner, A. (2022).** Analyzing the Effects and Applicability of Social Media Elements in Notification Systems in Large Interconnected Organisations. IARIA JOURNALS
- **Jakovljevic, I., Pobaschnig, M., Gütl, C., & Wagner, A. (2022).** Privacy Aware Identification of User Clusters in Large Organisations based on Anonymized Mattermost User and Channel Information. Proceedings of the 11th International Conference on Data Science, Technology and Applications - IARIA DATA ANALYTICS
- **Jakovljevic, I., Gütl, C., & Wagner, A. (2022).** Towards a Privacy-Aware Reproducible Machine Learning Pipeline for Open Data. In 4th International Open Search Symposium
- **Bobic, A., Jakovljevic, I., Gütl, C., Le Goff, J., & Wagner, A. (Accepted/In press).** Implicit User Network Analysis of Communication Platform Open Data for Channel Recommendation. In 9th International Conference on Social Networks Analysis, Management and Security - SNAMS 2022
- **Jakovljevic, I., Gütl, C., Wagner, A., & Nussbaumer, A. (2022).** Compiling Open Datasets in Context of Large Organizations while Protecting User Privacy and Guaranteeing Plausible Deniability. In Proceedings of the 11th International Conference on Data Science, Technology and Applications - DATA SciTePress - Science and Technology Publications
- **Jakovljevic, I., Wagner, A., & Gütl, C. (2021).** Applicability of Social Media Elements in Notification Systems in Large Interconnected Organisations. In SOTICS 2021: The Eleventh International Conference on Social Media Technologies, Communication, and Informatics
- **Jakovljevic, I., Russmann, S., Wagner, A., & Gütl, C. (2022).** A Proposal for Client Based User Profiles For Open Search in Large and Highly Connected Organisations. In Proceedings of the 3rd International Symposium on Open Search Technology: OSSYM 2021
- **Jakovljevic, I., Wagner, A., & Gütl, C. (2020).** Open Search Use Cases For Improving Information Discovery And Information Retrieval In Large And Highly Connected Organizations. In Proceedings of OSSYM 2020 CERN European Orga-

nization For Nuclear Research

- Papst F., Saukh O., Römer K., Grandl F., **Jakovljevic, I.**, Steininger F., Mayerhofer M., Duda J., & Egger-Danner C. (2019). Embracing Opportunities of Livestock Big Data Integration with Privacy Constraints. In Proceedings of the 9th International Conference on the Internet of Things (IoT '19). Association for Computing Machinery, New York, NY, USA
- **Jakovljevic, I.**, Wagner A., Gütl C., Pobaschnig M., & Mönnich A. (2022). CERN Anonymized Mattermost Data (Version 1). Zenodo.

2 Background and Related Work

In this section, an overview of the current state of privacy-aware recommender systems and open data-based information systems is provided, while highlighting key findings and contributions from previous studies. Gaps or limitations in the existing literature have been identified, together with the outline and motivation for this research. The aim of this section is to situate this research within the broader context of the field and to demonstrate the relevance and originality of our contributions.

The following sections are based on, supported by and taken from the work published in the following publications:

- **Jakovljevic, I.**, Pobaschnig, M., Gütl, C., & Wagner, A. (2022). Privacy Aware Identification of User Clusters in Large Organisations based on Anonymized Mattermost User and Channel Information. Proceedings of the 11th International Conference on Data Science, Technology and Applications - IARIA DATA ANALYTICS
- **Jakovljevic, I.**, Gütl, C., & Wagner, A. (2022). Towards a Privacy-Aware Reproducible Machine Learning Pipeline for Open Data. In 4th International Open Search Symposium
- Bobic, A., **Jakovljevic, I.**, Gütl, C., Le Goff, J., & Wagner, A. (Accepted/In press). Implicit User Network Analysis of Communication Platform Open Data for Channel Recommendation. In 9th International Conference on Social Networks Analysis, Management and Security - SNAMS 2022
- **Jakovljevic, I.**, Gütl, C., Wagner, A., & Nussbaumer, A. (2022). Compiling Open Datasets in Context of Large Organizations while Protecting User Privacy and Guaranteeing Plausible Deniability. In Proceedings of the 11th International Conference on Data Science, Technology and Applications - DATA SciTePress - Science and Technology Publications
- **Jakovljevic, I.**, Russmann, S., Wagner, A., & Gütl, C. (2022). A Proposal for Client Based User Profiles For Open Search in Large and Highly Connected Organisations. In Proceedings of the 3rd International Symposium on Open Search Technology: OSSYM 2021
- **Jakovljevic, I.**, Wagner, A., & Gütl, C. (2020). Open Search Use Cases For Improving Information Discovery And Information Retrieval In Large Anhighly Connected Organizations. In Proceedings of OSSYM 2020 CERN European Organization For Nuclear Research

2 Background and Related Work

- Papst F., Saukh O., Römer K., Grandl F., **Jakovljevic, I.**, Steininger F., Mayerhofer M., Duda J., & Egger-Danner C. (2019). Embracing Opportunities of Livestock Big Data Integration with Privacy Constraints. In Proceedings of the 9th International Conference on the Internet of Things (IoT '19). Association for Computing Machinery, New York, NY, USA

2.1 Information Generation and Consumption in Large Organisations

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic et al., 2020).

Every day, people send emails, take photos, record videos, create documents, and use different data and information generation techniques. (Forbes, 2018). The amount of data that humans produce is increasing as the world increasingly uses data. In 2017, humans were estimated to generate 33 trillion gigabytes of information per minute, with 90% of the total amount of global data generated between 2015 and 2017 (flexfibre, 2019). This increased influx of information has created challenges in the management of large amounts of information. On average, a user in an organisation sends and receives 128.8 emails per day, of which up to 24% contain additional information in the form of attachments (Tsipenyuk & Crowcroft, 2017; The Radicati Group, 2019). In addition to email, companies use face-to-face meetings, radio and television communications, mobile communication, electronic communication, and written communication as a means of sharing knowledge and generating information (Williams, 2019). According to Intel Group research, big data is associated with organisations that generate a median of 300 Terabytes (TB) of data each week. This study also showed that one third of the organisations analysed work with more than 500 TB of data per week. These organisations process non-structured data sources and aim to integrate data processing methods to transform data into structured form (Center, 2012). In the case of CERN, only LHC generates 10 Gigabytes (GB) per second or 500 TB in 19 hours. The data produced by the collider is used for research, reports, visualisation, communication, and more (Smith & Wagner, 2020).

User studies analysing user information consumption behaviour have been a prominent area of research, reflecting the growing importance of understanding how individuals interact with information in various contexts. Over the years, numerous studies have explored and analysed user behaviours, preferences, and motivations when it comes to searching, retrieving, and using information. These studies have employed

2.1 Information Generation and Consumption in Large Organisations

diverse methodologies, including surveys, interviews, observations, and experiments, to gain insight into the intricacies of information seeking behaviour.

One of the first widely recognised models for understanding the process of information-seeking was Ellis's model, known as the Information Search Process (ISP) model. It consists of four main stages: initiation, selection, exploration, and formulation. The ISP model provides a valuable framework for understanding the sequential and iterative nature of information seeking behaviour. It highlights the importance of various stages, from the initial recognition of an information need to the formulation of knowledge, in guiding individuals through the complex process of seeking and utilising information effectively (Folorunso, 2015).

Research done by Singh & Hsu (2007) evaluated the effectiveness of a novel approach for information retrieval and exploration, that supports human-machine synergy by identifying semantic correlations in the retrieved data and facilitating direct interactions between the users and the data. The first study aimed to assess the efficacy of the proposed novel approach compared to commonly used commercial search systems. Participants received 20 information goals and were assigned to find the corresponding information. The proposed system outperformed commercial systems in terms of time and number of clicks required for information retrieval. The second experiment focused on information discovery and clustering capabilities. Participants rated the proposed system higher in terms of ease of information discovery and understanding of the information structure. The importance of different interaction modalities was analysed in the third experiment, with text-clustering (47% participants prefer) being the most preferred, followed by spatial- display/interaction (19% participants prefer), temporal-display/interaction (12% participants prefer), and media-display/interaction (22% participants prefer). A brief study was conducted to measure improvement in user performance over time, with the proposed system showing the most significant reduction in task completion time. The user survey indicated that the proposed novel system received the highest rating for finding the desired information and exploring relevant topics (Singh & Hsu, 2007).

In Jokela et al. (2015) the practises and challenges of using and combining multiple information devices are evaluated. The research involved 14 participants and utilized diaries, interviews, and qualitative analysis. The most common devices, such as smartphones, computers, tablets, and home media centres, were studied. Through analysis of 123 real-life multi-device use cases, several usage patterns emerged, including sequential use, resource lending, related parallel use, and unrelated parallel use. The study also identified three levels of decision making involved in device selection and usage: acquiring, making available, and selecting devices. While participants desired seamless integration among their devices, they encountered practical difficulties, particularly in cross-ecosystem device interactions. Common challenges included connecting and transferring information, incompatible content formats, and

2 Background and Related Work

limited availability of applications and services across devices. Current support for multidevice use primarily focuses on Sequential Use and Resource Lending, with little support for Related Parallel Use. Participants expressed a need for improved content access between devices, suggesting direct sharing between devices as a potential solution. Additionally, the study highlighted the limitations of smartphones and tablets in replacing computers for demanding tasks due to factors such as limited text entry capabilities, imprecise pointing devices, and restricted multitasking. The findings also indicated that multi-device ecosystems could prolong the lifespan of old devices by assigning them specialised roles. The study's qualitative approach provides valuable insights into current practises, motivations, and needs related to multi-device use, while suggesting the importance of quantitative studies to validate and generalise the results.

In D. Zhang et al. (2004), a study was conducted on user information-seeking behaviour within a portal system. The study used existing framework and model definitions to analyse user behaviour. It classified information-seeking modes such as customisation, searching, browsing, communicating, monitoring, and service requesting. The researchers examined user access to functions and services, identified long-term information seeking patterns, and explored relationships between user groups and these patterns. The findings suggest that web-based portal systems can be suitable for studying information-seeking behaviour related to occupational tasks. The study also highlighted the diverse usage of information channels and the importance of communication and service requesting during the information-seeking process. The study acknowledges limitations such as the lack of insight into users' mental perspectives during information seeking and the possibility of users accessing other applications outside of the portal system. To address these limitations, the researchers propose conducting interviews, surveys, and observing users in person. Furthermore, future research will focus on exploring searching behaviour, studying long-term patterns as the portal system evolves, examining user-librarian interaction, and considering the impact of computing skills and multitasking on information seeking. The study emphasises the need for more research on how portal systems influence user information-seeking behaviour, especially as portal technology continues to improve.

A study conducted at the Nigerian Institute of Social and Economic Research (NISER), involving 58 active social sciences scholars examined information-seeking behaviour, with a focus on their preference for information sources, electronic resource usage patterns, and frequency of library visits. The results indicate that journals and books were the most preferred information sources for research needs. Most of the participants regularly used electronic information resources. This study also identified variations in the patterns of use of electronic resources according to academic rank and age range. The participants demonstrated regular usage of the web and email

2.1 Information Generation and Consumption in Large Organisations

for information gathering purposes. The study highlights the importance of aligning information resources and services with user requirements and the need for continuous research to enhance knowledge and support future developments in this area. Suggestions include adopting a user-centred approach in library services, conducting regular information-seeking behaviour studies, enhancing access to electronic resources, and using social media platforms to facilitate communication and collaboration between scholars (Folorunso, 2015).

Research done by Rouibah et al. (2009), investigates the organisational factors and human motivations that impact information systems and information technology usage and user satisfaction in an Arabic country. The study develops a research model linking three organisational factors (top management support, availability of training, and user involvement) to information systems and information technology usage and end-user satisfaction. The model is tested with 382 users and the results reveal that perceived usefulness has the strongest influence on usage and user satisfaction. The support of top management is the most significant organisational factor, followed by the availability of training and user participation. The study provides practical implications for practitioners seeking to improve system usage and user satisfaction. The model developed in this study demonstrates higher variance compared to previous similar studies and offers valuable insight into factors that affect system usage and user satisfaction.

In Han et al. (2020), the behaviour of the data workers during data curation activities (e.g., information seeking patterns and strategies) and how these behaviours relate to their performance was investigated. A data-driven experiment using eye tracking technology and an iPython Notebook-based platform was conducted among 39 participants. The participants consisted of undergraduate and master's students with Python knowledge. They had to identify data quality issues in a dataset by writing Python code and adding tags to indicate the errors. The study collected a multimodal dataset, including behavioural log entries, data quality issue tags, and gaze position data points. The results show that searching external resources, such as StackOverflow, is a common action that can improve performance. Providing sample code within the system and highlighting underlying data were effective in supporting data exploration. The most common reasons for triggering external search actions were seeking assistance in code writing or debugging and searching for relevant code to reuse. The findings provide insights into interaction patterns and information resource usage in data curation tasks, with implications for the design of domain-specific information retrieval systems. The study suggests the importance of presenting a data view and relevant code snippets internally, while also using external resources for more detailed examples and explanations. The research highlighted the need for customised data interaction interfaces based on user expertise and recommended integrating a code search interface to assist data workers.

2 Background and Related Work

In Reddy & Jansen (2008), limited understanding of collaborative information behaviour (CIB) in organisational settings is analysed, since most existing research focuses on individual information behaviour (IIB). The study presents the findings from two empirical studies investigating CIB, highlighting key differences from IIB in terms of communication patterns, triggers, and the role of information retrieval (IR) technologies. The results emphasise the importance of communication in CIB for sharing information and providing context during searches. Complex information needs act as triggers for collaborative behaviour, and IR technologies play a supporting role by enabling access to diverse information sources. Recognizing the growing importance of collaboration in organisations, the paper advocates for a shift in research focus from IIB to CIB. Drawing insights from healthcare teams, the authors develop a model of CIB based on behavioural and contextual dimensions, illustrating the interplay between IIB and CIB. The system aims to facilitate communication, automatic search based on ongoing dialogue, and real-time information sharing.

A survey-based investigation involving 200 employees from public and private sector organisations analysed the impact of gender, age, and income on employees' Internet usage. Regarding gender, the study found that male and female employees had similar distributions in their daily Internet usage. However, female employees tended to use the Internet more than males for communication/email/ chat. These findings align with previous studies suggesting a closing gender divide in Internet usage. Although some studies have reported gender differences in Internet usage, cultural influences and sample variations can contribute to the contradictory findings. The study also revealed that age significantly influenced employees' daily Internet usage and usage for information access/downloading/entertainment. Surprisingly, older employees (over 40 years) reported a higher daily Internet usage compared to younger employees. This finding contradicts some studies suggesting that younger individuals are the majority of Internet users. It may be attributed to the experience and opportunities available to older employees. However, the study noted that both younger and older employees showed minimal interest in using the Internet for electronic services. In terms of income, the study indicated a significant relationship between income and daily internet use. Higher-income employees tended to use the Internet for more than three hours daily compared to the lower-income group. This may be due to the affordability of Internet services and the higher qualifications among higher-income employees (Akman & Mishra, 2010).

Boland et al. (2012) focusses on the usage patterns and analysis of a research networking system (RNS) in the context of facilitating interdisciplinary scientific collaboration. RNSs have emerged as a promising approach to connect scientists with similar research interests and boost productivity. However, limited data and methods exist to understand how scientific professionals adopt and interact with RNSs. To understand this issue, a usage-logging program to track RNS use was created, recording

2.1 Information Generation and Consumption in Large Organisations

login activities, search queries, profile lookups, and user information. The analysis of the collected data revealed insights into the frequency of RNS usage, frequently searched topics, query types, user distributions, query complexity, and temporal patterns. The findings indicate that faculty members engage more in informational tasks, while research scientists and scientific administrators perform more navigational tasks. Navigational tasks refer to search queries performed to locate specific individuals, organisations, or resources within a system. Informational tasks involve search queries that are primarily focused on retrieving information on specific topics or concepts. The study highlights the need for personalised support and user-specific interfaces in RNSs to cater to the distinct information needs of different user types. The results contribute empirical knowledge and a methodology for studying RNS usage, providing a foundation for improving existing RNSs and fostering transdisciplinary scientific collaborations through these platforms.

Based on a study focussing on information consumption (Uda et al., 2018), for older adults with a high frequency of information system use, such as library systems, their previous experience significantly influenced their behaviour and perception. They did not change their search strategy, even when faced with difficulties. Similarly, older adults with low usage frequency but a preference for a search method (e.g. keyword search, browsing, etc.) showed a strong influence of previous experience on their behaviour and perception during the experiment. They adhered to their search strategy despite encountering difficulties. It was observed that older adults with low usage frequency and no preference for a search method changed their search methods between different questions, indicating the impact of the difficulties they experienced during the experiment on their behaviour and perception. Similar to those with high usage frequency, they also recognised the convenience of information search tools. The results highlighted the need for older adults to become familiar with the search tools available and suggested implementing an orientation program to enhance proficiency in using information search tools (Uda et al., 2018).

One of the main issues with such large amounts of information in organisations is how to retrieve information effectively and find information and messages when needed. This requires not only an efficient search index within the organisation, but also access to a great variety of information outside the organisation. Many tools have been created for these use cases; most of them are used simultaneously (M.-C. Lee, 2016). The tasks of searching for relevant information through files in email conversations or searching for specific pieces of information within conversations are becoming increasingly difficult due to the rapidly increasing information generation in organisations. This highlights the need for tools that can support efficient and effective information retrieval in large organisations (M.-C. Lee, 2016).

2.2 User Profiling and Profiles

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Russmann, et al., 2022; Bobic et al., 2022).

As mentioned in the previous section, the amount of information available to users is increasing exponentially. This increase in the amount of information causes information overload, difficulty understanding an issue, and lowers the effectiveness of making decisions when provided with too much information on an issue. This has led to an increase in the demand for personalised approaches for information retrieval and navigation. To address the demand for personalised information retrieval, systems must collect useful information about users, filter out unnecessary information, and recognise additional information that is of possible interest to them (Gauch et al., 2007).

A user profile, used for personalization within services, is a direct or indirect representation of information about an individual user. User profiles are essential for intelligent applications. The common content of the user profiles are user interests, user knowledge, background and skills, user goals, user behaviour, user interaction preferences, user characteristics and user context (Schiaffino & Amandi, 2009; Ghosh & Dekhil, 2009). User information can be divided into structure and unstructured. Structured user information refers to data that has a defined and consistent format, such as a user's name, date of birth, gender, or other personal attributes. This type of data is typically stored in a database or spreadsheet where it can be easily searched, sorted, and analysed. On the other hand, unstructured user information refers to data that does not have a pre-defined structure, and may be more difficult to analyse. Examples of unstructured data can include topic keywords, machine learning models, semantic networks, and other types of data that do not fit neatly into predefined categories. Unstructured data are typically more difficult to process and analyse than structured data but can provide valuable insights into user behaviour and preferences. (Yeung et al., 2008).

User profiling implies inferring unobservable information about users from observable information about them, that is, their actions (Zukerman & Albrecht, 2002). Creating user profiles for new users is a time-consuming process and can be done by manual or automatic user creation. One of the first user modeling or user profiling systems relied on knowledge bases that were manually created based on conclusions from observations about users. Early plan recognition systems used hand-crafted plan libraries to assume user intentions or preferences from his/her actions (Carberry, 2001).

Knowledge bases were built by thoroughly analysing several reoccurring instances of a problem that were assumed to be representative of this problem. The process of creating these knowledge bases is a resource-intensive process, and the resulting

knowledge bases are not adaptable or extendable. These drawbacks of the manual process of creating knowledge bases were also applicable in the area of artificial intelligence and became known as the knowledge bottleneck problem. This problem was also improved with the arrival of technologies that enabled innovative applications such as automatic message forwarding, document editing assistance, or user-recommended items. These applications can produce large volumes of data that are often corrupted by noise. Corrupting the data, together with the possible lack of user cooperation, also increases the uncertainty associated with making predictions from observed data. Building a knowledge base manually that is representative of many data points is clearly more labour intensive than building such a knowledge base when only a few examples are being considered. Additionally, when data are corrupted, the difficulty of this task and the possibility of errors increase.

2.2.1 Users Information Collection

The following paragraphs are based on, supported by, and taken from the work published in the following publications:

- Jakovljevic, Igor, Russmann, Stefan, Wagner, Andreas, and Gütl, Christian - A Proposal for Client Based User Profiles For Open Search in Large and Highly Connected Organisations

Collecting user information is considered the initial phase of user profiling. It involves the systematic collection and analysis of data related to users. This phase is crucial for developing a comprehensive understanding of users and their needs, which can inform the creation of personalised experiences or recommendations (Gauch et al., 2007). User information can be derived from various sources, each contributing distinct information of user behaviour, preferences, and interactions. These sources can be classified into two main categories: active user-provided information and passive data collection methods (Lu et al., 2021; Maher et al., 2019).

Active user-provided data encapsulates information that users consciously and voluntarily share with digital platforms. This is also often called explicit user feedback, which is based on user input of personal information. Demographic information such as birthdays, marriage status, occupation, or other personal information represents the type of information collected by explicit user feedback. The drawback of explicit user information collection is that it costs the user time and requires the user to participate and engage in the information collection (Lu et al., 2021). Popular methods for collecting Active User-Provided information are shown in Table 2.1.

Passive data collection techniques involve unobtrusive monitoring of user actions and interactions with digital platforms to implicitly collect and aggregate user data. These methods, shown in Table 2.2, yield observational insights with minimal direct

2 Background and Related Work

Method	Description
Personal Profiles	Users create online profiles with personal details such as full names, birthdates, gender, and contact information, providing a foundational source for identity verification. These profiles serve as digital identities, enabling personalised experiences and tailored content recommendations (Jakovljevic, Russmann, et al., 2022).
Preferences and Interests	Users articulate their preferences and interests by selecting choices in product categories, content genres, and favourite items, contributing to content recommendation engines (Qiu & Cho, 2006).
Feedback and Ratings	User-contributed feedback, reviews, and ratings provide valuable insights into their experiences, opinions, and sentiments regarding products, services, or content. Users actively engage with platforms by leaving reviews on purchased items, rating the quality of services, or providing feedback on content they have consumed (Poo et al., 2003).
User-Generated Content	Users generate diverse content types, including text, images, videos, and audio, which reveal their interests, behaviours, and contributions to the platform's ecosystem. Whether through social media posts, blog articles, or multimedia uploads, users actively create content that reflects their experiences, creativity, and viewpoints(Y. Li et al., 2018).
Surveys and Questionnaires	Users participate in surveys and questionnaires, offering structured data on preferences, habits, demographics, and satisfaction levels, informing targeted marketing and product development. These surveys can be designed to collect specific information, such as age, income, buying habits, and product preferences(Eke et al., 2019).

Table 2.1: Methods for Collecting Active User-Provided Information

user input. Early methods of implicit user information collection that are still used are user navigation tracking for web page suggestions. The main disadvantage of this method is that it raises privacy concerns for the user, which motivates the user to avoid explicit information collection. In practise, implicit information is more likely to be used because it does not require explicit user participation to collect information (Jakovljevic, Russmann, et al., 2022).

The data collected are then analysed using various statistical techniques, such as clustering, factor analysis, and regression analysis, to identify patterns and relationships between different variables. This analysis can help identify important user information, such as activity levels. This information can later be used to create user-centric systems such as systems that help users discover relevant information (Brusilovsky & Tasso, 2004; Jakovljevic, Russmann, et al., 2022).

Overall, user information collection is a critical phase in user profiling, as it enables

Method	Description
Cookies and Tracking	Cookies, small pieces of data stored on a user's device, are used to track user interactions with websites. They collect information such as browsing history, session duration, and clicked links. This passive method helps websites remember user preferences, improve user experience, and deliver targeted advertisements based on past behaviour (M. Ali et al., 2015).
Device Sensors	Mobile devices and smart appliances contain sensors (e.g., GPS, accelerometer) that passively collect data. GPS sensors track users' locations, while accelerometers detect device movement. These data sources enable location-based services, fitness tracking, and context-aware applications, all collected passively without user input (Kayacik et al., 2014).
Server Logs	Web servers passively record user interactions through server logs. These logs capture IP addresses, URLs visited, user-agents, and timestamps. Server logs are essential for website maintenance, security, and analytics. They provide insights into user traffic patterns, errors, and potential security threats, all without requiring user action (Rafter & Smyth, 2001).
App Usage Analytics	Mobile apps and software applications often include analytics tools that passively collect data on user behaviour within the application. This data includes screen views, button clicks, and time spent on specific features. App developers use this information to optimise user interfaces, fix bugs, and enhance the overall user experience (T. Li et al., 2023).
Passive Social Media Tracking	Social media platforms passively track user activity, including likes, shares, comments, and post views. This data is used to personalise content feeds, suggest connections, and deliver targeted ads. Users generate this passive data as they engage with social media, offering platforms insights into their preferences and interactions (Alarcón-del Amo et al., 2011).
IoT Device Data	Internet of Things (IoT) devices passively collect data from sensors, cameras, and other sensors embedded in everyday objects. These devices gather information about the environment, user habits, and device status. For example, smart thermostats passively record temperature preferences and occupancy patterns to optimise heating and cooling (Hernández-Álvarez et al., 2020).

Table 2.2: Methods for Collecting Passive Data

2 Background and Related Work

organisations to gain a deeper understanding of their users, which can inform the creation of more effective and personalised user experiences (Gauch et al., 2007).

2.2.2 User Profiling Methods

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Russmann, et al., 2022).

User profiling is a process that involves user information collection and analysis of the collected user data (behaviour, characteristics, preferences, etc.) to create a comprehensive profile of their interests and needs. User profiling is often used by information retrieval systems to personalise the content and services they provide to users, based on their explicit and inferred interests and preferences (Jakovljevic, Russmann, et al., 2022).

To achieve this goal, several methods are used for user profiling in information retrieval systems, including ontological user profiling, collaborative user profiling, knowledge-based user profiling, behaviour-based user profiling, and user filtering based on folksonomies and personomies. These methods employ various techniques, such as clustering, factor analysis, and regression analysis, to identify patterns and relationships between different variables and create a user profile that accurately represents their interests and preferences (Dickinson et al., 2003; Adomavicius & Tuzhilin, 2005; Delgado et al., 2002).

Overall, the methods used for user profiling in information retrieval systems enable personalised content and services that improve the user experience and satisfaction (Jakovljevic, Russmann, et al., 2022).

Behavior Based User Profiling

Behaviour-based user profiling refers to the process of creating a profile of a user's interests, preferences, and needs based on the analysis of their behaviour. It can include the collection and analysis of user search queries, browsing history, and social media activity. One of the key approaches to behaviour-based user profiling is the use of advanced machine learning and analytics techniques to analyse and interpret user data. These techniques identify patterns and trends in user behaviour to make predictions or recommendations based on these patterns. Behaviour-based user profiling can be particularly effective when combined with other user profiling forms, such as knowledge-based, as it can provide additional context and insights about the user (Jakovljevic, Russmann, et al., 2022; Yeung et al., 2008).

Collaborative User Profiling

Collaborative user profiling refers to the process of creating user profiles by aggregating and analysing the information and behaviour of multiple users. Collaborative user profiling is based on the assumption that similar users have similar preferences that can be inferred from the collective data and behaviour of the group. This can involve the creation of individual profiles that are based on the collective behaviour of the group or the creation of a shared user profile that represents the interests and preferences of a group of users. Collaborative user profiling is often used by online platforms and applications to personalise the content and services they provide users, based on the collective interests and preferences of a particular community or network (Abel et al., 2010; Liang, 2019).

Knowledge Based User Profiling

Knowledge-based user profiling refers to creating a profile of a user's interests, preferences, and needs to be based on the explicit representation and use of domain knowledge. This can involve the use of domain-specific concepts and relationships to represent and classify user data, as well as the use of reasoning and inference techniques to derive additional insights about the user. The use of structured knowledge representations, such as ontologies or taxonomies, represent and classify the interests and preferences of users in a precise and consistent manner, resulting in more sophisticated and accurate user profiling. This can facilitate the integration and interoperability of user profiles between different systems and applications (Delgado et al., 2002).

Ontological User Profiling

One of the key features of ontological user profiling is the use of structured knowledge representations, such as ontologies, to represent and reason about user interests and preferences. An ontology describes a number of specific terms and the relationship between them. The advantage of ontologies in computer science is that it enables concepts to be formalised, that is, it makes concepts readable for both computers and humans. In addition to the meanings of the terms, derivatives can also be described based on individual expressions (Felden & Linden, 2007). Ontologies can be used to represent and classify the interests and preferences of users in a precise and consistent manner. This can enable more sophisticated and accurate profiling of users and can also facilitate the integration and interoperability of user profiles between different systems and applications (Sieg et al., 2007). In general, ontological user profiling has the potential to improve the accuracy and effectiveness of recommendation and personalization systems, but it is important to carefully consider the privacy and ethical

implications of such systems (Felden & Linden, 2007; Sieg et al., 2007).

2.2.3 Short-Term and Long-Term Interests

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic et al., 2020).

Short- and long-term interest-based user profiling can lead to filter bubbles by creating a personalised information environment that reinforces users' existing beliefs and interests.

Short-term interests are goals or objectives of immediate or immediate-to-short-term importance. They evolve as users continuously interact with items, systems, or their surroundings. For example, users may establish new interests after discovering a hobby. Meanwhile, users may also gradually lose interest if they stop interacting with related items (Y. Zheng et al., 2022).

Long-term interests, on the other hand, refer to goals or objectives that are more distant or of long-term importance. Long-term interests are less susceptible to change-based recent interactions and are instead based on a comprehensive understanding of user preferences, making them more stable. In essence, long-term interests can be deduced from the entire history of user interactions (Y. Zheng et al., 2022).

Although some people give precedence to the achievement of short-term objectives, others prioritise long-term goals. The significance of both short-term and long-term interests is crucial, as they serve as drivers for individuals and influence their priorities and decision-making processes. However, the relative weight assigned to short-term versus long-term interests may differ depending on the context and the individual's goals and values (Jakovljevic et al., 2020; Y. Zheng et al., 2022).

2.2.4 Filter Bubbles

User profiling and filter bubbles are two interconnected concepts that are becoming increasingly relevant in today's digital world. As user profiling focusses on gathering and analysing data about individuals to understand their preferences, behaviours, and interests, it can contribute to the creation of filter bubbles by personalising information based on users' past behaviours and preferences.

Filter bubbles refer to the phenomenon in which individuals are exposed to a narrow range of views and information due to algorithms and filters used to customise content. These algorithms often prioritise content that matches the user's preexisting beliefs and preferences, resulting in the creation of "echo chambers", where individuals only see a

limited range of perspectives. This can lead to the reinforcement of existing beliefs and biases and to reducing exposure to new or dissenting perspectives. It can also limit the exposure of individuals to disagreeing viewpoints (Bruns, 2021; Geschke et al., 2019).

Staying within filter bubbles often occurs through a combination of conscious choices and algorithmic influences. Some individuals consciously consume information sources that align with their beliefs and preferences. They actively stay within their comfort zones, reading articles, following social media accounts, or watching news channels that reinforce their views. This deliberate selection of content creates a self-imposed filter bubble, where contrary opinions are intentionally avoided. On the other hand, algorithms play a crucial role in sustaining filter bubbles, often without users' explicit consent. Platforms utilise complex algorithms that analyse users' past behaviour and engagement patterns. These algorithms then prioritise content that matches users' existing preferences, effectively protecting them from different perspectives. Even users who may initially seek out varied information can find themselves within a filter bubble as algorithms continuously curate their content based on their interactions (Geschke et al., 2019).

Filter bubbles are closely tied to user profiles, as they rely on user data for personalised content. These profiles, built on user preferences and behaviours, form the basis for algorithms to tailor information, resulting in filter bubbles where users predominantly encounter content aligned with their profiles (Bruns, 2021; Geschke et al., 2019). Filter bubbles are the subject of considerable research and discussion, with some studies suggesting that they can negatively affect public opinion formation, political polarisation, and social cohesion. However, there is an ongoing debate about whether filter bubbles actually exist and their possible consequences. Some researchers argue that the influence of filter bubbles may be overestimated and that individuals still have the ability to search for a variety of sources of information if they choose to do so. However, filter bubbles are still an important and influential concept in the study of online communication and media (Saeed, 2022; Bruns, 2021).

2.2.5 Cognitive and Search Bias

Cognitive bias is a deviation from rationality in decision making due to certain assumptions that individuals unconsciously apply when processing information. Cognitive biases can lead to illogical interpretation, incomplete perceptions, evaluations, or decisions, and are influenced by various factors such as context, experience, culture, and personality (Haselton et al., 2015).

Search bias, on the other hand, refers to systematic patterns of deviation from norm or rationality in the search for information, whereby the search process itself may be biased in some way. This can occur when individuals are exposed to a narrow range of information due to the algorithms and filters used by online platforms to personalise

2 Background and Related Work

content, leading to the creation of “filter bubbles” where individuals are only exposed to a limited range of perspectives. Search bias can also occur when individuals are influenced by their own preconceptions or biases and only seek information that aligns with their preexisting beliefs (Grimmelmann, 2011).

Search bias is the manipulation of search results by search engines or other information retrieval systems, leading to unequal or biased access to information. Various factors, such as algorithms, data sources, user behaviour, and commercial or political interests, can deliberately or unintentionally cause search bias (Grimmelmann, 2011).

Cognitive bias and search bias can result from oversights in user profiling that focus too much on personalization rather than generalisation. Recommendations based on past behaviour and interests can reinforce cognitive biases by only presenting information that confirms a user’s existing beliefs, while filtering out opposing viewpoints leading to a lack of diversity in the user information, ultimately restricting their exposure to new ideas and perspectives. Additionally, search bias can occur when a search algorithm uses personalization data to prioritise search results over others, promoting particular viewpoints or sources.

2.2.6 Privacy Challenges in User Profiling

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author’s earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic et al., 2020; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).

Data privacy is an important aspect when it comes to sharing organisational data. It is necessary to protect persons, institutions and organisations (data subjects) following laws and ethical rules during the life cycle of data (collecting data, processing and analysing data, publishing and sharing data, preserving data and reusing data). Privacy issues in user profiling refer to concerns that arise when personal data is collected, analysed, and used for and by services. These privacy issues can include:

- **Collection of personal data** - Information retrieval systems and recommendation engines can collect and retain personal data, including, but not limited to, search queries, browsing history, and demographic information, to build user profiles. Storing personal data in databases or servers increases the risk of security breaches, which can lead to unauthorised access or theft of personal information. Personal data collected for one purpose may be used for another without the user’s consent. This can lead to unintended consequences, such as targeted advertising or discrimination.
- **Misuse of personal data** - Personal data can be used for surveillance purposes, such as tracking individuals’ movements or monitoring their online activity,

which can infringe on privacy rights. Data, such as social security numbers or credit card information, can be used by malicious actors to commit identity theft or financial fraud.

- **Lack of transparency** - The extent to which personal data is collected from users and how it is used to generate search results or recommendations may not be well understood by them. This lack of transparency can make it difficult for users to understand how their data are being used and to exercise control over their privacy.
- **Lack of control** - Users may have limited control over the way their personal data are used for search or recommendation purposes and therefore may not possess the ability to opt out of such usage. This lack of control can leave users vulnerable to the misuse of their data.

2.3 Information Retrieval and User Information Navigation

Information retrieval (IR) can be defined as the process of finding elements (e.g., documents) from within large collections of an unstructured nature (e.g., text) that satisfy an information need. Specifically, in computer science, it is the process of retrieving information system resources that are relevant to an information need from a collection of those resources (Baeza-Yates & Ribeiro-Neto, 1999; Mooers, 1960).

2.3.1 Brief History of Information Retrieval

The origins of IR can be traced back to ancient library systems, such as the Library of Alexandria¹ in the 3rd century BC. These libraries used cataloging and classification systems to organise scrolls and books, enabling users to locate specific texts (Mooers, 1960). Later, with the invention of the printing press and followed by the industrial revolution, the amount of available information increased exponentially. Libraries struggled to manage this inflow, leading to a need for more efficient retrieval methods. In the mid-20th century, researchers developed Boolean retrieval models that used logical operators (AND, OR, NOT) to filter and retrieve documents based on user queries (Manning et al., 2008). The Vector Space Model (VSM), introduced in the 1960s, represented documents and queries as vectors in a multidimensional space. This allowed for more sophisticated ranking of documents based on their relevance to a query (Mooers, 1960). The birth of the internet in the late 20th century brought about a new era in IR. Search engines like Google, based on complex algorithms and link analysis, revolutionised the way we find information online. Today's IR systems

¹https://en.wikipedia.org/wiki/Library_of_Alexandria

2 Background and Related Work

are highly sophisticated, using natural language processing, machine learning, and user behaviour analysis to provide personalised and context-aware search results (Baeza-Yates & Ribeiro-Neto, 1999).

The evolution of IR was driven by multiple factors such as the change in information need, technological advances, and a growing understanding of how users interact with information.

2.3.2 Information Retrieval Methods

IR is a rapidly growing field of computer science that has been extensively studied and developed over the past several decades. The main objective of IR is to provide users with relevant and useful information from large collections of data, including text, images, and multimedia content (Manning et al., 2008; Baeza-Yates & Ribeiro-Neto, 1999). There are diverse methods for finding and accessing relevant information from different types of repositories. The most used methods are searching and browsing and the choice between these methods relies on the information need and the characteristics of the repository.

Searching in the context of information retrieval (IR) is a formalised process characterised by the formulation of a user query, typically comprising keywords, phrases, or search terms, followed by its submission to a search system. This system, leveraging various algorithms and techniques, retrieves and subsequently presents a ranked list of documents or items derived from a stable repository, which may encompass a library catalog, a web index, or another structured collection of information resources. The aim of this process is to fulfill specific user information needs by providing a curated list of documents that best align with the query. Searching is well-suited for situations where users have well-defined information needs and expect a precise answer or set of documents (Baeza-Yates & Ribeiro-Neto, 1999).

Browsing is a user-centric exploratory approach to information retrieval that involves navigating through a collection of information resources, typically in a non-linear manner. Unlike searching, where users explicitly formulate queries to retrieve specific information, browsing is often more useful when users have less structured or evolving information needs, allowing them to discover information as they interact with a repository. Instead of retrieving specific documents, browsing allows users to interact with the repository's structure or content and discover relevant information as they explore. This approach is often seen in digital libraries, where users navigate through a collection of documents, or in online marketplaces, where users explore various products (Mooers, 1960).

In some cases, the choice between searching and browsing may be influenced by the stability of the repository. In these repositories, the information is well organised and remains moderately consistent over time. When dealing with a relatively stable

2.3 Information Retrieval and User Information Navigation

repository, such as a traditional library catalog or a static database, browsing is often the preferred method. In more dynamic environments where the repository's content is constantly changing or evolving, searching and filtering are the preferred techniques. Filtering involves defining specific criteria related to a topic, and the repository adapts to these criteria. For example, in webshop platforms, users can create filter queries for products based on attributes like price, type, or ratings, and the displayed products change dynamically based on the selected query filters. This approach allows users to interactively narrow down their options and explore information in a more tailored way (Baeza-Yates & Ribeiro-Neto, 1999; Manning et al., 2008).

2.3.3 Information Retrieval Process

The IR process can be broken down into several stages, including document indexing, query processing, and ranking. In the indexing stage, documents are analysed and structured in a way that facilitates efficient retrieval based on user queries. Query processing involves interpreting user queries and identifying relevant documents based on their content and context. Finally, ranking involves ordering the retrieved documents according to their relevance to the user's query (Manning et al., 2008).

The IR process has been extensively studied and many models and algorithms have been proposed to improve its effectiveness. IR models can be classified into four types: probabilistic models, algebraic and logical models, information theoretic models, and Bayesian models (Amati, 2009). For example, the Vector Space Model (VSM) is a well-known model used for ranking documents based on their similarity to a query. The VSM represents both documents and queries as vectors in a high-dimensional space, where the similarity between them is measured using the cosine similarity function (D. Lee et al., 1997). Other popular models and algorithms used in IR include the Okapi BM25 algorithm, which uses a probabilistic framework to rank documents based on term frequency and inverse document frequency, and the Language Modeling approach, which models the probability of generating a query from a document and ranks documents based on this probability (Whissell & Clarke, 2011).

The information retrieval process and information systems are intrinsically connected as IR methods and technologies form the foundation of effective information systems. An information system, in its essence, is a structured and organised framework designed to facilitate the storage, retrieval, and management of information to meet user needs (Baroudi et al., 1986).

Ives et al. (1983) define User Information Satisfaction (UIS) as a measure of user satisfaction with the underlying information system to meet user information requirements. User participation and feedback in system development lead to greater user satisfaction with user information and increased system usage (Baroudi et al., 1986). UIS has been identified as one of the main objectives of information systems; in the absence of better

2 Background and Related Work

evaluation criteria, it is used to evaluate an information system. Joshi et al. (1986) identify five significant factors that impact user information satisfaction in information systems: quality of information products, attitude towards staff and services, level of knowledge and participation of the user, fairness in the allocation of information and related resources to different user groups, and impact of the information system on the work environment of the user. According to Shirani et al. (1994) the identification of strong user expectations regarding the characteristics of the system would be a critical step to ensure higher user satisfaction, making the system more desirable and usable for the user.

Information retrieval is the foundation of information systems, and their combined objective is to optimise UIS by providing efficient and relevant access to information resources. Effective information systems employ IR techniques to ensure that users can search, browse, and filter information to meet their specific needs and preferences, ultimately enhancing their satisfaction with the system's performance.

Recent advances in machine learning and deep learning have also had a significant impact on IR. For example, deep neural networks have been used to learn distributed representations of documents and queries that can be used for efficient retrieval and ranking. In general, IR is a complex and interdisciplinary field that has many practical applications in areas such as information management, data mining, and Natural Language Processing (NLP). With the growth of big data and the growing need for efficient information retrieval and management, IR is expected to continue to be a rapidly evolving and important area of research and development (Manning et al., 2008).

2.3.4 User Information Navigation

User information navigation is a concept in information retrieval that refers to the process by which users find relevant information by navigating and interacting with information resources. The goal of user information navigation is to enable efficient and effective access to information resources, allowing users to quickly and easily find the information they need (Rist, 2009).

User information navigation involves multiple activities and techniques, including browsing, searching, filtering, and sorting. Browsing refers to exploring information resources by navigating through a predefined structure or browsing interfaces, such as a table of contents or a site map. Searching involves querying information resources using keywords or other search terms to retrieve relevant documents. Filtering and sorting involve manipulating search results to refine or prioritise them according to various criteria, such as relevance, popularity, and others (Reddy B et al., 2018; Ingwersen, 1992).

Various factors, such as user goals, preferences, knowledge, and design and structure

2.3 Information Retrieval and User Information Navigation

of information, influence user information navigation. Effective user information navigation requires a detailed understanding of the information resources and users that access them and involves the development of user-centered navigation interfaces and techniques (Manning et al., 2008; Amati, 2009; D. Lee et al., 1997).

User information navigation has been extensively studied in the field of information retrieval, with a range of models and algorithms developed to improve navigation and search efficiency. For example, the PageRank algorithm, developed by Google, is a well-known algorithm used to rank web pages based on their relevance and importance. Other popular navigation models include the HITS algorithm, which identifies central locations and authorities in a network of web pages, and the TF-IDF model, which ranks documents based on the frequency of keywords in the document and throughout the corpus (Manning et al., 2008; Amati, 2009; D. Lee et al., 1997).

2.3.5 Subscriptions and Publishing Systems

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Gütl, & Wagner, 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).

In computer science, subscriptions and publishing systems refer to the technologies and processes used to create, manage, and distribute digital content to users through subscription-based models. Such systems are based on several middle-ware applications to receive the publications and deliver the matching publications to the subscribers. These tools include content creation tools, content management systems (CMS) and distribution platforms that allow content providers to publish their content to a wider audience (Jacobsen, 2009; Hafeez et al., 2018). Subscriptions and publishing systems play a critical role in selective information dissemination, particularly in digital environments. These systems are designed to enable users to receive content updates, notifications, or publications from various sources based on their preferences. Selective dissemination of information (SDI) refers to the process of identifying and providing relevant information to specific users, according to their interests or needs. SDI systems are often used by libraries, information centers, and other organisations to provide customised support to users and help them access the information they need more efficiently (Hensley, 1963). SDI systems also called information systems typically operate by collecting information about the interests and needs of users and using this information to identify and provide relevant information to them. This can be done through various channels, including email alerts, text messages, or personalised web pages. SDI systems may also incorporate feedback mechanisms that allow users to adjust their preferences or request additional information (Ferguson & Hebels, 2003).

2 Background and Related Work

Research on SDI systems has explored a variety of topics, including the effectiveness of different SDI strategies, the impact of SDI on user satisfaction and information literacy, and the ethical and privacy implications of SDI (Hensley, 1963; Ferguson & Hebel, 2003; Chatterjee, 2017).

The two main elements of subscriptions and publishing systems are subscriptions and publishing concepts. Subscriptions are user-initiated actions in which they express interest in receiving content from a particular source, such as a website, blog, or online service. The process typically involves the following components: user subscription request, subscription management, notification preferences, and content delivery. User Subscription Request includes the process of the user expressing interest in receiving updates or content from a specific source by subscribing to it. Subscription Management includes providing options for users to manage their subscriptions, allowing them to choose the frequency of updates, select specific topics or categories of interest, or unsubscribe when they no longer wish to receive content. Subscribers can often specify their notification preferences, such as receiving email notifications, mobile app push notifications, or alerts within a web-based dashboard. Once a user subscribes to a source, the publishing system responsible for that source will periodically send updates or content to the subscriber based on their preferences (?). Publishing relates to actions aimed at creating, managing, and delivering content to subscribers. Authors, editors, or content creators use the publishing system to create and format content. This content can include articles, blog posts, videos, images, or any digital media. Publishing systems offer tools for organising and managing content, including categorisation, tagging, and archiving. This helps maintain a structured content repository. They often include analytics tools to track user engagement, content performance, and subscriber behaviour. User feedback mechanisms, such as comments and ratings, can also be integrated (Fung et al., 2010).

An example of a subscription and publishing system is a notification system. There are many different implementation versions of notification systems. The most commonly used are push notification systems for mobile phones, desktop status notification systems, browser-based notification systems, vehicle information systems, and others. Popular examples of notification systems include iOS ¹ and android push ² notification, messaging applications such as Facebook Messenger³, social media platforms notifications (e.g, twitter, facebook, etc.). Notification systems attempt to communicate important information to users effectively without creating unwanted intrusion into current user tasks. Selecting important information for the user is a difficult task. A study of 400+ participants has shown that users are not satisfied with the notifications they receive from notification systems because they do not express the user's current

¹https://en.wikipedia.org/wiki/Apple_push_notification_service

²https://en.wikipedia.org/wiki/Push_technology

³https://en.wikipedia.org/wiki/Messenger_software

2.3 Information Retrieval and User Information Navigation

interest. This leads to users ignoring most notifications from these systems. Besides determining what is relevant information for the user, an essential concern in notification systems is the display of notifications without a significant interruption of users' main tasks (Pradhan et al., 2017; Mehrotra & Musolesi, 2017). Visual implementations of notifications that are typically not a user's main attention priority are called secondary displays. Users willingly sacrifice brief interruptions from their primary task to view information of interest on these secondary displays ?. There are several ways to display notification messages, and the state-of-the-art can vary depending on the specific context in which the notifications are being used. Some common options for displaying notifications include using pop-up windows or banners on a computer or mobile device, using LED or visual indicators on hardware devices, or using in-app notifications within a mobile or web application. These technologies can be effective in alerting users to important information or events in a timely way and noticeable without being too disruptive. In general, the state-of-the-art for displaying notification messages is constantly evolving, and there are many different technologies and approaches that can be used to effectively alert users to new information (Pradhan et al., 2017; Mehrotra & Musolesi, 2017).

According to Sahami Shirazi et al. (2014); Mehrotra & Musolesi (2017), notifications have an important role in real-time information delivery helping users consume information. The side effects of notifications are that they can be disruptive if they demand users' attention at inconvenient moments. Despite the disruptive nature of notifications, users decide to use them because of their benefit in providing relevant information. In this context, notification systems can be beneficial and attempt to aggregate the previously mentioned information from different sources (e-mail clients, news portals, messaging platforms, and others) and deliver it to the user in the form of notifications (D. S. McCrickard et al., 2003). In addition to providing information aggregation and notification delivery, notification systems enable notification management (e.g., selecting which applications are allowed to send notifications), reducing the need of the user to constantly interact with different applications (Pielot et al., 2014). The success of a notification system depends on the accuracy of supporting the user with information between tasks, while simultaneously enabling utility by providing access to additional information (D. McCrickard et al., 2003). Notification systems attempt to keep users informed by balancing the amount of valuable information provided and the disruption caused by the information. It is necessary to find means to coordinate the delivery of notifications from multiple applications across multiple devices or/and display only relevant information at a glance. By bringing together multiple sources of notifications, the user can determine the importance of a notification and reduce the level of distraction (Weber et al., 2015). To address this challenge, intelligent notification mechanisms that monitor and learn users' behaviour have been proposed. These mechanisms maximise users' receptivity to information by determining the

2 Background and Related Work

right time and context for delivering notifications. In (Mehrotra & Musolesi, 2017), the significance of comprehending user preferences and usage patterns in notification systems to enhance the way users consume information is analysed. By gaining insight into users' preferences, such as the right time and context for delivering notifications, the effectiveness of information delivery can be improved. This understanding allows for the development of intelligent notification mechanisms that can adapt to users' behaviour, maximising their receptivity to the delivered information. The importance of considering user preferences and usage in designing notification systems for a more efficient and personalised user experience is emphasised.

2.4 Machine Learning

Machine Learning (ML) stands as a dynamic scientific field, delving deep into algorithms and statistical models that empower computer systems to accomplish designated tasks autonomously. This innovative discipline focuses on the development of computational techniques enabling machines to discern intricate patterns and draw insightful inferences from vast datasets. Unlike conventional programming, where tasks are explicitly defined, ML algorithms learn from data, refining their performance over time (Bishop, 2006, 2007).

2.4.1 Brief History of Machine Learning

The foundations of machine learning trace back to statistical concepts introduced in 1940, pioneered by eminent figures such as Thomas Bayes, Pierre-Simon Laplace, and Andrey Markov. The late 1940s witnessed the birth of the first manually operated computer system, called ENIAC, which initially earned the title "numerical computing machine" due to its intensive numerical computation capabilities (McCartney, 1999). This groundbreaking invention sparked the vision of creating a machine capable of emulating human thinking and learning processes. Arthur Samuel, in the 1950s, developed a computer program capable of playing checkers through simple learning and memorisation techniques, coining the term "Machine Learning" in 1952 (Wiederhold & McCarthy, 1992). Thereafter, Frank Rosenblatt integrated Donald Hebb's brain cell model with Samuel's machine learning breakthrough to devise the perceptron (Rosenblatt, 1958). The perceptron was regarded as the first successful neurocomputer, the perceptron marked a significant leap in machine learning. Over time, machine learning evolved through various algorithms and structures, including the nearest neighbor algorithm and artificial neural networks. These advancements have driven machine learning into its current state, finding applications in various domains such as predictive maintenance, fraud detection, recommendation, dynamic pricing, and NLP.

2.4.2 Main Challenges Of Machine Learning

According to Géron (2017) the main challenges of machine learning are related to issues with the ML model and the data used to create the model. They include insufficient quantity of training data, poor quality data, irrelevant features, and over- or under-fitting the training data.

Insufficient Quantity of Training Data

An insufficient quantity of training data is a common problem in machine learning where the dataset used to train a model is not large enough to capture the underlying patterns in the data. Lack of sufficient data can cause the machine learning model to be biased, inaccurate, or prone to overfitting, which occurs when the model is too complex and fits the training data too closely, leading to poor generalisation performance on new data (J. Li et al., 2019).

Poor-Quality Data

Poor-quality data is a common problem in machine learning, where the data used to train a model are incomplete, inconsistent, noisy, or biased. Poor quality data can negatively impact the performance of a machine learning model and can lead to inaccurate or biased predictions. Incomplete data have missing values or features, making it difficult for a model to learn patterns in the data. Inconsistent data is not uniform in its structure or format, which can cause errors in data processing and analysis. Noisy data contain errors or outliers, which can affect the accuracy of a model's predictions. Biased data do not represent the true distribution of the population, which can result in bias towards certain groups or outcomes (Chen et al., 2021).

Irrelevant Features

Irrelevant features are features in a dataset that do not contribute to the predictive power of a machine learning model. Including them in a model can negatively impact its performance, as it can introduce noise and reduce its ability to learn patterns in the data. Irrelevant features can arise in several ways, such as through data collection, data preprocessing, or feature selection. Irrelevant features may be generated during data preprocessing, such as using scaling or normalisation techniques. Furthermore, they may be selected during feature selection, intentionally or unintentionally, due to factors such as bias or incomplete knowledge of the problem (Géron, 2017; Chen et al., 2021).

Overfitting or Underfitting the Training Data

Overfitting and underfitting are common problems in machine learning that occur when a model cannot generalise to new, unseen data. Overfitting occurs when a model is too complex and fits the training data too closely, leading to poor generalisation performance on new data. It occurs when a model has too many parameters relative to the size of the dataset or when trained for too many epochs. Overfitting occurs when a model has high accuracy in the training data but poor accuracy in the new data. Underfitting is the opposite of overfitting and occurs when a model is too simple and does not capture the underlying patterns in the data, leading to poor performance on both the training and the new data. It also occurs when a model is too constrained, such as having too few parameters, or is not trained for enough epochs. Underfitting occurs when a model has low accuracy in both training data and new data (Géron, 2017; Chen et al., 2021).

Hallucinations

Hallucinations refer to a significant challenge related to the generation of artificial data or content by AI models, often neural networks, that appears convincing but is entirely fabricated. Hallucinations occur when a machine learning model generates data or information that it believes to be accurate, even when it does not correspond to reality. These false creations can manifest in various forms, such as language models generating coherent and contextually relevant text that seems factual but is entirely fictional. Another example is image or video generation models producing realistic-looking visuals of objects, scenes, or people that do not exist in the real world. Hallucinations are often difficult to detect, since the generated content aligns with the patterns learned by the model during training (De Pierrefeu et al., 2018).

2.4.3 Popular Machine Learning Approaches

Machine learning approaches are organised according to the type of learning. These algorithms can be described with several different types of learning, including supervised machine learning algorithms, unsupervised machine learning algorithms, semi-supervised machine learning algorithms, reinforcement machine learning algorithms, multi-task learning algorithms, ensemble learning algorithms, and neural networks (Bishop, 2006; Mahesh, 2019). Figure 2.1 illustrates this division of popular machine learning approaches.

Supervised learning algorithms are a type of machine learning algorithm that utilises a set of labeled training data to learn a mapping function between the input variables and the corresponding output variables. Creating a predictive model that can accurately predict output variables for new unseen input data (Shalev-Shwartz & Ben-David,

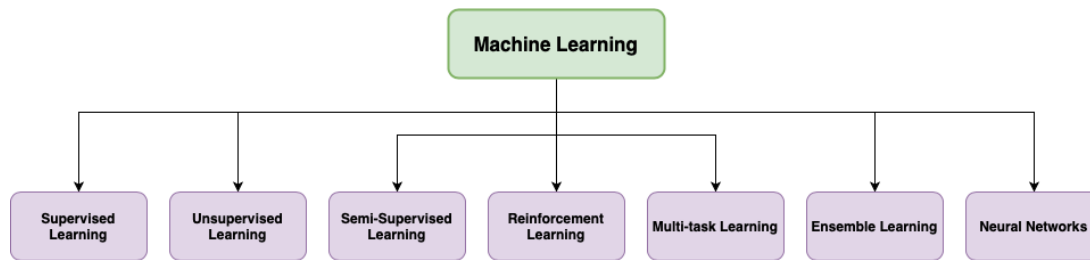


Figure 2.1: Popular Machine Learning Approaches

2014). Popular supervised machine learning techniques include Linear Regression, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Naive Bayes, k-Nearest Neighbours (k-NN), Gradient Boosting, Neural Networks (Deep Learning), Extreme Gradient Boosting (XGBoost), and others (Cunningham et al., 2008).

Unsupervised learning algorithms are a class of machine learning algorithms that facilitate the discovery of patterns and relationships within a dataset without labeled data. In unsupervised learning, the goal is to find hidden structures or groups without prior knowledge of the output variable. Unsupervised learning algorithms are a class of machine learning algorithms that facilitate the discovery of patterns and relationships within a dataset without labeled data. In unsupervised learning, the goal is to find hidden structures or groups without prior knowledge of the output variable (Bishop, 2006). Popular unsupervised machine learning techniques include K-Means Clustering, Hierarchical Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), PCA (Principal Component Analysis), ICA (Independent Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), Autoencoders, Isolation Forest, Self-Organising Maps (SOMs), Mean Shift Clustering and others (Alloghani et al., 2020).

Semi-supervised machine learning algorithms combine the functionalities of supervised and unsupervised learning approaches to create a mapping function or a classifier for a specific problem. Reinforcement learning problems involve learning to take the appropriate action to maximise a numerical reward signal to solve a problem (Sutton & Barto, 1998). Popular semi-supervised learning algorithms include Self-training, Multi-view Learning, Co-training, Tri-Training, Expectation-Maximization (EM), Label Propagation, Label Spreading, Transductive SVM (TSVM), Bootstrapping, and S3VM (Semi-Supervised Support Vector Machines) (Van Engelen & Hoos, 2020).

Reinforcement Learning is a type of machine learning algorithm where an agent learns to make sequences of decisions by interacting with an environment. It's often used in scenarios where an agent takes actions in an environment to maximise a cumulative reward signal. The agent learns through trial and error, receiving feedback in

2 Background and Related Work

the form of rewards or punishments (Oh et al., 2020). Popular Reinforcement Learning algorithms are Q-Learning, Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO), Actor-Critic, Deep Deterministic Policy Gradient (DDPG), Advantage Actor-Critic (A2C), Monte Carlo Tree Search (MCTS), Asynchronous Advantage Actor-Critic (A3C), and Soft Actor-Critic (SAC) (Oh et al., 2020).

Multi-task learning is a machine learning paradigm where a model is trained to perform multiple related tasks simultaneously. The idea behind MTL is that learning multiple tasks together can help improve the performance of each individual task by allowing them to share and leverage common features or representations (Y. Zhang & Yang, 2018). Popular multi-task learning algorithms include Multi-Task Lasso (MTLasso), Multi-Task Elastic Net (MTEN), Multi-Task Gaussian Processes (MTGP), Deep Multi-Task Learning (Deep MTL), Neural Multi-Task Learning (Neural MTL), Multi-Task Convolutional Neural Networks (MTCNN), Multi-Task Recurrent Neural Networks (MTRNN), Multi-Task Reinforcement Learning (MTRL), Transfer and Multi-Task Learning (TMTL), and Multi-Task Learning with Shared Knowledge (MTLSK) (Ruder, 2017).

Ensemble learning is a machine learning technique that combines the predictions from multiple models to produce a more accurate and robust prediction. The idea behind ensemble learning is to leverage the diversity among different models to improve overall prediction performance. It can be especially effective when individual models have different strengths and weaknesses or when they are based on different algorithms (Cunningham et al., 2008). Popular ensemble learning algorithms include Bagging (Bootstrap Aggregating), Random Forest, AdaBoost (Adaptive Boosting), Gradient Boosting, XGBoost (Extreme Gradient Boosting), LightGBM, CatBoost, Stacking, Voting Classifiers/Regressors, and Bootstrapped Ensembles (Cunningham et al., 2008; Dong et al., 2020).

Neural networks are a class of machine learning models inspired by the structure and function of the human brain. They consist of interconnected nodes, or artificial neurons, organised into layers. Information flows through these layers, with each neuron processing input data and passing it to the next layer. Neural networks are capable of learning complex patterns and relationships in data, making them suitable for a wide range of tasks, including image recognition, natural language processing, and more (Lawrence, 1993). Popular neural network algorithms include Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Gated Recurrent Unit (GRU), Autoencoders, Generative Adversarial Networks (GANs), Transformer, Self-Organizing Maps (SOMs), and Boltzmann Machines (Tkáč & Verner, 2016; Krogh, 2008).

2.4.4 Machine Learning Pipelines

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Gütl, & Wagner, 2022).

A ML pipeline is a sequence of stages that processes and transforms data, trains a ML model, and applies the model to make predictions on new data. The pipeline is designed to automate the end-to-end process of developing and deploying ML models to create a reliable, scalable, and reproducible system for solving real-world problems using ML (Jakovljevic, Gütl, & Wagner, 2022). Microsoft, Amazon, IBM, and other cloud providers have launched ML as a service that reduces costs, time, and risk of building ML infrastructure by offering premade generic ML tools. These tools recreate specific steps of the ML pipelines, allowing easy implementations of ML algorithms (De Cristofaro, 2021).

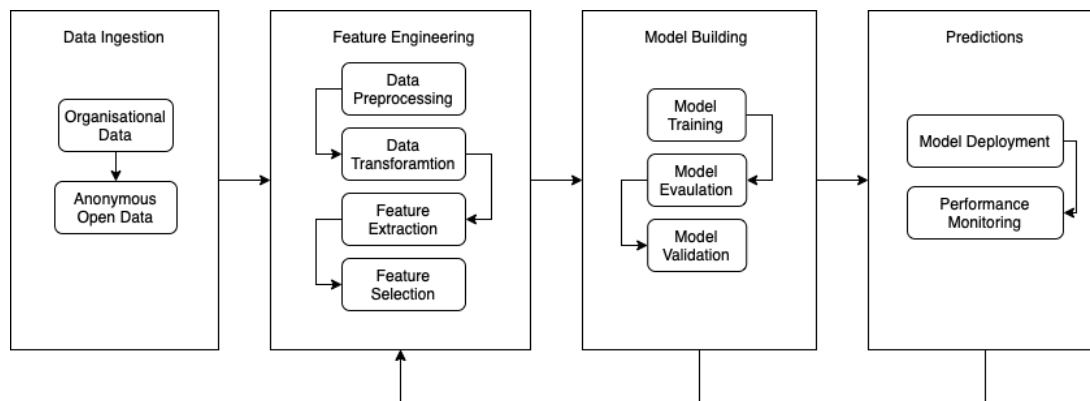


Figure 2.2: Open ML Pipeline Steps, image taken from (Jakovljevic, Gütl, & Wagner, 2022)

Based on previous research, the basic steps and substeps of ML pipelines have been identified and synthesised. These steps shown in Figure 2.2 are data entry, feature engineering, model building, and predictions.

Data Ingestion

Data Ingestion is the process of moving data from one or more sources to a destination where they can be stored and further analysed. Proper data ingestion ensures that the machine learning model receives high-quality data, leading to better model performance and more accurate predictions. In addition to the collection and aggregation

2 Background and Related Work

process, the data ingestion process can include anonymisation steps. Where private attributes are removed or anonymised to protect user privacy. The results of this step can also be used to publish the collected data (e.g. organisational data) as open data (Shawi et al., 2019; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022; Jakovljevic, Gütl, & Wagner, 2022).

Feature Engineering

Feature Engineering is a crucial process in ML in which raw data is selected, manipulated, and transformed into features that are suitable for ML algorithms. This process generally involves four main steps. Firstly, the data are preprocessed to remove any outliers and noise. Next, the preprocessed data are transformed to fit a specific data type and/or format. Following this, features are extracted from the transformed data using various feature extraction methods. Finally, the most appropriate features are selected for the given data and the task (Sugimura & Hartl, 2018; A. Zheng & Casari, 2018; Jakovljevic, Gütl, & Wagner, 2022).

Model Building

To create a ML model, it is necessary to learn and generalise the training data and apply that knowledge to new data. Model building has three main stages: model training, model evaluation, and model validation (Luo, 2016; A. Zheng & Casari, 2018; Shawi et al., 2019; Jakovljevic, Gütl, & Wagner, 2022).

1. **Model Training** - It is the initial step of the model-building process; in this step an ML algorithm/s are selected for model building. Training data are fed to the selected algorithms to help identify and learn adequate values for all the attributes involved (Luo, 2016; Shawi et al., 2019).
2. **Model Evaluation** - This step focusses on finding the most suitable model representing the input data and determining how the chosen model will perform in the future. Model performance evaluation is not done with training data because it can generate overoptimistic and overfitted models. To avoid these problems, evaluation methods use data from the initial dataset that are not used for model training (Luo, 2016).
3. **Model Validation** - Refers to the process of confirming that the model achieves its intended objective. This involves the confirmation that the model is predictive under the requirements of its intended use (Luo, 2016; A. Zheng & Casari, 2018).

If the models created in this phase do not perform as expected, it is possible to go back to Feature Engineering and use the knowledge gathered to improve and select better features.

Predictions

Once the ML model is created, evaluated, and validated, the next step is to make it available to end users. It involves the deployment of the model to applications or an endpoint for use. Additionally, it is essential to continuously monitor the model's performance to gather information for future development and improvements. If the monitoring phase indicates that the model is not performing as expected, it is possible to return to the Model Building or Feature Engineering phase to improve the ML model. Data collected during the monitoring phase can be used as additional information to create a better ML model or to select features from the old model. By monitoring the performance of the model and improving it, we can ensure that the ML model continues to provide the desired results over time (Luo, 2016; Shawi et al., 2019; A. Zheng & Casari, 2018; Jakovljevic, Gütl, & Wagner, 2022).

2.5 Recommender Systems

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Gütl, & Wagner, 2022; Bobic et al., 2022).

Recommender System (RS) a subclass of information filtering systems that use data on an individual's preferences, interests, and behaviours to suggest items that the user may prefer. They help users in decision-making by providing personalised recommendations. These systems are widely used by online platforms and applications, such as e-commerce websites, music streaming services, and news websites, to suggest products, songs, or articles to users (Schafer et al., 1999; Bobic et al., 2022).

2.5.1 Brief History of Recommender Systems

Recommender systems, also known as recommender systems, have a rich history that has evolved over several decades. These systems play a crucial role in helping users discover content, products, or information that aligns with their preferences and interests. The roots of recommender systems can be traced back to the 1960s when researchers began exploring the idea of automated information retrieval and recommendation. Early work in this field focused on library catalog systems and information retrieval techniques, laying the foundation for future developments. A pivotal moment arrived in the 1990s with the advent of collaborative filtering techniques, including user-based and item-based filtering (Baeza-Yates & Ribeiro-Neto, 1999). Collaborative filtering methods, such as user-based and item-based filtering, gained prominence. In 2006, the

2 Background and Related Work

company Netflix started a series of competitions called the Netflix Prize. It was an open competition for the best collaborative filtering algorithm to predict user ratings for movies, based on previous ratings without any other information about the users or movies. The participants produced algorithms that improved the recommender system by as much as 10% per year. Netflix offered a substantial reward for improving its movie recommendation algorithm, leading to innovations in collaborative filtering and matrix factorisation techniques (J. Zhang et al., 2020). The 2000s saw the emergence of content-based recommender systems, which considered user profiles and item features to make personalised recommendations. Additionally, hybrid recommender systems, combining collaborative filtering and content-based approaches, gained popularity. The 2010s marked the integration of deep learning techniques into recommender systems, enabling the modeling of complex user-item interactions. Neural collaborative filtering and deep recommendation models offered enhanced personalization (Schafer et al., 1999; Herlocker et al., 2004; Pazzani & Billsus, 2007; Capdevila et al., 2016; Bobic et al., 2022).

Some examples of companies building state-of-the-art RS include Netflix, Amazon, and Spotify. Netflix uses a hybrid recommender system that combines collaborative filtering with content-based filtering and other techniques, to provide personalised movie and TV show recommendations to users. Amazon uses a hybrid recommender system that combines collaborative filtering with content-based filtering and other techniques, in order to provide personalised product recommendations to users. Spotify uses a collaborative filtering approach to provide personalised music recommendations to users, based on their listening history and the listening habits of other users (Bobic et al., 2022).

According to Aftab & Ramampiaro (2022), a recommender system is expected to perform three important steps:

- Understand the user's needs, interests, and preferences
- Identify the items that satisfy the user with respect to the user's needs, interests, and preferences
- Rank the suggested items according to the user's preference structure

In general, recommender systems play a crucial role in providing personalised recommendations to users, and the use of different algorithms allows for more precise and varied recommendations. There are three main categories of recommendation algorithms: Collaborative filtering (CF), Content-based filtering (CBF), and Hybrid recommendations (Schafer et al., 1999; Herlocker et al., 2004; Pazzani & Billsus, 2007; Capdevila et al., 2016; Bobic et al., 2022).

2.5.2 Collaborative Filtering

Collaborative filtering bases the prediction of user preferences on similarities to other users by collecting and analyzing large amounts of information about the behaviour, activities, and preferences of these users (Herlocker et al., 2004). This approach uses data about the preferences and behaviours of a group of users to make recommendations to an individual user. Collaborative filtering algorithms identify patterns and trends in the data and use them to make recommendations based on the collective behaviour of the group. CF can be categorised into two main methods as user-based collaborative filtering (memory-based) and item-based collaborative filtering (model-based) (Shi et al., 2014). The main steps of collaborative filtering are as follows:

1. **Data Collection:** The first step involves collecting and aggregating user-item interaction data. This data typically includes user ratings, reviews, purchase history, clicks, or any form of user engagement with items in a system. The data is organised into a user-item matrix, denoted as R , where R_{ui} represents the user's u interaction with item i (Shi et al., 2014).
2. **User or Item Similarity Calculation:** Collaborative filtering depends on the calculation of similarity scores between users or items. User-based collaborative filtering measures how similar two users are in terms of their interactions with items, while item-based collaborative filtering measures the similarity between items based on user interactions. Common similarity metrics include cosine similarity, Pearson correlation, and Jaccard similarity (Hu et al., 2008).
3. **Neighborhood Selection:** For user-based collaborative filtering, for user u , a set of similar users is identified based on the calculated similarity scores. In item-based collaborative filtering, similar items for the items the user has interacted with are chosen. These groups of similar users or similar items are called neighbourhoods. The neighbourhood size can vary, but it typically includes a fixed number of nearest users or items (Bobic et al., 2022).
4. **Recommendation Generation:** To generate recommendations for a target user, for user-based collaborative filtering, the items that the similar users in the neighbourhood have liked or interacted with but are unknown to the target user are considered for recommendation. In item-based collaborative filtering, the items that are similar to those the target user has already shown interest in are considered (Shi et al., 2014).
5. **Ranking and Filtering:** The recommended items are often ranked based on relevance scores, derived from different factors, such as user-item interactions, user reviews, or metadata. A defined set of items that have high ranking are provided to the target uses as recommendations. Additionally, filtering may be applied to remove items the user should not see, such as items they have previously rated or interacted with (Bobic et al., 2022).

2 Background and Related Work

CF experiences difficulties when handling new users or items for which there is limited or no historical interaction data. Since it relies on user-item interactions, it cannot provide meaningful recommendations until provided with feedback data. These systems may struggle to provide diverse recommendations since they tend to recommend items similar to those a user has interacted with, potentially leading to a filter bubble (Bobic et al., 2022; Shi et al., 2014,?).

2.5.3 Content-Based Filtering

Content-based filtering RS analyze the description of items to identify those that are of interest to the user (Pazzani & Billsus, 2007). This approach uses data about the characteristics of items to make recommendations to a user. Content-based filtering algorithms identify items that are similar to items that the user has liked in the past and use these to make recommendations. The main steps of Content-Based Filtering are as follows:

1. **Item Representation:** Each item in the system is represented as a set of descriptive attributes or features. These features are used to characterise the content or properties of the items. For example, in a movie recommender system, item features could include genres (e.g., action, romance), actors, directors, and plot keywords. Item I_j is represented as a feature vector $X_j = (x_{j1}, x_{j2}, \dots, x_{jm})$. Here, x_{ji} represents the value of the i – th feature for item I_j (Renaud-Deputter et al., 2013).
2. **User Profile Creation:** To personalise recommendations, a user profile is created based on their historical interactions or explicitly provided preferences. The user profile is essentially a feature vector that summarises the user’s preferences, derived from the features of items they have interacted with or liked. User U_i has a feature vector representing their preferences: $P_i = (p_{i1}, p_{i2}, \dots, p_{im})$. The user’s feature vector P_i is computed based on the features of items they have engaged with (Pazzani & Billsus, 2007).
3. **Similarity Computation:** The similarity between the user profile and each item’s feature vector is used to produce the similarity score. This similarity score measures how well an item matches the user’s preferences based on shared features. Common similarity measures include cosine similarity, dot product, or other distance metrics. The similarity score between user profile P_i and an item I_j is $sim(P_i, X_j) = \frac{P_i \cdot X_j}{\|P_i\| \|X_j\|}$. Where the \cdot represents the dot product, and $\|P_i\|$ and $\|X_j\|$ are the Euclidean norms of the user profile and item feature vector, respectively (Bobic et al., 2022).
4. **Ranking and Recommendation:** The system ranks items based on their similarity scores with the user profile. Items with higher similarity scores are considered more relevant to the user and are recommended accordingly (Bobic et al., 2022).

5. **Filtering and Presentation:** The recommendation list may undergo additional filtering to remove items that the user should not see, such as items they have previously interacted with. The final ranked list of recommended items is presented to the user (Bobic et al., 2022).

2.5.4 Hybrid Recommender Systems

CBF can provide personalised recommendations even when there is limited user interaction data. However, it relies heavily on the quality and relevance of item features and may not suggest relevant recommendations. Hybrid approaches that combine Content-Based Filtering with Collaborative Filtering or other techniques are often used to enhance recommendation quality (Pazzani & Billsus, 2007; Bobic et al., 2022). Hybrid recommendation algorithm combines collaborative filtering and content-based filtering (Capdevila et al., 2016).

A hybrid system incorporates two or more recommendation techniques, such as Collaborative Filtering (CF), Content-Based Filtering (CBF), and Matrix Factorisation (MF). This can be done using weighted averaging, stacking, or other fusion methods. The fusion process combines the strengths of individual techniques. The hybrid system generates recommendations by combining predictions or scores from multiple recommendation techniques. This step typically involves ranking items based on the combined recommendation scores. The combination of collaborative and content-based approaches in hybrid recommender systems offers improved recommendation quality, robustness, and coverage while adapting to various recommendation scenarios, making them versatile solutions for recommendation tasks (Pazzani & Billsus, 2007; Bobic et al., 2022).

Hybrid recommender systems aim to address several limitations of both CF and CBF by combining their strengths. One such is the mitigation of the cold start problem by using content-based methods to provide recommendations for new users or items. Hybrid systems can introduce diversity by incorporating CBR considering item characteristics, ensuring users receive diverse recommendations. By combining the strengths of CF and CBF, hybrid systems often achieve higher recommendation accuracy and can provide more personalised and relevant suggestions (Pazzani & Billsus, 2007; Bobic et al., 2022).

2.5.5 Main Challenges Of Recommender Systems

Recommender systems have improved how users interact with information, providing personalised recommendations that enhance user experiences. However, these systems confront several complex challenges and ethical considerations. Data sparsity and the cold start problem restrict their ability to provide accurate suggestions when data

2 Background and Related Work

is limited. Scalability is essential for expanding datasets and user bases. Lastly, the collection and utilisation of user data for recommendation purposes raise significant privacy concerns (Guo et al., 2017; Lika et al., 2014; Patel et al., 2017; Milano et al., 2020).

Data Sparsity and the Cold Start Problem

Data sparsity is a common challenge of recommender systems which arises from having a large amount of users and items but no inartistic connection between them. In scenarios where users have rated only a few items, their profiles lack comprehensive information to accurately capture their preferences. This limitation restricts the ability to provide relevant recommendations since many recommender techniques depend on creating user recommendations based on similarity in user profiles. When users have limited interaction history, determining their true preferences becomes challenging, potentially leading to the formation of inaccurate recommendations (Guo et al., 2017).

The cold start problem is the result of data sparsity and in recommendation algorithms refers to the challenge of making recommendations or predictions for new users or items that have little or no prior data associated with them. The sparsity of information available for these new users or items makes it difficult for the system to make accurate recommendations. In the case of new users, the system lacks information about their preferences, interests, and behaviours to provide personalised recommendations. This problem can significantly affect the performance of recommender systems, as the availability of sufficient data is crucial for accurate recommendations. Therefore, finding solutions to the cold start problem is an essential aspect of the research of recommendation algorithms (Lika et al., 2014).

According to Lika et al. (2014); Panteli & Boutsinas (2023), there are several approaches that can address the cold start problem in recommender systems, such as:

- **Pre-populating data** - This approach uses external sources, such as social media profiles or databases, to automatically create initial profiles for new users or items in an information retrieval or recommender system.
- **User or item profiling** - This approach involves using metadata or other available information about new users or items to create initial profiles. By analysing the characteristics or attributes of users or items, such as their age, gender, or content metadata, it can infer their preferences or interests.
- **Collaborative filtering** - This approach utilises the preferences and behaviours of existing users to make recommendations for new users. Collaborative filtering algorithms analyse data patterns and trends to provide recommendations based on the collective behaviour of user groups.

- **Content-based filtering** - This approach utilises content-based filtering algorithms that identify the characteristics of items and recommend new items based on similarity to items previously liked by the user. The algorithm analyses the features or attributes of the item and uses similarity metrics to make recommendations.

Scalability

Scalability is a critical aspect of recommender systems as is defined as the ability to handle growing data volumes efficiently. Recommender systems aim to provide real-time responsive user experiences, making scalability a crucial challenge when creating these systems. Scalability involves designing databases and data structures capable of handling the growing data volume without compromising system performance. Since users expect real-time or near-real-time recommendations across various applications and devices, it is necessary to ensure that recommendation algorithms can generate personalised suggestions quickly. Scalability is not only a matter of processing and handling data but also algorithmic efficiency. Algorithms must be optimised to provide high-quality recommendations within acceptable timeframes, even when dealing with extensive datasets (Patel et al., 2017).

User Privacy and Ethics

User privacy is a major concern in the field of recommender systems, closely intertwined with ethical considerations. Accurate recommendations often require the collection of extensive user data, including demographics and location information. Privacy risks can occur across multiple stages of recommendation systems. Initially, during data collection and sharing, which may happen without explicit user consent. Later, there is the threat of data leaks or de-anonymisation attempts once datasets are stored. Moreover, privacy risks extend beyond data collection; observing generated user recommendations can enable external parties to infer sensitive information, a significant privacy concern (Patel et al., 2017).

These privacy challenges underscore the exposure to risks related to rights violations, such as unfair targeting and manipulative techniques. Ensuring ethical recommender systems requires addressing issues like inappropriate content, data breaches, bias in recommendations, content censorship, and unequal treatment. To create an effective recommendation system, it is necessary to strike a balance between personalisation and user privacy, guided by ethical principles that uphold individual rights while delivering valuable user content (Milano et al., 2020).

2.5.6 Recommender Systems Evaluation Methodologies

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Gütl, & Wagner, 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022; Igor et al., 2023).

RS evaluation is a critical aspect of recommendation algorithms. Evaluating the performance of a recommender system helps determine whether the system provides accurate and relevant recommendations. One of the challenges in evaluating recommender systems is the choice of evaluation metrics. Statistical Accuracy Metrics (SAM) and Decision Support Accuracy Metrics (DSAM) are the two types of metrics commonly used to evaluate the performance of RS (Bobic et al., 2022).

SAM aim to measure the difference between the predicted and actual values of the recommendations. These metrics, such as mean absolute error (MAE) and root mean squared error (RMSE), are suitable for evaluating the predictive accuracy of the recommender system. However, they do not take into account the relevance of recommended items for the user. Therefore, DSAM are used to assess the effectiveness of recommended items based on user preferences. DSAM include recall, precision, and the F1 score. These metrics assess the relevance of recommended items to user preferences and are suitable for a list-wise evaluation approach for recommendations, which considers the order and relevance of the recommended items (Aftab & Ramampiaro, 2022; Hu et al., 2008).

Each evaluation methodology has its own advantages and disadvantages, and the choice of methodology will depend on the goals and requirements of the system. For example, offline evaluation is typically less expensive and faster than online evaluation or user studies but may not provide insight into how the system performs in real-world scenarios. On the other hand, online evaluation or user studies can provide valuable insights into how users interact with the system, but may be more time-consuming and expensive. In conclusion, the appropriate selection of evaluation metrics and methodology is crucial for accurately assessing the performance of a recommender system. By selecting appropriate metrics and methodology, it is possible to gain a comprehensive understanding of the system's performance and identify areas for improvement. This, in turn, can lead to the development of more effective and efficient recommender systems that better meet the needs and preferences of users.

Decision Support Accuracy Metrics

Figure 2.3 illustrates the main elements used for classification evaluation. True positive values are values when the actual and predicted conditions are positive. False positive values are states in which the predicted value is positive, but the actual value is

negative. The true negative value indicates that the actual and predicted conditions are the same and both are negative. The state in which the actual condition is positive, but the prediction is negative, is called false negative values (Igor et al., 2023).

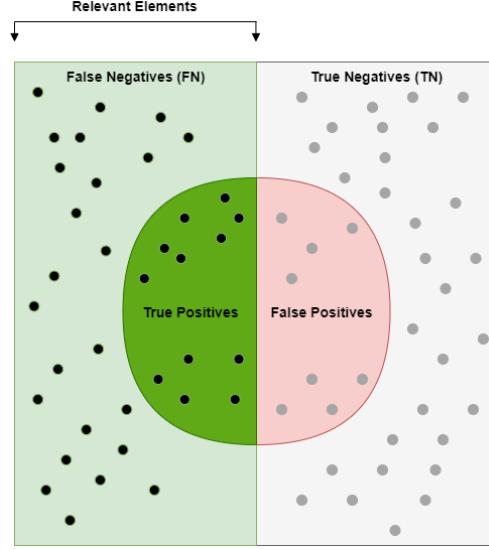


Figure 2.3: Classification Evaluation Representation taken from (Igor et al., 2023)

The actual positive values (AP), as seen in Equation 2.1, refer to the number of true positives (TP) together with the number of false positives (FP).

$$AP = TP + FP \quad (2.1)$$

The actual negative values (AN), as seen in equation 2.2, refer to the number of false positives (FP) together with the number of true negatives (TN).

$$AN = FP + TN \quad (2.2)$$

Predicted positive values (PP), as seen in equation 2.3, refer to the number of true positives together with the number of false negatives.

$$PP = TP + FN \quad (2.3)$$

The predicted negative values (PN), as seen in equation 2.4, refer to the number of false negatives along with the number of true negatives.

$$PN = FN + TN \quad (2.4)$$

2 Background and Related Work

Sensitivity describes the ratio of correct predictions to all actual positive conditions and is calculated as shown in equation 2.5.

$$sensitivity = \frac{TP}{AP} \quad (2.5)$$

The specificity describes the ratio of correct rejections to all actual negative conditions and is calculated as shown in equation 2.6.

$$specificity = \frac{TN}{AN} \quad (2.6)$$

Precision describes the ratio of correct predictions from all positively predicted values and is calculated as shown in equation 2.7.

$$precision = \frac{TP}{PP} \quad (2.7)$$

According to Ye et al. (2012), the f score is a measure of the accuracy of the prediction. It is the harmonic mean between precision and sensitivity and is calculated as shown in Equation 2.8.

$$f_{score} = 2 \frac{precision * sensitivity}{precision + sensitivity} \quad (2.8)$$

Normalized Discounted Cumulative Gain

Normalized Discounted Cumulative Gain (NDCG) is a commonly used metric to evaluate the effectiveness of RS. Although many evaluation methods focus on measuring the difference between predicted and actual values, NDCG captures the relevance and impact of the order of placement of the recommended items. It is particularly important, as research has shown that the position of individual recommendations can have a significant influence on user decision making processes (Pan et al., 2007).

NDCG is calculated based on the cumulative gain and the discounted cumulative gain, which is then normalised to produce a score between zero and one. Cumulative gain measures the relevance of the recommended items, with higher relevance resulting in higher scores. The discounted cumulative gain takes into account the position of the recommendation in the list, giving higher weights to items that appear higher up in the list. Normalising the score ensures that it is comparable across different lists with varying numbers of items (Pan et al., 2007).

Receiver Operating Characteristics Curve

Receiver Operating Characteristics (ROC) curve is a graphical representation of the performance of a binary classifier system, such as a recommender system or a diagnostic tool, over a range of decision thresholds. The curve plots the true positive rate

(sensitivity) against the false positive rate (1-specificity) for different decision thresholds. ROC curve analysis is a widely used method for evaluating the performance of binary classifiers. The curve provides a visual representation of the trade-off between sensitivity and specificity for different thresholds, allowing the selection of a threshold that balances the two (Obuchowski, 2003; Park et al., 2004).

Area Under the Curve

The area under the ROC curve (AUC) is often used as a summary statistic to measure the overall performance of the classifier. The AUC provides a measure of the classifier's ability to correctly distinguish between positive and negative classes across a range of cutoff values. It takes into account the trade-off between sensitivity and specificity, which are two critical measures of the classifier's performance. Sensitivity is defined in Equation 2.5, while specificity is defined by the Equation 2.6. AUC ranges from 0 to 1, with a value of 0.5 indicating random chance and a value of 1 indicating perfect classification. A higher AUC value indicates that the classifier is better able to distinguish between positive and negative classes. Therefore, the AUC is a widely used performance metric for classifiers and is particularly useful in applications where the cost of false positives and false negatives is different (Obuchowski, 2003; Park et al., 2004).

Statistical Accuracy Metrics

The mean absolute error (MAE) and the root mean squared error (RMSE) are commonly used statistical accuracy metrics to evaluate the performance of predictive models. Both metrics quantify the degree of deviation between predicted and actual values. The MAE represents the average absolute difference between the predicted and actual values, while the RMSE represents the square root of the average squared differences between the predicted and actual values. These metrics are widely used in various fields, including machine learning, economics, and engineering, to evaluate the accuracy of predictive models and to compare the performance of different models (Chai & Draxler, 2014).

Mean Absolute Error

The MAE is calculated from the average of the sum of the absolute amounts of the individual differences between the observed and predicted values. MAE gives a basic overview of the error behaviour of the model and does not weight statistical outliers higher. It is therefore not possible to say how serious the errors found are, but it is easier to interpret as a result of values that are more in context of the data (Chai & Draxler, 2014).

2 Background and Related Work

$$MAE = \sum_{i=1}^D |x_i - y_i| \quad (2.9)$$

Root Mean Square Error

RMSE is calculated from the average of the sum of the squared differences between the observed and predicted values. In addition, the square root of this result is formed. In contrast to MAE, RMSE is much more sensitive to outliers. If an outlier has a stronger influence on the desired result, it is advisable to use the RMSE. Another advantage of RMSE is that, unlike MAE, the absolute value is not formed in the calculation. This facilitates further mathematical methods such as the use of cost functions or the calculation of gradients, i.e. a derivation of a function with more than one input variable (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{d_i - f_i}{\sigma_i} \right)^2} \quad (2.10)$$

Statistical Accuracy Metrics Versus Decision Support Accuracy Metrics

Statistical accuracy metrics and decision support accuracy metrics are two types of evaluation metrics used in different contexts. Statistical accuracy metrics, such as MAE and RMSE, are widely used to measure the accuracy of models that make predictions or forecasts. In contrast, decision support accuracy metrics evaluate the effectiveness of recommender systems. These metrics assess how well the system supports decision making, rather than how accurate the predictions or forecasts are. Both statistical accuracy metrics and decision support accuracy metrics are important to evaluate the performance of different types of models and systems. However, it is crucial to choose the appropriate metrics based on the specific context and goals of the model or system evaluated (Obuchowski, 2003; Park et al., 2004; Igor et al., 2023).

2.6 Open Science Principles

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported

by and taken from the work published in the following publication (Jakovljevic, Gütl, & Wagner, 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022; Igor et al., 2023; Jakovljevic, Russmann, et al., 2022; Jakovljevic et al., 2020).

The concept of open data and open information has a long history in the scientific community, dating back to the Enlightenment era, when philosophers and scholars began advocating for the free exchange of ideas and knowledge. In the twentieth century, the open science movement emerged as a response to the increasing specialisation and fragmentation of scientific knowledge and called for greater collaboration and transparency in the scientific enterprise (Ramachandran et al., 2020; Jakovljevic et al., 2020). Open data, open information, and open science principles are fundamental components of a transparent and reproducible scientific research ecosystem. Open data refers to making research data openly available to the public, to improve accessibility, collaboration, reuse of scientific findings, and foster reproducibility (Ramachandran et al., 2020).

Reproducibility can be defined as the ability to replicate a model that produces the same result as the original model given the same input data (Sugimura & Hartl, 2018). It is also essential to promote open and accessible research, the use of robust experimental workflows, and to allow researchers to quickly convert ideas into practise while reducing unintentional errors (Pineau et al., 2020). Similarly to the principles of reproducibility, a movement to conduct science transparently by making code, data, scientific communications, and any other research artifact publicly available and easily accessible over the long term is called Open Science (Antony & Salian, 2021). Open science is a broader concept that includes open data and other practises, including open access publishing, open peer review, and open educational resources. By promoting the sharing of research results, open science principles aim to foster greater transparency and trust in the scientific process, and ultimately accelerate scientific discovery and innovation (Ramachandran et al., 2020). Furthermore, open science principles encourage the sharing of scientific knowledge and resources, fostering the development of new research ideas and collaborations. Open science also promotes the use of preregistration, which involves publicly registering a research plan before data collection, to reduce the likelihood of data manipulation and increase transparency. In addition, open science aims to promote diversity, equity, and inclusion in research by making scientific knowledge more accessible to marginalized communities and promoting the participation of diverse voices in research. In general, the adoption of open science principles can improve the quality and impact of scientific research, while also promoting greater public trust and participation in science (Vrouwenvelder & Stall, 2023).

In recent years, the open data movement has gained further momentum with the rise of open data initiatives and platforms, such as the Open Science Data Cloud, the European Open Science Cloud, and the Research Data Alliance. These initiatives aim to

2 Background and Related Work

promote the availability and accessibility of scientific data and to encourage the use of open data for research and innovation. Many scientific journals and funding agencies now require researchers to make their data and methods openly available to facilitate review and replication of their work. In general, open data and open information have played a key role in promoting transparency, collaboration, and progress in the scientific community.

2.6.1 Open Data and Open Information

Open data is the term used to describe freely available data that anyone can use for analysis and research (Antony & Salian, 2021). Open information refers to the idea that certain information should be freely available to the public, without restrictions on access or usage. This concept is often linked to the principles of transparency, accountability, and democracy. There have been different initiatives for collaboration based on open data and open information, such as the previously mentioned Netflix Prize, OpenStreetMap, the CERN Open Science Initiative, Open City Initiatives, and more. All of these collaboration projects faced a common issue. Sharing of data that contain identifies, quasi-identifies, and sensitive attributes. In addition to these issues, political factors, such as structures, regulations, and ways of working, become challenges to data sharing even within an organisation (Runeson et al., 2021; Antony & Salian, 2021).

Open information refers to the idea that certain information should be freely available to the public, without restrictions on access or usage. This concept is often linked to the principles of transparency, accountability, and democracy. Open search refers to the ability to search and access information from a wide range of sources, without limitations or bias (Jakovljevic et al., 2020). An important event in the history of open data in science was the creation of the Open Data Protocol (OData) in 2007, which allowed data to be shared and accessed over the Internet in a standardised way. The Open Data License (ODL) was introduced in 2010, providing a legal framework to make data openly accessible (Staunton et al., 2021).

2.6.2 Open Innovation

Open innovation emphasises the importance of combining external knowledge and expertise in an innovation process. This approach is different from traditional closed innovation, which relies solely on internal research and development efforts. The open innovation paradigm is often associated with collaborations with external parties, such as customers or academic institutions. By leveraging external resources, organisations can develop new products, processes, and business models, increasing competitiveness and growth. Furthermore, the combination of open information, open search, and

open innovation can contribute to the creation of an open and inclusive society, where knowledge and ideas can be shared and built upon for the benefit of all (Jakovljevic et al., 2020; Jakovljevic, Russmann, et al., 2022).

Institutional knowledge is an amalgam of experiences, processes, data, skills, values, and information from a business employee. It defines a company's history and can contain important trends, projects, and perspectives that define a company's history and can include crucial trends and perspectives. Proper documentation and sharing of institutional knowledge throughout the organisation ensure that the organisation can address different problems and daily routines (*Institutional knowledge: What it is & how to use it*, 2022). However, if a problem cannot be solved using the organisation's current routines, the organisation is induced to innovate through the development of new knowledge. Organisations are looking for alternative solutions to problems when existing procedures and processes do not produce results that align with organisational goals (Huizingh, 2011a). The drive to use external sources of information for innovation and the extension of knowledge drive organisations to be open and distribute information to facilitate innovation (Lopez-Vega et al., 2016a; Helfat, 2006). Open innovation can also be defined as the use of purposeful input and output of knowledge to stimulate internal innovation and increase the demand for external use of innovation, respectively (West et al., 2014). An important idea in open innovation is that organisations should utilise the search for information outside of their organisation. The search for external knowledge is quite complicated and challenging, involving difficulties such as tacitness, complexity, rivalry, and indivisibility of knowledge, which may not be helpful for its discovery and transfer (March, 2008).

Overall studies indicated that the greater the search for knowledge, the higher the level of innovation in the organisation. This has led to the conclusion that the development of information openness can stimulate innovative activities, the creation of innovative approaches, and increased performance (Cruz-González et al., 2014; Lopez-Vega et al., 2016b). Previous studies have been inclined to demonstrate the benefits of openness in organisations. Recently, studies have begun to stress the downsides of openness, which would justify why not many organisations have adopted the concept of openness (Huizingh, 2011b; Jakovljevic et al., 2020). One of the main disadvantages of openness in an organisation is the risk of losing competitive advantage and leaking private information (Isfandyari-Moghaddam, 2015).

2.6.3 Importance of Privacy in Open Data

Privacy is a fundamental human right protected by laws and regulations and refers to an individual's ability to control the collection, use, and disclosure of their personal information. In the context of open data, privacy is important because the release of personal information can lead to a range of negative consequences, such as identity

2 Background and Related Work

theft, discrimination, and stigma. Additionally, privacy breaches can undermine trust in open data initiatives and discourage individuals and organisations from sharing their data in the future. Therefore, it is crucial to ensure that open data are released in a way that protects the privacy rights of individuals (Huston et al., 2019).

When distributing sensitive information, it is important to follow guidelines, which can be given by countries, public entities, or even organisations themselves. Most of these guidelines require a level of privacy and anonymisation for sensitive data (Personal Data Protection Commission Singapore, 2018; Antony & Salian, 2021; Van Schalkwyk & Verhulst, 2017). Privacy-preserving and anonymisation methods use some form of data transformation. Naturally, such methods reduce the granularity of the representation and remove information. This results in a loss of effectiveness for data management, data processing methods and algorithms created from these data (Barbaro & Jr., 2006).

2.6.4 Open Data Initiatives

Open Data Repositories (ODR) are structures, whether academic or non-academic, that host data and allow free access to them. Examples of open repositories are Zenodo, arXiv, CiteSeerX, UK Data Archive and Figshare (Costa et al., 2021).

Open Data Ecosystems (ODEs) are an emerging concept for data sharing under public licences in software ecosystems. A study by (Runeson et al., 2021), interviewed 27 participants from 22 different private companies and public authorities on conceptual ideas about ODE. Their qualitative analysis of data and interviews concluded that the value of ODE lies in the data they produce and in the collaboration around the data. Furthermore, they concluded that identifying data (e.g. identifiers and quasi-identifiers) is challenging from a legal point of view, and liability issues are also unclear. Trust in the data and governance of an ODE is also a challenge.

As seen in the previous sections, there is a need for organisations to share data and/or make them publicly available. To correctly open internal organisational data, it is necessary to assess potential risks, evaluate if the data contain sensitive information, determine to which ODR to distribute the data, evaluate which licence to use for data sharing, and more.

2.7 Related Work

This section provides an overview of the research on privacy in machine learning, with a specific focus on privacy challenges and solutions in recommender systems and community detection. Additionally, this section also highlights the importance of open science, open data, and open information in advancing the field of privacy-preserving machine learning.

2.7.1 Overview of Existing Approaches for Privacy-Aware Machine Learning in Large organisations

With the increasing demand for data-driven insights, organisations face the challenge of leveraging sensitive data while ensuring compliance with privacy regulations and protecting individual privacy. Privacy-preserving machine learning has become popular due to increasing concerns over data privacy. These approaches offer solutions to balance the need for data analysis with the preservation of privacy (Q. Zhang et al., 2021). An overview of existing approaches and frameworks for the integration of privacy into machine learning within the context of large organisations is presented in the following sub sections.

Privacy-Preserving Data Collection and Storage

Secure and ethical data handling represents the foundation for privacy-aware machine learning. One approach to ensure that the data within organisations is kept decentralised and that only model updates are shared instead of raw data is federated learning. Additionally, privacy-centric storage solutions, including secure multi-party computation and differential privacy, ensure protected storage and retrieval of sensitive data (Abadi et al., 2016).

Depending on the type of data, different methods for anonymization can be applied (Pramanik et al., 2021; Ergüner Özkoç, 2021). Past research indicates that the most used methods for anonymisation of datasets are:

- **Randomization Methods** - add noise to data to conceal the attribute values of records. The added noise is large enough so that individual records cannot be recovered (Ergüner Özkoç, 2021).
- **Cryptographic Approaches** - are based on applying a cryptographic function over data that is presented in raw format. This raw data is also called plaintext. Applying a cryptographic function to plaintext produces ciphertext. It is hard to reproduce the original raw data, from the ciphertext. This is why it is used for anonymisation of identifying data (e.g., names, addresses, etc.) (Sousa et al., 2021).
- **k-anonymity** - follows the idea that the release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals. Let $T(A_1, \dots, A_n)$ be a table and Q_T be the quasi-identifiers associated with it. T is said to satisfy k-anonymity if and only if for each quasi-identifier $Q \in Q_T$ each sequence of values in $T[Q]$ appears at least with k occurrences in $T[Q]$ (Samarati & Sweeney, 1998).
- **l-diversity** - is an improvement to k-anonymity and aims to mitigate possible defects of k-anonymity like homogeneity and background Knowledge attacks. In

2 Background and Related Work

homogeneity attacks, all the values for a sensitive attribute within a group of k records are the same. Even if the data is k -anonymized, the value of the sensitive attribute for that group of k records can be predicted exactly (Aggarwal & Yu, 2008)..

- **t-closeness** - While k -anonymity protects against identity disclosure, it does not protect in general against disclosure of a sensitive attribute corresponding to an external identified individual. t -Closeness is another extension of k -anonymity which tries to solve this issue. t -closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of a sensitive attribute in the overall dataset (N. Li et al., 2007).
- **Partition Based Privacy** - For an aggregate function $f : D \rightarrow R$, a dataset D with n records of n individual users, and a privacy preference $\phi = (\epsilon_1, \dots, \epsilon_n)$ ($\epsilon_1 \leq \dots \leq \epsilon_n$), where ϵ is the privacy parameter. Let $\text{Partition}(D, \phi, k)$ be a procedure that partitions the original dataset D into k partitions $(D_1 \dots D_k)$. The partitioning mechanism is defined as $PM = B \left(DP_{\epsilon_1}^f(D_1), \dots, DP_{\epsilon_k}^f(D_k) \right)$ where $DP_{\epsilon_i}^f$ is any target ϵ_i -differentially private aggregate mechanism for f , B is an ensemble algorithm (N. Li et al., 2007)

Encrypted Machine Learning and Privacy-Preserving Machine Learning Algorithms

The need to provide personalised and evolving artificial intelligence (AI) services, such as voice assistant, word suggestion, facial recognition, and smart video feeds, has increased the amount of data generated. In most of these applications, machine learning models are refined by continuously feeding new user data (as features) and their feedback (as labels). However, these data, such as type history, web access logs, and frequently visited locations, often include sensitive and private information (H. Zheng et al., 2020). Due to these risks, the initial work on mitigating privacy risks during the machine learning process focusses on privacy challenges and risks associated with the ML models. Commonly used methods for protecting sensitive data in ML processes are Local Differential Privacy and Federated Machine Learning (Liu et al., 2021; H. Zheng et al., 2020).

Privacy-aware machine learning often depends on encryption methods that are applied to raw data that allows executing computations on encrypted data without exposing sensitive information. This enables organisations to outsource data analysis while at the same time preserving data confidentiality (Yagisawa, 2015).

Machine learning models themselves can be designed with privacy preservation in mind. Private aggregation of teacher ensembles (PATE) and federated learning enable model training without revealing individual-level data, making them suitable for large-scale privacy-aware applications. Federated learning enables on-device training over distributed networks consisting of a large number of devices. It can ensure privacy

by keeping the data localised and performs model training and aggregation on client devices or edge servers (Yao et al., 2019). Privacy-aware machine learning in large organisations requires techniques that can preserve the privacy of sensitive data while still allowing for effective machine learning. Differential privacy, federated learning, homomorphic encryption, and secure multiparty computation are the most used techniques for privacy-aware machine learning in large organisations (Jain et al., 2018; Yao et al., 2019).

Differential privacy is a widely recognised privacy-preserving framework that aims to provide rigorous privacy guarantees. It involves injecting carefully calibrated noise into data or model updates to prevent the disclosure of sensitive information (Jain et al., 2018). Differential Privacy can be understood as a randomised function k which is applied to document collections or query results before their public release. According to Sousa et al. (2021), for all subsets S in the range of k , and a document collections D and D' differing on at most one element, k provides ϵ -differential privacy if:

$$Pr[k(D) \in S] \leq \exp(\epsilon) Pr[k(D') \in S]$$

Ethical Machine Learning and Regulatory Compliance

In the area of privacy-aware machine learning for large organisations, ethical considerations and following regulatory frameworks and instructions play crucial roles. These aspects are important for building trust with users and ensuring responsible machine learning practises. Compliance with regulations like the General Data Protection Regulation (GDPR) remains essential for large organisations handling user data.

Handling bias and discrimination in machine learning models is a major ethical problem. Fairness-aware machine learning techniques mitigate biases that result from biased training data or algorithmic decisions. These methods ensure equitable outcomes for all user groups, regardless of their demographic characteristics. Additionally, techniques like adversarial debiasing, reweighted loss functions, and disparate impact analysis are actively used to identify and correct bias (Zafar et al., 2017).

Model transparency and interpretability are essential for the ethical aspect of machine learning. XAI methods provide insights into how machine learning models make their decisions. Approaches such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) offer interpretability by highlighting the key features that influence a model's output. By making these decisions interpretable, organisations can enhance trust and accountability in their machine learning systems (Ribeiro et al., 2016).

In summary, ethical machine learning practises and adherence to regulations are integral components of privacy-aware machine learning in large organisations. By employing fairness-aware techniques, implementing XAI methods, and ensuring compliance with data protection regulations like GDPR, organisations can foster trust,

2 Background and Related Work

transparency, and ethical responsibility in their machine learning endeavors while safeguarding user privacy.

2.7.2 Privacy in Recommender Systems

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Gütl, & Wagner, 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022; Igor et al., 2023).

The following sections are based on, supported by, and taken from the work published in the following publications:

- **Jakovljevic Igor**, Gütl Christian and Wagner Andreas - Towards a Privacy-Aware Reproducible Machine Learning Pipeline for Open Data
- **Jakovljevic Igor**, Gütl Christian, Wagner Andreas and Nussbaumer Alexander - Compiling Open Datasets in Context of Large Organizations while Protecting User Privacy and Guaranteeing Plausible Deniability.

Privacy has become a crucial concern in our digital age, where vast amounts of personal data are generated and collected by various entities. Privacy-based information retrieval and recommender systems aim to address these concerns by minimising the collection and use of personal data, while still providing the benefits of personalised recommendations (Jeckmans et al., 2013; Himeur et al., 2022).

The key aspects of privacy in recommender systems are outlined and described below:

- **Data Collection and Storage:** Recommender systems collect user data, including browsing history and preferences, which raises concerns about data privacy. It is essential to strike a balance between collecting sufficient data for effective recommendations and respecting user privacy (Jeckmans et al., 2013; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).
- **User Profiling:** User data is used to create detailed profiles, potentially compromising user privacy if not handled properly. Implement strict data access controls and consider user consent for detailed profiling (Himeur et al., 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).
- **Data Anonymisation:** Anonymising user data is crucial for privacy protection, but it must be done effectively without degrading recommendation quality. Research and employ advanced anonymization techniques such as differential privacy (Jeckmans et al., 2013; Himeur et al., 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).

- **Personalised vs. Privacy-Preserving Recommendations:** Personalisation is a core feature, but it must coexist with privacy preservation. Explore privacy-preserving recommendation algorithms that balance personalisation and privacy (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).
- **User Consent:** Obtaining user consent for data collection and processing is fundamental. Clearly communicate data usage policies and provide users with the option to opt-in or opt-out (Barocas et al., 2017).
- **Third-Party Data Sharing:** Sharing user data with third parties requires safeguards to prevent misuse. Ensure that user data shared with third parties is protected and used responsibly (Barocas et al., 2017).
- **Algorithmic Transparency:** Users want to understand how recommendations are generated. Promote algorithmic transparency to build user trust and address concerns about data usage (Grimmelmann, 2011).
- **Filter Bubbles and Echo Chambers:** Recommender systems can unintentionally create filter bubbles, limiting diverse perspectives. Implement algorithms that balance user preferences with exposure to diverse content (Grimmelmann, 2011).
- **Regulatory Compliance:** Compliance with privacy regulations is essential. Stay updated with privacy laws like GDPR or CCPA and ensure that your system adheres to them (Barocas et al., 2017).
- **User Control:** Providing users with control over their data is crucial. Enable users to review, modify, or delete their collected data (Jeckmans et al., 2013).

There are several approaches that can be used to design privacy-based information retrieval and recommender systems, anonymisation, pseudonymization, differential privacy, and user-controlled privacy. Anonymisation is a popular approach that involves removing or masking personal identifiers from the data used for recommendation or personalization. This method prevents personal data from being linked to specific individuals, thus safeguarding their privacy (Jeckmans et al., 2013; Himeur et al., 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022). Pseudonymization, on the other hand, replaces personal identifiers with substitute identifiers or pseudonyms that can be used to track the data of specific users without revealing their identities. This approach allows for analysis and personalization while maintaining user privacy (Jeckmans et al., 2013; Himeur et al., 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022). Finally, user-controlled privacy is an approach that grants users autonomy over their personal data by allowing them to control the collection and use of their data for recommendation or personalization. This can be achieved through transparent communication of data collection processes and options for opting out of specific data processing or collection activities (Jeckmans et al., 2013; Himeur et al., 2022).

These privacy-based information retrieval and recommender systems provide a significant benefit to society by protecting individual privacy, while still allowing personalization and recommendation. They are essential to building trust between

2 Background and Related Work

individuals and organisations, thereby enabling a more transparent and secure digital ecosystem.

2.7.3 Privacy in Community Detection

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Pobaschnig, et al., 2022).

In the field of community detection, privacy is a topic that has not received sufficient attention, as noted by Javed et al. (n.d.) in their paper on Open Issues and Future Trends. Unfortunately, data breaches continue to expose large amounts of personal information. For example, in April 2021, a Facebook data breach resulted in the leakage of private data, including the names, dates of birth, and locations of more than 500 million users (Abrams, n.d.). Similarly, in the case of LinkedIn, data from more than 500 million users was scraped and offered for sale (Canales, n.d.). These incidents illustrate how even large corporations can easily lose private data. Users can only limit the amount of data they provide to a service to reduce the risk of data leakage. However, even if users are cautious about what they share, specific attributes or behaviours can still be inferred by their friends. In fact, using publicly available information, such as liked pages on social networks, it is possible to correctly infer the city in which a user lives for 57% of all users in a real-world data set, increasing to more than 90% when using confidence estimation (Gong & Liu, n.d.). Although there are mitigations, such as the framework proposed by Salamatian et al. (n.d.), which gives users advice on how to distort their public data to prevent inference attacks, these approaches have their own limitations, such as scalability issues. However, researchers have proposed several techniques that can be used to improve privacy in community detection, such as differential privacy and k-anonymity.

Privacy-preserving analytics refers to a set of methods aimed at collecting, measuring, and analysing data in a way that respects the privacy rights of individuals. These techniques enable data-driven decision making while giving individuals control over their personal data. It is important to note that restricting access to data can hinder support for different types of data analysis. However, by adopting approaches that limit information in the data, such as removing personal identifiers and sensitive information that can lead to individual identification, the data can be made available without compromising privacy. In recent years, there has been increasing interest in developing privacy-preserving data mining algorithms to ensure the confidentiality of data while still enabling data analysis (Jakovljevic, Pobaschnig, et al., 2022).

2.7.4 Importance of Open Science, Open Data, and Open Information

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic, Pobaschnig, et al., 2022; Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022; Jakovljevic, Russmann, et al., 2022; Jakovljevic et al., 2020).

Open data, open information, and open innovation are crucial to the advancement of society in numerous ways. They promote transparency and accountability, allow equal access to information and resources, and encourage diversity and inclusivity. Moreover, they foster innovation, serve as a catalyst for progress, and contribute to economic growth.

One significant advantage of open data and open information is that they provide individuals and organisations with access to vast amounts of information that may not be available through traditional channels. This information can be used to develop new products, services, and insights that benefit society. Open innovation also plays a critical role in promoting diversity by encouraging the participation of external parties with unique perspectives and expertise. By creating an inclusive environment, open innovation fosters the creation of new products, processes, and business models.

Furthermore, open data, open information, and open innovation facilitate the expansion of knowledge, which serves as a catalyst for innovation and progress. They allow individuals and organisations to build on existing knowledge and expand their horizons, leading to the development of new ideas and solutions. Finally, they contribute to economic growth by enabling the creation of new products and services and increasing the efficiency and productivity of organisations. By sharing knowledge and ideas, businesses can work more effectively and efficiently, leading to increased profitability and growth. In general, open data, open information, and open innovation are necessary for the creation of an open and inclusive society, where knowledge and ideas can be shared and built for the benefit of all.

2.8 Summary

The second chapter of this research discusses various background and related work on information generation in large organisations. It covers user profiles, information retrieval and navigation, machine learning, recommender systems, and open science principles.

The first topic is information generation in large organisations. This section discusses the challenges and importance of information generation in large organisations. Fur-

2 Background and Related Work

thermore, the concept of user profiling is explained, and the section describes how users' information is collected and the methods used for user profiling. In addition, challenges related to user profiling, including filter bubbles, cognitive and search bias, and privacy concerns, are discussed. Furthermore, the application of machine learning to information retrieval and user profiling is examined, including a brief history of machine learning and the primary challenges associated with the field. Popular machine learning approaches, such as collaborative and content-based filtering, are examined, along with their drawbacks and benefits. Moreover, the section delves into recommender systems, which employ machine learning algorithms to provide personalised recommendations to users. Collaborative filtering, content-based filtering, and hybrid systems, as well as challenges related to the cold start problem and the evaluation methodologies of recommender systems, are discussed. Finally, the principles of open science are explored that promote transparency and collaboration in research. The importance of open data and open innovation in research is discussed, along with the privacy concerns associated with open data and initiatives that address these concerns.

The issue of privacy in machine learning and recommender systems has become a crucial concern in the current digital age. Machine learning models rely heavily on user data such as web access logs, type history, and frequently visited locations to provide personalised AI services. The increasing use of ML in personalised AI services, such as facial recognition and voice assistant, generates and uses considerable amounts of sensitive and private data. These data often contain sensitive and private information. The initial work on mitigating privacy risks during the ML process focusses on privacy challenges and risks associated with ML models.

In addition, privacy-preserving IR and RS exist and aim to handle privacy concerns by minimising the collection and use of personal data while still providing personalised recommendations. Several approaches, such as anonymisation, pseudonymisation, differential privacy, and user-controlled privacy, are used. These techniques could allow researchers and developers to access and analyse sensitive data without compromising the privacy of individuals or organisations. They are essential to building trust between individuals and organisations, enabling a more transparent and secure digital ecosystem. For these methods to be useful, it is essential to determine how important privacy and privacy-preserving concepts are for users or organisations.

The field of community detection has not paid enough attention to privacy and this has led to large amounts of personal information being exposed by data breaches. Although users can limit the data they provide to services to reduce the risk of data leakage, even if they are cautious, attributes or behaviours such as city or age can be inferred from their contacts (e.g., friends, work associates, neighbours). To improve privacy in community detection, researchers have proposed various techniques, such as differential privacy, k anonymity, and privacy-preserving analytics. These techniques aim to protect individual privacy while still allowing the measurement and analysis of

data in a way that respects the privacy rights of individuals.

Privacy concerns associated with ML and RS have led to the need for the development of novel systems proposals that respect user privacy. These systems are based on open data and principles to instil confidence in users and promote transparency. The main questions for building these systems are how sensitive information can be used in a privacy-preserving way for personalisation and how well they compare to traditional IR and RS systems.

Such systems should show how user data are used and focus on using anonymous data and small amounts of user data. In general, privacy should be a crucial consideration when designing any AI, ML, or data-driven system. Such systems would help build trust between individuals and organisations and enable a more transparent and secure digital ecosystem.

Overall, this chapter aims to provide a comprehensive overview of the different fields and concepts related to information generation and retrieval, highlighting the challenges and opportunities associated with each.

3 Analysis Of User Behavior

In the age of information systems, user behaviour analysis has become increasingly important for organisations to understand their users' preferences and improve their overall experience. The aim of this chapter is the analysis of user behaviour at CERN, one of the largest and most complex scientific organisations in the world. Specifically, we examine three different user studies conducted at CERN, including the CERN Newcomer Analysis, the CERN IT Department User Survey Analysis, and the CERN User Information Consumption Analysis. The CERN User Information Consumption Analysis study was created as a central part of this chapter. This study was designed to collect information related to the patterns of information consumption among CERN users. In the final section of this chapter, we discuss the findings and insights gained from these three studies, including commonalities and differences between user behaviour, preferences, and work habits at CERN. Through these studies, our primary objective is to deepen the understanding of user behaviour in large organisations, with a strong emphasis on gaining insights that will not only inform future strategies for enhancing user experience and satisfaction but also refine the methodologies and approaches used in the development of information and recommendation systems, ultimately driving significant improvements in their effectiveness and user-centric design.

The following sections are based on, supported by, and taken from the work published in the following publications:

- **Jakovljevic, I., Gütl, C., & Wagner, A. (2023).** Exploring Information Consumption Patterns Among Users in Large Organisations: A Survey Analysis of CERN Users. MDPI - Multidisciplinary Digital Publishing Institute, [In Review]
- **Jakovljevic, I., Gütl, C., Wagner, A., & Nussbaumer, A. (2022).** Compiling Open Datasets in Context of Large Organisations while Protecting User Privacy and Guaranteeing Plausible Deniability. In Proceedings of the 11th International Conference on Data Science, Technology and Applications - DATA SciTePress - Science and Technology Publications

3.1 Contribution

This chapter is the initial step of the research's "Data Gathering and Data Analysis" phase and addresses the research question: "How do users behave and consume

3 Analysis Of User Behavior

information in large, highly connected organisations?” It focusses on analysing user behaviour at CERN and the primary contribution of this chapter provides insight into the needs and preferences of users in large organisations, especially regarding information consumption. By analysing two previous CERN surveys, the newcomer behaviour and IT department survey, valuable insights related to changes in user preferences over time are gathered. Additionally, conducting and reporting on a large-scale study that focusses on extracting insights into the ways users consume information is part of this chapter. The significance of this chapter extends beyond CERN, since it offers practical insights into improving user experience and engagement in similar organisations.

3.2 CERN IT Department User Survey Analysis

The IT Department User Survey Analysis is an investigation into current IT department users and their work habits, including the tools they use for work and their preferred methodologies. This analysis is critical to improving the IT department’s overall effectiveness in supporting the needs of its users. The user community at CERN is very diverse, consisting of more than 30,000 external and internal computer users from more than 15 departments, with different work habits and methods of working. The main concerns of the survey are how users collaborate, which devices and software applications they use devices (Jones, 2017).

The survey began by collecting data on several aspects of CERN computer users, including their use cases, working preferences, computing choices, software usage, and preferred IT-supported services. The objective was not to evaluate new tools or recommend any particular hardware or software but to provide reports to help the IT department improve its services and recommend certain tools for specific scenarios. The survey was intended for all CERN computer users, including external users, and all departments and experiments were invited to take part. To gather data, a combination of interviews, automated data gathering, and an online survey in English and French were used. The survey covered several areas, including user computing environments, hardware, work locations, communication preferences, collaboration preferences, coordination preferences, software development usage, and discovering information. User profile questions were also asked, such as the user’s work type, age, and organisational unit, which helped with finer correlations. To make the survey more user-friendly and to encourage participation, the researchers refined the initial survey, which was too large and could have put off many participants. As a result, they published a 94-question survey that could be completed by users within 5 to 10 minutes. In October 2016, a survey was launched to gather input from computer users at CERN. By the end of December, over 1100 users had already participated. However, upon analysis, it was discovered that some departments had low participation rates.

3.2 CERN IT Department User Survey Analysis

In February 2017, a second push was made to encourage more responses. Finally, the survey was closed in March, with a total of 1233 completed responses received (Jones, 2017).

Based on the survey results, it was found that out of the 1233 participants, 1216 disclosed their age range. As seen in Figure 3.1, the majority of participants were between the ages of 31 to 40, with 356 participants (29.3%) falling into this age range, followed by 303 participants (24.9%) between the ages of 18 to 30. There were 264 participants (21.7%) between the ages of 41 to 50, 204 participants (16.8%) between the ages of 51 to 60, 67 participants (5.5%) between the ages of 61 to 70, and 22 participants (1.8%) over the age of 70. It was also observed that 17 participants chose not to disclose their age and remained anonymous.

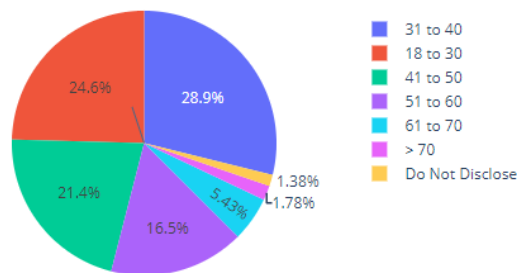


Figure 3.1: What is your age range?

Furthermore, the survey indicated that out of the 1233 participants, 712 participants (57.7%) agreed to be contacted for further information, while 520 participants (42.3%) did not consent to further contact, as described in Figure 3.2.

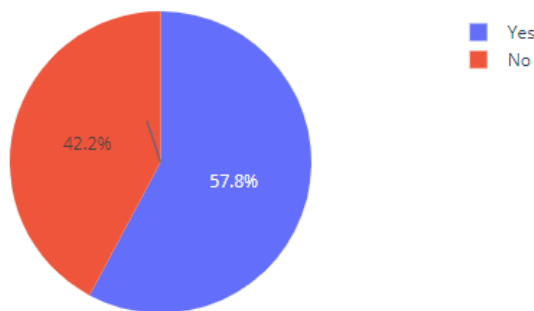


Figure 3.2: Can we contact you for further information?

These findings could be used to guide the design of services and resources at CERN,

3 Analysis Of User Behavior

with a particular focus on the preferences and needs of different age groups.

Based on previous definitions, big data organisations are organisations that produce more than 500 TB of data per week. In the case of CERN, just the Large Hadron Collider generates 10 GB per second. Data produced by the collider is used for research, reports, visualisation, in communication and more. As seen in table 3.1, CERN employees also use various methods of communication and data sharing. Communication methods used at CERN are chat, email, face-to-face, phone, SMS, social media, where people used email the most followed with face-to-face meetings Jones (2017).

Method of Communication	User Count
Email	1216
Face-To-Face	1015
Chat	585
Phone	575
SMS	170
Social Media	111

Table 3.1: IT Department User Survey Report: Which methods do you use to communicate with colleagues? Taken from (Jones, 2017)

Apart from these different communication methodologies, CERN users have different hardware at their disposal, ranging from mobile phones, tablets, laptops to desktop PCs Jones (2017). The majority (584 participants, 48.6%) reported owning a desktop computer, followed by 421 participants (35.0%) owning a laptop. A smaller proportion of respondents reported owning a smartphone (9.5%), cellphone (5.1%), or tablet (1.8%). It is important to note that the survey did not include questions regarding the number of devices owned by each participant, and it is possible that some individuals may have owned multiple types of devices. Nevertheless, these results provide insight into the prevalence of different types of devices among CERN employees and users. Additionally, the participants were asked about how often they work using Desktops, Laptops, Mobile Phones, Tablets, and Virtual Machines. The results of the survey indicated that Desktops were the most frequently used device for work, with 650 participants reporting that they use them often. Laptops were also frequently used for work, with 864 participants reporting that they use them often. In contrast, Mobile Phones were reported to be used often by only 441 participants, while Tablets were reported to be used often by only 71 participants. Furthermore, the survey results showed that 396 participants never use Desktops for work, whereas 756 participants never use Virtual Machines. Finally, 173 participants rarely use Laptops, while 346 participants rarely use Mobile Phones for work.

This also illustrates the magnitude of information generated by such an organisation. Even though a lot of information is generated by organisations, information is only

3.2 CERN IT Department User Survey Analysis

useful when it is stored and organised in a way that it can be used at an appropriate time. Organisations see big data and big data analytics as a high priority in their IT environments. They rated improving data analytics capabilities as very important and improving data analytics capabilities as one of the most important tasks to face.

Information is shared, transferred, and transformed with remarkable velocity nowadays. This has created a need for efficient knowledge management for the information to be useful and accessible. Sharing knowledge in organisations can help build a competitive advantage if the knowledge is managed adequately. In literature, knowledge sharing is also called voluntary dispersion of acquired experiences and skills and it makes employees capable to develop values, skills, and competencies. With the appearance of the internet and social media, information management and knowledge sharing have become more accessible, but shortages within these systems have also been emphasized. Shortages like misinformation, fake news, and dissemination of false information. According to the CERN user survey, a high percentage of users share their knowledge and retrieve information first with the use of email, then face-to-face meetings, and ultimately with the use of the IT infrastructure (CERNSearch or CERN-Website). Mostly used to share documents with colleagues is email, followed by cloud storage Jones (2017). The distribution of the usage of these methods for document and knowledge sharing can be seen in table 3.2.

Methods for sharing documentation	User Count
Via Email	899
Indico	479
Afs	455
CernBox	377
Dfs	318
Twiki	315
Other	237
Edms	196
Sharepoint	192
Cds	144
Eos	118
Onedrive	44
No	25

Table 3.2: IT Department User Survey Report: How do you share documents with colleagues? Taken from (Jones, 2017)

In recent years, CERN has been working on the concept of open science, which uses the principles of open data and open search to make the research done at CERN more visible and accessible to the general public. This CERN solution for open science

3 Analysis Of User Behavior

has made it easier to access research papers and to easily validate the information presented by those papers. The goal of open innovation and open information is to increase accountability, transparency, and provide new and efficient services. Before the information is made public, in general cases, it is anonymised with the intent to remove any personal identifiers from the data. With a higher degree of anonymisation, the less useful the information is. This fact gives leverage to the deanonymisation of the data. Cases like the Massachusetts health data leak and Netflix credit card backtracing are just a few that represent the danger of open data. For organisations, not being open to information is the key to obtaining a competitive advantage. CERN does not have the risk of losing the competitive advantage as other more business-oriented organisations, but it has to maintain a level of anonymisation to protect its employees. Since most of the information at CERN is shared via email or face-to-face meetings, it is understandable that the risk of deanonymisation of data is prevailing. The survey shows that a large percentage of surveyed users do not prefer to use social media and avoid sharing personal information on the workspace. Of the 1,232 users, 913 do not use social networking services for work, and 55 users encrypt their files for security and privacy reasons (Jones, 2017).

3.2.1 Conclusion

The sample size of 1233 responses gives a 99% confidence level and a margin of error of 3.5%, meaning that the survey data would be within $\pm 3.5\%$ of the actual results in 99 out of 100 surveys. Of the total participants, 459 were members of the personnel. This sample size gives a 95% confidence level and a margin of error of 4%.

In conclusion, the results of the survey showed that the CERN computer community uses various devices that run different operating systems. Microsoft Windows is the most common operating system for desktop computers, but users reported using laptops more often and predominantly with MacOS. Over 20% of participants used non CERN-supported operating systems. Email remains the preferred method of communication, with 85% of users saying that they use their CERN email account. However, over 20% forward their email onto another email system. The survey revealed that users have strong preferences for hardware and software, making it impossible to propose closed solutions for service delivery. The survey results listed in the report can be used as a reference to help IT-CDA members improve on their services.

3.3 CERN New Comers Analysis

The previous survey found a correlation between the survey results and age groups, which prompted the organisation to investigate the differences between those arriving at CERN and established CERN users, as well as any new trends in the user community.

3.3 CERN New Comers Analysis

The CERN Newcomers Analysis is a study that aims to determine the technological and work habits of new employees. This analysis provides insights into how new employees adapt to the unique work environment at CERN and their preferred methods of communication and collaboration (Jones, 2020).

This new survey was launched in the spring of 2018 and aimed to find out the choices and preferences of CERN's younger generation. The survey consisted of 20 questions and was made available to new comers who followed the HR onboarding process. It was not mandatory, and responses were kept confidential. The results of the survey covered one year up to April 2020, and previous study results were used for comparison. The survey results presented in May 2019 revealed that new-comers to CERN preferred to use laptop computers, with most having Windows operating systems. However, there was a trend towards more macOS and Linux users. The survey also found that 95% of new-comers had smartphones, with a trend towards Android devices. The most popular web browser was Chrome, and very few users preferred Safari and IE. Email remained the preferred method of communication, with Gmail being the most popular email service. There was an upward trend in the use of chat. The survey also found an increase in the use of instant messaging, with WhatsApp being the most popular service.

The survey revealed that Microsoft Office was the most used Office suite. Most newcomers were familiar with CERN-supported version control systems; most of them were using GIT with a small number of users using SVN. Email remained the most used method for sharing documents, and Google Drive was the preferred method for cloud storage, with hardly anyone using iCloud. For issue tracking, users preferred Git-based and Jira software. Online calendars were used by 50% of the users for scheduling meetings, with Google being the most popular online calendar. WordPress remained the best-known CMS for new-comers, with not many SharePoint users. Python and C++ were the most popular computer programming languages, with only 5 Perl users. Notepad++ remained the most popular editor, followed by Visual Studio Code.

Over 80% of survey respondents said they used social media, mostly Facebook, but not many used it for work purposes. User profiles were similar to the previous survey results, with most users being engineers and under 30 years old.

Most of the users use the CERN webpage and meetings for information discovery about CERN related information and events Jones (2020). This behavior of searching for information within a specific group of users facilitates the creation of information bubbles. In large organisations, users have to be trained to use a vast number of tools that have a similar purpose. This is also accurate for CERN, besides the tools for information discovery of data in their departments, about CERN and events at CERN, users use software for technical documentation, version control software, blogging software, survey software, video software, project management software, wiki software, and multiple hardware components. This generation of large amounts

3 Analysis Of User Behavior

of data complicates the process of keeping track of communication and information, not only for internal purposes but also for external. Enabling individuals inside and outside the organisation, groups, and organisations to consume and share the generated information is a basis point for enabling open innovation and increasing the value of the produced information. Taking into consideration the large amount of information that is continuously being generated, it is understandable that innovative methods for organisation, aggregation, storing, and use of this information are needed. Organisations have to solve the problem of cost-effecting storage of information and avoid information loss while taking into account privacy concerns. To make the stored information available and useful, methods for instant and reliable information retrieval and information discovery are needed. These methods have the goal to enable the validation of new information, improvement of access to existing information, improvement of existing processes, creation of new processes, and promotion of innovation and transparency. In conclusion, the survey provided insights into the work practises and preferences of new-comers to CERN over the period of 12 months. The survey results were very similar to previous year's results, with some clear trends.

3.3.1 Conclusion

In conclusion users in large and highly connected organisations, such as CERN, prefer the use of emails and face-to-face meetings as their main communication and knowledge sharing method. Most of CERN users use email for information discovery about their department and peer group Jones (2020). Emails and face-to-face offer confidentiality and protect sensitive information, but limit the ease of access to information. When trying to access documents via email, it is necessary to execute at least two operations, to send an inquiry to an individual and wait for an extended period for feedback, depending on the availability of the other person. Face-to-face meetings can both reduce or extend the time needed to retrieve relevant information. All meeting participants have to be available for the meeting to take place, which needs time to organise. On the other hand, meeting in large organisations results in meeting minutes, which is a quick summary of the meeting. These produced documents can then be inspected and information can be retrieved easily. The drawback of face-to-face meetings is that a part of the information produced in meetings can get lost due to inefficient documentation practises. The documentation produced at face-to-face meetings is restricted to the group of individuals that attended the meeting. This brings difficulties in sharing information with individuals outside of the work-group or meeting group and retrieving information. With the option of searching open information, the wait duration for the feedback and information is reduced, it is not necessary to arrange meetings and wait for another individual to execute queries. Identical queries can be executed multiple times with no additional expenditure of

3.4 CERN User Information Consumption Analysis

resources. In order for the information produced in organisations to be usable by open search tools, it is necessary to structure and normalize it for search. This leads to a more reliable approach to information organisation and standardisation in organisations and the possibility to share the information in a more manageable way not only between organisation members but also outside of the organisation. Search ready open information can be grouped by interest groups, filtered by users according to their interests, recommended to users that share similar interests or shared between similar organisations.

3.4 CERN User Information Consumption Analysis

The study was designed to gain insight into the information consumption patterns and preferences of users in large organisations, focussing on CERN user groups. A descriptive user study was conducted, using a mixed method design, combining quantitative and qualitative data collection techniques to obtain comprehensive information about the research objectives. The study involved the distribution of a survey to a diverse sample of CERN users, providing a comprehensive understanding of their behaviours and attitudes towards various aspects of information consumption. The study was conducted in accordance with the data privacy rules of CERN, ensuring the protection of participants' personal information and adherence to ethical standards.

This study aims to provide valuable insights into information consumption user preferences to improve the information consumption experience for users across various domains. The main research questions are:

- **R3.1:** What are the time intervals, device usage and medium preferences of users in large organisations for accessing both work-related and personal-related information?
- **R3.2:** What are the similarities and differences in work-related and personal-related information consumption?
- **R3.3:** How do users in large organisations perceive and engage with personalised recommendation systems, specifically notification systems?
- **R3.4:** Which challenges do users in large organisations encounter while consuming and rediscovering previously accessed information, and which factors contribute to these difficulties?

3.4.1 Setting and Instruments

The study was created with the use of a DRUPAL online survey tool as one of the data collection instruments. It was also used to ensure the privacy and confidentiality of the participant data. Furthermore, the survey underwent a rigorous development

3 Analysis Of User Behavior

process following an iterative approach, incorporating valuable insights and inputs from domain experts at CERN as seen in Table 3.4. Table 3.3 provides an overview of the five main survey sections.

Section Title	Number of Questions	Question Types
Introduction	0	No Questions
Demographics	4	Single Choice
Work Related Information Consumption	11	Single Choice, Multiple Choice, Likert Scale
Personal Information Consumption	11	Single Choice, Multiple Choice, Likert Scale
Notification System Questions	6	Single Choice and Likert Scale

Table 3.3: Survey Sections and Question Details

The "Introduction" section provided essential information to participants such as the purpose of the survey, clearly stating the research objectives and explaining how the collected data would contribute to the study. Additionally, it included a privacy notice, outlining the commitment to protecting participants' privacy and maintaining the confidentiality. Participants were informed about the measures taken to ensure data security and were assured that their personal information would be handled in compliance with data protection regulations. The "Demographics" section was designed to collect essential demographic data from users, including gender, age range, job position, and type of contract. The section "Work-Related Information Consumption", aimed to understand the services, devices, and mediums participants use to stay informed about work-related topics. It explored the timing and frequency of accessing work-related information, capturing specific intervals and the duration of the engagement. The fourth section, "Personal Information Consumption", focused on investigating the services, devices, and mediums participants employ to gather information regarding their personal and work-related interests. Similarly to the previous section, it delved into the timing and frequency of accessing personal information, including specific times of day, days of the week, and duration of the engagement. Lastly, "Notification System Questions," included questions related to participants' willingness to adopt a notification system to receive updates on both work-related and personal interest topics.

3.4.2 Procedure

Table 3.4 provides an overview of the various phases involved in the survey implementation process and describes the characteristics of the participants at each stage.

3.4 CERN User Information Consumption Analysis

Phase	Phase Description	Phase Goals	Participants Description
Pre-Study Initial Validation	The survey was distributed to a group of 10 expert users within the organisation to validate the questionnaire and make necessary adjustments based on user feedback.	<ul style="list-style-type: none"> - Validate the questionnaire - Incorporate user feedback 	10 expert users within the organisation were invited and participated
Pilot Survey	A pilot survey was conducted among a work group to identify and resolve any ambiguities or issues with the survey instrument.	<ul style="list-style-type: none"> - Identify and resolve issues - Refine the survey instrument - Collect data from a larger user set 	150 users within the IT department were invited to participate, 56 participated in the pilot survey
Study CERN-wide Survey	The survey was disseminated to a broader audience of 17,000 CERN users to gain a comprehensive understanding of user perceptions and experiences.	<ul style="list-style-type: none"> - Gather comprehensive data - Capture user perceptions and experiences - Capture diverse perspectives 	17000 CERN users were invited to participate in the survey, 767 participated
Post-Study Data Processing	After survey completion, data was carefully processed to ensure anonymity and privacy. Personal identifiable information was removed or anonymised.	<ul style="list-style-type: none"> - Ensure anonymity and privacy - Remove personal identifiable information 	767 Participants

Table 3.4: Survey Phases and Participants Description

The subsequent sections will present the findings and analyses derived from the survey data, offering more detailed information about the patterns of information consumption and preferences of CERN users.

To ensure the integrity of the survey and the protection of participant information, the

3 Analysis Of User Behavior

design and implementation of the survey were reviewed and approved by department management. This step aimed to ensure that the survey adhered to ethical guidelines and complied with relevant privacy regulations.

3.4.3 Results and Discussion

The results section presents a comprehensive and detailed analysis of the survey data, providing insight into the information consumption patterns among users at CERN. By examining the relationships between demographic characteristics, device usage, preferred mediums, and information consumption behaviours, we aim to contribute to a deeper understanding of how individuals at CERN engage with information, facilitating the development of customised information delivery strategies to meet their diverse needs and preferences.

Demographic Data

Participant characteristics play a crucial role in research studies, as they contribute to the diversity and representativeness of the sample. Out of the 17,000 CERN-affiliated users invited to participate in the survey, a total of 767 respondents completed and submitted the questionnaire, making the participant response rate 4.511%. This subset of respondents represents a diverse cross-section of the CERN community, spanning various scientific disciplines and age ranges. Table 3.5 describes the basic demographic characteristics of the participants.

Participant Characteristics	Female	Male	I prefer not to say	Total
Gender	180 (23.47%)	535 (69.75%)	52 (6.78%)	767
Preferred English	153 (85%)	481 (89.9%)	49 (94.2%)	683 (89%)
Preferred French	27 (15%)	54 (10.1%)	3 (6.8%)	84 (11%)
Age Range: 18-25	20 (11.11%)	44 (8.23%)	5 (9.7%)	69 (8.996%)
Age Range: 25-30	41 (22.72%)	84 (15.7%)	4 (7.7%)	129 (16.818%)
Age Range: 31-40	51 (28.34%)	115 (21.5%)	4 (7.7%)	170 (22.164%)
Age Range: 41-51	36 (20%)	104 (19.44%)	3 (5.77%)	143 (18.644%)
Age Range: 51-61	27 (15%)	120 (22.43%)	8 (15%)	155 (20.208%)
Age Range: 61-70	3 (1.67%)	39 (7.29%)	3 (5.77%)	45 (5.867%)
Age Range: 70+	1 (0.55%)	27 (5.04%)	0 (0%)	28 (3.65%)
Age Range: I prefer not to specify	1 (0.55%)	2 (0.37%)	25 (48.36%)	28 (3.65%)

Table 3.5: Demographic Characteristics of Study Participants

3.4 CERN User Information Consumption Analysis

Among the respondents, a higher proportion identified as male (535 participants; 69.75%) compared to female (180 participants; 23.47%). Additionally, a small percentage (52 participants; 6.78%) preferred not to disclose their gender.

Language preference is a fundamental aspect of communication and information consumption. In this survey, the majority of respondents (683 participants) indicated English as their preferred language. This finding aligns with the widespread use of English as the dominant language in scientific and research environments. Furthermore, a smaller group of participants (84) expressed a preference for French, highlighting the linguistic diversity within the CERN community. A higher proportion of males (89.9%) chose English as their preferred language, while a smaller percentage of males (10.1%) preferred French. Females displayed a slightly lower preference for English (85%) but a higher preference for French (15%).

Analysing the age distribution of participants provides insight into the generational composition of the user base. Survey respondents spanned a wide range of age groups, with notable concentrations in the 31-40, 41-51 and 51-61 brackets. This distribution suggests a significant presence of people with mid-career experience and expertise at CERN. Furthermore, participants from younger age groups, such as 18-25 and 25-30, as well as more senior people aged 61 and over, were also represented, contributing to the diversity of perspectives and experiences. Most of the male participants were in the 31-40 age bracket, while the female participants were more in the 51-61 age bracket. Participants who did not share their gender information also predominately did not want to specify their age.

Exploring the affiliations of the participants provides insight into their roles and responsibilities within the CERN organisation. The survey revealed a diverse range of affiliations, the most common being Users (USER) and Staff (MPE). Some participants chose not to disclose their affiliation, while others held unique roles such as retired or guest professors. This diverse affiliation landscape underscores the multidisciplinary and collaborative nature of the work at CERN.

Table 3.6 presents the distribution of participants based on their nature of work at CERN, including gender-specific percentages and the proportion of individuals who preferred not to disclose their information. The data provides insight into the representation of different work categories within the CERN community.

The largest category of participants is "User" (USER), comprising 36.11% female participants, 38.69% male participants, and 15.38% participants who preferred not to disclose their gender. The "Staff" category (MPE) is the second largest, with 31.66% females, 34.20% males, and 32.69% participants who preferred not to disclose their gender. The "Fellow" (MPE) category consists of 12.77% females, 9.53% males, and 1.92% participants who preferred not to disclose their gender. In the "Doctoral Student" (DOCT) category, 5.55% are females, 4.48% are males, and 3.84% chose not to disclose their gender. It is also visible that most of the individuals that did not want to share their

3 Analysis Of User Behavior

demographic information, also did not share their nature of work information. Other categories such as "Technical Student" (TECH), "Project Associate" (PJAS), "Trainee" (TRNE), and "Cooperation Associate" (COAS) each comprise a small percentage of participants across genders, with some participants choosing not to disclose their information.

Nature of Work at CERN	Female	Male	I prefer not to say
User (USER)	65 (36.11%)	207 (38.69%)	8 (15.38%)
Staff (MPE)	57 (31.67%)	183 (34.20%)	17 (32.69%)
I Prefer Not To Say	6 (3.33%)	6 (1.12%)	21 (40.38%)
Fellow (MPE)	23 (12.77%)	51 (9.53%)	1 (1.92%)
Technical Student (TECH)	5 (2.77%)	13 (2.43%)	3 (5.76%)
Trainee (TRNE)	3 (1.66%)	2 (0.37%)	0
Project Associate (PJAS)	3 (1.66%)	12 (2.24%)	0
Administrative Student (ADMI)	3 (1.66%)	0	0
Cooperation Associate (COAS)	2 (1.11%)	3 (0.56%)	0
Doctoral Student (DOCT)	10 (5.55%)	24 (4.48%)	2 (3.84%)
Visiting Scientist (VISC)	1 (0.55%)	19 (3.55%)	0
Summer Student (SUMM)	1 (0.55%)	1 (0.18%)	0
Boursier TTE	1 (0.55%)	0	0
Scientific Associate (SASS)	0	3 (0.56%)	0
Apprentice (APPR)	0	1 (0.18%)	0
Early Graduate	0	1 (0.18%)	0
GRAD	0	1 (0.18%)	0
Guest Professor (GPRO)	0	1 (0.18%)	0
I prefer not to say	0	1 (0.18%)	0
Industrial Liaison Officer	0	1 (0.18%)	0
Master degree student	0	1 (0.18%)	0
Postdoc affiliated with ATLAS experiment	0	1 (0.18%)	0
Retired	0	1 (0.18%)	0
Special staff	0	1 (0.18%)	0
VIA	0	1 (0.18%)	0
Total Participants	180	535	52

Table 3.6: Participant Nature of Work at CERN

3.4.4 Comparing Work Related and Personal Information Consumption

Table 3.7 and 3.8 show that in both contexts (personal and professional), email stands out as the most prominent information source, with a high percentage of participants across all gender groups relying on it. This suggests that email remains a widely adopted and versatile tool for information consumption and generation, regardless of the setting.

Information Source	Female	Male	I prefer not to say	Total
E-mail	177 (98.333%)	522 (97.57%)	51 (98.077%)	750 (97.53%)
Mattermost	103 (57.222%)	263 (49.159%)	34 (65.385%)	400 (52.07%)
Twiki	63 (35.0%)	144 (26.916%)	14 (26.923%)	221 (28.81%)
Meetings	136 (75.556%)	374 (69.907%)	39 (75.0%)	549 (71.53%)
CERN Bulletin	80 (44.444%)	206 (38.505%)	23 (44.231%)	309 (40.27%)
Other	21 (11.66%)	50 (9.34%)	6 (11.53%)	77 (10.03%)
Total Participants	180	535	52	

Table 3.7: Utilization of Information Sources for Work-Related Information Retrieval and Consumption

It is also noticeable that users do not use email to the same degree while consuming information for personal related topics as for work-related topics. For personal information consumption most of the participants selected that they use other informational sources than the predefined selection of E-mail, Mattermost, Twiki, Meetings, or CERN Bulletin.

Information Source	Female	Male	I prefer not to say	Total
E-mail	67 (37.22%)	215 (40.18%)	17 (32.69%)	299 (38.98%)
Mattermost	0%	0%	0%	0
Twiki	0%	0%	0%	0
Meetings	18 (10.00%)	52 (9.72%)	4 (7.69%)	74 (9.64%)
CERN Bulletin	0%	0%	0%	0
Total Participants	180	535	52	

Table 3.8: Preferred Information Sources for Personal Usage

Figures 3.3, 3.4, 3.5 present the preferred devices for personal usage among female participants, male participants, and those who prefer not to disclose their gender,

3 Analysis Of User Behavior

respectively. The tables showcase the frequency of device usage categorized into five options: very frequently used, frequently used, occasionally used, rarely used, and very rarely used.

In Figure 3.3 for female participants, the most commonly preferred device is the mobile phone, with 126 participants indicating very frequent usage. The laptop follows with 64 participants reporting very frequent usage. Desktop PCs and tablets are also frequently used by female participants, but with lower frequencies compared to mobile phones and laptops.

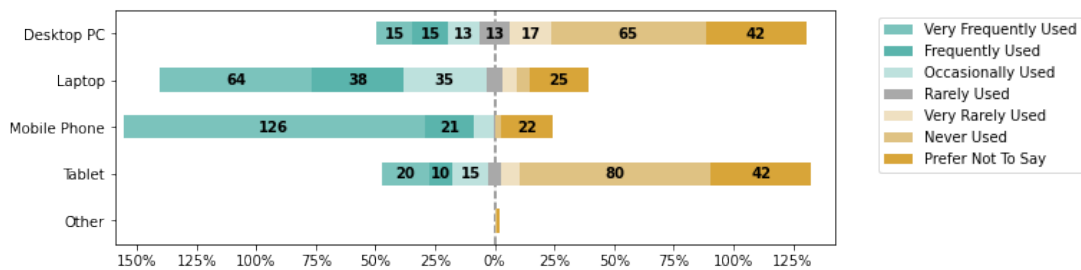


Figure 3.3: Preferred Device for Personal Usage (Female)

For male participants, as seen in Figure 3.4, the preferred device for personal usage is also the mobile phone, with 286 participants (40.18%) reporting very frequent usage. The laptop is the second most popular device, with 235 participants (33.09%) indicating very frequent usage. Desktop PCs and tablets are also commonly used by male participants, but with lower frequencies compared to mobile phones and laptops.

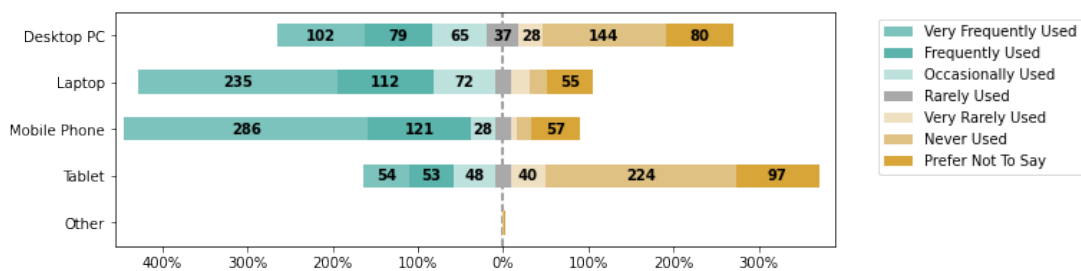


Figure 3.4: Preferred Device for Personal Usage (Male)

Among participants who prefer not to disclose their gender, as seen in Figure 3.5, the pattern remains consistent, with the mobile phone being the most preferred device, followed by the laptop. However, the overall frequencies of device usage are lower in this group compared to female and male participants.

3.4 CERN User Information Consumption Analysis

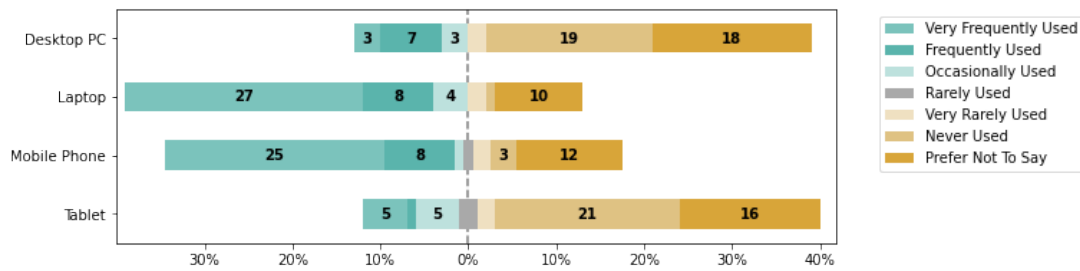


Figure 3.5: Preferred Device for Personal Usage (Preferred not to say)

Figures 3.6, 3.7, 3.8 illustrate how the participants selected their preferred device for work-related information consumption. When comparing the usage patterns across genders, it is observed that among female participants (Figure 3.6), laptops are frequently used, followed by mobile phones and desktop PCs. Tablets are less commonly used. Male participants (Figure 3.7) also exhibit similar preferences, with laptops being the most frequently used device, followed by mobile phones and desktop PCs. However, male participants tend to use tablets more often compared to female participants.

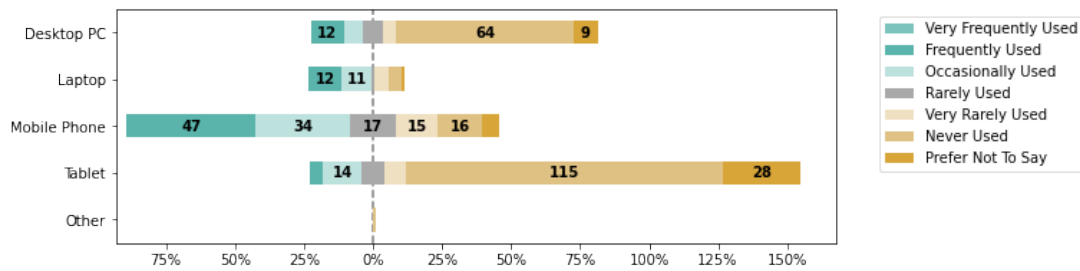


Figure 3.6: Preferred Device for Work-Related Information Consumption (Female)

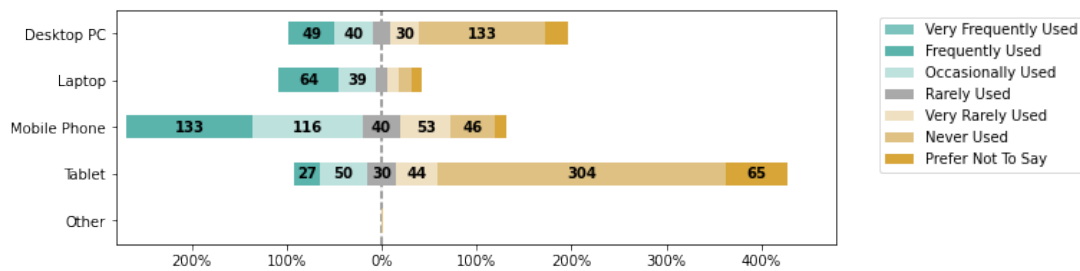


Figure 3.7: Preferred Device for Work-Related Information Consumption (Male)

3 Analysis Of User Behavior

In the case of participants who prefer not to disclose their gender (Figure 3.8), laptops and mobile phones are the most preferred devices, with 11.54% and 15.38% of participants using them frequently, respectively. Occasional usage of laptops and mobile phones is reported by 7.69% and 23.08% of the participants, respectively.

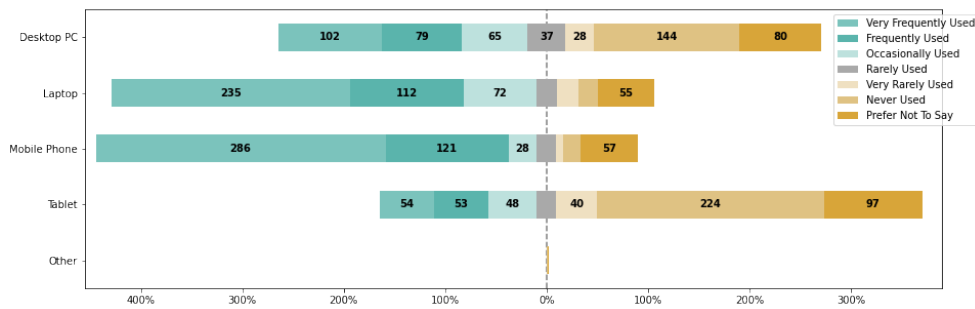


Figure 3.8: Preferred Device for Work-Related Information Consumption (Preferred not to say)

When comparing the preferred devices for personal usage and work-related information consumption, various commonalities and differences can be observed. For personal usage, for male and female participants, the most frequently used device is the mobile phone. It is very frequently used by 70.00% of females and 53.46% of males. Following that, laptops are the second most popular device, with 35.56% of females and 43.93% of males frequently using them. Desktop PCs are also commonly used, but their usage is lower compared to mobile phones and laptops. In terms of gender differences, females show a higher preference for desktop PCs for personal usage, with 8.33% very frequently using them compared to 5.77% of males. However, laptops and mobile phones are more commonly used by males, with higher percentages across all usage frequencies. When it comes to work-related information consumption, the usage patterns differ. Desktop PCs are the most frequently used devices for both genders, although their usage is relatively low compared to personal devices. Males show a higher preference for desktop PCs, with 9.16% frequently using them compared to 6.67% of females. Laptops and mobile phones are also used for work-related purposes, but their usage percentages are lower compared to personal usage. Tablets are particularly popular for work-related information consumption, with higher percentages across all usage frequencies. Females tend to use tablets more frequently for work-related purposes compared to males.

Figures 3.9, 3.10 and 3.10 present the preferred mediums for personal usage among female participants, male participants, and individuals who prefer not to disclose their gender. In Figure 3.9 the most preferred medium for female participants is textual content, with 103 participants expressing a strong preference for it. Audio and video mediums are also popular among female participants, followed by verbal

3.4 CERN User Information Consumption Analysis

communication and a combination of multiple mediums.

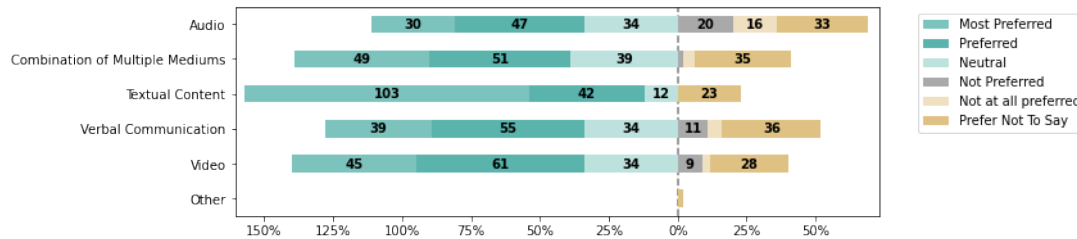


Figure 3.9: Preferred Mediums for Personal Usage (Female)

Figure 3.10 shows statistics for male participants where the preferred mediums for personal usage are similar to those of female participants. Textual content remains the most preferred medium, with 318 participants indicating a high preference. Other preferred mediums include a combination of multiple mediums, video, and verbal communication.

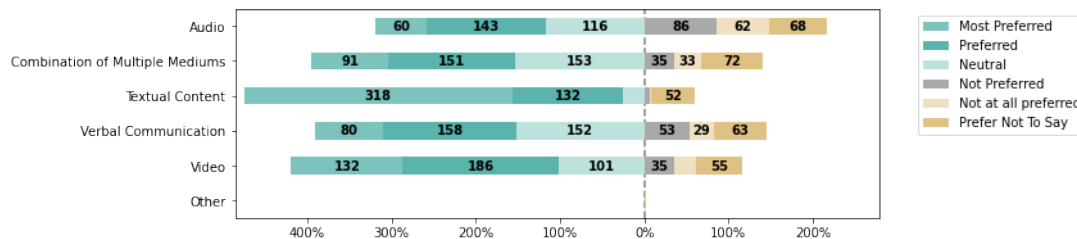


Figure 3.10: Preferred Mediums for Personal Usage (Male)

In Figure 3.11 for participants who prefer not to disclose their gender, the preferred mediums for personal usage include textual content, audio, and video. However, the number of participants in this category is relatively smaller, resulting in lower frequencies across all preferred mediums.

The Preferred Mediums for Personal Usage for both females and males show similar trends in their preferences. For both genders, textual content is the most preferred medium, with 57.22% of females and 59.44% of males indicating their preference for it. The combination of multiple mediums and video content is the second most preferred option for both groups. However, there are some notable differences. Verbal communication is the second most preferred medium for males, at 14.95% and for females, at 21.67%. On the other hand, audio is the least preferred medium for both genders, but females (18.33%) are more likely to dislike it compared to males (12.71%). Interestingly, the "Prefer Not To Say" group shows a significant variation in their

3 Analysis Of User Behavior

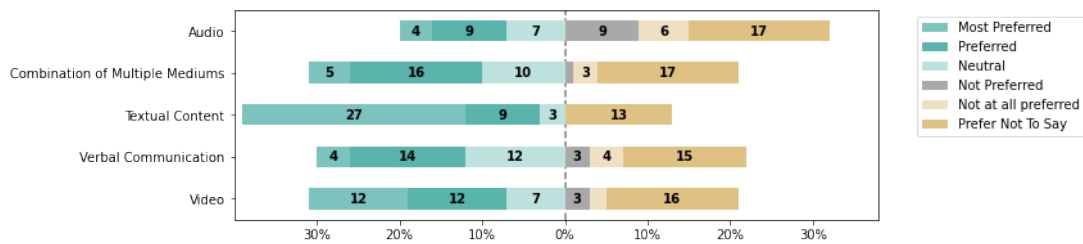


Figure 3.11: Preferred Mediums for Personal Usage (Preferred not to say)

preferences compared to females and males. They tend to prefer textual content the most (51.92%) and are less interested in video content. For Work-related Information Consumption the patterns are consistent with the preferences for personal usage. Textual content is the most preferred medium for both genders, with 57.22% of females and 59.44% of males indicating their preference for it. There are slight variations between the genders. Males show a higher preference for verbal communication (29.53%) as their second choice, while females prefer the combination of multiple mediums (28.22%). Both genders express a lower interest in audio as their least preferred medium, with females slightly more disinterested than males. Both genders have a similar hierarchy of preferences for different mediums, with textual content being the most preferred for both personal and work-related information consumption. However, there are slight differences in the order of preferences and levels of interest in certain mediums between females, males, and the "Prefer Not To Say" group.

Figure 3.12 shows the preferred mediums used by female participants to gather information about work-related topics. The data reveals the distribution of preferences across various mediums. Textual content emerges as the most preferred medium, followed by verbal communication and audio. The combination of multiple mediums and video are favored by a substantial number of female participants as well. The preferences for other mediums are relatively low in comparison.

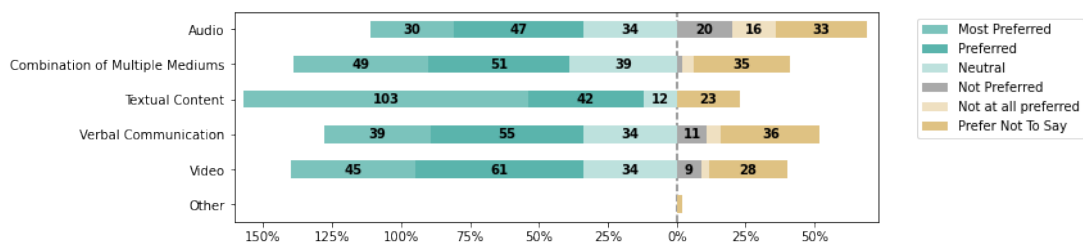


Figure 3.12: Preferred Mediums for Work-Related Information Consumption (Female)

Figure 3.13 focuses on the preferred mediums for work among male participants.

3.4 CERN User Information Consumption Analysis

Similar to the female participants, textual content is the most preferred medium, with 318 male participants favoring it. Video and the combination of multiple mediums are also highly favored among males, with 132 and 91 participants respectively. Verbal communication and audio receive significant preference as well. The preferences of male participants align closely with those of female participants, with similar patterns observed across the different mediums.

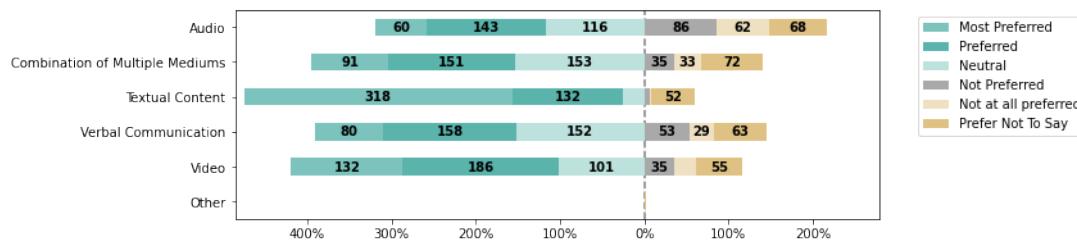


Figure 3.13: Preferred Mediums for Work-Related Information Consumption (Male)

Figure 3.14 represents participants who prefer not to disclose their gender. Among this group, textual content remains the most preferred medium, with 27 participants selecting it. Video and the combination of multiple mediums also receive considerable preference, as indicated by the responses of 12 and 5 participants respectively. The preferences for verbal communication and audio are relatively lower compared to the other mediums.

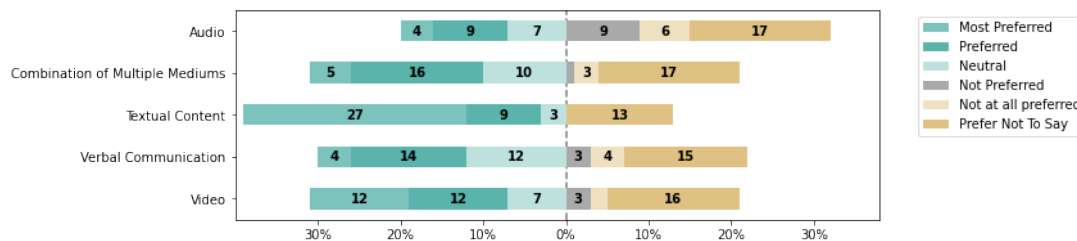


Figure 3.14: Preferred Mediums for Work-Related Information Consumption (Preferred not to say)

The findings from this survey shed light on the device usage and medium preferences of CERN users when it comes to consuming personal-related information. The prevalence of laptops and mobile phones as the primary devices, along with the preference for textual content, audio, and video mediums, highlight the diverse ways in which CERN users engage with and seek information related to their personal interests. These insights can inform the development of targeted communication strategies and platforms that cater to the specific information needs and preferences of CERN's user community when it comes to personal-related topics.

3 Analysis Of User Behavior

The provided data in Table 3.9 illustrates the frequency of news portal usage for obtaining information on topics related to personal interests, categorized by gender and participant preference. The participants' responses varied across different frequency options.

	Female	Male	I preferer not to say
Prefer Not To Say	21 (11.67%)	49 (9.16%)	12 (23.08%)
I do not read news portals	13 (7.22%)	14 (2.62%)	1 (1.92%)
A couple of times a day	55 (30.56%)	268 (50.09%)	16 (30.77%)
Once a day	43 (23.89%)	118 (22.06%)	17 (32.69%)
Once a week	12 (6.67%)	21 (3.93%)	1 (1.92%)
Once every few days	36 (20.0%)	65 (12.15%)	5 (9.62%)
Total Participants	180	535	52

Table 3.9: How often do you read news portals to get informed about topics related to personal interests

Table 3.10 provides valuable insights into the frequency at which participants engage with news portals to obtain work-related information. The table explores the reading habits across different gender categories, namely Female, Male, and participants who preferred not to disclose their gender.

	Female	Male	I prefer not to say
I do not read work-related announcements	16 (8.89%)	49 (9.16%)	7 (13.46%)
A couple of times a day	34 (18.89%)	106 (19.81%)	10 (19.23%)
Once a day	38 (21.11%)	139 (25.98%)	16 (30.77%)
Once a week	38 (21.11%)	100 (18.69%)	8 (15.38%)
Once every few days	54 (30.0%)	141 (26.36%)	11 (21.15%)
Total Participants	180	535	52

Table 3.10: How often do you read news portals to get informed about work-related topics?

When comparing the preferred intervals for reading news portals to get informed about personal interests and work-related topics, both males and females show similar patterns in their preferred reading intervals. The most common choice for both genders is to access news portals a couple of times a day, with 30.56% of females and 50.09% of males selecting this option. Once a day is the second most popular interval for both genders, chosen by 23.89% of females and 22.06% of males. Once every few days is another common choice, selected by 20% of females and 12.15% of males. Additionally, a smaller percentage of participants prefer not to disclose their gender, with 30.77% also opting for a couple of times a day. For work-related topics, the preferred intervals for reading news portals display a similar distribution among genders. A couple of

3.4 CERN User Information Consumption Analysis

times a day remains the main choice, with 18.89% of females, 19.81% of males, and 19.23% of participants who preferred not to disclose their gender. Once a day is the second most popular interval for all genders, chosen by 21.11% of females, 25.98% of males, and 30.77% of those who preferred not to say. Once a week and once every few days are also selected by participants from all genders, albeit with varying percentages.

When comparing the preferred intervals for reading news portals to get informed, commonalities and differences were identified. The majority of participants, regardless of gender, prefer to check their personal email "1-3 Times A Day" on weekdays. This interval is chosen by 52.78% of females, 38.13% of males, and 48.08% of participants who preferred not to disclose their gender. Similarly, the interval "1-3 Times A Day" is the most preferred for checking personal email on weekends. This choice is consistent across genders, with 56.67% of females, 44.86% of males, and 46.15% of participants who preferred not to disclose their gender opting for this interval. When it comes to work-related email, there is a notable difference compared to personal email habits. The preferred interval for checking work email on weekdays is "More than 12 Times a Day." This choice is made by a significant majority of participants, with 62.78% of females, 54.95% of males, and 42.31% of participants who preferred not to disclose their gender. On weekends, there is a shift in the frequency of checking work email. A significant proportion of participants, including 36.11% of females and 22.62% of males, mentioned that they do not check their work email on weekends. Among those who do check work email on weekends, the interval "1-3 Times A Day" remains the most common choice. It is selected by 41.67% of females, 39.63% of males, and 44.23% of participants who preferred not to disclose their gender. Additionally, a smaller percentage of participants opt for higher intervals such as "3-6 Times A Day" and "6-12 Times A Day" when checking work email on weekends. In summary, there is a general trend of checking personal email with a moderate frequency of "1-3 Times A Day" on both weekdays and weekends. However, the pattern shifts when it comes to work email, with a significant majority checking it "More than 12 Times a Day" on weekdays and a reduced frequency on weekends. These findings suggest that individuals prioritise work-related communication more frequently during weekdays, while weekends provide a break from constant work email checking.

When it comes to work-related information consumption, individuals tend to be more focused and task-oriented. Work-related information is typically obtained through professional sources such as colleagues, industry reports, and online databases. In contrast, personal information consumption is often more diverse and multifaceted, involving a wider range of sources such as social media, news outlets, and personal networks.

In terms of content, work-related information consumption revolves around acquiring knowledge and skills specific to one's profession or industry. This often involves staying up-to-date with the latest research, industry trends, and best practises. On the

3 Analysis Of User Behavior

other hand, personal information consumption covers a broader spectrum of interests, including news, entertainment, hobbies, and personal development.

3.5 Conclusion

Despite the differences, there are also several similarities in the way people consume information for work and personal purposes. Both work and personal information consumption rely heavily on digital platforms and technologies. With the proliferation of smartphones and other mobile devices, individuals can access information conveniently and instantaneously for both work and personal needs.

Moreover, critical thinking and information evaluation are vital skills that apply to both work and personal contexts. Individuals need to assess the reliability, credibility, and relevance of information, whether it is for work-related research or personal decision-making. The ability to discern accurate information from misinformation is crucial in both domains.

A considerable number of participants reported difficulties in rediscovering previously accessed personal-related information. This finding suggests that users face challenges in organising and retrieving personal interest information, which may hinder their ability to engage with their hobbies, sports, or other areas of personal interest effectively. This aligns with previous studies that have highlighted the importance of information organisation and retrieval in personal information management. Future research could focus on developing strategies or tools to assist users in organising and retrieving personal-related information more efficiently.

The findings of this study have several implications for the design and development of information delivery systems at CERN. Understanding the device usage and medium preferences of users can inform the creation of user-friendly platforms that cater to their specific needs. Furthermore, the insights into the attitudes towards personalised recommendation systems highlight the importance of considering user preferences and concerns in the implementation of such systems.

3.6 Summary

This chapter covers a comprehensive investigation into various dimensions of user behaviour within the setting of large organisations. It contains three key studies, each providing information on distinct aspects of user engagement.

The IT Department User Survey Analysis focused on determining the preferences, habits, and needs of IT department users, revealing their behavioural patterns and demands. This analysis was carried out with a sample size of 1233 responses, providing a confidence level of 99% with a margin of error of 3.5%. Specifically, it focused on

459 personnel members. This analysis unveiled valuable insights into the preferences of the CERN computer community. In particular, it highlighted the usage of various devices with different operating systems. Microsoft Windows emerged as the most common operating system for desktop computers, while laptops, primarily with MacOS, were favoured. Additionally, more than 20% of the participants reported using non-CERN supported operating systems. Email was identified as the preferred method of communication, with 85% of users using their CERN email accounts. However, more than 20% forwarded their email to another system. The survey highlighted strong user preferences for hardware and software, making it challenging to propose closed solutions for service delivery. These results, detailed in the report, serve as a reference for IT members to improve their services.

The CERN New Comers Analysis provided insight into newcomer behaviour within CERN over a 12-month period. The study revealed that large and highly connected organisations, such as CERN, prefer email and face-to-face meetings as primary communication and knowledge-sharing methods. Although these methods offer confidentiality and protect sensitive information, they also impose limitations on ease of access. Face-to-face meetings can vary in their impact, either by expediting or extending the time required to access relevant information. Documentation produced during such meetings may lack efficiency in information sharing beyond the meeting participants. This poses challenges in sharing information outside of specific groups or retrieving it efficiently. To mitigate these challenges and improve information sharing, open search tools were considered a viable option. Open information can be structured and normalised for search, leading to more efficient information organisation, standardisation, and external sharing. This approach can enable information grouping, filtering, recommendation, and sharing among users with similar interests or between similar organisations. The study also highlighted that CERN users primarily rely on the CERN website and meetings for information discovery on CERN-related topics. However, the wide range of tools and platforms used for various purposes within an organisation complicates communication and information tracking. Innovative methods for information organisation, aggregation, storage, and retrieval were identified as necessary to address these challenges and promote open innovation.

The CERN User Information Consumption Analysis investigated aspects of information consumption, with a focus on both personal and work-related topics. Interestingly, both males and females exhibited similar patterns in their preferred information consumption intervals. For personal use, mobile phones have become the most frequently used device for both genders. Laptops followed as the second most popular choice, with desktop PCs also in common use. However, desktop PCs showed a higher preference among females for personal usage. On the contrary, for the consumption of work-related information, desktop PCs were the most used devices. The analysis also revealed variations in email checking habits. Users tended to check their personal

3 Analysis Of User Behavior

email "1-3 Times A Day" on both weekdays and weekends. However, when it came to work-related email, a significant majority of participants checked it "More than 12 Times a Day" on weekdays, reflecting a heightened emphasis on work-related communication during the workweek. On weekends, there was a decrease in the frequency of checking work-related email, and a notable portion of participants abstaining from such checks. Regarding the content of information consumption, the study highlighted the distinctions between work-related and personal information. Work-related information typically revolved around acquiring knowledge and skills pertinent to one's profession or industry, necessitating access to research, industry trends, and best practises. In contrast, personal information consumption spanned a broader spectrum of interests, including news, entertainment, hobbies, and personal development. Despite these differences, the analysis highlighted several commonalities. Both work and personal information consumption heavily relied on digital platforms and technologies, emphasising the importance of online sources in today's information landscape. Furthermore, critical thinking and information evaluation were crucial skills for users in both contexts, highlighting the need to assess information for credibility and relevance.

These findings collectively contribute to answering several research questions. The study revealed preferences and trends in device usage and access times for information. It revealed distinctions in device usage and preferences between work-related and personal information consumption. The findings provide insight into how users in large organisations engage with different communication methods. The study indicated challenges in organising and retrieving personal information, determining factors that contribute to these difficulties. These findings are significant for the design and development of information delivery systems at CERN, emphasising the importance of user preferences and concerns in system implementation.

4 Data Analysis and Exploration

This chapter introduces the Data Lift framework and analyses its efficacy through its application to the CERN Mattermost dataset. Framework phases are introduced, which include purpose definition, data collection, risk assessment, transformation, evaluation, and publication. The CERN dataset published via the Data Lift Framework is analysed, capturing its essence through data metrics and statistical insights. The network analysis clarifies intricate user-channel and user-user connections, unveiling interaction dynamics. Furthermore, the chapter delves into community detection, identifying emergent social structures within the dataset. By unifying theory and practise, this chapter reinforces the importance of structured data exploration in revealing valuable insights within organisational datasets.

The following sections are based on, supported by, and taken from the work published in the following publications:

- **Jakovljevic, I., Gütl., C., Wagner., A., & Nussbaumer., A. (2022).** Compiling Open Datasets in Context of Large Organizations while Protecting User Privacy and Guaranteeing Plausible Deniability. In *Proceedings of the 11th International Conference on Data Science, Technology and Applications - DATA SciTePress - Science and Technology Publications*

4.1 Contribution

This chapter represents essential work on the "Data Gathering and Data Analysis" step of the thesis research methodology. It contributes to answering the question of how sensitive organisational data can be collected for reproducible research and development in a privacy-preserving way. It introduces the innovative DataLift framework designed to facilitate the responsible and secure publication of organisational data as open data. Through the application of this framework, the chapter demonstrates a practical approach to making previously closed and sensitive data accessible for research purposes. By showcasing the implementation of DataLift on CERN data, it establishes a model for enhancing the transparency and reproducibility of research, thereby addressing the first research question. Another critical aspect explored in this chapter relates to the importance of privacy and privacy-preserving concepts for employees within large organisations. The user study conducted at CERN provides

a valuable empirical basis for understanding how employees perceive and prioritise privacy concerns when organisational data is leveraged for research and development. The study's findings provide deep insights into the attitudes and perspectives of employees regarding the use of their data, aligning with the second research question.

4.2 Large Organisations and Sensitive Data

Large organisations generate a median of 300 terabytes (TB) of data per week. Data are generated from the use of various methods of communication (chat, email, face-to-face, phone, SMS, social media) between organisation members, data sharing tools, internal processes, different hardware units (mobile phones, tablets, laptops, etc.) and more. The release of data generated by large organisations as open data has been a great source of information for researchers, facilitating innovation and advances in various areas and encouraging collaboration to bring new technology, knowledge, and capabilities to solve problems (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).

Public services, like non-governmental organisations (NGO) or governmental organisations (GO), have recognised the benefits and competitive advantages of open data concepts. Open data initiatives in these organisations have resulted in greater availability of data, improved efficiency and effectiveness, improved decision making, increased transparency, accountability, citizen participation in NGOs, and the creation of economic and social value (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022). The reuse of open government data will create billions of euros in economic value (European Commission and Directorate-General for the Information Society and Media, 2002). One example is the move of the Obama administration to increase access to government data, by launching data.gov, to increase the visibility and legitimacy of government data (Van Schalkwyk & Verhulst, 2017).

In addition to the benefits of sharing organisational data, there are also risks and drawbacks, such as the exposure of sensitive and private information, if not properly shared (J. Zhang et al., 2020; Navarro-Arribas et al., 2012). An example of exposing sensitive data is the Netflix Prize. It was an open competition for the best collaborative filtering algorithm to predict user ratings for movies, based on previous ratings without any other information about the users or movies. The participants produced algorithms that improved the recommendation system by up to 10% per year (J. Zhang et al., 2020). In 2007, researchers were able to identify individual users by matching Netflix data sets with movie ratings from the Internet Movie Database (IMDB), leading to the cancelation of the competition due to privacy concerns (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022).

Based on the previous examples, it is evident that data privacy is an important aspect when it comes to sharing organisational data. It is necessary to protect people, institutions and organisations (data subjects) following laws and ethical rules during

the data life cycle (collecting data, processing and analysing data, publishing and sharing data, preserving data, reusing data) (Ergüner Özkoç, 2021).

Many different organisations such as the European Union, PDPC Singapore, CERN, and others have created guidelines for sharing data. This research focusses on defining a framework and extracting best practises from the previously mentioned guidelines and existing frameworks for publishing organisational data as Open Data.

4.3 Data Lift Framework

Based on (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022), relevant tasks have been identified, adapted, and synthesised into a framework to transform organisational data into open and sharable data. Figure 4.1 illustrates the main steps of this framework including 'Define the Purpose And Scope of Data', 'Data Classification', 'Risk Assessment', 'Data Transformation' and 'Anonymization, Evaluation, and Publishing'.

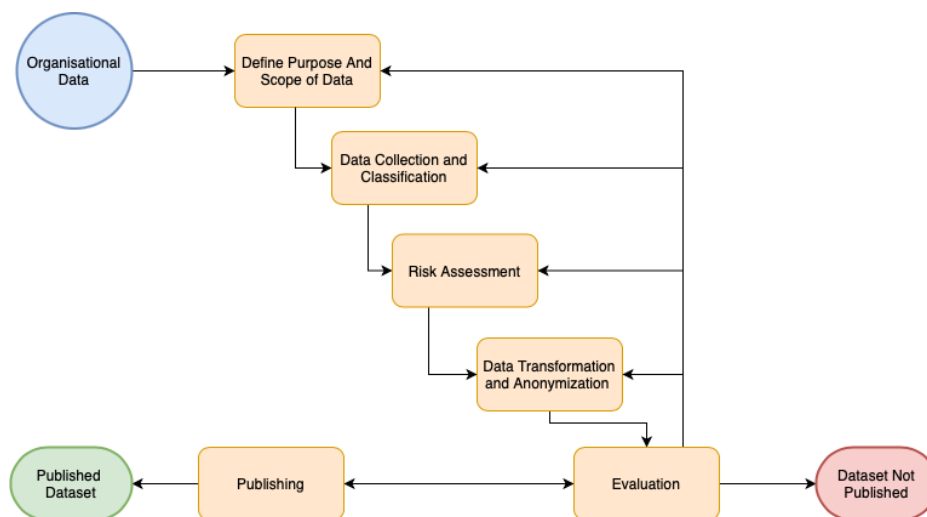


Figure 4.1: DataLift Framework Steps - Organisational Framework for Open Sourcing Data (taken from (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022))

4.3.1 Define Purpose And Scope of Data

The first step to compile organisational data into open data is determining why and for which data should be distributed and for how long. Answering the following questions allows one to formulate a clear and concise plan with a specified purpose and scope for the data:

4 Data Analysis and Exploration

- Q1: What is the intention of its collection and processing?
- Q2: Which type(s) of data is being processed (e.g. machine information, user input data, user data, sensor information, etc.)?
- Q3: To which audience (public or internal organisational shareholders) will the data be distributed?
- Q4: What is the data retention period (e.g. GDPR suggestion is 6 years, indefinitely for internal data)?
- Q5: Where applicable, are there details regarding transfers of data (e.g. what are the necessary actions before moving the data to a different repository or a new governing body)?

4.3.2 Data Collection and Classification

In the second step, it is required to collect the data and evaluate and classify the data according to the level of sensitive attributes. Before collecting the data, it is necessary to determine which data format (e.g. pdf, CSV, XML, JSON, etc.) to use and where to temporarily securely store the data before publishing. The selection of the format and storage depends on the organisational requirements. Data can be collected from multiple sources such as newly generated data or data from another internal or external source, which implies that it can even be in different formats. The data collection step focuses on aggregating data from different formats and transforming them into a single predetermined format.

Metadata information, such as datatypes, should be assigned to data attributes. It is recommended that the value of the property is taken from a well-governed set of resource types, such as the Dublin Core Metadata Initiative (DCMI) ¹. In addition to basic datatype information, metadata should include additional up-to-date information, such as context, qualities, and meaning of each attribute of the dataset. Based on table 4.1, it is necessary to assign privacy classification classes that follow the definitions for Identifiers (ID), Quasi-identifiers (QID), Sensitive attributes (SA), and Insensitive attributes (IA), to data attributes from the mentioned dataset.

Under QID we understand identifiers that combined identify an entity but separated do not, an example is the combination of gender, birth dates, and postal codes. Data attributes such as address, salary, political affiliation, and more can be classified as sensitive attributes since they disclose sensitive information about an individual in the case that the individual is identified. Elements that cannot be linked to individuals, such as the count of entities or publicly available data, can be classified as non-sensitive data.

Data containing IDs, QIDs or SAs are classified as sensitive data and require additional transformations for publication, while data with IAs do not need additional

¹<http://dublincore.org/documents/dcmi-terms/section-7>

Type	Description
Identifiers (ID)	Information that uniquely and directly identifies individuals (e.g. full name, driver licence, and social security number, etc.)
Quasi-identifiers (QID)	Identifiers that, combined with external data, lead to the indirect identification of an individual (e.g. gender, age, date of birth, zip code, etc.)
Sensitive attributes (SA)	Contains data that is private and sensitive to individuals, such as sickness and salary (e.g. medical records, bank records, etc)
Insensitive attributes (IA)	Contains general and non-risky data that are not covered by other attributes (e.g. web site visits, number of likes of a post, etc.)

Table 4.1: Privacy data classification (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022)

data transformation. The result of this step is aggregated data that has been stored in a unified predetermined format together with corresponding metadata information (in the form of a metadata repository or a data dictionary), that has also been classified based on data attributes into sensitive and non-sensitive data.

4.3.3 Risk Assessment

In this step, the organisational risk of disclosing data and the risk of data disclosure to the individual is assessed. The main risk assessment dimensions are the strategic, the economic, the legal, and the technical and organisational.

Strategic Dimension

The goal of this dimension is to determine the strategic risks of releasing data. Organisations have different types of data, that can contain sensitive and private information, but also data that lacks any harmful organisational information. Data such as machine or sensor data that do not contain any sensitive information could be freely available, without damaging the organisation. However, some company data contain sensitive attributes such as employee date of birth, address, personal identification numbers, and more. Table 4.1 describes data attribute classification based on the level of sensitive information that it contains. If the data contain any ID, QID, or SA it is necessary to analyse the data from different strategic perspectives (organisational learning and growth, organisational processes, user perspectives, etc.). One such perspective is the organisational reputation; for example, providing data with these attributes, without any access limitations or protection, can damage the reputation of an organisation or its members. Another strategic perspective is competition risk,

4 Data Analysis and Exploration

does releasing such data put the organisation at risk from competitors (e.g. abuse of methods used in an organisation or user poaching).

Economic Dimension

When analysing data from an economic perspective, it is necessary to estimate the economic risks of sharing the data. Data produced in companies is often a byproduct or day to day workflows in organisations, which makes sharing data a low expense process. However ensuring secure usage of the correct data and data governance can be a costly process. With the appearance of big data analytics, sharing huge data repositories free of charge, can result in a economic loss for organisations, since the produced data contains economic value. Contrary, many studies have also indicated that the development of information openness can stimulate innovative activities, the creation of innovative approaches, greater performance, and greater economic benefits for organisations.

Legal Dimension

Legal issues like data licencing, sensitive user information regulations (GDPR), storing and distributing regulations, and others that have legal implications have to be analysed. The objective is to determine the legal risks that arise from open sourcing data and possible ways to mitigate them. One example is determining the usage of proper open licences (MIT, Apache, etc.) for certain types of organisational data (primarily non-sensitive and anonymous data) and the legal risks of these licences.

Technical and Organisational Dimension

When analysing the data, it is necessary to determine the technical and organisational implications of releasing organisational data. Organisational implications such as the question if it is necessary to invest additional staff members to maintain the dataset, how difficult is it to gather the data from different sources within the organisation, what are the necessary technical skills needed to release the data, are there staff members that are capable of performing the release without exposing the organisation to risks (e.g. data loss and security leaks). On the other side, technical implications are how to publish the data, anonymise the data, ensure data quality, low error rate, machine readability, and continuity of access.

Quantitative Risk Rating (QRR)

Quantitative Risk Rating (QRR) is a tool to calculate a numerical value for existing and future risks by estimating the level of risk with the use of the likelihood of the risk arising and the severity of the injury.

The steps for the calculation of the quantitative risk rating are:

1. Assign a likelihood to a risk dimension from the following values Highly Probable, Probable, Possible, Unlikely, and Rare (e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations is highly probable).
2. Assign a severity of injury value to a risk dimension from the following values

4.3 Data Lift Framework

	Very Low	Low	Medium	High	Very High
Highly Probable	Moderate	High	Severe	Severe	Severe
Probable	Moderate	High	High	Severe	Severe
Possible	Minor	Moderate	High	High	Severe
Unlikely	Minor	Moderate	Moderate	High	High
Rare	Minor	Minor	Minor	Moderate	Moderate

Figure 4.2: Risk Dimension Evaluation Matrix (Jakovljevic, Gütl, Wagner, & Nussbaumer, 2022)

Very Low, Low, Medium, High, or Very High (e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations has a very high severity of injury value).

3. Allocate data attributes that contribute to the risk dimension.(e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations; user attributes such as address, DOB, personal identification number can be assigned to this risk dimension)
4. Based on the likelihood, severity, and the table in figure 4.2 determine a risk level value (Minor, Moderate, Major or Severe) for the risk dimension (e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations; based on the table in figure 4.2 and the previously stated likelihood and severity, the risk level for this strategic dimension is Severe).

After each dimension has been assigned a risk level, it is necessary to calculate the risk level for the whole dataset and each data attribute. For the dataset, the risk level is the average of all risk levels of each dimension (e.g. if the dataset has only the strategic and economic risks, then the risk level of the dataset is the average of those two dimensions). For each data attribute, the average of each occurrence in a risk dimension is calculated. For example, if a data attribute is assigned to the strategic dimension only then the risk level of the attribute is the risk level of the dimension. Additionally, if the data attribute is assigned to the strategic and economic dimensions, the risk level of the data attribute is the average risk level of these dimensions.

The steps for the calculation of the quantitative risk rating are:

1. Assign a likelihood to a risk dimension from the following values Highly Probable, Probable, Possible, Unlikely, and Rare (e.g. Strategic Dimension - Combining

4 Data Analysis and Exploration

multiple sensitive data attributes to identify individuals in organisations is highly probable).

2. Assign a severity of injury value to a risk dimension from the following values Very Low, Low, Medium, High, or Very High (e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations has a very high severity of injury value).
3. Allocate data attributes that contribute to the risk dimension.(e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations; user attributes such as address, DOB, personal identification number can be assigned to this risk dimension)
4. Based on the likelihood, severity, and the table in figure 4.2 determine a risk level value (Minor, Moderate, Major or Severe) for the risk dimension (e.g. Strategic Dimension - Combining multiple sensitive data attributes to identify individuals in organisations; based on the table in figure 4.2 and the previously stated likelihood and severity, the risk level for this strategic dimension is Severe).

After each dimension has been assigned a risk level, it is necessary to calculate the risk level for the whole dataset and each data attribute. For the dataset, the risk level is the average of all risk levels of each dimension (e.g. if the dataset has only the strategic and economic risks, then the risk level of the dataset is the average of those two dimensions). For each data attribute, the average of each occurrence in a risk dimension is calculated. For example, if a data attribute is assigned to the strategic dimension only then the risk level of the attribute is the risk level of the dimension. Additionally, if the data attribute is assigned to the strategic and economic dimensions, the risk level of the data attribute is the average risk level of these dimensions.

4.3.4 Data Transformation and Anonymization

This step focuses on determining which method to use for data anonymization and transformation. Table 4.2 aggregates the knowledge produced from the literature study and previous steps (Section 4.3.2 and Section 4.3.3).

Based on the quantitative risk level calculated in Section 4.3.3 and the level of data sensitivity of the data attributes, the anonymization and transformation methods are determined for the data attributes. According to the literature survey mentioned in Section 4.3.2, data attributes containing sensitive information should be anonymized using randomization, cryptographic methods, k-anonymity, l-diversity, and t-closeness methods. In the case that they pose a highly severe risk, these attributes need to be removed from the dataset. For sensitive attributes that pose a low or moderate risk, simple anonymization methods, such as randomization, are suggested. For attributes with a moderate or high risk level, it is suggested to use k-anonymity, l-diversity, or t-closeness methods for anonymization. The drawbacks and benefits of k-anonymity,

	Data Sensitivity Level	Risk
None	Non-Sensitive	Low
Randomization	Non-Sensitive / Sensitive	Low-Moderate
Cryptographic	Sensitive	Low-High
k-anonymity l-diversity t-closeness	Sensitive	Low-High
Remove	Non-Sensitive / Sensitive	High-Severe

Table 4.2: Data Anonymization Methods Matrix

l-diversity, or t-closeness methods are mentioned in Section 4.3.2 and can be used to select the appropriate method. It is also worth mentioning that applying more complex methods of anonymization is possible on data attributes with low-risk levels, but it results in a waste of resources and loss of information.

The purpose of applying the correct method is to ensure privacy for sensitive information, reduce information loss, and reduce the effort necessary to publish non-sensitive and low-risk information.

4.3.5 Evaluation

After the data transformation and/or data anonymization step mentioned in Section 4.3.4, it is necessary to evaluate the resulting dataset. The main question to answer in this step is:

- **RQ4.1:** How to compile organisational datasets into open data and guarantee anonymization?
- **RQ4.2:** What are the benefits and drawbacks of the proposed framework for organisations?
- **RQ4.3:** Does the newly created dataset mitigate risks (e.g. privacy risks, strategic risks, economic risks, legal risks, etc.) and fulfills the purpose and scope defined in the initial step of the framework (Section 4.3.1)?

For the evaluation of the data, it is necessary to review the dataset on quality, check the data on timeliness and consistency, evaluate the dataset on the use of standards, and evaluate the dataset on technical openness. To precisely answer the mentioned main question and to ensure that the dataset confides to the minimal requirements for publishing, it is necessary to evaluate the features mentioned in Table 4.3.

If the resulting dataset does not mitigate risks and fulfills the purpose and scope defined earlier, it is necessary to iterate back to a previous step. This can be a return to

4 Data Analysis and Exploration

Feature	Description
Discoverability	Existence of descriptors and metadata to ensure that the dataset can be discovered
Format	Machine-readable non-proprietary data file format to support data reading and processing
Validity	Clearly defined time of validity of the data and defined scope of data validity
Quality	Clearly defined data attributes without missing values
Granularity	Defined data granularity with low levels of aggregated data

Table 4.3: Features for Dataset Evaluation

selecting a new method for data transformation, establishing a new purpose or/and scope for the data, or reevaluating the risk. Otherwise, the data is ready for the publishing step. Based on the quantitative risk rating of the legal dimension mentioned in 4.3.3, the Open Source Licence¹ needs to be carefully selected for the dataset. It is important that the licence does not expose the data to unnecessary risk while being in line with features mentioned in Table 4.3.

As shown in Figure 4.1, if the evaluation yields that the data prepared for publishing does not conform to the standards, it is necessary to determine which step of the framework could mitigate the problems and return to that stage. If it is impossible to mitigate the problem, the dataset is not published and the process is discontinued. Besides the publishing step, the evaluation step is the only step from which you can navigate back to a previous step to minimise problems or to change the scope and goal of data publishing.

4.3.6 Publishing

Before publishing, it is necessary to prepare descriptive information about the data in human-readable and machine-readable formats to ensure discoverability and validity. It should contain a comprehensive description (e.g., sources, entities, metadata information, licence information, etc.), a self-explanatory title, privacy declaration, contact information and the information related to the scope and purpose defined in 4.3.1. When adding metadata to the dataset and data attributes, the use of standardized metadata schema by public bodies such as the W3C Data Catalog Vocabulary (DCAT) is recommended to ensure the dataset respects the FAIR principles. To ensure that

¹<https://opensource.org/licences/category>

4.4 Application of the Data Lift Framework on CERN Mattermost Dataset

issues detected with the dataset are mitigated in time, it is necessary to determine the contact entity (e.g. person, organisation, group, etc.). The role of this entity is to monitor the published dataset and ensure that it conforms to the predefined scope and goal of the dataset while adhering to the FAIR principles, ensuring that the data are discoverable and valid. The next step consists of determining the correct ODR for publishing and preparing data governance rules. The final step is publishing the data with all meta-information to an ODR. As mentioned previously, the publishing step results in a published dataset, but it is necessary to continue monitoring the dataset. In the case that issues are detected in the publishing step, it is necessary to return to the evaluation step to determine if the dataset issues can be mitigated or if the issues are critical to the level that the publication has to be canceled.

4.4 Application of the Data Lift Framework on CERN Mattermost Dataset

This section focuses on the evaluation of the framework and the evaluation of the framework application. Additionally, the focus is to answer the following research questions:

- What are the benefits and drawbacks of the proposed framework for organisations?

The framework was applied to the CERN Mattermost Dataset, which contains information about teams, channels, message threads, user messages, user reactions, basic user information and entity relationship information (Jakovljevic, Wagner, Gütl, Pobaschnig, & Mönnich, 2022).

One part of the evaluation consisted of a structured interview containing open-ended questions. Six in-person expert interviews were conducted, where the participants were young professionals between the ages of 25 and 29, working at CERN in the IT department as full-stack developers. The participants needed to read and analyse all steps of the framework. Afterward, they were presented with the results of the framework applicable to the CERN dataset. The results were discussed and they had to describe the drawbacks and benefits, general feedback, and statements about the framework and its use within organisations by answering the following open-ended questions per step:

- Did you find any issues with the current step? If yes, what are they?
- Do you see any positive aspects of this step? If yes, what are they?
- Do you have any general comments about this step?

Additionally the participants were asked to answer the following questions:

4 Data Analysis and Exploration

- What is your general opinion about the Framework?
- What are possible improvements to the framework, ideas, future work or additional steps?

The second part of the evaluation consisted of a user survey, where the participants were first asked to read and analyse all steps of the framework. The survey then focused on determining the level of expertise the participants had with open data, publishing organisational data, and privacy. Subsequently, users were requested to provide demographic data. Each step of the framework needed to be evaluated for possible issues and benefits, additionally, the participants were asked to rate the level of understanding of the step. The survey ended with two general open-ended questions:

- What is your general opinion about the Framework?
- What are possible improvements to the framework, ideas, future work or additional steps?

4.4.1 Expert Interview

The participants had a positive opinion of the framework and data produced by it. All participants agreed that that it is a well structured framework with several iteration loops to re-evaluate the data before publishing.

Define the Purpose And Scope of Data

Interview participants expressed that this step was understandable and provided a solid start to the process of open-sourcing data. The defined questions are clear and objective, while allowing for generalization. The participants also stated that some of the points made seemed more related and focused on why the organisation is gathering data than on open-sourcing the data.

Data Collection and Classification

Although the participants stated that this step provides clear and understandable introductions on how to classify and identify private/public information, concerns were identified. Concerns like specification on what seems to be possible subjectiveness. Private and sensitive seems very tied with common sense that might vary between organisations and people that might try to apply this framework.

Risk Assessment

According to the participants, this step contains understandable descriptions of all relevant dimensions required for risk assessment, with sufficient descriptions to understand the goal. It was also stated that this step was the least understandable but also

4.4 Application of the Data Lift Framework on CERN Mattermost Dataset

the most complex. There are whole departments dedicated to risk assessment. It was also stated that companies/organisations find gray areas with risk assessments, and without strict and careful guidelines, this can lead to data abuse.

Data Transformation and Anonymization

The participants identified that the risk matrix and method selection matrix provides a simple but effective way to determine which data and how to anonymize. They also recognized that other methods could also be used for anonymization and that specific cases could not be covered with the use of the previously mentioned tools.

Evaluation

Even though the participants rated this step as crucial and highly beneficial, it was suggested, to avoid any sort of bias, that some evaluation parameters should be defined during the first step. The evaluation parameters should be clearly defined and presented as a checklist or a set of specific questions to answer.

Publishing

The participants stated that this step contains very clear points on how to publish the data, from elements (description, title, etc.) to references of a vocabulary. They also stated that having a source to consult makes it easier to follow the whole process of open-sourcing data.

4.4.2 User Study

In total, 11 individuals from CERN and Graz University of Technology participated. Of the participants, 10 (90.9%) were male and 1 (9.1%) female, with 9 (81.81%) holding a Master's degree and 2 (18.19%) holding a Bachelor's Degree, and all 11 participants were in the age range of 25-34 years old. In general, the respondents were mainly relatively young and educated males.

The participants were also asked about their general knowledge related to open data, publishing organisational data and privacy, which generally showed a moderate level of technological orientation and knowledge.

Based on the input of the participants as seen in Table 4.4, all steps of the framework are easy to understand. On average, each step was rated as very good or easy to understand, with steps such as "Risk Assessment" and "Data Transformation and Annonimization" being the most difficult steps of the framework, and "Data Collection and Classification" being the easiest step of the framework to understand.

4 Data Analysis and Exploration

	Poor	Fair	Good	Very Good	Excellent
Define the Purpose And Scope of Data	0 (0%)	0 (0%)	1 (9.1%)	6 (54.5%)	4 (36.4%)
Data Collection and Classification	0 (0%)	0 (0%)	1 (9.1%)	7 (63.6%)	3 (27.3%)
Risk Assessment	1 (9.1%)	0 (0%)	0 (0%)	9 (81.8%)	1 (9.1%)
Data Transformation and Annonimization Evaluation	0 (0%)	2 (18.2%)	1 (9.1%)	7 (63.6%)	1 (9.1%)
	0 (0%)	1 (9.1%)	1 (9.1%)	7 (63.6%)	2 (18.2%)
Publishing	0 (0%)	0 (0%)	3 (27.3%)	4 (36.4%)	4 (36.4%)

Table 4.4: User Rating of Understandability per Framework Step

Define the Purpose And Scope of Data

Based on the answers of the participants, the general opinion is that the step contains defined objectives while allowing space for generalization, and enabling even non-expert users to understand it. The main drawbacks that were identified by the participants are: Following such a framework might be a huge endeavor for companies; There should be an initial description of how the data should be treated (defined data lifecycle); The questions focus on data rather than open-sourcing, some points that are made are more related on why is the organisation gathering data than why is it publishing the data.

Data Collection and Classification

This step has been evaluated as overall good and informative. Defining the type of data seemed clear and well described, making it uncomplicated to identify what is and is not sensitive data. Although it lacks some clarification on what can be subjective interpretations. The classification of data seems to be tied with common sense that might vary between organisations and people. Participants recommended including an annex/document where there is a better distinction between what is private and public, sensitive or not, or tools that can help perform such collection/classification.

4.4 Application of the Data Lift Framework on CERN Mattermost Dataset

Risk Assessment

The participants agreed that this step contains understandable descriptions of all relevant dimensions required for risk assessment, with sufficient descriptions to understand the goal of each step. It serves as a great template for risk assessment in this context and makes it simpler to understand the risk areas of publishing data. However, there is no objective interpretation of what is "very low" to "very high" risk and could lead to invalid risk assessments.

Data Transformation and Anonymization

The step was informative for the participants and in direct relation to the data classification step. The selection of the correct anonymization methods that best fit the risk was clear and the explanations for the selection were well-defined. Some participants raised the issue of de-anonymization attacks.

Evaluation

This step provides a clear and easy-to-understand explanation of how to evaluate the results of the previous steps and simplifies the process of determining possible side effects of publishing data. It was also suggested to avoid bias and mistakes to predetermine evaluation steps (or part of them) during the initial step. Additionally, the mentioned standard in this step should be followed with annexes, explaining them in detail to make this step more transparent.

Publishing

The main characteristics of this step, according to the participants, are close connections with open standards and schemas, which are the pillar of open-sourcing data. Contains very clear points on how to publish the data, from elements (description, title, etc.) to references of a vocabulary that can be used. It was suggested also to create descriptive information about the data in the initial step and validate them only in the publishing step to make the whole process easier and to have a source to consult. Additionally, it should be clearly stated that once the data is published on the internet it is almost impossible to delete.

Main Improvements

The main improvement that was suggested by the participants is to include concrete use-cases and helpful tools to give users a better sense of how to practically use this framework.

4.5 Summary

This chapter is an important part of the research, as it provides answers to questions related to handling and leveraging sensitive organisational data while upholding privacy and ensuring reproducibility.

In response to the first research question, "How can organisational datasets be compiled into open data while ensuring anonymisation?" the chapter introduces and describes six phases of the DataLift framework, consisting of 'Define the Purpose and Scope of Data', 'Data Classification', 'Risk Assessment', 'Data Transformation' and 'Anonymisation, Evaluation, and Publishing'. Within this methodical presentation, the process of transitioning sensitive data into open data is explained, with a particular emphasis on the critical element of anonymisation. This chapter, therefore, lays the groundwork for a crucial element of the overarching research question: "How can sensitive organisational data be harnessed for reproducible research and development?" The findings highlight the impact of the DataLift framework in compiling open data organisational datasets while protecting data privacy and security.

For the second research question, "What benefits and drawbacks does the proposed framework offer organisations?" the chapter conducts an exhaustive analysis of the CERN dataset published via the DataLift Framework. The proposed framework for data publication offers both benefits and drawbacks for organisations, as highlighted in the user study conducted to assess in Section 4.4.2. One significant advantage of the framework is its clarity and ease of understanding. According to the participants, all the steps within the framework were rated as very good or excellent in terms of understandability. This implies that the framework is accessible to a wide range of users, including those without extensive technical expertise, making it a valuable resource for organisations. Another notable benefit is the inclusion of a comprehensive risk assessment step. The participants found that the framework provides clear descriptions of the dimensions required for risk assessment. This objective risk assessment can help organisations make informed decisions about data publication, ensuring that potential risks are carefully considered. The framework also offers guidance on data transformation and anonymisation, which was well received by participants. It simplifies the process of selecting appropriate anonymisation methods based on the classification of data and associated risk levels, addressing a critical aspect of data preparation. Additionally, the framework's emphasis on standardisation and adherence to open data principles is advantageous. It aligns organisations with open standards and schemas, promoting compatibility and interoperability with other open data initiatives and platforms.

Based on Section 4.4.2 one of the main drawbacks observed by the participants is the potential resource-intensive nature of implementing the entire framework. Some organisations, especially those with limited resources, may find it challenging to follow all prescribed steps effectively. The participants also highlighted the need

for an initial description of how data should be treated in the data lifecycle, which is currently lacking in the framework. This omission leaves organisations without clear guidance on data handling from the outset. Subjective interpretations of what constitutes sensitive data in the "Data Collection and Classification" step were another concern, this ambiguity could lead to privacy vulnerabilities in the produced results.

The third research question, "Are the newly created datasets effective in mitigating risks (e.g. privacy risks, strategic risks, economic risks, legal risks, etc.) and do they fulfil the initially defined purpose and scope of the framework?" delves into the topic of risk mitigation and dataset utility. Based on the evaluation and assessment of the DataLift framework through Sections 4.4.1 and 4.4.2, the participants generally found the framework helpful in mitigating the risks associated with data publication. The inclusion of a comprehensive risk assessment step was particularly appreciated, as it provided a structured approach to identify and evaluate potential risks. The participants agreed that the framework offers clear descriptions of relevant dimensions for risk assessment, enabling organisations to understand and address risk areas effectively. However, a notable drawback highlighted by the participants was the absence of objective criteria for defining risk levels, such as "very low" to "very high." This lack of clarity could potentially lead to variations in risk assessments and undermine the effectiveness of risk mitigation strategies. Addressing this problem by providing specific criteria for the determination of risk level is a suggested improvement. The step related to data transformation and anonymisation was well received by participants. They found it informative and practical, facilitating the selection of appropriate anonymisation methods based on data classification and associated risk levels.

In summation, the chapter on "Data Analysis and Exploration" offers insights into the utilisation of sensitive organisational data. Its findings not only enrich the development of the DataLift framework but also provide a comprehensive understanding of the balance between data openness, privacy preservation, and the reproducibility of research and development. These findings improve existing research, underlining the importance of privacy-aware data management and its implications for the development of personalised recommendation and information retrieval systems.

5 Applicability of Social Media Elements in Notification Systems

In today's digital age, social media platforms have emerged as prominent avenues for social interaction, capturing the attention of a significant portion of Internet users. However, the extensive amount of information presented on these platforms often leads to information overload, where users struggle to process the overwhelming influx of content. This chapter delves into the applicability of Social Media Elements (SME) in notifications as a potential solution to mitigate the negative effects of information overload. Through an investigation of the integration of SMEs into notification systems, this chapter aims to describe improvements in the credibility and clarity of notifications with the use of SMEs. In addition, this chapter aims to open up new possibilities for employing social media applications in notifications and other domains where users' cognitive abilities and the detrimental effects of information overload are of concern.

The following sections are based on and supported by the work published in the following publications:

- **Jakovljevic, I.,** Gütl, C., & Wagner, A. (2022). Analyzing the Effects and Applicability of Social Media Elements in Notification Systems in Large Interconnected Organisations. IARIA JOURNALS
- **Jakovljevic, I.,** Wagner, A., & Gütl, C. (2021). Applicability of Social Media Elements in Notification Systems in Large Interconnected Organisations. In SOTICS 2021: The Eleventh International Conference on Social Media Technologies, Communication, and Informatics

5.1 Contribution

This chapter plays an important role in the research methodology of the thesis, particularly within the "Data Gathering and Data Analysis" step. In this phase of the research, the primary aim is to collect relevant data, analyse them rigorously, and draw meaningful insights that contribute to addressing the core research question: "How can sensitive information be used in a privacy-preserving way for the creation of a personalised recommendation and information retrieval system, and what is the performance of such a system compared to traditional systems?"

5 Applicability of Social Media Elements in Notification Systems

This chapter is designed to explore the integration of social media elements within notification systems, focussing on their applicability and effectiveness. By examining the inclusion of social media elements, the research seeks to gather valuable data related to user interactions, preferences, and behaviours within these systems. The analysis of these data points serves as a critical component in understanding how sensitive information can be used while preserving user privacy.

Finally, the insights generated from this chapter will contribute significantly to shaping the privacy-preserving recommendation and information retrieval system. By evaluating the impact of social media elements, the research aims to provide valuable recommendations and findings that can inform the design and implementation of systems that balance personalization, privacy, and performance. This, in turn, will provide substantial contributions to the overarching research question and enhance the understanding of the intricate dynamics between user data, recommendation systems, and privacy preservation.

5.2 Social Media Elements

Based on the analysis of social media sites and extensive research on various aspects of social media, including hashtags, microblogs, content approval/disapproval, user groups, user-to-user relationships, and user identity, the most common elements have been carefully identified, summarized, and organised. These key elements have been compiled and presented in Table 5.1, offering a comprehensive overview of the fundamental components that shape social media platforms and their functionalities. This table serves as a valuable resource for understanding the core elements of social media and their significance in the realm of information technology and computer science (Correa & Sureka, 2011; F. F. Li et al., 2020; Kietzmann et al., 2011).

Microblogs

Similar to microblog posts, notifications are messages displayed to the users with the intent to share information. These messages can contain information from different applications (e.g., email subject and part of email text, new message alert). Based on the above description, it can be concluded that notifications share similarities to microblog post entries. However, unlike notifications that do not contain much additional information in their visual representation, microblog posts can contain aspects of social media, such as hashtags, group information, content approval/disapproval, and others. These elements allow the users to determine the importance and validity of a post. Aspects such as the number of individuals that have shared, liked, or approved the post, topics related to the shared post, and the type of individuals that have interacted with the

post are of crucial importance to assess the value of the post and the information within (Efron, 2010; Zamberi et al., 2018).

Social Media Element	Description
Hashtags	A hashtag is a type of metadata tag used on social networks to help users find resources with a specific theme or content.
Microblogs	Microblog services allow users to post and share short textual messages that are then propagated to an audience, which can then quickly interact with the posts and between each other.
Content approval / disapproval	Social cues that send signals of social appropriateness or social acceptance of content to the content creator. Examples of these social elements are Likes, Retweets, Reactions, and more.
User Groups	User groups represent the extent to which users can form communities and subcommunities. The more 'social' a network becomes, the larger the group of friends, followers, and contacts.
User-to-User Relationship	User-to-user relationships express the extent to which users can relate to each other (e.g., friendships on Facebook or Followers on Twitter).
User Identity	It represents the degree to which users expose their identities on social media sites. It includes exposing information such as name, age, gender, profession, location, and other user identifiable information.

Table 5.1: Summary of main social media elements taken from (Jakovljevic, Gütl, & Wagner, 2022)

An example of a microblogging service is Twitter, one of the largest microblogging services with more than 300.000 posts generated daily. Twitter is classified as a social network because individuals can communicate and connect with each other to form social groups on Twitter. They form social groups by following each other or following trending hashtags and/or topics (Kwak et al., 2010).

Hashtags

A hashtag is a metadata tag type that is used on social networks to help users find resources with a specific theme or content. The content of hashtags can be dynamically

5 Applicability of Social Media Elements in Notification Systems

generated or user-generated and can only consist of letters, digits, and underscores. Hashtags are iconic features that enable easy retrieval of connected resources. They are also used to construct a personal word/hashtag vector space of a user by examining the users' linguistic expression. Hashtags are inserted into the existing user word vector space using co-occurrence information and evaluated to determine whether the newly constructed vector space represents the personal linguistic expression of the individual. These methods intend to represent individuals by learning about potential representations using hashtags. Different methods for learning semantic representations exist. Some of them are Word2Vec, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation, and Recurrent Neural Network Language Model (RNNLM). Besides identifying and representing user characteristics, hashtags are used to connect similar resources, by assigning tags to provide contextual information (Jakovljevic, Gütl, & Wagner, 2022; Bieniasz & Szczypiorski, 2019).

Hashtag Retrieval is an information retrieval methodology which aims to retrieve relevant hashtags for a given query from a collection of resources. Besides retrieval, an interesting topic for notifications is hashtag automatic annotation, where hashtags are generated based on content, with the goal to classify the content per topic. Automatic tag recommendation or annotation can improve the efficiency of text-based information retrieval systems. Due to the nature of hashtags it is possible to extract correlations between resources from different systems by exploring their hashtag representations (Correa & Sureka, 2011).

Content approval/disapproval

There are several approaches to provide a system with the necessary user feedback information. IR systems use explicit information through user feedback or implicit information through user monitoring to determine user interests. Unobtrusive user interaction monitoring identifies content potentially interesting for users, without interfering with the user's normal work activity. Monitoring systems also leverage heuristics to deduce negative examples from observed behavior. Providing and receiving feedback is also a fundamental component of participation in social media. In addition, the popularity of social media has enabled the use of rich user information from Facebook and other social networks to predict users' latent traits for recommendation. Based on the study mentioned above, users have expressed a need for more personalization in notifications; integrating likes or dislikes into a notification system as a means of collecting feedback from the user related to notifications could be beneficial for improving the satisfaction rate of users (Jakovljevic, Gütl, & Wagner, 2022; Pradhan et al., 2017; Sedhain et al., 2014).

User Groups

A widely discussed relationship group metric is the Dunbar Number, proposed by Robin Dunbar in 1992. He theorized that people have a cognitive limit that restricts the number of stable social relationships with other people to about 150. Social media platforms have recognized that many communities grow well beyond this number and offer tools that enable users management of memberships ?. The assumption that the vocabulary used to discuss a topic stays similar between different user communities and does not vary significantly over time directly suggests that it is possible to compute the overlap of topics of two or more communities. This community similarity can connect communities from different social networks (e.g., Facebook), facilitate information sharing between communities, and extract community interest. Furthermore, user groups and group behavior information infer social cues, including group information (e.g., number of people with the same interests who approved a notification or executed a specific action) in notifications could increase the credibility and information dissemination of notifications (Lorenz-Spreen et al., 2018; Jakovljevic, Gütl, & Wagner, 2022).

User-to-User Relationship

The type of relationships users form between each other determines what information exchanges between them. For example, when users form professional relationships online, the information exchanged between them will have professional content and high value, compared to friendly relationships where the information is of a different nature. User relationship information could be used in notification systems to determine the character of information presented to the user Kietzmann et al. (2011); Jakovljevic, Gütl, & Wagner (2022).

5.3 Research Study

5.3.1 Study Design

To investigate the impact of Social Media elements on notifications and notification systems, a user study was conducted to simulate user interactions with notifications and present various types of notifications to participants. To quantify participant engagement with different notification types and measure their interactions, the study aimed to stimulate participants' interest in cognitive tasks (such as reading and comprehending articles) while displaying notifications. The intention was for participants to perceive the notifications as unrelated to the evaluation, ensuring unbiased results. Furthermore, the study aimed to assess participants' comprehension of new concepts,

5 Applicability of Social Media Elements in Notification Systems

their preferences, and dislikes. The evaluation encompassed several components, including a user demographics and general knowledge questionnaire, the execution of predetermined tasks, rating the difficulty and information provided by the system following task completion, questions based on the System Usability Scale (SUS) and the Computer Emotion Scale (CES), a questionnaire for rating the importance of Social Media elements, and a feedback questionnaire.

The user study focused on providing insights and answering the following research questions:

- **RQ5.1:** Which elements of social media can be integrated into notification systems to display understandable and valuable notifications at a glance without explicitly disturbing the user?
- **RQ5.2:** Would users prefer to receive notifications with integrated social media elements like hashtags, topic keywords, source information, rating by other users, and groups information?
- **RQ5.3:** How do users react to notifications with additional information (hashtags, user group information, content approval/disapproval, and social media posts)?
- **RQ5.4:** Which emotions do users experience when receiving notifications with and without this additional information?

5.3.2 Settings and Instruments

The user study was conducted online, involving participants from Graz University of Technology and high schools in Austria and Kosovo. Employing an AB study design, the participants were divided into two groups, namely Group A and Group B. Each participant was instructed to independently perform a series of diverse tasks. Upon completing the tasks, they were prompted to fill out the aforementioned questionnaires and specific survey questions.

General Questionnaire

The general questionnaire contains questions listed in Table 5.2 that aim to identify the value and effects of additional information in notifications on the user. This questionnaire aims to provide insights to RQ5.1 and RQ5.2, by explicitly asking the participants about how they perceive SMEs in the notifications they received.

Article Feedback

The article feedback contains questions listed in Table 5.3 aimed to resolve if the users were able to determine and evaluate if the presented articles were fake or not. Since the survey was an AB survey, it is possible to use the feedback from this questionnaire

Question
GQ1: Did you find the additional information in the notification valuable?
GQ2: When I received notifications with additional information I was more confident in the notification?
GQ3: Rank the additional information by importance
GQ4: It was easier to understand the notification when I had additional information in the notification?
GQ5: Did the notification break your concentration while executing the task?

Table 5.2: General Questionnaire Questions

to evaluate whether notifications with additional information help determine the truthfulness of articles and how users react to notifications with additional information.

Question
Q1: Do you think that the article "Friends Reunion" is Fake or Real?
Q2: Do you think that the article "Instagram for Children" is Fake or Real?
Q3: Do you think that the article "People live in a 3D-Printed House" is Fake or Real?
Q4: Do you think that the article "3 Reasons Why You Should Stop Eating Peanut Butter Cups!" is Fake or Real?
Q5: Do you think that the article "Us Bacon Reserves Hit 50 Year Low" is Fake or Real?

Table 5.3: Article Feedback Questions

Computer Emotion Scale (CES)

The CES is used to assess the emotions of the participants, as it provides one of the most scientific ways to evaluate emotions. Anger, anxiety, happiness, and sadness are the emotions evaluated by the CES. The scale was used to answer RQ5.4 by determining the emotional influence of notifications on the participants, since it provides one of the most scientific ways for emotion evaluation Kay & Loverock (2008).

System Usability Scale (SUS)

The SUS is used to measure the Ease of Use (EOU) of a system. It consists of ten items designed to assess EOU on a 100-point scale. Since its creation in the 1980s, SUS has been extensively used in Human-Computer Interaction (HCI) research and practice to evaluate Information Technology (IT). It consists of a ten-item attitude likert scale that

5 Applicability of Social Media Elements in Notification Systems

gives a global view of subjective assessments of usability scale. It is used to determine whether participants would prefer to receive additional information notifications, which is directly correlated with RQ5.2 and RQ5.3. It provides a trustworthy evaluation tool for usability testing (Bangor et al., 2009).

SME Importance Rating

The SME Importance Rating was used to determine which SMEs from Table 5.4 were important for understanding and perceiving notifications, participants were asked to rate the importance of the elements mentioned above, from 1 (not important at all) to 5 (Very Important).

Social Media Element	Usability in Notification Systems
Hashtags	Quick access to topic information; Enables instant classification of notifications by topic; Linking external information to the notification
Microblogs	Social Media Posts provide information representation ideas for notification due to their similarity; Content Sharing does not have a direct use in notifications
Content approval/disapproval	Provide a way for the user to express interest
User Groups	Provide additional information and credibility of information based on the opinion of a group of users
User-to-User Relationship	Provide different types of additional information based on relationships with different users

Table 5.4: Usability of social media elements in Notification systems, taken from (Jakovljevic, Gütl, & Wagner, 2022)

Article Classification Questionnaire

It was used to determine if users concluded that the read article was a real article or fake news. After reading all articles and receiving notifications related to the articles, participants were asked which articles they thought were fake news articles and which were real articles.

The study was created using the CoDiS Survey Tool Jakovljevic (n.d.). The CoDiS Survey Tool is a web based evaluation tool which tracks and analyses participants' behavior while presenting specific assignments, displaying custom notifications, and displaying questionnaires to the participants. Figure 5.1 displays how the user interface of the CoDiS Survey Tool displays the articles, notifications and tasks for the participants. The user interface consists of a task view, where the participants can read

5.3 Research Study

the tasks related to the current article, seen in the top section of the image, above the title of the article and below the progress element. Below the article text are the user interaction buttons, that enable the user to execute article specific actions (e.g. share article on facebook, comment on article, etc.).

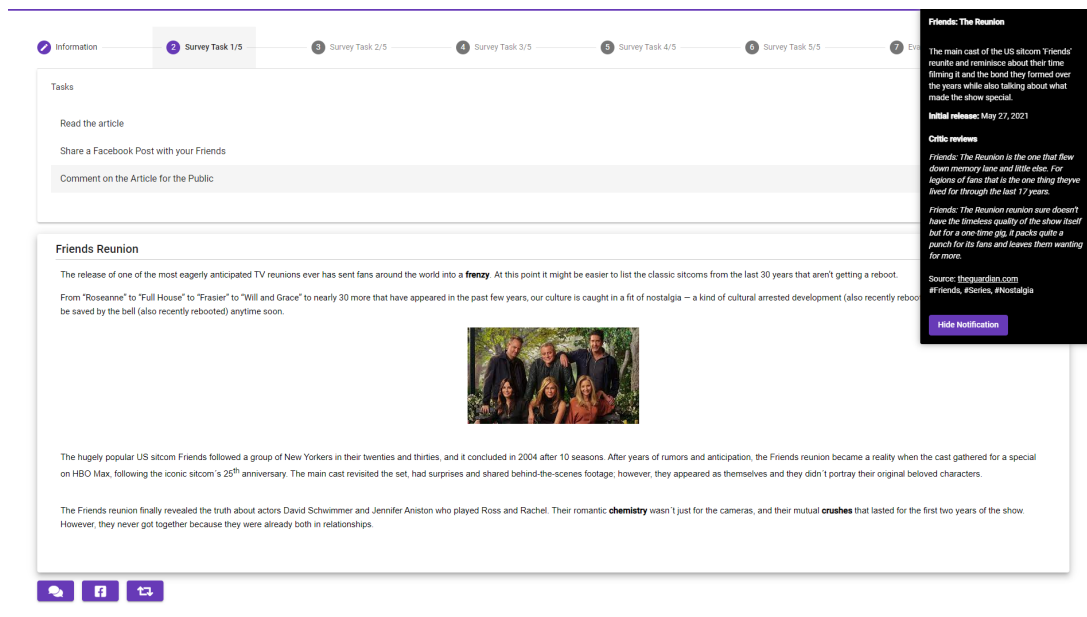


Figure 5.1: Codis Survey Tool User Interface

5.3.3 Procedure

Participants were asked to read articles mentioned in Table 5.5 and execute predefined tasks (share articles, comment on the article, and more).

#	Title	Is Fake
1	Friends Reunion	No
2	People live in a 3D-Printed House	No
3	Instagram for Children	No
4	US Bacon Reserves Hit 50 Year Low	Yes
5	3 Reasons Why You Should Stop Eating Peanut Butter Cups!	Yes

Table 5.5: Article Title and Validity

5 Applicability of Social Media Elements in Notification Systems

As the participants completed these tasks, the notifications related to the articles were displayed. Notifications were displayed as part of the CoDiS Survey Tool as web elements that appear when the user starts reading an article. Depending on the user group, these notifications were either with additional information or without additional information. The additional information included hashtags, user group information, and social media post formatting. This additional information integrates all selected social media elements from the previous chapter.

Figure 5.2 displays a standard notification instance with only the notification text and action button. While Figure 5.3 previews a notification with additional information.

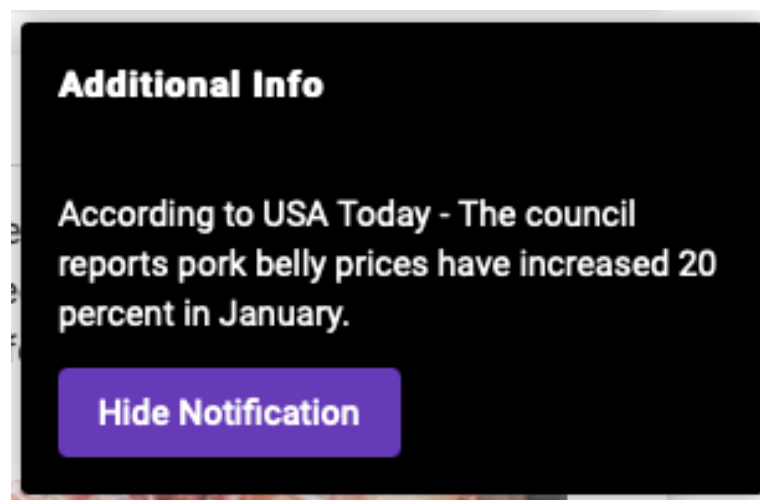


Figure 5.2: Simple Notification Information Display

Hashtag display is marked with the number 1 on the figure, while the number 2 marks group information (e.g. number of readers that validated and/or shared the article). The notification text source is enumerated with number 3, and the notification text is marked with the number 4.

5.3.4 Study Participants

The participant target groups for the study were high school and university students. In total, 215 individuals were asked to participate and only 35 completed the study. The age of the participants ranged from 15 to 34 years old, with 57.14% of the participants in the range from 15 to 20 years, 25.72 % in the range 20-25, 14.289 % in the range 25-30 and 2.85% in the range above 30 years old. Female participants made 28.57% of the total amount of participants, while male participants made 71.43%. As stated previously,

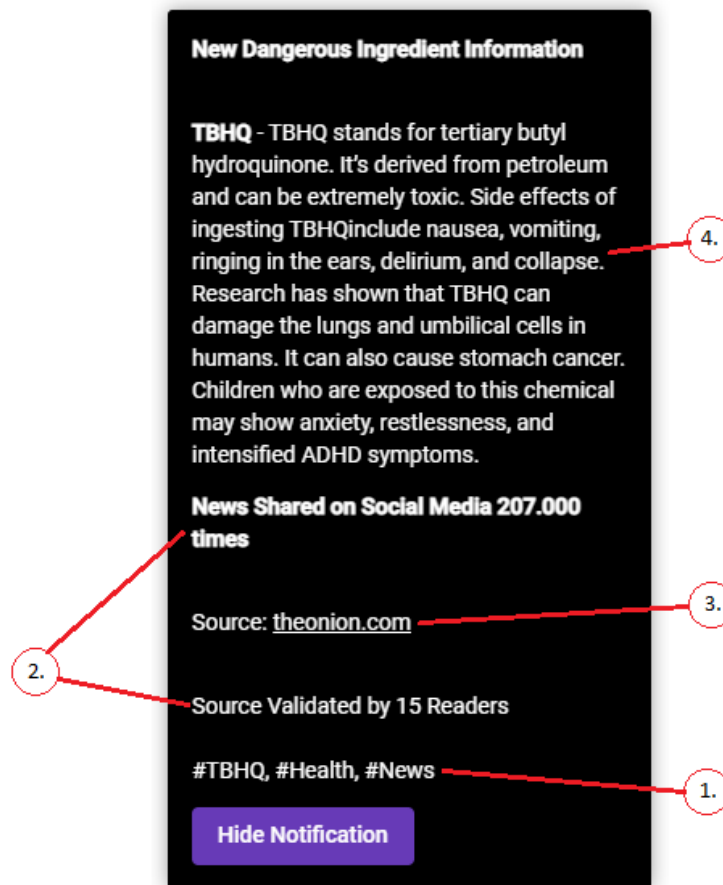


Figure 5.3: Notification with Additional Information Display

the study was designed as an AB study, which is why the participants were divided into two groups (Group A and Group B). The purpose of this division is to reduce bias between users. Both groups received the first article with additional information notifications. The purpose of this was to create a control article and familiarize the users with this type of notifications. Group A received simple notifications on even-numbered articles, while group B received them on odd-numbered articles. After the participants finished reading the articles and the article-related tasks, they had to complete an evaluation.

5.4 Findings and Discussion

In this section, we present a comprehensive overview of the findings and engage in a discussion of the implications from the study's results. To begin,

As seen in Table 5.6, 85.71% confirmed the assumption that additional information is valuable to notifications. The additional information in notifications has increased the value of notifications to the user based on the answers to GQ1 from Table 5.2.

Question	Yes	No
Did you find the additional information in the notification valuable?	30 (85.71%)	5 (14.29%)
When I received notifications with additional information I was more confident in the notification?	21 (60.00%)	14 (40.00%)
It was easier to understand the notification when I had additional information in the notification?	27 (77.14%)	8 (22.86%)
Did the notification break your concentration while executing the task?	21 (60.00%)	14 (40.00%)

Table 5.6: Questionnaire Results, taken from (Jakovljevic, Gütl, & Wagner, 2022)

Participants had more confidence in notifications with additional information compared to standard notifications. Table 5.6 shows that 60% of the participants voted that additional information increases the confidence of the notifications. Based on GQ3 from Table 5.2 77.14% of the participants stated that they find it easier to understand notifications with additional information. With 60% of the participants answering with "Yes" to GQ5 of Table 5.2, we can confirm that the notifications break user concentration, which validates the results of previous research (Jakovljevic, Gütl, & Wagner, 2022).

The success of notification systems is dependent on the information and information type they convey to the user. Access to internal and external information together with aggregation of different data sources is one of the main factors that increase the transparency, innovation, and productivity of large organisations. The survey participants agree with this as shown in Table 5.7. It reveals that users are predominantly concerned with the content and source of notifications. It implies that adding additional information to validate the content and source increases their value to users. The results in Table 5.7 also validate our proposal that formatting notifications as social media posts could improve the information presented to the user, since the content was formatted to be similar to a social media post. Contrary to our research, the group information (e.g., "22 readers validated text", "50 of your friends liked this article", "Expert A confirmed the validity of this post") was not ranked as highly important by the participants.

5.4 Findings and Discussion

Additional Information	Very Important	Not at all important
Information Source	10 (33.33%)	1 (3.33%)
Hashtags	2 (6.67%)	4 (13.33%)
Content of the Notification	10 (33.33%)	1 (3.33%)
Group or Reader Validation Info (e.g., "22 Readers Validated Text")	1 (3.33%)	10 (33.33%)
Notification Position	7 (23.33%)	14 (46.67%)

Table 5.7: Additional Information Ranking by Importance, taken from (Jakovljevic, Gütl, & Wagner, 2022)

As seen in Figure 5.4, the distribution of SUS answers reveals that most of the users agree or strongly agree with questions 2, 4, 6, 8, and 10 of the SUS (Bangor et al., 2009) while disagreeing or strongly disagreeing with the rest of the questions. It is also visible that questions related to the negative rating of the system contain a significant portion of neutral answers, compared to questions focused on positive system ratings. These results indicate that the participants have formulated a positive opinion about notifications with additional information and that they would use a system with this feature, while not explicitly agreeing that there are negative aspects in such a system.

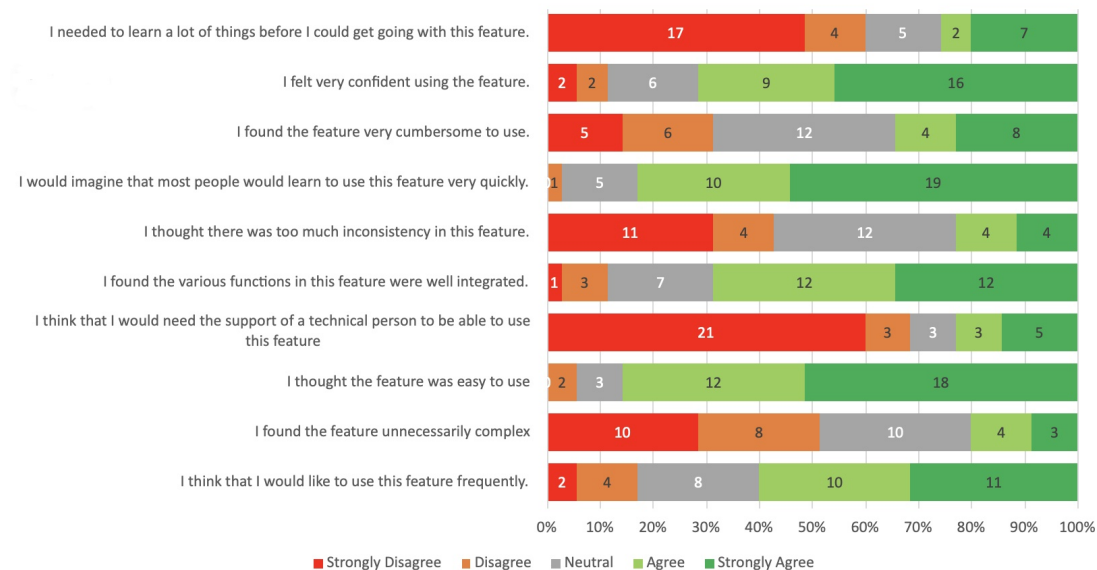


Figure 5.4: System Usability Scale Detailed Results

Due to the large number of neutral responses, the average rating of the SUS scale is 69.78. This is slightly above the limit of 68 set by Bangor et al. (2009) as the value that

5 Applicability of Social Media Elements in Notification Systems

is the minimum for a usable system. Based on the SUS results, we can infer that users would prefer to use a social media notification system.

The result of the CES is shown in Table 5.8, the table contains a list of feelings a participant has experienced. The CES shows that users were happy most of the time executing tasks and receiving notifications, while none of the time experiencing sadness, anxiety, and anger. According to Table 5.8, the emotion anxiety has the lowest score because most users rated it with "none of the time" followed by sadness and anger.

	None of the Time	Some of the Time	Most of the Time	All of the Time
Happiness	20.95%	31.43%	24.76%	22.86%
Sadness	68.57%	24.29%	4.29%	2.86%
Anxiety	69.29%	18.57%	8.57%	3.57%
Anger	65.71%	21.90%	7.62%	4.76%

Table 5.8: Computer Emotion Scale Answer Emotion Distribution

The emotion rated the best was "Happiness", where the majority of users answered with "Some of the Time", "Most of the Time" or "All of the Time". These results do not correlate with previous studies in which users experienced negative emotions and stress while receiving notifications.

As part of the evaluations, participants had to determine which articles were fake and which were real. The results of this evaluation are presented in Table 5.9. Besides the first article ("Friends reunion"), the participants could not distinguish fake from real. Only 57.93% of the cases were the articles correctly classified. Participants who received notifications with additional information classified articles with a 6.61% greater accuracy.

Article Name	Fake News	Real Article
Friends Reunion	2 (6.90%)	27 (93.10%)
Instagram for Children	13 (44.83%)	16 (55.17%)
People live in a 3D-Printed House	15 (51.72%)	14 (48.28%)
3 Reasons Why You Should Stop Eating Peanut Butter Cups!	14 (48.28%)	15 (51.72%)
Us Bacon Reserves Hit 50 Year Low	13 (44.83%)	16 (55.17%)

Table 5.9: Article Validity Evaluation Results

As previously mentioned the participants were asked to rate the importance of the additional information, Figure 5.5 describes the result of the importance classification. The source of the information and the content of the notification were the elements

5.4 Findings and Discussion

that were rated very important, while the position of the notification was classified as not at all important. Hashtags and group information were elements that were rated as important, leaning more toward slightly important than fairly important. In conclusion, users appreciated the information sources and content of the notifications more than the position of the notifications or the validation information from the user group, while the hashtags received a neutral rating.

With an emphasis on SME applications in notification systems, this research study concludes by demonstrating the potential of social media aspects in several fields. Preliminary analysis shows how the selected social media elements could improve user satisfaction and the significance of information in notification systems. Additionally, it is observable that additional SME information does not enhance the effects of information overload but improves the perception and understanding of notifications.

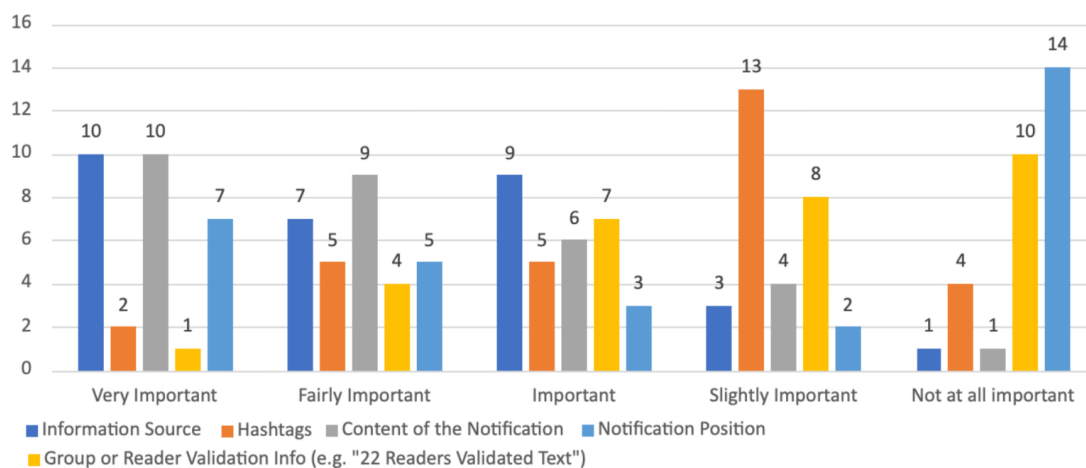


Figure 5.5: Notification Element Ranking

Based on SUS it was determined that a system that uses SME for the enhancement of notifications is considered a usable system. Not all SMEs' use cases were investigated due to time restrictions and the limited number of SMEs considered. Additional SMEs, other media and devices for notification display could be considered for review in future work. Future research may also examine how users respond over time to messages that provide additional information. This could allow us to evaluate the SME studied and other SMEs that were unable to participate in this study more effectively. This may allow for a better evaluation of the analysed SME and additional SMEs that could not be part of this study. Tracking user reactions for longer periods to different combinations of SMEs in notifications could lead to a novel approach to the use of SMEs within notification systems. The survey results could provide the initial

steps toward new use cases of Social Media applications in notifications and other disciplines that deal with users' cognitive ability to process information or disciplines where information overload is considered detrimental. In conclusion, there are several advantages to integrating social media elements into notification systems. Using social media elements for notifications can help ensure that important information is quickly and widely distributed. This can be especially useful in emergencies, where time is of the essence. This research shows potential improvements to notification systems with the use of SME.

5.5 Summary

This chapter aims to contribute to answering the overarching research question: "How can sensitive information be harnessed in a privacy-preserving manner for the creation of a personalised recommendation and information retrieval system, and how does the performance of such a system compare to traditional approaches?". Unlike the previous chapters, this chapter focusses on investigating how users react to additional information (e.g. sensitive information) by providing insights to a series of research questions (RQs) that are linked to the integration of social media elements into notification systems and their influence on users' experiences.

The first research question ("Which elements of social media can be integrated into notification systems to display understandable and valuable notifications at a glance without explicitly disturbing the user?") delves into the aspects of social media element integration within notification systems. It attempts to identify specific elements drawn from the area of social media that can be integrated into notification systems. The primary aim here is to improve the comprehensibility and overall value of notifications, while ensuring that user experience is not negatively affected. This question underscores the importance of striking a harmonious balance between traditional notification styles and emerging social media elements.

Building upon the foundation of the first question, the second research question ("Would users prefer to receive notifications with integrated social media elements like hashtags, topic keywords, source information, rating by other users, and groups information?") shifts the focus to the end user. It investigates the preferences of users concerning the delivery of notifications that include integrated social media elements. These elements include hashtags, topic keywords, source information, ratings provided by other users, and group information. This investigation offers insight into users' viewpoints, shedding light on their attitudes and readiness to receive notifications enhanced by these novel social media components.

The third research question ("How do users react to notifications with additional information (hashtags, user group information, content approval/disapproval, and social media posts)?") attempts to understand user interactions with notifications.

Specifically, it explores how users respond to notifications that are enriched with additional information gleaned from social media. This additional information comprises elements such as hashtags, user group affiliations, indicators of content approval or disapproval, and snippets of social media posts. The source of the information and the content of the notification were the elements that were rated as very important, while the position of the notification was classified as not at all important. Hashtags and group information were elements that were rated as important, leaning more towards slightly important than fairly important. In conclusion, the users appreciated information sources and content of the notifications more than the position of notifications or user group validation information, while hashtags received a neutral rating. By studying user reactions, this question aims to illustrate the ways in which individuals engage with notifications that extend beyond the traditional paradigm.

The final research question ("Which emotions do users experience when receiving notifications with and without this additional information?") investigates the emotional aspects of social media elements and information for notifications. This investigation spans notifications delivered both with and without the added context provided by social media-related elements. By examining emotions, this question seeks to determine user experiences insights in the context of notifications.

Through research and answering these research questions, this chapter offers a comprehensive understanding of the role and impact of integrating social media elements within notification systems. The overarching objective is to mitigate the challenges posed by information overload and concurrently improve user satisfaction and the overall relevance of the information contained in notifications. Although the findings point toward the potential benefits of integrating social media aspects into notification systems across various fields, the chapter remains aware of its limitations. These include time constraints and deliberate focus on a limited subset of social media elements. In summary, this chapter serves as a stepping stone toward the exploration of new use cases for social media applications in notifications and other domains grappling with information processing and overload.

Ultimately, it contributes to the development of privacy-aware and effective recommendation and information retrieval systems. By understanding how social media elements impact user experiences with notifications, this chapter explores the privacy aspects of such integrations. Insights into how sensitive information can be used in a privacy-preserving manner are provided. Since privacy considerations are crucial in an era where data privacy and security are major concerns, the insights from this chapter are highly relevant to the broader goal of developing recommendation systems. Recommendation systems are dependent on user interactions and preferences to provide tailored content. By studying user reactions to notifications enriched with social media elements, this research provides insights into the design of recommendation algorithms that can take advantage of these elements while respecting user privacy.

6 Requirements and System Design

Within this chapter, the current CERN notification system, accompanied by its encompassing libraries and services, is analysed and described. Based on previous chapters and this investigation, an assessment of both functional and non-functional requirements is performed. These requirements are used to determine technologies to use for the prototype of the recommendation system. Finally, the outline of a prototype concept and a corresponding conceptual architecture are described.

The following sections are based on and supported by the work published in the following publications:

- **Jakovljevic, I., Gütl, C., & Wagner, A. (2023).** Privacy-Preserving Collaborative Filtering: Evaluating a Machine Learning Recommender System in a Large Interconnected Organization. In 5th International Open Search Symposium
- **Jakovljevic, I., Gütl, C., & Wagner, A. (2022).** Towards a Privacy-Aware Reproducible Machine Learning Pipeline for Open Data. 4th International Open Search Symposium
- **Jakovljevic, I., Russmann, S., Wagner, A., & Gütl, C. (2022).** A Proposal for Client Based User Profiles For Open Search in Large and Highly Connected Organisations. In Proceedings of the 3rd International Symposium on Open Search Technology: OSSYM 2021
- **Jakovljevic, I., Wagner, A., & Gütl, C. (2020).** Open Search Use Cases For Improving Information Discovery And Information Retrieval In Large And Highly Connected Organizations. In Proceedings of OSSYM 2020 CERN European Organization For Nuclear Research

6.1 Contribution

This chapter represents the first step of the "Design and Development" phase of the research methodology. It plays a central role in converting the extensive insights gained from chapters 3, 4 and 5 into actionable requirements and system design. Building on the valuable findings and research gaps identified in those chapters, this chapter provides the foundation for the extension of the notification system initially introduced in the Chapter 1. The research gap identified during the exploration of the existing literature and user behaviour analysis was the foundation for architectural choices and

design considerations. The suggested recommendation system extends the notification system taking into account terms of privacy-awareness, personalization, and user engagement. This chapter bridges comprehensive research insights with practical implementation of an innovative, privacy-conscious, and effective recommendation and information retrieval system.

6.2 CERN Notifications System

The existing system operates on a subscription model for notifications, where users opt to receive updates from various notification channels (Ormancey et al., 2022). A simplified architecture of the notification system can be seen in Figure 6.1.

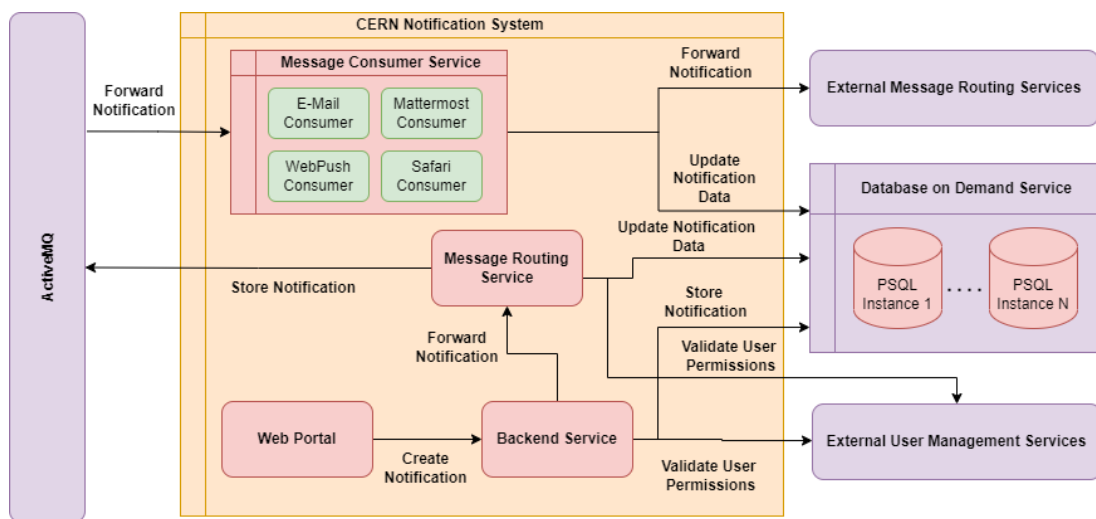


Figure 6.1: Simplified CERN Notification System Architecture Overview based on (Ormancey et al., 2022)

As outlined in Chapter 1, a range of services and applications, including monitoring systems, email, newsletter, and user management services, are used frequently at CERN. In response to this complexity, the notification system aims to integrate various services and libraries within CERN. Additionally, this section introduces key architectural components of the CERN notification system and describes concepts such as channels and notifications.

6.2.1 Web Portal

The web portal component of the system serves as the user interface for a range of essential functions. This front-end application, developed using React JS¹, offers users the capability to create, manage, and explore channels and preferences. Additionally, it provides the functionality to compose and access notifications, while also enabling the management of channel memberships. Figure 6.2 illustrates the primary interface of the web portal, showcasing the main page that users interact with. This portal acts as a central hub for users to seamlessly navigate through various features, ensuring efficient communication and information sharing within the system.

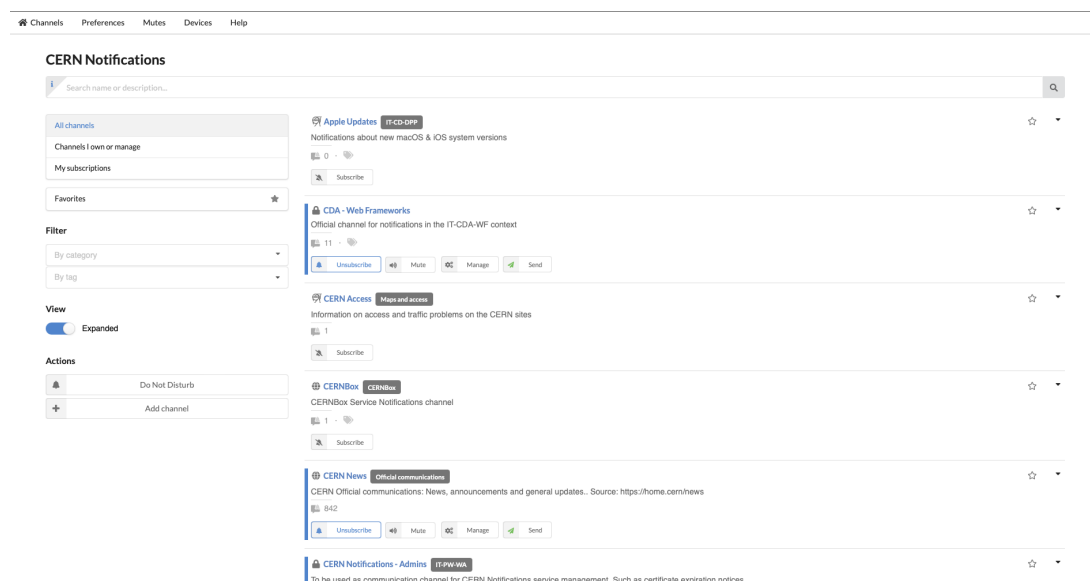


Figure 6.2: Web Portal Main Page

6.2.2 Backend Service

This component is responsible for Create, Read, Update, and Delete (CRUD) operations on the PostgreSQL² (PSQL) database, which are coupled together with complex business logic. It ensures the effective management and manipulation of data stored within the database. Moreover, the backend service establishes communication with an external user management system to validate users and obtain crucial user information. This interaction is essential for maintaining a secure and authenticated user environment.

¹<https://react.dev>

²<https://www.postgresql.org>

within the system. Beyond these functionalities, the backend service facilitates seamless communication with the message routing component which is a key component for directing notifications to the appropriate devices and users, ensuring efficient and accurate message delivery. The backend service enables interactions between various system components, optimizing the overall functionality of the system.

6.2.3 Message Routing Service

The Message Routing Service is a crucial component of the system, developed using Python¹ and utilising the specialised CERN library known as MegaBus² for streamlined communication with the ActiveMQ³ messaging platform. This component is designed with a dual-purpose functionality that significantly enhances the notification system's efficiency and reliability. The primary role of the Message Routing Service is to ensure the accurate distribution of notifications to the ActiveMQ environment, directing notifications to the appropriate queues. Furthermore, the Message Routing Service plays a crucial part in upholding the quality and integrity of the notifications. It retrieves necessary information from external databases and user management services to validate the notification information. This process verifies the accuracy and relevancy of the messages before they are sent to ActiveMQ. By implementing these validation measures, the component contributes to maintaining the system's reliability and overall effectiveness.

6.2.4 Message Consumer Service

The Message Consumer Service is a component within the system, comprising of four dedicated consumers elements. These consumers communicate with the ActiveMQ component to retrieve notifications that have been organised within the messaging queues. Each consumer within this service is designed for specific user-device-service combination. Leveraging user and device information, these consumers make informed decisions regarding the appropriate external services to which the notifications must be dispatched. The purpose of the Message Consumer Service is to bridge the gap between the system's internal notification generation and the external platforms where users interact. These external services encompass a diverse array of communication channels, including email services, Mattermost hooks, web push endpoints, and more. By selecting the correct service for each notification, the Message Consumer Service ensures that the right information reaches the right user, on the right device, through the appropriate external channel.

¹<https://www.python.org>

²<https://gitlab.cern.ch/push-notifications/python-megabus>

³<https://activemq.apache.org>

6.2.5 External Services

ActiveMQ

ActiveMQ is a robust and versatile message-oriented middleware that serves as a fundamental component in modern distributed computing systems. Designed to facilitate seamless communication between disparate software applications, ActiveMQ employs the publish-subscribe and message queuing paradigms to enable efficient and reliable data exchange. At its core, ActiveMQ functions as a broker, mediating the exchange of messages between producers and consumers within a networked environment. Messages, which encapsulate data and instructions, are generated by producers and subsequently transmitted to designated destinations, referred to as queues or topics. Queues offer point-to-point communication, ensuring that each message is consumed by a single recipient, while topics enable publish-subscribe communication, allowing multiple consumers to receive the same message. ActiveMQ's intricate infrastructure ensures message persistence, thereby safeguarding data integrity and fault tolerance. It effectively handles message buffering, acknowledgment mechanisms, and redelivery protocols to enhance system reliability. Furthermore, ActiveMQ supports a variety of communication protocols, including open standard protocols such as MQTT, Stomp, and AMQP, accommodating diverse system requirements. In scientific contexts, ActiveMQ finds extensive utility across domains such as high-performance computing, data-intensive research, and distributed sensor networks. Its role in orchestrating seamless, asynchronous communication between software components makes it an indispensable tool for facilitating data exchange, information dissemination, and cooperative computation within intricate computational frameworks.

Database on Demand Service

A database on demand service is a specialised service that provides users with the ability to create and manage databases as needed, without the necessity of manual provisioning or extensive administrative tasks. This service allows users to request and configure databases based on their specific requirements, such as database type, size, and performance specifications, all through a streamlined and user-friendly interface. Operated within the context of CERN, the database on demand service is an integral component of the organisation's infrastructure. It empowers researchers, scientists, and staff members to effortlessly create and manipulate databases, tailored to their research needs or operational requirements. By abstracting the complexities of database administration, the service ensures that users can focus their efforts on utilising the database for its intended purpose, whether it be data storage, retrieval, or analysis. The database on demand service optimises resource utilisation within CERN's computing ecosystem. It eliminates the need for manual intervention and long

lead times associated with traditional database provisioning, fostering a more agile and responsive environment.

External User Management Services

External user management services refer to specialised systems or platforms that are utilised to centrally manage user identities, access privileges, and authentication across various applications, services, and resources. These services provide a unified approach to user management, allowing organisations to streamline user authentication, authorisation, and profile management processes. Within the context of CERN, external user management services play a crucial role in maintaining a secure and efficient digital environment. These services facilitate the management of user accounts, access controls, and authentication mechanisms for a wide range of applications and systems utilised by CERN’s research community, staff, and collaborators.

6.2.6 Notification System Main Activity

Figure 6.3 illustrates the step-by-step process for the “Send Notification” activity within the Notification System. This activity involves the initiation and distribution

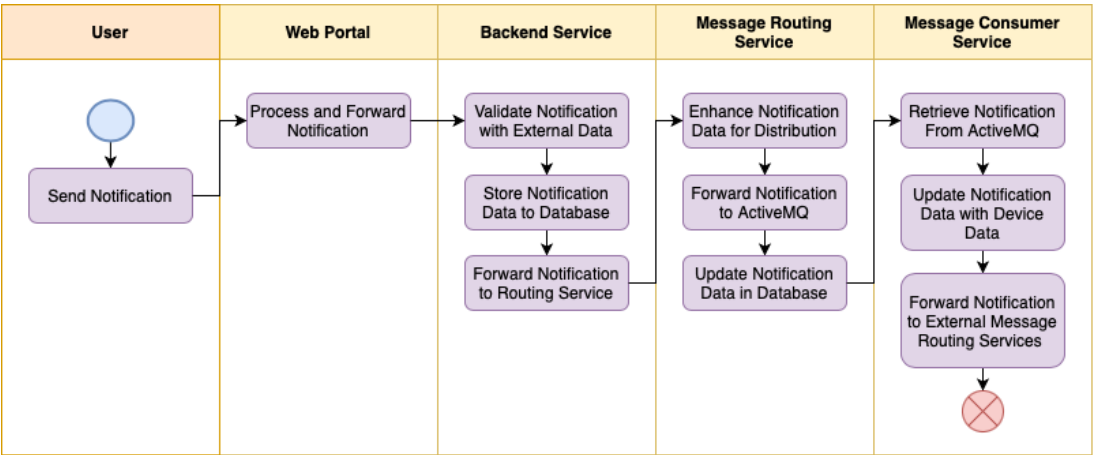


Figure 6.3: Notification System Send Notification Activity

of notifications to targeted users and devices. The figure provides a comprehensive overview of the sequence of actions that occur when a notification is generated and subsequently delivered to users through various communication channels.

6.3 Requirements

In this section, the requirements that guide the development are discussed. Requirements can be understood as instructions that define how the system should work. There are two types: functional and non-functional requirements. Functional requirements describe the specific functions the system should be able to do. They outline the features and actions that users will use and interact with. Non-functional requirements focus on how well the system performs. These include factors like speed, security, and user-friendliness. By understanding and defining these requirements, a clear path for building a system that meets the research needs is described. This section will provide insights into how these requirements shape the design and development process (Dabbagh et al., 2015).

The analysis of the notification system, as conducted throughout this study, was tightly aligned with the research questions initially established in Chapter 1, where it was stated that to enhance the functionality of the Notifications project, the integration of a recommendation system is crucial. Such a system would empower the project to proactively suggest content, updates, and events that are highly relevant and potentially interesting to individual users. This proactive approach can boost user engagement by offering tailored and timely suggestions. Users will not only receive information they explicitly subscribe to but also discover content and events they might have otherwise overlooked. Furthermore, the system should be created as a standalone component serving as an additional module of the notification system.

Chapter 3 provides insights to understanding the preferences and expectations of users, by determining the type of information users desire, the devices they prefer for receiving notifications, and their medium of choice. These insights directly inform the functional requirements of the recommendation system by defining what content and formats should be delivered to users. Based on the studies from Chapter 3, it was concluded that CERN users would like to use a system that provides personalised work-related information to users. The results also emphasised that it was important to the users that the recommendation system does not invade their privacy and that it does not over-recommend information, with a majority of participants agreeing that if the system would send more than 6 recommendations, send duplicate recommendations, or send recommendations without clear sender information that they would ignore such a system.

Chapter 2 introduced background information about user profiling, machine learning, recommender systems, and best practices related to these topics. Based on this chapter, it was concluded that a modern recommendation system should be based on reproducible machine learning pipelines and shareable open data. Additionally, it was stated that recommendation systems have to provide unique lists of suggestions that are tailored to the needs, preferences and interests of the users. For large organisations it has become important to include privacy preserving principles and adhering to

transparency.

Chapter 4, the handling of sensitive data within organisations is analysed. It explores the possibilities of sharing and deploying such data while closely monitoring user reactions to these data-sharing practices. These insights shape non-functional requirements related to data security, privacy, and ethical considerations. They ensure that the recommendation system respects user privacy and adheres to legal and ethical standards.

Chapter 5 contributes to the requirements by determining how users react to additional information integrated into notifications. By examining user responses and their interpretation of this supplementary content, it informs the functional requirements related to notification content enrichment. It helps determine what elements are most effective in enhancing user engagement and understanding.

As a result of this comprehensive investigation spanning multiple chapters, a valuable set of functional and non-functional requirements has been derived. These requirements are instrumental in guiding the subsequent development phase, ensuring the creation of a recommendation system that is both effective and purposeful. The functional requirements, encompassing the core functionalities of the recommendation system, have been documented in Table 6.1.

Functional Requirements
The recommendation system must exclusively use channel membership information for generating recommendations.
The recommendation system must provide a comprehensive list of relevant recommendations tailored to each user's preferences and interests.
The recommendation system must avoid suggesting channels to users if they are already subscribed to those channels.
If a user is new to the system or if there are insufficient relevant channels to suggest, the recommendation system must present a curated selection of popular channels.
The recommendation system must maintain a periodic update cycle to refresh and enhance the set of recommendations available to users. This ensures that users are consistently provided with up-to-date suggestions aligned with their evolving interests and preferences.

Table 6.1: Functional Requirements for the Recommendation System.

These requirements encompass the system's operational aspects, such as generating recommendations based solely on channel membership information, tailoring recommendations to users' preferences, avoiding redundant channel suggestions, presenting popular channels when necessary, and maintaining a regular update cycle for relevant suggestions. On the other hand, the non-functional requirements, delineating the essential qualities and characteristics of the recommendation system, have been detailed

in Table 6.2.

Non-Functional Requirements
The recommendation system must be architecturally segregated from the existing notification system to ensure modularity and maintainability.
The recommendation system should seamlessly integrate with the current notification system, leveraging and enhancing its existing functionalities.
The recommendation system must consistently generate the same set of recommendations for a given user and context, ensuring reproducibility and reliability.
While handling user data, the recommendation system must adhere to stringent privacy protocols to protect user information.
The recommendation system must be developed using a programming language that is already established within the organisation, ensuring compatibility and familiarity.
The recommendation system should have an adaptable design, allowing for straightforward integration into various notification systems.
The recommendation system must support effortless expansion, enabling the addition of new features or integrations as the need arises.
The recommendation system must be designed with evaluation in mind, providing mechanisms and interfaces to assess its performance, accuracy, and user satisfaction effectively.

Table 6.2: Non-Functional Requirements for the Recommendation System.

These non-functional requirements include the architectural segregation of the recommendation system to ensure its modularity, seamless integration with the existing notification system, consistent generation of reproducible recommendations, stringent privacy protocols while handling user data, compatibility with an established programming language, adaptability and expandability to accommodate future enhancements, and finally, a thoughtful design that facilitates comprehensive evaluation of its performance, accuracy, and user satisfaction.

Together, these functional and non-functional requirements form the foundation upon which the recommendation system will be built, ensuring its alignment with the objectives of the study and its effectiveness in enhancing the notification experience for users.

6.4 Conceptual Architecture

Building upon the analysis of the system's requirements, the comprehensive examination of the notification system itself, and specific research topics, this section presents a well-considered proposal for a recommender system architecture that integrates with

6 Requirements and System Design

the existing notification framework. In response to the functional and non-functional requirements, the conceptual architecture aims to provide a holistic solution that not only addresses the identified challenges but also aligns with the overarching goals of this research. The proposed architecture aspires to enhance the notification system's efficacy and user experience. Key components, interaction flows, and underlying principles of the conceptual architecture are explained in this section. A clear understanding of how the proposed architecture integrates into the notification system will be described.

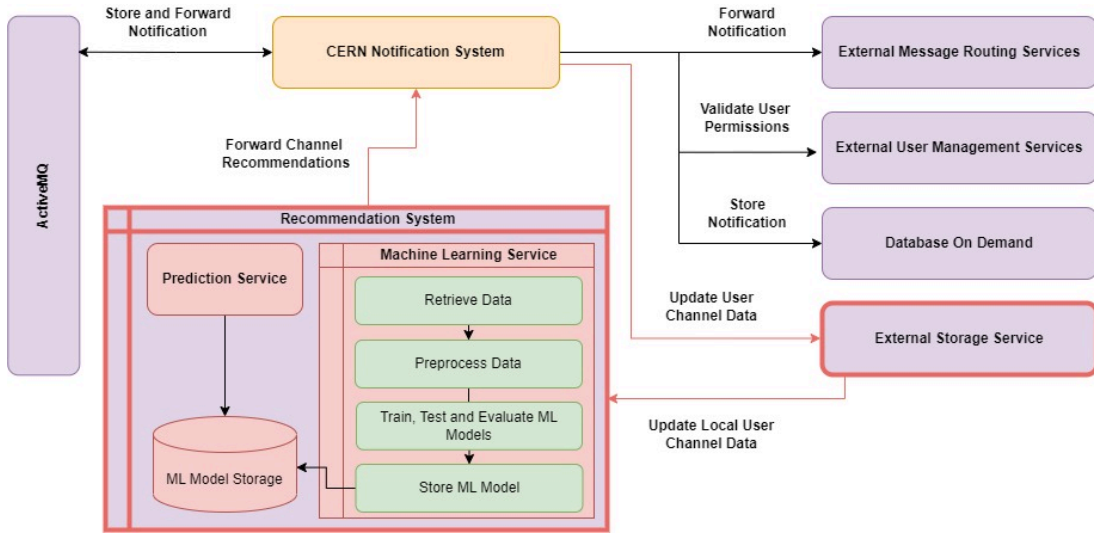


Figure 6.4: Conceptual Architecture for the Recommendation System

The conceptual architecture depicted in Figure 6.4 provides an overview of the recommendation system structure.

6.4.1 Recommendation System

The recommendation system functions as an external, specialised component that communicates with the notification system. The recommendation system contains three main components: the prediction service, the machine learning service, and the ML model storage.

At the core of this system is the prediction service that serves ML models. It is the link between the ML models stored in the ML model storage and the user-channel interactions within the notification system. This operation involves retrieving the requisite ML models from the ML model storage and applying these models to user-channel data. Through this process, personalised recommendations are generated, reflecting the unique preferences and needs of each user.

The machine learning service produces ML models, based on anonymised user-channel relationships. This process is envisioned as a reproducible process, where the result is not only the ML model but also necessary steps, data, algorithms and evaluation results. These results are published as open data to ensure that all recommendations can be validated and additionally used in a scientific concept.

The ML model storage component has an important role in the recommendation system's architecture. It serves as the repository for various ML models, created and stored for future use by the machine learning service.

Together, these components form the backbone of the recommendation system, orchestrating data processing, machine learning, and user interactions to deliver valuable, personalised recommendations to users within the organisation.

6.4.2 External Storage Service

An additional external storage service has been integrated into the architecture, to facilitate sharing and managing anonymous channel-user relationship data between the recommendation and the notification system. This data repository proves indispensable for constructing and training machine learning models on anonymised data to ensure user privacy and data protection. This service enables external and internal users to view and download all the data used for the creation of an ML model, together with the ML model and its metadata.

6.5 Technology Decisions

Building upon the functional and non-functional requirements outlined in Section 6.3 and the conceptual architecture discussed in Section 6.4, this section presents potential technological solutions for implementing the recommendation system.

6.5.1 Machine Learning Service

Numerous frameworks and libraries are available to oversee the lifecycle of machine learning. During the assessment for a suitable machine learning service, MLflow¹, Flyte², MLRun³, and Kubeflow⁴ were investigated. Table 6.3 provides an outline of important attributes that influenced the decision-making process. Kubeflow and MLRun exhibited nearly identical feature sets, but Kubeflow was opted for, mainly due to its robust documentation and comprehensive dashboard resources.

¹<https://mlflow.org>

²<https://flyte.org>

³<https://www.mlrun.org>

⁴<https://www.kubeflow.org>

6 Requirements and System Design

	Kubeflow	MLFlow	Flyte	ML Run
Open Source	Yes	Yes	Yes	Yes
Language	Python	Python	Python	Python
Documentation	Very Good	Good	Poor	Good
Tracking and Versioning	Yes	Yes	Yes	Yes
Pipeline Orchestration	Yes	No	No	Yes
Model Deployment	Yes	No	No	Yes
Scheduler	Yes	No	Yes	No
Dashboard	Yes	Yes	Yes	Limited Functionality

Table 6.3: Software packages for building reproducible machine learning applications. Taken From "Towards an Open Data based Privacy-Aware Reproducible Machine Learning Pipeline", by Jakovljevic, Wagner, & Gütl (2022)

6.5.2 ML Model Storage

The storage provider requirements were to enable data sharing and efficient data scaling while ensuring data security, additionally it should be compatible with selected solutions for Machine Learning and Prediction service. Choices like Amazon Web Services' Simple Storage Solution (S3)¹ or S3-compatible MinIO² offer secure and reliable storage solutions, similar to folders. Due to the fact that the desired storage solution should be open source and because MinIO offers a smooth integration with Kubernetes and ability to handle growing data needs make it a fitting choice. This decision aligns with the need for effective data scaling, and robust access controls.

6.5.3 External Storage Service

For the external storage service, a strategic choice was made to use a S3 storage service provided by CERN. The decision aimed to minimise the implementation complexities, as S3 could seamlessly integrate into the notification and recommendation systems without the need for additional customization. This approach allows for immediate use of S3 and its capabilities, reducing the necessity for additional development efforts and takes advantage of the existing infrastructure provided by CERN .

6.5.4 Prediction Service

Various methods exist for serving ML models for recommendations and enabling recommendation queries.

¹<https://aws.amazon.com/s3/>

²<https://min.io>

Service	Description
TensorFlow Serving	TensorFlow Serving is an open-source library developed by Google for serving TensorFlow models in production environments. It offers efficient, high-performance serving of machine learning models with support for model versioning, batching, and monitoring.
PyTorch Serve (formerly Torch-Serve)	PyTorch Serve is an open-source model serving tool designed for PyTorch models. It provides features like model versioning, multi-model serving, and dynamic batching, making it suitable for deploying PyTorch-based machine learning models.
KFServing	KFServing is part of the KubeFlow project and is used for serving machine learning models on Kubernetes. It supports multiple machine learning frameworks, autoscaling, and canary deployments, making it suitable for scalable and production-ready ML serving.
Seldon Core	Seldon Core is an open-source platform for deploying and managing machine learning models on Kubernetes. It offers support for multiple machine learning frameworks, model explainability, and scaling, making it suitable for production deployments.
Clippier	Clippier is an open-source prediction serving system that can be used for deploying machine learning models in real-time applications. It supports model versioning, latency-aware routing, and A/B testing.

Table 6.4: Popular Machine Learning Model Serving Tools

One such service is KServe¹ which is a native component of KubeFlow. KServe is a model inference service that operates within Kubernetes environments. It is designed to facilitate the deployment and serving of machine learning models, allowing users to make predictions based on these models through API calls. KServe also supports the creation of custom predictors that extend beyond traditional machine learning models, broadening its applicability to various types of predictive tasks.

6.6 System Architecture

The system's architecture, depicted in Figure 6.5, resulted from the previous sections, based on the analysis of the existing notification system, earlier outlined requirements and expert technological decisions. A component was added to the system architecture that was not described in the conceptual architecture. This component is the Data Anonymization and De-Anonymization Service. The implementation of the recommendation system aligns with prior technological decisions. By harnessing the KServe

¹<https://github.com/kserve/kserve>

6 Requirements and System Design

module, it simplifies the deployment of Machine Learning (ML) models through KubeFlow Pipelines, thereby assuring the construction of replicable ML models. Following the use of open data in KubeFlow for the creation and assessment of ML models, the resulting model is stored in the MinIO object storage, serving the requirements of the KServe Prediction Service. In parallel, it is securely stored within CERN's S3 storage, alongside the input dataset together with metadata, guaranteeing their reproducibility.

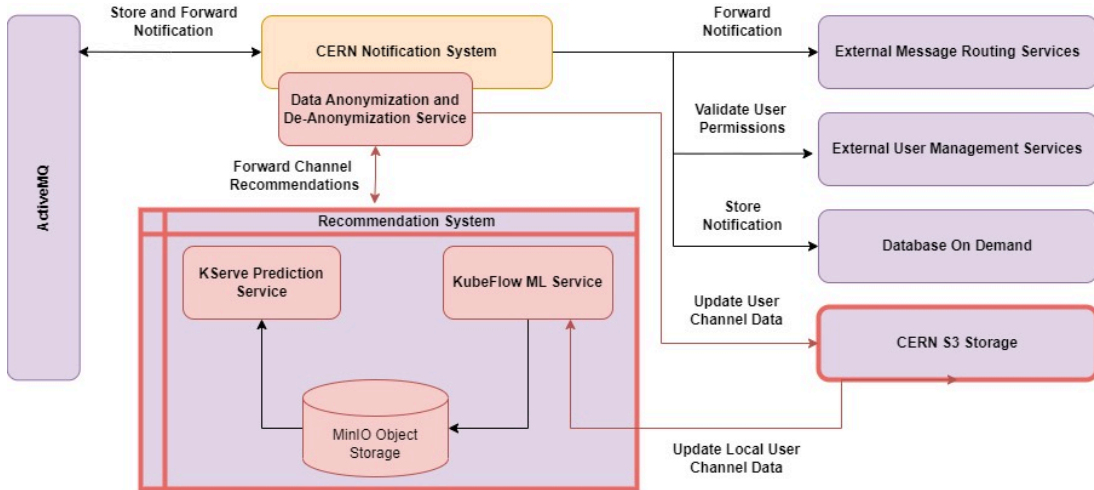


Figure 6.5: System Architecture for the Recommendation System

6.6.1 KubeFlow ML Service

Based on the technological description, KubeFlow was used to implement the Machine Learning Service from the conceptual architecture displayed in Figure 6.4. KubeFlow was leveraged to enable data retrieval and processing, machine learning model training and deployment, workflow orchestration, and ML model storage procedures.

6.6.2 KServe Prediction Service

The decision to use KServe for the Prediction Service within the Recommendation System architecture was driven by several key considerations and advantages. KServe's architecture is designed to handle high-throughput, low-latency inference workloads, making it well-suited for the needs of a Prediction Service. As user bases grow and recommendation requests increase, Serve can efficiently scale to meet demand, ensuring a responsive user experience. It provides robust model management capabilities, simplifying the deployment and serving of multiple machine learning models. This

was crucial for the Prediction Service, which needed to serve various recommendation models to different users based on their preferences and behaviors. Being part of the Kubeflow ecosystem, Serve benefited from a open-source community and ongoing development. This ensured access to updates, enhancements, and best practices, contributing to the long-term sustainability of the Prediction Service.

6.6.3 Data Anonymisation and De-Anonymisation Service

This component, driven by the Data Lift framework mentioned in Chapter 4, anonymises user data by extricating identifiable attributes while restoring the anonymised data to its original user context after an anonymised recommendation is received from the recommendation system, thereby enabling interaction of the notification system with the recommendation system.

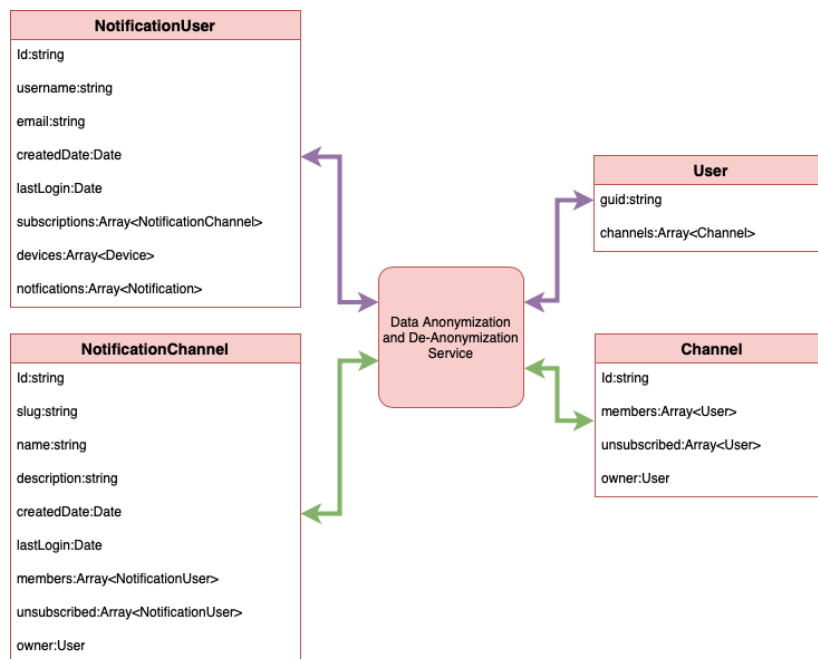


Figure 6.6: System Architecture for the Recommendation System

Figure 6.6 illustrates the main responsibility of this component. To ensure that the data used for creating ML models and recommendations respects user privacy, it was necessary to convert the real database objects into simplified objects. As the recommendation system is based on implicit data, the most important information was the connections between users and channels (subscriptions or unsubscriptions).

The user and channel objects were stripped of all PIDs and their identifiers were encrypted into hashed identities. This anonymised data was then stored as open data on the CERN S3 service to be used by the machine learning service. On the other side, when the recommendations were sent to the notification system it was necessary to decrypt these anonymous identifiers for users and channels in order to forward the recommendations to corresponding users. This service enables the notification system to export data without invading the privacy of the users.

6.6.4 MinIO Object Storage

MinIO is designed to be massively scalable, making it an excellent choice for storing machine learning models, which can grow in size as more data becomes available or as new models are developed. It ensures that the storage infrastructure can effortlessly expand to accommodate future requirements while offering high-performance object storage, optimised for read-intensive workloads. MinIO includes robust security features, including encryption at rest and in transit, access control policies, and compliance capabilities. These features align with the need to protect sensitive machine learning models and user data, ensuring data privacy and compliance with organisational and regulatory requirements. Additionally, it integrates with Kubernetes, which is a fundamental component of KubeFlow.

6.7 Summary

In this chapter, the focus was set on creating a software architecture for a recommender system that can seamlessly integrate it into the existing CERN notification infrastructure. The emphasis throughout this chapter was the prioritisation of user privacy while facilitating the core aspects of enabling research and development reproducibility. It began with an in-depth analysis of the existing CERN notification system. The current CERN notification system was analysed, allowing invaluable insights to be gained into its inner workings. This initial analysis served as the foundation for the subsequent steps, offering a clear understanding of the system's operations. Multiple technical and scientific factors to create an architecture for the recommendation system that can be integrated into the notification system while respecting user privacy and ensuring reproducible results were investigated. This was followed by the identification of functional and non-functional requirements, ensuring alignment with research objectives. These requirements were shaped by insights from previous chapters and research. They provided a clear road map, ensuring that the architectural design would align with the overarching research objectives. A conceptual architecture was proposed, depicting the possible integration of the recommendation system. This was supported

by informed technology decisions to enhance compatibility and performance. Ultimately, the chapter presented the final system architecture, representing the synthesis of analysis, requirements, and technological choices. This holistic approach provided a solid foundation for the subsequent stages of the development, implementation, and evaluation of the recommendation system.

The importance of this chapter plays a central role in addressing the critical research questions posed in Chapter 1. By examining the notification system and designing an architecture that places user privacy at the forefront, a balance between personalisation and privacy has been achieved. This balance enables not only an innovative privacy-aware recommender system, but also the generation of open data for reproducible research. In this chapter, the current CERN notification system, accompanied by its libraries and services, is analysed and described. Based on previous chapters and this investigation, an assessment of both functional and non-functional requirements is performed. These requirements are used to determine the technologies to use for the prototype of the recommendation system. Finally, the outline of a prototype concept and the corresponding conceptual architecture are described. An additional contribution of this recommender system is the ability to generate open data that can be used to support reproducible research and externally validate ML models used in the system.

7 Development

In this chapter, all details related to the implementation of the proposed system architecture from Chapter 6 are explained. First, the workflow of the recommendation system is explained to describe how the system created recommendations and what are all the necessary activities related to ensuring that user privacy is not breached. The second part of this chapter focusses on the implementation of machine learning aspects of the system (KubeFlow and KServe). Lastly, the integration with the notification system is described and illustrated. In the last section of this chapter, the main points and findings are summarised.

The following sections are based on and supported by the work published in the following publications:

- **Jakovljevic, I., Gütl, C., & Wagner, A. (2023).** Privacy-Preserving Collaborative Filtering: Evaluating a Machine Learning Recommender System in a Large Interconnected Organization. In 5th International Open Search Symposium
- **Jakovljevic, I., Gütl, C., & Wagner, A. (2022).** Towards a Privacy-Aware Reproducible Machine Learning Pipeline for Open Data. 4th International Open Search Symposium

7.1 Contribution

This chapter is the last step of the "Design and Development" phase of the research methodology. Chapter 6 was used to combine the knowledge from Chapters 4, 3, and 5 to create a conceptual architecture of the recommendation system for notifications, this chapter outlines all the necessary implementation steps that were taken to implement the architecture. This chapter specifically focusses on the implementation aspects of creating a personalised recommendation system in a privacy-preserving way. Furthermore, practical elements used to create open data from organisational data are introduced in this chapter, which were conceptualised in the chapters 3 and 6.

7.2 Recommendation System Workflow

Based on the literature review of Machine Learning pipelines (Chapter 2) and the requirements mentioned in Chapter 6, this workflow has been constructed to capture

7 Development

the process of generating personalised recommendations from user data. Through this section, the key stages and methodologies used for the creation of recommendations are explained, providing a comprehensive and systematic view of the intricate process that transforms raw user data into valuable, tailored suggestions. This diagram shown in Figure 7.1 outlines the step-by-step recommendation system process for the notification system. The process starts with KubeFlow collecting anonymised user data from the S3 storage, and illustrates all steps that are necessary to produce channel recommendations for the user. The process ends with the user approving or ignoring the recommendations.

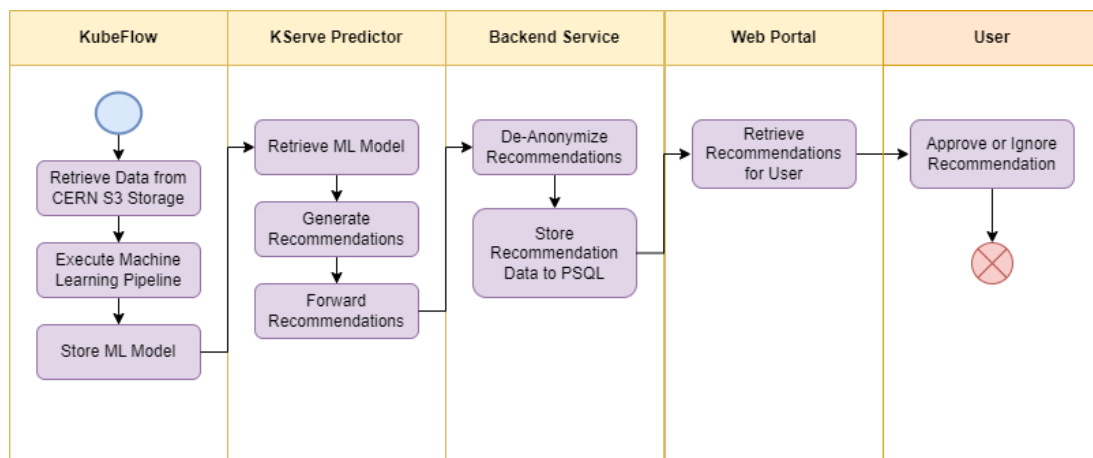


Figure 7.1: Workflow of the Recommendation System

Figure 7.3 is shows the structured sequence of steps inherent to the Kubeflow process, designed for the development of machine learning models dedicated to recommendation systems. Within this process, a division into four distinct sections is observed, each characterised by its unique role and objectives.

The first section of this process is dedicated to the task of obtaining and preparing data. In this phase, the primary emphasis is placed on the aggregation and preprocessing of raw data, which form the foundation of a recommendation system. The quality, relevance, and integrity of these data underlie the entirety of the recommendation process, making this phase extremely crucial.

Transitioning to the subsequent section, attention is directed towards the extraction and selection of features. Within this phase, efforts are made to identify and extract the most relevant features from the prepared data. These features enable our machine learning models to efficiently determine patterns, trends, and user preferences. The careful curation of these features is of principal significance in ensuring the precision and relevance of our recommendations.

The third section includes steps for the creation, evaluation, and validation of machine learning models. Within this phase, algorithms are selected, models are formulated, and their performance is subjected to evaluation.

Finally, the last section of the Kubeflow process addresses the necessity of storing machine learning models for future use. The storage step ensures that models remain reproducible, accessible, and deployable.

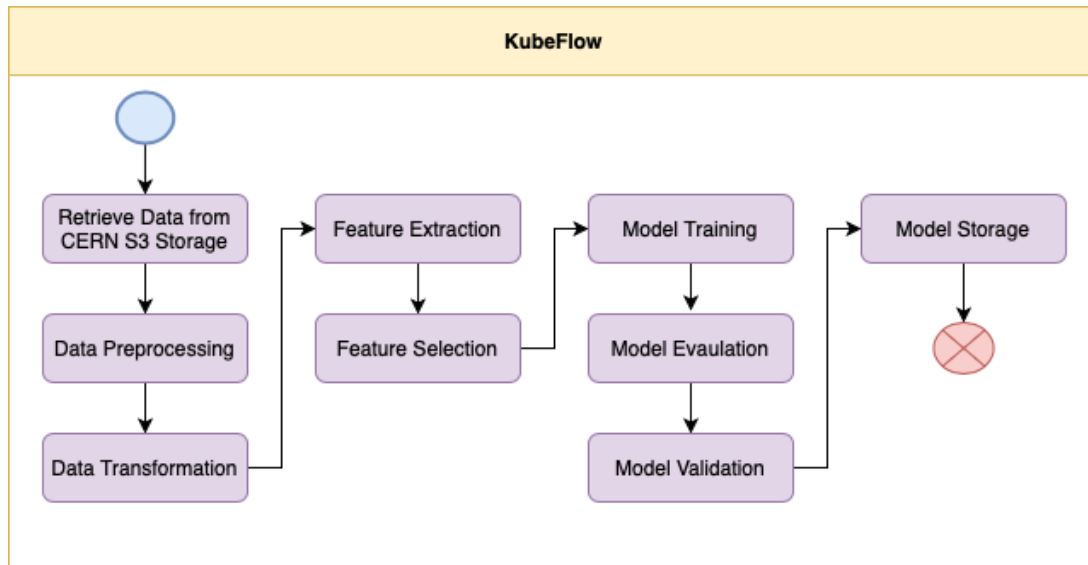


Figure 7.2: KubeFlow Process

7.3 Kubeflow

This section provides a comprehensive overview of the Kubeflow workflow (Figure) of the recommendation system, where each phase within the workflow is examined. Specific attention is devoted to outlining the code executed at each phase, ranging from data retrieval from external repositories to preprocessing, creating recommendations, and the subsequent construction of the pipeline.

Figure 7.3 describes the flow diagram provided by the Kubeflow documentation for the creation of ML pipelines with tools that can be used in each step. The experimental and production phases are the two main phases for developing and deploying an ML system that have been identified by the Kubeflow community.

In the experimental phase, the model is developed, tested, and updated based on initial assumptions to produce favourable results. This phase is compromised

7 Development

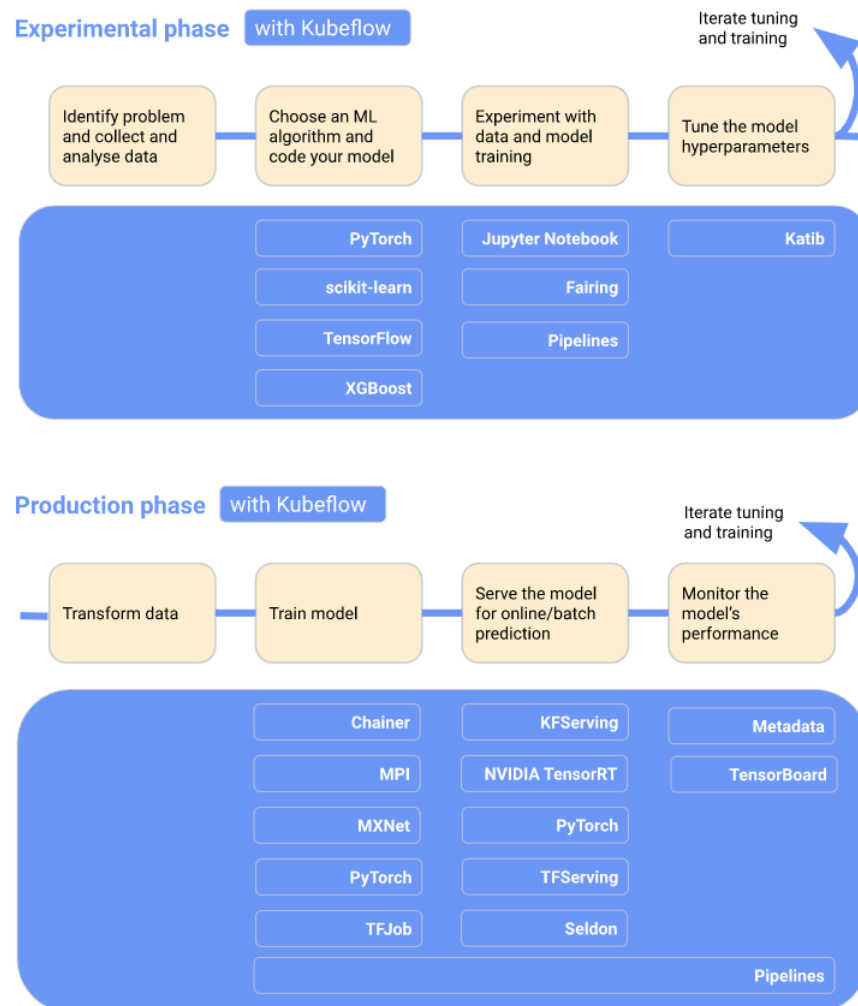


Figure 7.3: Kubeflow components in the ML workflow¹

of four steps. First, it is necessary to identify the ML problem. The data are then collected and analysed for training. Afterward, an ML framework and algorithm are selected. KubeFlow offers a selection of different frameworks, including PyTorch²,

²<https://pytorch.org>

scikit-learn¹, TensorFlow², and XGBoost³. Using the selected framework, the initial version of the model is produced, be experimenting with the data using a set of tools provided by KubeFlow (Jupyter Notebook⁴, Fairing⁵, KubeFlow Pipelines). The last step includes experimenting with the data and with model training and tuning the model's hyperparameters to ensure the most efficient processing and the most accurate results possible. The hyperparameters can be done with the use of Katib, which is a component of Kubeflow that automates hyperparameter tuning and optimisation for machine learning models on Kubernetes⁶ clusters.

In the subsequent production phase, the ML model is reproduced, trained, and deployed in the production phase, by executing a series of operations. These operations encompass the transformation of data into a format compatible with the requirements of the training system, with a crucial emphasis on maintaining consistency in the transformation process across both experimental and production phases to ensure model uniformity during both training and prediction. Additionally, this phase involves the training of the ML model, the provision of the model for online prediction or batch-mode execution, and the ongoing monitoring of the model's performance. The model training step can be done with the use of multiple tools provided, such as Chainer⁷, MPI⁸, MXNET⁹, Pytorch and TFJob. The models produced can be served to users with the use of multiple serving tools(the list of model serving tools can be found in Chapter 6). The results obtained from this monitoring process are integrated into the processes for model tuning or retraining, thereby facilitating continuous improvement and adaptation.

7.3.1 Kubeflow Pipelines

In Chapter 2, it was discussed that Kubeflow pipelines depict the workflow of a machine learning application. These pipelines are made up of separate components. Their execution sequence can be set and is represented as a graph. Each component is a standalone function, packaged as a Docker image. Input and output parameters, which may connect it with other components, might be included. Typically, it is recommended to pass only small amounts of data (such as settings or values) between components, while storage should be utilised for larger data volumes.

¹<https://scikit-learn.org/stable/>

²<https://www.tensorflow.org>

³<https://xgboost.readthedocs.io>

⁴<https://jupyter.org>

⁵<https://www.kubeflow.org/docs/external-add-ons/fairing/fairing-overview/>

⁶<https://kubernetes.io>

⁷<https://chainer.org>

⁸<https://www.kubeflow.org/docs/components/training/mpi/>

⁹<https://github.com/apache/mxnet>

Based on the system architecture detailed in Chapter 6 and the ML process workflows illustrated in Figures 7.3 and 7.3, the ML pipeline is constructed. The initial stages involve data retrieval and preprocessing. This is followed by a focus on feature engineering. Subsequently, the ML model undergoes training, evaluation, and validation, resulting in a recommender system model. Finally, the recommender system model is saved for future applications. Each step of the pipeline is explained in the following subsections.

Data Retrieval and Preprocessing

Data retrieval and preprocessing are the initial steps in Figure 7.1. This action corresponds includes "Retrieve Data from CERN S3 Storage", "Data Preprocessing" and "Data Transformation" steps from the mentioned figure. Listing 7.1 pertains to a python function dedicated to importing data from external sources and subsequently storing it into an S3 service. Initially, it establishes connections using the request module to fetch data from a specified external URL. The data fetched are then saved locally under a designated file name. Subsequently, this local file is uploaded to S3 compatible storage using the Minio library. Notably, the storage details including the host address, access and secret keys, and the desired bucket name are parameterised for flexibility and security.

```

1 def import_data():
2     import requests
3     from minio import Minio
4
5     req = requests.get("<EXTERNAL_DATA_URL>")
6     file_name = "<FILE_NAME>"
7
8     with open(file_name, "w") as file_obj:
9         file_obj.write(req.text)
10
11     minio_client = Minio("<HOST_ADDRESS>",
12                          access_key="<ACCESS_KEY>",
13                          secret_key="<SECRET_KEY>")
14     minio_bucket = "<BUCKET_NAME>"
15     minio_client.fput_object(minio_bucket,
16                              file_name, file_name)

```

Listing 7.1: Function for Importing Data from External Sources into a S3 Service

The second listing (referenced as 7.2) is centred around preprocessing the data obtained in the previous step. It starts by establishing a connection with the same S3-compatible storage using the Minio library, fetching the file, and then reading its

contents into a string variable. The function reads the data as CSV rows. The first entry in each row is identified as $user_i d$, and the subsequent entries represent $group_i d$. The function populates a dictionary (termed as 'groups') with each $group_i d$ associated with a list of $user_i ds$. The collated groups are then serialised and saved in a TOML (Tom's Obvious, Minimal Language) format using the toml library. Post-serialisation, this processed data file is uploaded back to the S3-compatible storage.

```

1 def preprocess():
2     import csv
3     import toml
4     from minio import Minio
5     from typing import Dict, List
6
7     minio_client = Minio("<_HOST_ADDRESS_>",
8                           access_key="<_ACCESS_KEY_>",
9                           secret_key="<_SECRET_KEY_>")
10    minio_bucket = "<_BUCKET_NAME_>"
11    file_name = "<_FILE_NAME_>"
12    minio_client.fget_object(minio_bucket, file_name, file_name)
13
14    raw_text = ""
15    with open(file_name, "r") as f:
16        raw_text = f.readlines()
17
18    user_channels: List[Dict[str, str]] = []
19    for row in csv.reader(raw_text):
20        user_channels.append(
21            {user_id: row[0], channel_id: row[1]})
22
23    user_channels_file = "<_GROUPS_FILE_NAME_>"
24    with open(user_channels_file, "w") as f:
25        toml.dump(user_channels, f)
26
27    minio_client.fput_object(minio_bucket,
28                             user_channels_file,
29                             user_channels_file)

```

Listing 7.2: Data Preprocessing

Both listings exemplify robust data management practices, from the initial data import to structured preprocessing, ensuring efficient and organized storage and retrieval.

ML Model Creation

ML model creation represents a combination of multiple steps from Figure 7.1. This action includes "Feature Extraction", "Feature Selection", "Model Training", "Model Evaluation", and "Model Validation" steps from the mentioned figure. The provided listing 7.3 presents the *model_training* function, which is dedicated to the creation and storage of a machine learning model.

```

1 def model_training():
2     import implicit
3     import pandas as pd
4     import numpy as np
5     import scipy
6     import json
7     from minio import Minio
8
9     # Initialize Minio client
10    minio_client = Minio("<_HOST_ADDRESS_>",
11                        access_key="<_ACCESS_KEY_>",
12                        secret_key="<_SECRET_KEY_>")
13    minio_bucket = "<_BUCKET_NAME_>"
14    file_name = "<_FILE_NAME_>"
15
16    # Fetch data from Minio and load the json
17    minio_client.fget_object(minio_bucket,
18                            file_name, file_name)
19    with open(file_name, 'r') as json_file:
20        channel_members = json.load(json_file)
21
22    # Create DataFrame
23    df = pd.DataFrame(channel_members)
24    df["index"] = df["channel_id"] + "-" + df["user_id"]
25    df.set_index('index', inplace=True)
26
27    # Filter channels with more than 5 users
28    allowed_channels = df
29                        .groupby(["channel_id"])
30                        .filter(lambda x: len(x) > 5)["channel_id"]
31                        .unique()
32    df = df[df["channel_id"].isin(allowed_channels)]
33
34    # Convert 'user_id' and 'channel_id' to category codes

```

```

35 df['u_id'] = df['user_id'].astype("category").cat.codes
36 df['c_id'] = df['channel_id'].astype("category").cat.codes
37 df['score'] = 1
38
39 # Create sparse matrix
40 sparse_item_user = scipy.
41     sparse.csr_matrix((df['score']
42         .astype(float), (df['c_id']
43         .astype(int), df['u_id'].astype(int))))
44
45 # Train the implicit model
46 regularization = 0.1
47 alpha_values = 40
48 factors = 150
49 iterations = 100
50 ml_model = implicit.als.AlternatingLeastSquares
51 model = ml_model(num_threads=4,
52     factors=factors,
53     regularization=regularization,
54     iterations=iterations)
55 model.fit(sparse_item_user)
56
57 # Save model and upload to Minio
58 model_file = "<_MODEL_FILE_"
59 model.save(model_file)
60 minio_client.fput_object(minio_bucket,
61     model_file, model_file)

```

Listing 7.3: ML Model Creation

The function commences by initializing a MinIO client using the provided host address credentials, facilitating both the retrieval of necessary data and the eventual storage of the trained model. Using this client, data is sourced from MinIO based on the designated bucket and file identifiers. Once fetched, this JSON-formatted data is loaded into the *channel_mmembers* variable. Subsequently, this loaded data is transformed into a pandas dataframe. To uniquely identify each entry, an index, crafted by concatenating *channel_id* and *user_id*, is devised and set for the dataframe. A significant preprocessing step involves filtering the dataset to retain only channels with more than five users, ensuring that the subsequent modelling process is informed by meaningful interactions. Further refinement involves the conversion of the *user_id* and *channel_id* columns into a categorical datatype, followed by their encoding into numerical category codes. This conversion is pivotal for the generation of the sparse matrix, a requisite for the implicit

modelling approach. Alongside, a 'score' column is introduced, uniformly set to a value of 1 for all entries.

A sparse matrix, representing user-item interactions essential for collaborative filtering methods, is then formulated using the `scipy` library. With the data appropriately structured, attention shifts to the modelling phase. An Alternating Least Squares (ALS) model from the `implicit` library is initialized with specific hyperparameters, including the number of factors, the regularization strength, and the iteration count. This model is subsequently trained using the previously crafted sparse matrix.

In the final stages, the trained model is saved to a predetermined file, which is then uploaded to the MinIO bucket. This ensures that the model is not only preserved but is also readily accessible for subsequent deployment or further analysis. In its entirety, this function encapsulates an end-to-end workflow, from data acquisition and preprocessing to modelling and storage.

Pipeline Creation

The pipeline creation action combines the previous two actions of gathering and preprocessing data and ML model creation into a process that can be easily reproduced. The first listing 7.4, focuses on creating individual components for the pipeline. A list of Python packages is declared, which includes modules like `"igraph"`, `"kfp"`, `"minio"`, `"requests"`, and `"toml"`. These packages are essential for the correct functioning of the components being defined. Three pipeline components are then created: one for importing data (`comp_import_data`), another for preprocessing the data (`comp_preprocess`), and a third for model training (`comp_model_training`). Each component is constructed from a specific function, such as `import_data`, `preprocess`, or `model_training`. To ensure each function has access to necessary dependencies, the list of packages (`packages_to_install`) is passed during the component creation process.

```

1 packages_to_install = ["igraph", "kfp", "minio",
2                       "requests", "toml"]
3
4 comp_import_data = components.
5     create_component_from_func(import_data,
6                               packages_to_install=packages_to_install)
7
8 comp_preprocess = components.
9     create_component_from_func(preprocess,
10                              packages_to_install=packages_to_install)
11
12 comp_model_training = components.
13     create_component_from_func(model_training,
```

```
14 | packages_to_install=packages_to_install)
```

Listing 7.4: Creating Pipeline Components

The second listing, 7.5, provides the actual pipeline definition. A pipeline named 'recommender' with the description 'Find recommendations for user.' is set up using Kubeflow Pipelines (KFP) DSL (Domain-Specific Language). The *comp_import_data* function is invoked, marking the initiation of the data importation process. The pre-processing step, *comp_preprocess*, is set to execute subsequent to the data importation step, ensuring that the raw data is in place and ready for refinement. Finally, the *comp_model_training* step is executed after the preprocessing step. This ensures the model training process is fed with the preprocessed data. The order of the steps is explicitly managed using the *after* method, confirming that the pipeline components are executed in the prescribed sequence.

```
1 @kfp.dsl.pipeline(name='recommender')
2 def pipeline():
3     step1 = comp_import_data()
4     step2 = comp_preprocess()
5     step2.after(step1)
6
7     step3 = comp_model_training()
8     step3.after(step2)
```

Listing 7.5: Pipeline Definition

In summary, these listings demonstrate a structured and organised approach to defining and sequencing a machine learning pipeline using Kubeflow. The process initiates with data importation, transitions into data preprocessing, and culminates in model training.

7.4 KServe Custom Predictor

A KServe predictor is an application (e.g. web service) designed to receive requests and provide responses based on the ML model it uses for recommendations. In this section, the structure of a custom predictor is elaborated, along with the necessary procedures for constructing it as a Docker image suitable for use by the inference service.

As mentioned in Chapter 6, KServe offers multiple pre-build models for model serving (e.g., TensorFlow, PyTorch, scikit). Since the recommender that we created does not use predefined ML models, it was necessary to create a custom KServe ML model. To create a custom model the *kserve.Model* class was used, seen in Listing 7.6. This class implements four main functions: *load*, *preprocess*, *predict* and *postprocess*, which are executed in the respective order. To use the custom KServe ML model it is

necessary to create a docker image. Buildpacks ¹ was used to create a simple docker image without the need to use dockerfiles and other docker logic.

```

1 import kserve
2
3 class CustomModel(kserve.Model):
4     def __init__(self, name: str):
5         super().__init__(name)
6         self.name = name
7
8     def load(self):
9         pass
10
11    def preprocess(self, request):
12        pass
13
14    def predict(self, request):
15        pass
16
17    def postprocess(self, response):
18        pass
19
20 if __name__ == "__main__":
21     model = CustomModel("recommender-model")
22     kserve.ModelServer().start([model])

```

Listing 7.6: KServe Custom Model Class

7.4.1 Custom KServe Predictor Implementation

Based on the custom predictor model class structure shown in Listing 7.6 the custom predictors for the recommendations are created. It was only necessary to implement two functions for the recommendations to work, the function to load data and the function to create predictions. In the load function, the predictor loads the necessary model that was created in the previous steps together with channel and user data. While the predict function parses the request, applies the ML recommendation model on the request and sends a response object with the recommended items to the caller of the endpoint.

In Listing 7.7, the code segment pertains to loading data within a KServe predictor. The code establishes a connection to a Minio object storage service using access and

¹<https://buildpacks.io/>

secret keys, accesses a specified bucket, and fetches relevant files, such as popular channels information and model data. The *implicit.RecommenderBase.load()* method is used to load the model. The code also retrieves and processes user-channel data from a TOML file.

```

1 def load(self):
2     minio_client = Minio("<HOST_ADDRESS>",
3                           access_key="<ACCESS_KEY>",
4                           secret_key="<SECRET_KEY>")
5
6     minio_bucket = "recommender"
7     popular_channels_file = "<POPULAR_CHANNELS>"
8     minio_client.fget_object(minio_bucket,
9                              popular_channels_file,
10                             popular_channels_file)
11
12     model_file = "<MODEL>"
13     minio_client.fget_object(minio_bucket,
14                              model_file,
15                              model_file)
16     self.model = implicit.RecommenderBase.load(model_file)
17
18     user_channels_file = "user_channels.toml"
19     minio_client.fget_object(minio_bucket,
20                              user_channels_file,
21                              user_channels_file)
22     with open(user_channels_file, "r") as f:
23         self.user_channels = toml.load(f)

```

Listing 7.7: Loading Data

In Listing 7.8, the code defines the predict method for the KServe predictor. This method accepts a request as a dictionary, extracts the user ID from the request, and specifies the maximum number of recommended items. The model's recommend function is utilised to generate recommendations based on the user's preferences and interactions with items. The code then returns a dictionary containing the recommended channels.

```

1 def predict(self, request: Dict) -> Dict:
2     user_id = request["request"]["user_id"]
3     max_recommendations = 10
4
5     recommendations = model.recommend(userid,
6                                       user_item_data[userid],

```

```

7         N = max_recommendations)
8     return {"recommended_channels": recommendations}

```

Listing 7.8: Prediction Method

7.5 CERN Notification System Integration

7.5.1 CERN Notification System Backend

To integrate the recommendation system with the CERN Notification System it was necessary to extend it with additional API Endpoints. These endpoints serve as the bridge between the recommendation system and the core notification infrastructure, enabling the interaction of these distinct and interconnected components. Table 7.1 outlines these endpoints, each accompanied by a descriptive summary of their functionalities.

API Call	Description
[GET] /recommendations/	Returns Channel Recommendations of the current User. This endpoint provides the channel recommendations for the current user, excluding those that have been ignored or already subscribed to. Access to this information is granted to authorised requesters.
[POST] /recommendations/	Creates a new channel recommendation as specified in the body parameter. This API call allows the creation of a new channel recommendation according to the information provided in the request body.
[PATCH] /recommendations/	Updates Channel Recommendation ignored date or subscribed date attributes.

Table 7.1: Recommendations API Calls

7.5.2 Notification and Recommendation System Data Communication

The data communication between the Notification and Recommendation Systems is a multifaceted process that combines data anonymisation, machine learning, and data deanonymization to deliver personalised recommendations while preserving sensitive information. In this workflow, the Notification System serves as the initial point of interaction with organisational data. It employs advanced anonymisation methods to transform sensitive data into open data, ensuring that individual identities

and confidential details remain protected. This anonymised data is then securely stored on an external service, maintaining a clear distinctiveness between sensitive and non-sensitive information.

When a recommendation request is triggered within the notification system, it activates the recommendation system to generate personalised suggestions. The recommendation system calls upon a machine learning model trained with the anonymised open organisational data. This model leverages the anonymised dataset to generate recommendations, without compromising the privacy of the original data.

The resulting recommendation is then sent back to the notification system via the Notification API. At this point, the notification system decodes the recommendation by performing data deanonymization. This process is executed with the goal of ensuring that the recommendation can be presented to the end user in a comprehensible and useful manner. This process, the system maintains a balance between personalization and data privacy.

7.5.3 CERN Notification System Recommendation Feedback Mechanism UI

Figure 7.4 shows a user interface (UI) displaying a list of channel recommendations, which is part of the CERN Notification System. This UI was created for users to provide feedback on the recommendations they receive. This UI implementation is an integral element for the recommendation feedback mechanism, introduced in Section 7.1.

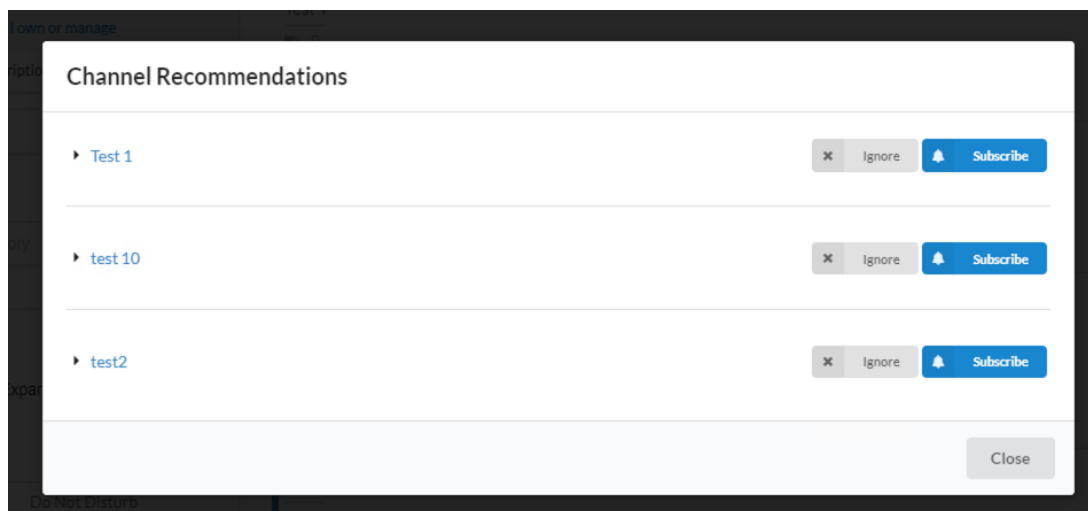


Figure 7.4: Recommended Item List UI for the CERN Notifications System

7 Development

The creation of this UI was driven by the imperative need to engage users proactively, enabling them to provide feedback on the accuracy and relevance of the suggested channels. In the interface, users are presented with a curated list of recommended channels with channel descriptions that aid in informed decision-making. Importantly, this interface enables the users to interact with each recommendation in one of two ways: to either "Ignore" or "Subscribe" to the recommendation. When a user opts to "Ignore" a suggestion, they signal that the recommendation does not align with their preferences or interests. In contrast, when a user chooses to "Subscribe" to a suggested channel, they explicitly acknowledge that the recommendation fits their preferences. These two distinct feedback mechanisms provide valuable information that serves two fundamental purposes.

First, the feedback collected serves as a basis for evaluating the effectiveness of the recommendation algorithm. Patterns emerging from user "Ignore" and "Subscribe" actions provide actionable data to assess the accuracy of generated recommendations. This evaluation loop is crucial for enhancing the precision of the recommendation algorithm over time.

Secondly, the aggregated feedback can be used for further algorithmic development. By analysing the patterns of user interactions with recommendations, the recommendation system can learn and adapt to users' evolving preferences. Subsequently, this iterative feedback-driven approach contributes to the creation of more effective recommendation algorithms that, in turn, amplify user satisfaction and engagement.

The interface displayed in Figure 7.4 is an important component in the recommendation feedback mechanism. By empowering users to express their preferences explicitly through "Ignore" and "Subscribe" actions, this UI enables not only the evaluation of existing algorithms, but also the enhancement and creation of new algorithms that align more closely with users' evolving needs.

7.6 Summary

In this chapter, the detailed process of creating a recommendation system and seamlessly integrating it into the CERN Notifications System is outlined. The chapter begins by providing an exploration of the recommendation system workflow, explaining each step involved in the generation of personalised suggestions. Particular emphasis is placed on the preservation of user data privacy throughout this process, ensuring that sensitive information is rigorously protected.

A significant highlight of this chapter is the utilisation of KubeFlow, a powerful platform used for the creation of machine learning models. This chapter offers a clear and insightful illustration of how raw data is methodically transformed into valuable recommendations, all while adhering to strict privacy standards. This step is crucial for achieving a balance between personalization and data security.

Moreover, KServe is thoroughly discussed, describing its role with KubeFlow for the generation of predictions within a privacy-preserving framework. Technical difficulties and best practises are detailed to ensure that sensitive information remains protected, while delivering recommendations.

The chapter also addresses the essential aspect of extending the CERN Notification System to seamlessly accommodate these innovative recommendations. This integration is of paramount importance in harnessing the power of personalization while ensuring that the system remains adaptable and user-centric.

Finally, the CERN Notification System Recommendation Feedback Mechanism UI is demonstrated, representing a tool that not only suggests channels but also empowers users to provide valuable feedback. This feedback loop is instrumental in the ongoing enhancement of the overall system.

Although chapter 6 combined the knowledge gleaned from previous chapters (4, 3, and 5) to formulate a conceptual architecture for the recommendation system, this chapter represents the practical representation of that concept.

The central focus throughout this chapter is on creating a personalised recommendation system in a way that preserves privacy. This involves planning and execution, ensuring that user data remain confidential and protected at all stages. Additionally, practical elements that transform organisational data into open data, based on the work produced in previous chapters (3 and 6) were introduced.

In the end, this chapter addresses the fundamental question of how sensitive information can be leveraged in a privacy-preserving manner for the development of a personalised recommendation and information retrieval system. Designing recommendation systems in this way ensures that the produced systems are reproducible, enabling easy validation if there are errors in the recommendation process. Furthermore, this approach is seamlessly aligned with the open data principles. Moreover, it provides an opportunity to share problems outside of the organisation since it is based on open data and reproducible pipelines, fostering transparency and external collaboration.

8 Evaluation

This chapter focusses on evaluating how recommendation algorithms perform on implicit datasets. This evaluation involves two types of evaluations. The first type focused on offline evaluations to determine how well different recommendation algorithms work with a custom open dataset. This custom dataset was created following the instructions in Chapter 4 and resembles the datasets used in publishing and subscription systems as described in Chapter 2. The results of this evaluation were used to select the best performing algorithm for an implicit data recommendation and cluster-based recommendation, which were evaluated in two user studies. The first user study evaluated the best-performing cluster-based algorithm and was distributed via the CERN email system. The second user study evaluated the best performing implicit data recommendation system and was distributed through a custom UI component in the notification system. This chapter highlights the strengths and limitations of these algorithms with implicit datasets and evaluates the real-life usage of such algorithms.

- **Jakovljevic, I.,** Gütl, C., & Wagner, A. (2023). Privacy-Preserving Collaborative Filtering: Evaluating a Machine Learning Recommender System in a Large Interconnected Organization. In 5th International Open Search Symposium
- **Jakovljevic, I.,** Gütl, C., & Wagner, A. (2023). Privacy-Preserving User Clustering: The Application of Anonymized Data to Community Detection in Large organisations. IARIA JOURNALS
- **Jakovljevic, I.,** Pobaschnig, M., Gütl, C., & Wagner, A. (2022). Privacy Aware Identification of User Clusters in Large Organisations based on Anonymized Mattermost User and Channel Information. Proceedings of the 11th International Conference on Data Science, Technology and Applications - IARIA DATA ANALYTICS
- Bobic, A., **Jakovljevic, I.,** Gütl, C., Le Goff, J., & Wagner, A. (Accepted/In press). Implicit User Network Analysis of Communication Platform Open Data for Channel Recommendation. In 9th International Conference on Social Networks Analysis, Management and Security - SNAMS 2022

8.1 Contribution

This chapter is the core component of the "Evaluation and Validation" phase of the research methodology. It addresses the following research question: "How can sensitive

information be used for privacy-preserving personalised recommendations, and how does such a system perform compared to traditional ones?”. The key contributions in this chapter come from two data-based evaluations. The Implicit Recommendation Algorithm Evaluation explores the possibilities of detecting user groups without compromising privacy and evaluates clustering algorithms on sparse, anonymised user data. The Implicit Recommendation Algorithm Evaluation examines detecting user groups without compromising privacy and evaluates clustering algorithms on sparse, anonymised user data. These evaluations identify the best algorithms for implicit data recommendations while protecting privacy. This chapter also includes two user studies. The first study examines the performance of implicit recommendation algorithms that have been integrated into the CERN notification system, measuring its real-world performance and user reactions. The second user study collects information from users interacting with a cluster-based implicit recommendation system and explores user reactions to email-based recommendations.

8.2 Data-based Evaluations

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author’s earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by, and taken from the work published in the following publication (Bobic et al., 2022).

Offline evaluation (Data-based Evaluation) involves assessing the performance of a recommendation system using historical data where various metrics are applied to measure the accuracy and quality of recommendations generated by the system.

As part of the data-based evaluation, two different assessments were carried out. The first one focused on analyzing popular recommendation algorithms for implicit data and examining the effect of extracting complex measures from implicit data on recommendations. The second evaluation centered around assessing clustering-based recommendation algorithms using implicit data.

8.2.1 Data Sources

The following data sources are used in both data-based evaluations, to enable a non-biased evaluation opportunity for different recommendation algorithms using implicit data and to ensure that the dataset used for offline evaluation is similar to the actual notification dataset.

The anonymised CERN Mattermost dataset (CDHF)¹ is analysed with the CERN Data Handling Framework² and various Python tools to extract complex factors used

¹<https://zenodo.org/record/6319684>

²<https://github.com/mpobaschnig/cdhf>

for an experimental RS (Jakovljevic, Pobaschnig, et al., 2022). The dataset represents a one-time SQL data snapshot of the Mattermost communication platform used by CERN employees between January 2018 and November 2021. It consists of five core connected entities described in Figure 8.2.



Figure 8.1: Simplified Diagram of the CERN Mattermost Dataset, taken from (Bobic et al., 2022)

In addition to entity properties included in CDHF, multiple additional properties listed in Table 8.2.1 were extracted. Building and organisational unit employee counts were extracted by counting the number of unique employees assigned to each building or unit. The employees with no building and organisational unit assigned were given the employee type external, while all other employees were considered internal. Multiple properties had to be extracted to enrich the channel entity. First, channels were assigned a public or a private label depending if they were extracted directly through the channel entity in the dataset or if they were extracted through a user property indicating to which channels a user belongs to respectively. Next, the active member counts were extracted by counting the number of channel members who sent at least one message to a particular channel. Finally, popular member counts were extracted by counting the number of channel members mentioned at least once in a specific channel.

Entity	Property	Value span
Building	Building employee count	\mathbb{N}_0
Organisational unit	Org. unit employee count	\mathbb{N}_0
Employee	Employee type	Internal, External
Channel	Channel type	Public, Private
Channel	Active member count	\mathbb{N}_0
Channel	Popular member count	\mathbb{N}_0

Table 8.1: Mattermost dataset entity properties were extracted for the purpose of calculating complex relevance factors, taken from (Bobic et al., 2022)

8.2.2 Data-based Evaluation of Implicit Recommendation Algorithms

This section is based on, supported by, and taken from the work published in the following publication (Bobic et al., 2022) in which **Aleksandar Bobic** and **Igor Jakovljevic** contributed equally. The majority of the contribution in this section was done by **Aleksandar Bobic**.

Study Design

The initial evaluation revolved around the identification, extraction and assessment of various measures from implicit data, with potential utility for recommender systems. Additionally, we employed four widely recognised recommendation algorithms to gauge their effectiveness in terms of both the implicit data and the significance of the extracted measures. Additionally, four popular recommendation algorithms were selected to assess their effectiveness with implicit data and the impact of the extracted measures.

Research Question 8.1: *How do traditional recommendation algorithms perform with implicit datasets, and what is the influence of enhanced metrics compared to simple ones on algorithm performance within the context of large organisations?*

Data Preprocessing

Multiple assumptions had to be made to extract the weighted connections between channels and their users, which will be later used for channel recommendations. These assumptions resemble those of related work (Gupte & Eliassi-Rad, 2012; Lattermann et al., 2009). The end result is a set of simple factors introduced in Table 8.2.2 which are then combined into the first complex factor presented in (8.1).

$$C_{u,c} = \alpha * TM_{u,c} + (1 - \alpha) * (\beta * M_{u,c} + (1 - \beta) * MT_{u,c}) \quad (8.1)$$

Three simple factors $TM_{u,c}, M_{u,c}, MT_{u,c} \in [0,1] \subset \mathbb{R}$. $TM_{u,c}$ are defined as the average ratio of the number of team t channels that user u belongs to ($NT_{t,u}$), to the number of all channels in team t (NT_t) plus the ratio of NT_t to the largest number of channels in a team ($\max(NT)$). $M_{u,c}$ is defined as the averaged ratio of messages u posted in c ($NM_{u,c}$) to the number of all messages in c (NM_c) plus the ratio of the number of users who wrote at least one message in c (A_c) to the number of all users in c (S_c) subtracted from 1. Finally, $MT_{u,c}$ is the average ratio of the number of mentions u received in c ($NMT_{u,c}$) to the number of all mentions in c (NMT_c) plus the ratio of the number of users that received at least one mention in c (P_c) to S_c subtracted from 1. All the factors introduced in this table belong to $[0,1]$. Team membership and

message in their simplest form contribute to the RS while the mention factor is not taken into account due to scarcity. In this first attempt of calculating complex measures, we assume independence and equal contribution of simple factors.

Since channels are grouped into teams that represent a particular topic or subgroup of CERN users with common interests, the first assumption takes this into account in the team membership factor $TM_{u,c}$ (see Table 8.2.2). If u belongs to many channels of the team t and t is very large; then these channels are more relevant to u and $TM_{u,c}$ has a higher value. The second assumption states that if the number of messages u sent to channel c represents a large percentage of messages in c and if only a small percentage of members of c are sending messages in c then this user has a greater interest in this channel and therefore the message factor $M_{u,c}$ will be higher. The third assumption states that if the number of mentions of u in c represents a large percentage of all mentions in c and if a small percentage of all c members are mentioned in c then c is strongly related to u and the mention factor $MT_{u,c}$ has a higher value.

Factors	Factor Formula
Team membership factor	$TM_{u,c} = 12 * (NT_{t,u}NT_t + NT_tmax(NT))$
Message factor	$M_{u,c} = 12 * (NM_{u,c}NM_c + 1 - A_cS_c)$
Mention factor	$MT_{u,c} = 12 * (NMT_{u,c}NMT_c + 1 - P_cS_c)$

Table 8.2: Factors for the Creation of Complex Measures

We use all values of $\alpha \in \{0,0.5,1\}$ and $\beta \in \{0,0.5,1\}$ to identify which of the previous simple factors or combinations of simple factors most realistically reflect the connection weight between u and c . The three values for α and β are used to represent states where the first factor is favoured more, the second factor is preferred more, or both factors are favoured equally. The results are later evaluated as part of the RS evaluation.

To identify user subgroups with potentially common interests, an implicit SN was created and analysed using a number of implicit signals computed with the following assumptions.

A connection between users is directional, since there could be popular users in a channel known to everyone¹. The user from whose perspective the measure of visibility is calculated is called the ego and is represented by u_e while their neighbours are represented by u_n . The more active u_n is in c ² the more visible they are to others in c and the more such channels u_n and u_e share, the more visible u_n is to u_e . If two users have a high visibility to each other in multiple channels, it indicates that they might

¹For example a channel may have an administrator that takes care of posting important information regularly in the channel and is also mentioned by many users in the channel when his or her input is needed.

²The activity of a user is measured by the mention and message factor.

8 Evaluation

share common interests.

Based on the above, the visibility $V_{u_n, c} \in [0, 1] \subset \mathbb{R}$ detailed in (8.2) represents the sum of $2 \div S_c$ where S_c is the number of users in channel c^1 , the mention MT_{c, u_n} and message M_{c, u_n} factors. $\eta \in \{0, 0.5, 1\}$ and $\gamma \in \{0, 0.5, 1\}$ are used to evaluate the factors' influence on the final score.

$$V_{u_n, c} = \gamma * 2 \div S_c + (1 - \gamma) * (\eta * M_{c, u_n} + (1 - \eta) * MT_{c, u_n}) \quad (8.2)$$

Finally if u_e and u_n work in the same building ($B_{u_n, u_e} \in \{0, 1\}$ is 1) and belong to the same organisational unit ($O_{u_n, u_e} \in \{0, 1\}$ is 1), the likelihood of them knowing each other greatly increases. This likelihood further increases with the decreasing numbers of users in the building and organisational unit, represented by $BS_{u_e} \in \mathbb{N}$ and $OS_{u_e} \in \mathbb{N}$ respectively. Since we only consider building and organisational units with at least two users the last two factors of the equation can have the maximum value of 0.5 each. These assumptions resulted in (8.3), a measure of the likelihood of u_e knowing u_n . The factors representing the previous assumptions are weighted by $\epsilon \in \{0, 0.5, 1\}$ in (8.3). S_{u_n, u_e} is the set of channels shared between u_e and u_n , N_{u_n, u_e} is the number of these channels and N_{u_e} is the number of channels u_e belongs to. Since the likelihood of users knowing each other through very large channels is small² and since these channels would create connections with very low weights between many users and introduce a lot of noise in the data they were not considered in (8.2) and (8.3).

$$V_{u_e, u_n} = 12 * \left(\epsilon * \sum_{c \in S_{u_n, u_e}} V_{u_n, c} N_{u_n, u_e} + (1 - \epsilon) * N_{u_n, u_e} N_{u_e} + O_{u_n, u_e} OS_{u_e} + B_{u_n, u_e} BS_{u_e} \right) \quad (8.3)$$

Using the permutations of factors β , γ , and ϵ multiple directed weighted implicit networks have been created where users are represented with nodes, and the directed edges between them are present if they have a channel in common and weighted using (8.3). Finally, the Leiden algorithm was used to identify user communities using the default configuration where a set of communities for network n is defined as CM_n . To identify networks that produced useful user clusters and could improve the performance when used in a RS, their modularity was calculated. All networks had a modularity between 0.45 and 0.6, with the weight between users being defined only by the inverse channel size performing the worst and the weight using only the mention factor (modularity 0.602) the best, followed closely by a weight with only the message factor (modularity 0.601). These values are comparable to values of real SN and could potentially indicate that the generated measures work very well. Once users are

¹Only channels with at least two users are considered, therefore the minimum value of S_c is 2.

²We defined very large channels as channels with more than 500 users.

separated in smaller groups through clustering a recommend system still has to be trained using weights between users and channels.

Finally, taking inspiration from approaches such as Gmail's recipient recommendation, another measure of channel relevance was extracted (Roth et al., 2010). This new weight $C_{u_e,c}^* \in [0, 1] \subset \mathbb{R}$ is defined by (8.4) and is a weighted normalised sum of connections between an ego u_e and one of their channels c through all of their neighbors u_n which also belong to c . $U_{u_e,c}$ represents all neighbors of u_e connected through channel c and $N_{u_e,c}$ represents the number of these neighbors. The more strongly visible users u_e shares c with, and the more visible u_n is in c , the higher the value of $C_{u_e,c}^*$ will be. It is assumed that V_{u_e,u_n} and $C_{u_n,c}$ contribute equally to the individual weights between u_e and c . These weights are also used instead of $C_{u,c}$ to evaluate the experimental RS.

$$C_{u_e,c}^* = 1/N_{u_e,c} * \sum_{u_n \in U_{u_e,c}} \left(12 * (V_{u_e,u_n} + C_{u_n,c}) \right) \quad (8.4)$$

In addition to the previously mentioned data processing steps, CF techniques require supplementary preparation and transformation of the CERN dataset. CDHF was used to extract user channel membership information from the data set. Outliers, such as excessively large channels (more than 2000 members), small channels (less than 4 members), and private channels, were discarded to reduce data noise and for privacy reasons.

Procedure

This section is based on, supported by, and taken from the work published in the following publication (Bobic et al., 2022) in which **Aleksandar Bobic** and **Igor Jakovljevic** contributed equally. The majority of the contribution in this section was done by **Igor Jakovljevic**.

The following CF methods have been selected because they work well with implicit datasets: Alternating Least Squares (ALS)(Hu et al., 2008), Bayesian Personalised Ranking (BPR)(Rendle et al., 2012), Logistic Matrix Factorization (LMF)(Johnson, 2014), and Neighbourhood Models (KNN)(Renaud-Deputter et al., 2013). These methods are based on matrix factorizations, which assign vectors of factors to both items and users based on item-rating patterns. The goal of these recommendation methods is to recommend channels to users, based on their channel membership and the memberships of similar users. The mentioned techniques required the creation of the following matrices: Rating Matrix, Preference Matrix, and User Preference Confidence Matrix. (Hu et al., 2008; Verstrepen et al., 2017; Renaud-Deputter et al., 2013; Takács et al., 2011; Rendle et al., 2012).

Let \mathbf{U} be a set of all Users and \mathbf{I} a set of all Channels, we define $\mathbf{R} \in \mathbb{R}^{|\mathbf{U}| \times |\mathbf{I}|}$ as the

8 Evaluation

ratings matrix and $\mathbf{L} \in \mathbb{L}^{|U| \times |I|}$ as the user channel link matrix. For explicit datasets, rating matrix values indicate the preference by user u of item i . For implicit datasets, values are calculated using observations of users actions. We used the number of user messages (*msg_count*) from user u to channel i as $r_{u,i}$. A value in the link matrix (\mathbf{L}) is 1 if a connection between a user and a channel exists, otherwise 0. Based on \mathbf{R} we created the user preference confidence matrix \mathbf{C} which measures the confidence in observing user preferences, was calculated as $c_{u,c} = 1 + \alpha \log(1 + r_{u,c}/\epsilon)$. Where α is the rate of confidence increase and ϵ is a value near zero (Hu et al., 2008).

Name	Description
Simple Measure Collaborative Filtering (SMCF)	Using user channel message count (<i>msg_count</i>) as rating values ($r_{u,i}$) and applying CF algorithms on user channel connections (\mathbf{L})
Complex Measure Collaborative Filtering (CMCF)	Using user channel measures ($\mathbf{C}_{u,c}$, $\mathbf{C}_{e,c}^*$) defined in section 8.2.2 as rating values ($r_{u,i}$) and applying CF algorithms on user channel connections (\mathbf{L})
Team and Organisation Grouping with SMCF (TSMCF and OSMCF)	Grouping Users based on their Team/Organisation connection (section 8.2.2) to produce Team ($\mathbf{L}_t^1 \dots \mathbf{L}_t^n$) and Organisation ($\mathbf{L}_o^1 \dots \mathbf{L}_o^n$) link vectors. Then applying SMCF to the produced vectors (\mathbf{L}_t^* , \mathbf{L}_o^*)
Team and Organisation Grouping with CMCF (TCMCF and OCMCF)	Grouping Users based on their Team/Organisation connection (section 8.2.2) to produce Team ($\mathbf{L}_t^1 \dots \mathbf{L}_t^n$) and Organisation ($\mathbf{L}_o^1 \dots \mathbf{L}_o^n$) link vectors. Then applying CMCF to the produced vectors (\mathbf{L}_t^* , \mathbf{L}_o^*)
Clustered Users with SMCF (CSMCF)	Applying SMCF on user communities (\mathbf{CM}_n) created from measures (\mathbf{V}_{e,u_n}) defined in section 8.2.2
Clustered Users with CMCF (CCMCF)	Applying CMCF on user communities (\mathbf{CM}_n) created from measures (\mathbf{V}_{e,u_n}) defined in section 8.2.2

Table 8.3: Recommender System Use-Cases and Definitions

Previous research mentioned in Chapter 2 indicates that applying collaborative filtering on sparse datasets would not produce acceptable results. To this end, explicit grouping and clustering methods were applied to identify organisational and interest groups in the dataset (Hu et al., 2008). The above-mentioned matrices were created to fit the use-cases from Table 8.3. After the creation of the user channel matrices, they were split into two groups. The first group contained 80% of user channel matrices data and was used to train CF models. The second group contained the remaining 20% and was used to evaluate the performance of the selected models (Wong, 2015).

After the extraction of optimal configuration parameters (Table 8.4) for CF algorithms and the definition of important use-cases, the previously defined simple and complex

user similarity metrics (Table 8.3) were used to train RS.

Name	Rate of Confidence Increase	Number of Training Iterations	Number of User Latent Factors
Value	40	150	100

Table 8.4: The most performing configuration was identified by analyzing 2201 different attribute combinations. The values of the most performing combination are similar to the ones in literature (Hu et al., 2008; Renaud-Deputter et al., 2013)

SMCF was used as the baseline configuration for CF to determine implicit user interests. We used SMCF since it uses simple features most similar to baseline features found in the literature (Shi et al., 2014). For TSMCF, OSMCF, TCMCF, and OCMCF explicit grouping of users to teams and organisations were used to execute CF algorithms on user channel memberships of the respective team or organisation group. For each \mathbf{o} from the set of all Organisations \mathbf{O} a new vector was created \mathbf{L}_o^n where \mathbf{n} is the id of the organisation. The elements of this vector are user ids of the users from the set \mathbf{U} that belong to the organisation \mathbf{O} . Respectively the team vectors \mathbf{L}_t^n were created. For CMCF and SMCF, the clustering metrics mentioned in section 8.2.2 were used to compartmentalise and apply CF algorithms per cluster.

Error-based metrics are used for measuring predictive accuracy, but are not suitable for a list-wise evaluation approach for recommendations since they treat all recommended items equally. For the evaluation of the CF models, rank-based metrics such as Area Under the Curve (AUC), Normalized discounted cumulative gain (NDCG), and Mean Average Precision (MAP) are more suitable (Aftab & Ramampiaro, 2022; Hu et al., 2008). A performant model should strive to have a high value (close to 1) for MAP, AUC, and NDCG. AUC values close to 0.5 indicate that the model does not perform well, since it does not distinguish between recommendations. Low MAP and NDCG indicate that the model does not identify relevant items and has a low rating quality for models (Aftab & Ramampiaro, 2022; Renaud-Deputter et al., 2013; Rendle et al., 2012).

Results and Discussion

Table 8.5 contains only the best-performing models and represents a portion of the results. In total 578 CF models were evaluated with different user confidence matrices, generated from simple and complex measures. For SMCF and CMCF 73 different confidence matrices were evaluated. Additionally each explicit grouping by team (TSMCF, TCMCF), organisation (OSMCF, OCMCF) and clustering (CSMCF, CCMCF) were evaluated with 72 confidence matrices.

8 Evaluation

Use Case	Algorithm	Rating Measure	AUC (Mean, Max)	Precision (Mean, Max)	NDCG (Mean, Max)
SMCF	ALS	<i>msg_count</i>	(0.6 , 0.6)	(0.33 , 0.33)	(0.33 , 0.33)
CMCF	ALS	$C(\alpha = 0; \beta = 0)$	(0.65 , 0.65)	(0.4 , 0.4)	(0.42 , 0.42)
OSMCF	BPR	<i>msg_count</i>	(0.68 , 1)	(0.36 , 1)	(0.28 , 1)
TSMCF	LMF	<i>msg_count</i>	(0.84 , 1)	(0.65 , 1)	(0.55 , 1)
OCMCF	BPR	$C(\alpha = 0.25; \beta = 0.25)$	(0.7 , 1)	(0.41 , 0.41)	(0.43 , 0.43)
TCMCF	LMF	$C(\alpha = 0; \beta = 0)$	(0.845 , 1)	(0.67 , 1)	(0.58 , 1)
CSMCF	BPM	<i>msg_count</i> $V(\beta = 0; \gamma = 0; \varepsilon = 1)$	(0.61 , 1)	(0.21 , 1)	(0.22 , 1)
CCMCF	BPM	$V(\beta = 0.5; \gamma = 0.5; \varepsilon = 0.5)$ $C(\alpha = 0.25; \beta = 0.25)$	(0.62 , 1)	(0.23 , 1)	(0.23 , 1)

Table 8.5: Experimental Results of Best Performing CF Algorithms with Optimal Configurations

The evaluation results indicate that the usage of custom user similarity measures in combination with methods to reduce data sparsity has a positive influence on the recommendation results. It is also visible that the usage of explicit data, such as explicit grouping, greatly improves recommendations in comparison to clustering approaches. An important conclusion to this evaluation is that it is possible to create a well-performing recommender system without using sensitive user data and invading user privacy. Our results also show that techniques such as BPR or LMF outperform ALS and KNN by a small margin, which is in line with previous research (Rendle et al., 2012). Due to data sparsity issues, our models did not reach the AUC values (0.96) presented in previous research, thus confirming their data sparsity concerns (Rendle et al., 2012). The decrease in data sparsity by grouping or clustering users yields better results, which is in line with the literature and the experimental CF results obtained from different datasets. Representing user preferences as implicit data is a difficult task. An influential factor in the performance of an implicit data-based CF recommender system is the validity and strength of the implicit data. Implicit data has the benefit of protecting user privacy, but devalues the quality and understanding of the data.

However, since all assumptions used to create custom measures are based on implicit behavioural data and are inspired by previous work, some of our assumptions might not accurately represent real-world behaviour. These assumptions could lead to an

inaccurate representation of connections between users and channels. As part of this exploratory work, we attempt to mitigate this by weighting each factor created based on our assumptions and testing multiple weight configurations. As part of future work, a meaningful sample of Mattermost users at CERN should be interviewed about their behaviour patterns and the accuracy of our custom recommender system recommendations. The interview results would enable us to understand better user behaviour in Mattermost in the context of large organisations and test our assumptions' accuracy.

8.2.3 Data-based Evaluation of Cluster Based Implicit Recommendation Algorithms

The following section builds upon previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by, and taken from the work published in the following publication (Jakovljevic, Pobaschnig, et al., 2022).

Study Design

In this section, the focus is shifted towards the assessment of various clustering algorithms and their applications on implicit data-driven recommendations. The evaluation process consists of a series of steps. First, a set of widely used clustering algorithms was detected, followed by the identification of a comprehensive list of metrics to assess their performance. Lastly, the algorithms were evaluated using the previously mentioned metrics on the Mattermost dataset.

Research Question 8.2: *How do cluster-based recommendation algorithms perform with implicit datasets in large organisations, and what is the influence of community and group information on enhancing recommendation accuracy and relevance?*

Preprocessing

As mentioned above, the data set includes 21231 users, 2367 teams, and 12773 channels. Most of the teams are relatively small, typically consisting of 1 to 18 members, with an upper limit of 41. However, teams with just one member essentially act as isolated nodes, not contributing to the creation of a graph structure. Consequently, these single-member teams can be excluded to better represent the data used.

Table 8.6 shows that most teams have memberships ranging from 2 to 23 users, with a maximum of 52 members. However, the upper limit of 52 is arbitrary and may not be the most suitable criterion for identifying meaningful communities. To explore this, various thresholds are considered for comparative analysis: 23, 52, 200,

8 Evaluation

Minimum	Lower Limit	Q_1	Median	Q_3	Upper Limit	Maximum
2	2	4	10	23	52	4512

Table 8.6: Recalculated five-number summary of members withing teams ignoring teams with one member.

500, 1000, and no threshold. These thresholds will help assess the impact of varying channel membership counts on the graph structure and, ultimately, identify the optimal threshold value.

Procedure

The preceding section introduced the essential components required for the evaluation of the chosen algorithms offered by igraph. On the basis of the assessment of data metrics, statistics, and the CDHF, graph structures are generated. Because various community detection algorithms rely on different underlying methodologies, each graph corresponding to different thresholds must be individually assessed to determine the algorithm that produces the most favourable outcomes. These communities serve as the basis for extracting collaborative teams formed by each user, with a focus on recommending the most popular teams to other users within the community.

The creation of a separate graph for each threshold involves utilising the relationships between users and the Mattermost channels to which they belong. Figure 8.2 visually depicts the concept of graph creation. For instance, when two users, denoted user 1 and user 2, are both members of channel 1, an edge between them is established with an initial weight of 1. If these same users share membership in additional distinct channels, the edge weight is incremented, as exemplified in channel 2, leading to the creation of new edges connecting them with other users.

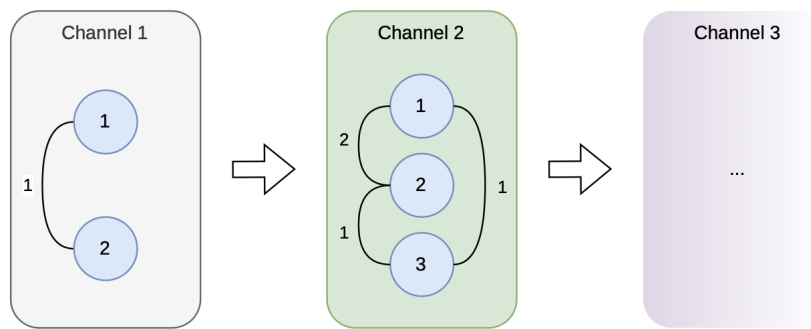


Figure 8.2: Demonstration of a Simplified Graph Creation Process

Table 8.8 lists several advanced clustering algorithms utilised to detect communities in the Mattermost dataset along with the corresponding evaluation results. The mentioned clustering algorithms were selected for evaluation because they were the commonly used algorithms for clustering and were also available in those of the igraph library.

The evaluation of recommendation algorithms employs a statistical approach that relies on a designated test set containing connections between users and channels. This test set serves as the foundation for assessing the quality of the recommendations generated by the algorithms. In essence, evaluation operates as a classification problem, where the objective is to compare and evaluate the actual state against the predicted state. In this context, the actual condition involves the presence or absence of edges between users and channels, whereas the predicted condition is the value made by the recommendation algorithms. The outcome of this comparison encompasses both positive and negative results. To provide a comprehensive evaluation, key metrics such as sensitivity, specificity, precision, and F-score will be utilised, enabling a thorough examination of the performance and effectiveness of the recommendation algorithms. In addition to these primary evaluation metrics, an analysis was performed to examine the impact of various parameters of community creation. The specific parameters under consideration are detailed in Table 8.7.

Parameter	Description
Upper User Threshold	The number of users above which teams are omitted for graph creation.
Modularity	Value representing the modularity of the community detection as explained in (Newman & Girvan, 2004).
Communities	Number of communities found within the graph.
Duplicates	Number of duplicates found between discovered communities and Mattermost teams. Duplicates occur when one Mattermost team matches multiple communities. This is only considered for Mattermost teams and not for organisational units since organisational units are not well captured.
Average Size	The average size of a community.
Average Runtime	The average runtime of the algorithm in seconds.

Table 8.7: Parameters and their description of the first table of each algorithm result.

In the initial evaluation step, the average edge weight of the graph for each threshold is calculated to get a general sense of the weight influence. Then, starting alphabetically with Fastgreedy, the algorithms provided by igraph are deployed on each graph of the different thresholds. The results of the algorithm run, as described in table

8 Evaluation

8.7, are analysed for each threshold. The edge weight over thresholds and various algorithms deployed over different thresholds and their results are analysed. Due to the randomness within several algorithms, multiple runs were performed, and the average and standard deviation over all iterations were calculated to get more precise results. The literature suggests the use of 100 iterations for the evaluation of community detection algorithms. However, from the measurements, 25 iterations are a good trade-off between precision and runtime, where the average change between iterations is below 0.5%.

To assess the similarity between the generated clusters and the actual clusters, a suitable similarity metric was required. Similarity is calculated using the Jaccard similarity of members between a discovered community and a Mattermost team (or organisational unit). For example, given the community discovered C and the Mattermost team T , the similarity is $J(C, T) = \frac{|C \cap T|}{|C \cup T|}$. The higher the number of common members between the two sets, the higher the similarity. The similarities are used with the five-number summary for comparison, as described in Table 8.6.

Results and Discussion

As seen in Table 8.8, seven of the ten community detection algorithms provided by igraph were consistently used on all thresholds. However, the Spinglass algorithm was only evaluated for thresholds 23 and 52 due to runtime constraints encountered with higher thresholds. The Edge Betweenness and Optimal Modularity algorithms were entirely excluded from consideration for all thresholds, given similar runtime concerns. The modularity values were consistently observed to be high, ranging between 0.86 and 0.76 between algorithms. In particular, the Fastgreedy, Leading Eigenvector, Multilevel, and Spinglass algorithms identified relatively fewer communities compared to their counterparts, despite achieving comparably high modularity scores. The average community size showed substantial variation. Algorithms like Fastgreedy, Leading Eigenvector, Multilevel, and Spinglass exhibited larger average community sizes and greater standard deviations, primarily due to their limited community discovery. Infomap, Label Propagation, and Walktrap algorithms discovered the most communities, generating 347 ± 1 , 454 ± 8 , and 379 ± 0 communities, respectively. The Leiden algorithm performed moderately, identifying around 120 ± 0 communities across different thresholds. Infomap and Label Propagation closely approximated the typical size of the Mattermost team, yielding average community sizes of 17 ± 10 and 13 ± 10 , respectively. Moreover, the Leiden algorithm's community discovery pattern was distinctive, particularly for threshold 23, where it detected fewer than 1200 communities. In terms of organisational units and communities uncovered, minimal similarities were discerned. Across algorithms, graphs with an upper threshold of 52 consistently demonstrated raised modularity values, ranging from 0.64 to 0.78. Additionally, the

Infomap, Label Propagation, and Leiden algorithms yielded average community sizes of 23 ± 20 , 21 ± 23 , and 7 ± 10 , respectively, closely mirroring the average team size within the Mattermost dataset.

Algorithm	Communities	Modularity	Minimum [%]	Q_3 [%]	Maximum [%]
Fastgreedy	44 ± 0	0.83 ± 0.00	8.03 ± 0.00	51.32 ± 0.00	91.67 ± 0.00
Infomap	347 ± 1	0.79 ± 0.00	25.95 ± 1.01	77.71 ± 0.22	100.00 ± 0.00
Label Propagation (teams)	454 ± 8	0.76 ± 0.00	21.75 ± 2.51	79.91 ± 0.46	100.00 ± 0.00
Label Propagation (channels)	475 ± 18	0.67 ± 0.01	0.45 ± 0.14	71.40 ± 0.96	100.00 ± 0.00
Leading Eigenvector	62 ± 0.00	0.77 ± 0.00	4.32 ± 0.00	54.95 ± 0.00	100.00 ± 0.00
Leiden	120 ± 3	0.86 ± 0.00	8.08 ± 0.00	58.57 ± 0.00	100.00 ± 0.00
Multilevel	52 ± 2	0.86 ± 0.00	7.85 ± 0.94	34.04 ± 4.81	100.00 ± 0.00
Spinglass	25 ± 0	0.84 ± 0.00	6.27 ± 0.42	12.65 ± 0.97	21.44 ± 2.60
Walktrap	379 ± 0	0.77 ± 0.00	11.24 ± 0.00	80.00 ± 0.00	100.00 ± 0.00

Table 8.8: Results including communities, modularity, and most important values of the five-number summary of similarities between Mattermost teams and found community with different algorithms at threshold 23. Values within columns represent mean and standard deviation over 25 iterations.

The average run-time of the algorithms exhibits significant variability, ranging from less than a second for several algorithms to nearly ten minutes for Spinglass. Among all algorithms tested, the best performance in terms of similarities between Mattermost teams and identified communities was delivered by Infomap, Label Propagation, and Walktrap. However, it is important to note that all algorithms reveal relatively low similarities between organisational units and the communities they identify. Consequently, the discovered communities tend to align more with user interests rather than with the organisational structure within CERN. Graphs with higher thresholds exhibit reduced modularity, implying lower clustering performance in all algorithms. In general, the number of detected communities decreases with the thresholds increase, except for the Leiden algorithm, which demonstrates alternating behaviour. The reduction in the number of communities with an increase in nodes and edges is typically indicative of added noise to the graph.

8 Evaluation

As the number of communities decreases, the average community size consequently increases. Variations in similarities between Mattermost teams and identified communities are substantial, influenced by the diverse methods used for community detection, which result in clusters of varying sizes. Overall, the similarities between organisational units and the discovered communities are consistently low, regardless of the algorithm or threshold used. The average runtime heavily varies between algorithms, from the fastest taking a few seconds to the slowest taking roughly 30 ± 5 minutes per run.

Despite the inclusion of random operations in certain algorithms, their influence on the outcomes is nearly negligible, with only Multilevel and Spinglass exhibiting minor differences between iterations at specific thresholds. In particular, the Leiden algorithm distinguishes itself by identifying numerous communities across all thresholds. Consequently, a high number of duplicates are encountered. The average community size remains consistently small across all thresholds, effectively representing the typical team size. However, the similarity between Mattermost teams and identified communities is comparatively lower in some algorithms, particularly with higher thresholds.

In conclusion, graphs with a threshold of 23 generally exhibit the best performance in terms of modularity, the number of detected communities, and the similarities between Mattermost teams and the identified communities. Specifically, the Label Propagation and Infomap algorithms demonstrated the best performance with threshold 23, while the Leiden algorithm excelled with threshold 52. The performance of graphs with threshold 23 closely approaches that of threshold 52. Nevertheless, for some algorithms like Leiden, performance is notably inconsistent, while for others like Label Propagation, it exhibits slight improvements. Table 8.8 gives an overview of the results of threshold 23.

Of all available algorithms, infomap, label propagation and random walk algorithms performed the best in terms of modularity, similarity, and communities, as shown in Table 8.8. The label propagation algorithm finds communities with slightly less similarity than the infomap algorithm, which performs best with respect to similarity measurement. However, it finds many and much more detailed communities.

Figure 8.3 represents the similarities of the users between the communities found and the Mattermost teams, and Figure 8.4 illustrates the results of 10 iterations as violin plots.

An upper threshold of 52 for the teams was used for this figure, as described later in this section. Of all detected communities, 75% have similarities above 47.79%, 50% have similarities above 61.18%, and 25% have similarities above 74.99%. Similarities are measured by comparing the discovered community with all Mattermost teams and counting the common members in both sets. The percentage value of the Mattermost team with the most common members is used.

Depending on the number of communities found, there might be overlaps such that

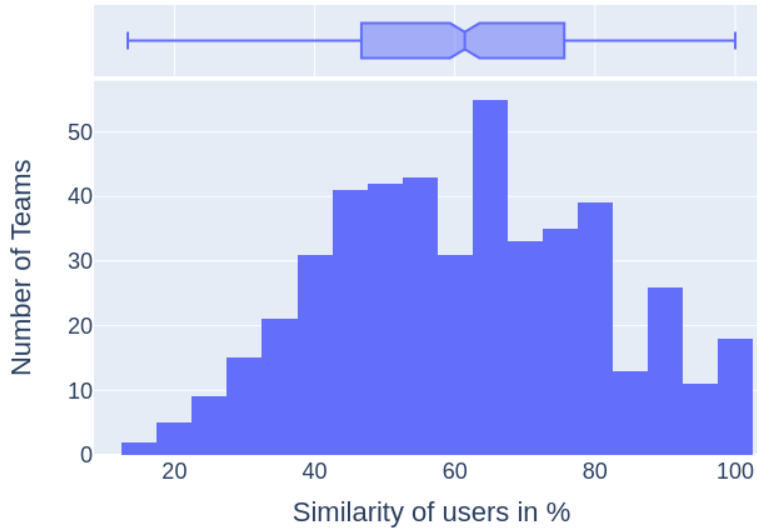


Figure 8.3: Sample run showing similarities of users between found communities and Mattermost teams, taken from (Pobaschnig et al., 2023)

one team fits multiple communities as the best match. This might be the case where the size of communities is smaller than the size of teams, such that communities form subgroups of the teams.

However, less than 0.01% of the communities discovered are matched against the same Mattermost team. The average size of the communities discovered is 20 ± 23 , the minimum is 2, the first quartile Q_1 is 6, the median is 13, the third quartile Q_3 is 26, and the maximum is 421. For the label propagation algorithm, two different methods were used to create the graph. The first method applies the threshold on team members so that teams with more than 52 members are left out. The second method applies the threshold on channel members where all channels with 52 or fewer members are used to build the graph. In table 8.9 the difference in nodes and edges between the methods with threshold 52 is listed.

The results of the label propagation algorithm on thresholds with both methods are presented. In comparison, method 2 already has lower modularity, lower communities, and a larger average size with considerable standard deviation. Method 2 has slightly lower values on the five-number summary with threshold 52 and somewhat higher values on the other thresholds. However, visually analysing the graphs, method 2 produces large clusters already with threshold 52, while method 1 keeps the clusters smaller and contained. This is also indicated by the number of communities found.

8 Evaluation

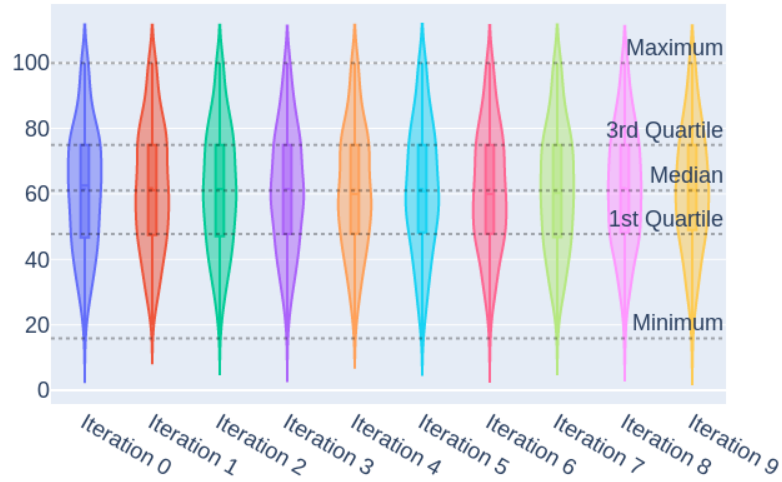


Figure 8.4: Similarities between discovered communities and Mattermost teams over iterations with threshold 52, taken from (Pobaschnig et al., 2023)

Generally, method 2 results in worse results as more users are considered that only participate in teams with more than 52 members, while the average team size is below this threshold.

Method	Nodes	Edges
Team Threshold	9520	151501
Channel Threshold	14540	447695

Table 8.9: Nodes and edges of graphs with threshold 52 created with team and channel method.

Table 8.10 describes the result of the clustering, showing the number of nodes, edges, and overall weight of the edges. The results of thresholds 23 and 52 are addressed individually, as these thresholds delivered the best results, while the other thresholds will be described as a group. However, they might also be grouped altogether, where practical. Expanding the threshold to higher values leads to the inclusion of channels with more users. Consequently, the edge weight between these large numbers of users is added, and the weight level difference within and outside emerged clusters is flattened. With higher thresholds, more users appear, reducing the overall weight within the graph.

Upper User Threshold	Nodes	Edges	Overall Weight
23	5881	46024	3.03 ± 2.43
52	9520	151501	2.94 ± 2.35
200	14906	809012	2.82 ± 2.25
500	17124	1909964	2.65 ± 1.88
1000	17948	3104814	2.53 ± 1.66
None	19682	15194697	2.44 ± 1.62

Table 8.10: Number of nodes, edges, and overall weight of the edges over different thresholds.

8.3 User Feedback-based Studies

In the online evaluation (User Feedback-based Evaluation) approach, the performance of the recommendation system is evaluated by directly collecting user feedback. Users' interactions with the recommended items are observed and analysed to determine the effectiveness and relevance of the recommendations.

8.3.1 User Feedback-based Study of Cluster Based Implicit Recommendation Algorithms

The following section builds on previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Igor et al., 2023)

Research Question 8.3: *How does the performance and user response vary in the context of the CERN notification system when employing a cluster-based implicit recommendation system, particularly focussing on user reactions to email-based recommendations, within an organisational setting?*

Study Design

In this study, we investigate the performance and user reactions with respect to a cluster-based implicit recommendation system within the context of the CERN notification system. The primary objective was to study the effectiveness of such a system within an organisation and how users respond to email-based recommendations. This study spanned a period of 40 days and served as a detailed exploration of the system's dynamics.

Settings and Instruments

Our study involved a group of active users of the CERN notification system. These participants, carefully selected for their participation in the notification system, were invited to participate in the survey. Data collection was carried out through email interactions with selected users. As seen in Figure 8.6, these emails included recommended channels, detailed descriptions of the recommendations, links to resources explaining how the recommendations were generated, and information on how participating in the survey would benefit the recommendation system. The study was carried out over a period of 21 days, during which participants were reminded every 7 days to complete the evaluation. Reminders included detailed explanations of the algorithms and the reasons for participating.

Procedure

The procedure of this study was a six-step process spanning 40 days, the phases of this process can be seen in Figure 8.5.

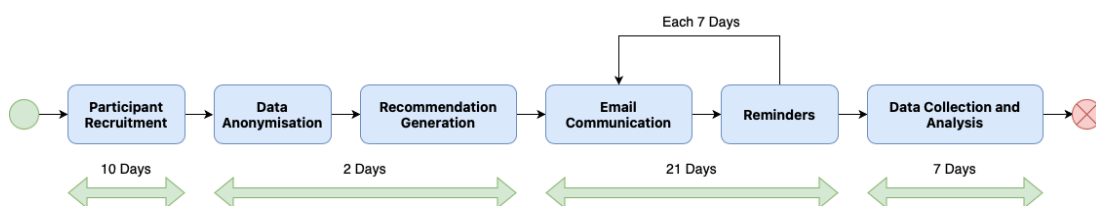


Figure 8.5: Procedure Steps of the Evaluation of Cluster Based Implicit Recommendations

In the participant recruitment step, initially 68 active notification system participants were contacted through a dedicated Mattermost channel. Of these, 15 participants from the IT department agreed to participate in the study.

For the data anonymisation step, user data from selected participants was collected from a PSQL database and converted to open data following the steps outlined in Chapter 4 to protect the privacy and confidentiality of the participants.

Machine learning models were used in the recommendation generation step in the open data produced to train and generate up to five channel recommendations for each of the 15 participants.

Participants were contacted through the notification system in the email communication, receiving emails containing recommendations lists, as seen in Figure 8.6. These emails provided clear explanations of how the recommendations were generated, the purpose of the study, its impact on the recommendation system, and the instructions for providing responses.

[Community Based Channel Recommendations Experiment] - Channel Recommendation

Suggested Channels - Recommendation Evaluation

You are receiving this email since you are part of a simple evaluation of community-based privacy-aware channel recommendations.

This evaluation should not take longer than **5 minutes**.

Please do the following tasks to complete the evaluation:

1. Have a look at each of the suggested channels (including previous notifications)
2. Determine which of the channels is interesting to you
3. After you have determined the interesting channels for you, please send an email to [<receiver.email>](#) with the list of interesting channels.

You might find the following channels interesting (please do not subscribe for now):

- [<Recommended Channel 1>](#)
- [<Recommended Channel 2>](#)
- [<Recommended Channel 3>](#)
- [<Recommended Channel 4>](#)
- [<Recommended Channel 5>](#)

Additional Information

How will this help out?

Your feedback will be used to statistically determine the accuracy and recall of the used algorithms. The evaluation is also a crucial part of a master's thesis and is important for its completion

Hints

If you do not have access to the channels, please do the following:

1. Go to the [Notifications Portal](#)
2. Log in with your main user account
3. Try to open the links for this email again

Want to read more about the actual algorithms used?

Algorithms used to recommend these channels respect your privacy, no personally identifiable data was used to create the ML models.

[Collaborative Filtering for Implicit Feedback Datasets](#)

[Applications of the conjugate gradient method for implicit feedback collaborative filtering](#)

[Privacy Protected Identification of User Clusters in Large Organizations based on Anonymized Mattermost User and Channel Information](#)

Figure 8.6: Recommendation Email Sample with Anonymous Data

8 Evaluation

Reminders were sent every 7 days to participants who had not submitted responses, urging them to do so and providing the option to share reasons for any delays. After the 21-day data collection period ended, the responses were collected and organised into an Excel file for further analysis.

Study Participants

A total of 68 active notification system users were initially contacted. Of the initial contacts, 15 users agreed to participate in the survey. Ultimately, 13 of these participants provided their responses. As seen in Table 8.11, among the participants, there were 10 males and 3 females. The age distribution was as follows: 5 participants were aged 20-25, 3 were aged 35-40, 1 was aged 50-55, 2 were aged 25-30, and 1 was aged 30-35.

Age Group	Male	Female
20-25	3	2
25-30	1	1
30-35	1	0
35-40	3	0
50-55	1	0

Table 8.11: Participant Demographics of the Cluster Based Implicit Recommendation Evaluation

Results and Discussion

The presented table (Table 8.12) provides a detailed breakdown of the results obtained from the user evaluation. In this evaluation, users were categorised into two distinct groups: active and inactive, according to their interaction history with the notification system. Active users are those who have consistently engaged with the system over a period of time, while inactive users represent newcomers who have had limited or no interaction with the notification system. This classification of user activity status was determined by analysing data extracted from the internal statistics of the notification system's database.

The analysis of user responses yields several noteworthy insights. First, it is essential to consider the precision of channel recommendations. On average, the precision for relevant channels is 50%, which is equivalent to the average precision for irrelevant channels. This balance suggests that, on average, users found an equal number of recommended channels to be both relevant and irrelevant.

An interesting observation is that there are no significant differences in the accuracy of the relevance of the channel when comparing active and inactive users. This finding suggests that the performance of the recommendation system, in terms of channel

8.3 User Feedback-based Studies

User	Active	Relevant Channels	Irrelevant Channels	Precision
1		3	2	0.60
2		1	4	0.20
3		3	2	0.60
4	Yes	1	4	0.33
5		1	4	0.20
6		4	1	0.80
7		1	4	0.20
8		2	3	0.40
9	Yes	3	2	0.60
10	Yes	5	0	1.00
11	Yes	1	4	0.20
12		4	1	0.80
13		3	2	0.60

Table 8.12: User Response Evaluation

precision, remains consistent regardless of whether users have a history of sustained interaction with the system or are newcomers. This lack of differentiation in precision between the two user groups prompts further investigation of the underlying mechanisms of the recommendation algorithms and the factors influencing channel relevance.

However, several crucial considerations must be taken into account in the interpretation of these findings. As seen in Table 8.13, these results may be influenced by various contextual factors that emerge from both the background research and similar studies conducted in the field.

Limitations	Explanatory Notes
Limited Participation and Interaction	The number of participants and interactions within the notification system is low, since the system is new and is used mostly by users in the IT Department.
Departmental Restriction	Given that the evaluation mainly focused on the IT department, it is plausible that many users were already subscribed to the relevant notification channels.
Channel Diversity	The current notification system exhibits limited channel diversity. Expanding this diversity by incorporating other departments and diversifying the range of notification channels is crucial to achieving a more comprehensive and representative evaluation.

Table 8.13: Study Limitations

8 Evaluation

Notably, we can anticipate that these results are subject to influence from factors such as the organization's specific characteristics, the evolving dynamics of user engagement within the IT department, and the inherent limitations of the notification system's current user base. Therefore, while these findings undoubtedly offer valuable insight into the system's performance, they should be regarded as a preliminary assessment, subject to refinement and further investigation. This becomes particularly important as the notification system expands its user reach and diversifies its channels.

These findings should be perceived as an initial assessment of the prototype's performance. For a more comprehensive evaluation, it is essential to wait for the full-scale adoption of the notification system by multiple departments, leading to increased user engagement and interaction. The insights gained from this study serve as a valuable foundation for future assessments, offering valuable initial impressions of the prototype's functionality and usability.

8.3.2 User Feedback-based Study of Implicit Recommendation Algorithms

The following section builds on previous research by the authors and expands on its findings. While it draws heavily from the author's earlier work, it also incorporates new insights and analysis. Additionally, the following sections are based on, supported by and taken from the work published in the following publication (Jakovljevic et al., 2023; Pobaschnig et al., 2023).

Research Question 8.4: *How do users respond to recommendations delivered through the notification system, as opposed to email, in the context of a large organisational setting when employing an implicit data recommendation system?*

Study Design

Our research aimed to evaluate the performance of a recommendation system based on implicit data within the context of large organisations. Specifically, we aimed to answer the research question: "How well does a implicit data recommendation system perform in large organisations?" This study adopted an observational approach, focussing on the real-life usage of the recommendation system within the organisation.

Settings and Instruments

The study focused on the evaluation of the performance of an implicit data recommendation system in the context of a large organisation. It was conducted exclusively within the notification system, with particular attention given to a dedicated channel known

as "CERN Notifications - Recommendations." This channel was used to invite users to subscribe and receive personalised recommendations. An observational methodology was followed, allowing valuable insights to be gained into how the recommendation system operated in a real world setting. This observational approach enabled understanding of user interactions and system functionality without disrupting the natural user experience. The primary instrument for data collection throughout the study was the notification system. This platform served as both a means of delivering recommendations to users and collecting their feedback. Data collection procedures were initiated by generating personalised recommendations through the notification system. Subsequently, these recommendations were communicated to users through the Web portal of the notification system. Importantly, users were actively encouraged to engage with these recommendations and provide feedback, ensuring a continuous flow of user-generated data. Structured user participation was maintained through regular notifications to users. Specifically, reminders were sent to users who had subscribed to the "CERN Notifications - Recommendations" channel or had chosen to ignore recommendations, motivating them to participate actively and share their feedback regarding the recommendations they received. Data privacy and security were important in the study, with strict adherence to data protection policies. Only user interactions within the recommendation part of the notification system were recorded. All relevant data, including user interactions and recommendation feedback, were recorded within the notification system. The study was carried out over a two-month period. During this time, users were consistently engaged through the notification system at regular intervals.

Procedure

Figure 8.7 provides an illustrative summary of the steps of the study procedure.

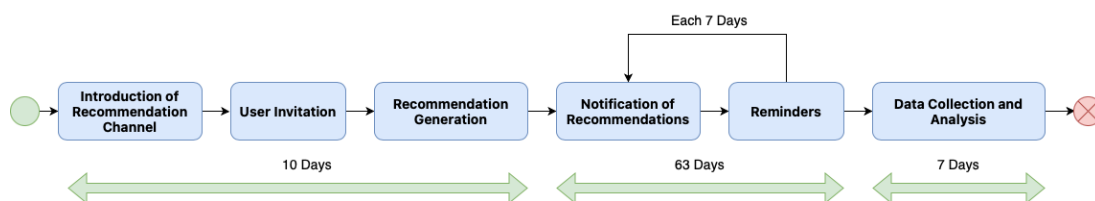


Figure 8.7: Procedure Steps of the Evaluation of Implicit Data Based Recommendations

The organisation of the study started with the introduction of a dedicated channel within the organisation's notification system. This channel, named "CERN Notifications - Recommendations", served as the central platform for disseminating recommendations to users.

8 Evaluation

Users were actively encouraged to participate by subscribing to the newly created “CERN Notifications - Recommendations” channel. These invitations were extended through the Mattermost channel within the notification system service, to ensure that users were made aware of the opportunity to receive personalised recommendations.

Personalised recommendations were generated for the 31 users who opted to join the “CERN Notifications - Recommendations” channel. Figure 8.8 displays the notification sent to participants to alert them that they received channel recommendations. These recommendations were created using the best-performing recommendation algorithm for implicit data recommendation.

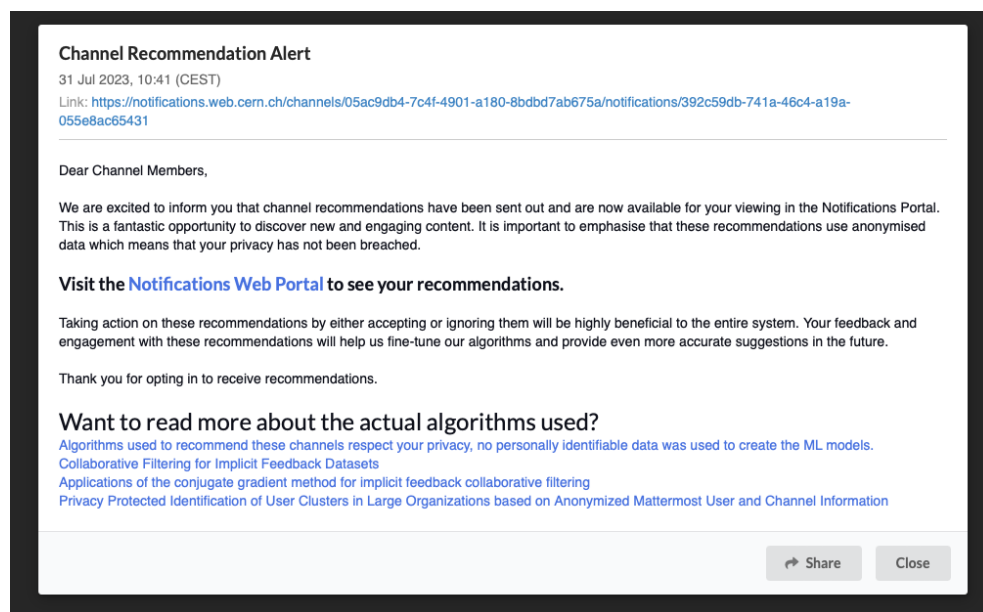


Figure 8.8: Recommendation Explanation Notification

To ensure that users were informed of the availability of new recommendations, notifications were sent through the notification system’s web portal. This encouraged users’ access and interaction with the recommendations.

Users who had subscribed to the recommendation channel or had not yet provided feedback received reminders every 7 days. These reminders served to encourage their participation in the study.

The last step of the study was data collection and analysis, where the data was exported from the notification system into Excel files. Personal and sensitive information about users was removed.

Study Participants

The participant selection process, detailed in Table 8.14, involved reaching out to 87 users through a Mattermost channel, inviting them to participate in the study. It is noteworthy that a total of 31 users subscribed to the recommendations. This level of user engagement is encouraging, considering that participation in research studies can often be challenging to elicit. The fact that the number of users who subscribed and provided feedback matches indicates a high level of interest and follow-up among the participants. However, the limited scope of the study within the IT department may have influenced these numbers. As the notification system expands its reach to other departments, future studies should explore participation dynamics in a broader organisational context.

Total Users Contacted	87
Users who Subscribed to Recommendations	31
Users who Provided Feedback	8

Table 8.14: Participant Selection

The demographics of the participants in our study, as shown in Table 8.15, indicate an interesting gender distribution. Among the 31 users who actively engaged with the recommendation system, 21 were male, while 10 were female. This distribution reflects certain gender disparities that are common in the tech industry, where males often outnumber females. However, it is important to note that our study's limited sample size may not be fully representative of the organisation's overall demographics.

Gender	Number of Users
Male	21
Female	10

Table 8.15: Demographics of Participants

Results and Discussion

In this section, we provide an analysis of the survey results presented in Table 8.3.2, which explored the use of sensitive information within a privacy-preserving personalised recommendation and information retrieval system. It is important to note that our survey had a limited response rate of 25.80%, with only 8 out of 31 invited participants providing information. Our analysis of user engagement within the recommendation system showed that when presented with recommendations, participants would complete all tasks of rating them. Among the participants, User 7 emerged as

8 Evaluation

the most engaged, actively subscribing to 5 recommendations. In contrast, Users 1, 2, and 6 displayed mixed engagement patterns, combining subscriptions with ignored occasional recommendations.

User	Ignored Count	Subscribed Count	Average Time to Ignore	Average Time To Subscribe	Total Recommendations	Precision
1	3	1	22 days	22 days	4	0.25
2	1	3	1 day	1 day	4	0.75
3	3	0	1 day	No Data	3	0.00
4	2	1	3 days	3 days	3	0.33
5	2	1	10 days	10 days	3	0.33
6	3	1	1 day	1 day	4	0.25
7	0	5	No Data	3 days	5	1.00
8	0	1	No Data	2 days	1	1.00

Table 8.16: User Response Evaluation

Precision, a crucial metric for evaluating recommendation accuracy, was employed to assess user satisfaction with the system's output. The computed average precision across user responses yielded a value of 0.48. This value implies that the system's recommendations exhibit a moderate level of accuracy, falling within a range that signifies neither exceptionally high nor exceptionally poor recommendation quality. In particular, the calculated average precision surpasses the expected average precision values derived from the statistical evaluation, hinting at potential strengths in the recommendation engine. Users 3 and 4 recorded a precision score of 0.00, indicating a complete lack of satisfaction with the recommendations they received. This stark contrast is seen in Users 7 and 8, who attained a perfect precision score of 1.00, indicative of their profound satisfaction and the high relevance of the recommendations to their preferences. Meanwhile, users 1, 2, 4 and 6 displayed precision scores in the range of 0.25 to 0.75. This variability highlights the importance of fine-tuning recommendations to align with the unique preferences and expectations of individual users.

Comparing the performance of our privacy-preserving system with traditional recommendation systems is a challenge given the limited sample size and the absence of a control group. To draw robust conclusions regarding performance, a more extensive study encompassing a diverse user population and benchmarking against established standards would be necessary.

8.4 Evaluation and Study Findings

Research Question 8.1: *How do traditional recommendation algorithms perform with implicit datasets, and what is the influence of enhanced metrics compared to simple ones on algorithm performance within the context of large organisations?*

The results indicate that custom user similarity measures and methods to reduce data sparsity had a positive influence on recommendation results, aligning with the goal of evaluating enhanced metrics. Furthermore, the evaluation showed that explicit data, such as explicit grouping, significantly improved recommendations compared to clustering approaches, which was another aspect of the goal. The evaluation also demonstrated that it is possible to create a highly performing recommender system without using sensitive user data and invading user privacy, contributing to the evaluation of algorithm performance in the context of large organisations. Furthermore, the comparison of techniques such as BPR and LMF with ALS and KNN aligns with the goal of assessing the performance of traditional recommendation algorithms with implicit data sets. However, it is important to note that the evaluation acknowledged data sparsity as a challenge that prevented the models from reaching certain AUC values presented in previous research. This suggests that while progress has been made, there may still be room for improvement to achieve the highest levels of recommendation performance within the context of large organisations.

Research Question 8.1: *How do traditional recommendation algorithms perform with implicit datasets, and what is the influence of enhanced metrics compared to simple ones on algorithm performance within the context of large organisations?*

The evaluation effectively evaluated multiple community detection algorithms across different thresholds, providing insight into their performance in terms of modularity, number of detected communities, and similarities between discovered communities and Mattermost teams. The results indicate that the modularity values were consistently high, and several algorithms, such as Infomap, Label Propagation, and Walktrap, showed promising performance in terms of identifying communities. The study also highlighted that the communities identified by these algorithms tended to align more with user interests than the organisational structure within CERN, which is an important aspect of evaluating the impact of community and group information on recommendations. The discussion also provided detailed information on the average community size, variations in similarities between Mattermost teams and identified communities, and the run-time performance of the algorithms. Overall, the study achieved its goal of assessing the performance of cluster-based recommendation algorithms within large organisations and investigating the impact of community

8 Evaluation

and group information on improving recommendations. The results and discussion provide valuable insights into the strengths and limitations of these algorithms in this context. However, it is important to note that while the goal was achieved in terms of evaluating algorithm performance, the study also raised questions about the alignment of identified communities with organisational units. This suggests potential areas for further research and refinement of the recommendation system to better align with organisational structures.

Research Question 8.3: *How does the performance and user response vary in the context of the CERN notification system when employing a cluster-based implicit recommendation system, particularly focussing on user reactions to email-based recommendations, within an organisational setting?*

The study successfully conducted a user evaluation, categorising users into active and inactive groups, and assessing the precision of channel recommendations. The results revealed that, on average, users found an equal number of recommended channels to be relevant and irrelevant, suggesting a balance in the precision of channel recommendations. Additionally, there were no significant differences in channel relevance precision between active and inactive users, indicating that the recommendation system's performance remained consistent regardless of user activity status. However, the discussion also highlighted several important limitations and contextual factors that may have influenced the results, including limited user participation and interaction, departmental restrictions, and channel diversity. These limitations suggest that the findings should be considered preliminary and subject to refinement as the notification system expands its user base and channel diversity.

Research Question 8.4: *How do users respond to recommendations delivered through the notification system, as opposed to email, in the context of a large organisational setting when employing an implicit data recommendation system?*

The study successfully analysed user engagement with the recommendation system and presented findings on user behaviours, such as ignoring and subscribing to recommendations. It also calculated the precision metric to assess recommendation accuracy, with an average precision value of 0.48, indicating a moderate level of accuracy in the system's recommendations. The study provided insight into individual user responses, with some users showing high satisfaction (precision score of 1.00) and others displaying lower satisfaction (precision scores ranging from 0.00 to 0.75). However, the study acknowledged several limitations, including a limited survey response rate and the absence of a control group or a comparison with traditional recommendation systems. These limitations make it challenging to draw robust conclusions regarding the system's performance compared to traditional systems.

8.5 Summary

In the initial phase of this research, offline evaluations were conducted to assess the performance of various recommendation algorithms using anonymised data from Mattermost and the CERN notification system. Recommendation algorithms were applied to implicit data in this phase. Subsequently, the best performing algorithms were selected for further evaluation across these datasets. Two studies were conducted for online evaluations: the first study focused on graph-based recommendation algorithms customised for the CERN notification system, while the second study examined traditional recommendation algorithms optimised for implicit data. Insights into the performance and suitability of these algorithms in real-world scenarios were aimed at being provided by these studies.

The results of the first data-based evaluation (Table 8.5) highlighted the effectiveness of custom user similarity measures and methods to reduce data sparsity in improving recommendation results. Additionally, the usage of explicit data, such as explicit grouping, significantly improved recommendations compared to clustering approaches. The findings suggest that a highly effective recommender system can be created without compromising user privacy, and techniques like BPR or LMF outperformed ALS and KNN, aligning with previous research.

However, data sparsity remained a challenge, preventing models from reaching high evaluation values. It was observed that reducing data sparsity by grouping or clustering users yielded better results, consistent with the existing literature. Representing user preferences as implicit data remains challenging due to the quality and understanding of the data.

In the evaluation of cluster-based implicit recommendation algorithms, various community detection algorithms were analysed at different thresholds. It was observed that the choice of algorithm and threshold significantly influenced the number of detected communities, their sizes, and the modularity of the resulting clusters. Algorithms like Infomap, Label Propagation, and Walktrap performed well in terms of modularity and similarity to Mattermost teams.

In the user feedback-based study of cluster-based implicit recommendation algorithms, user engagement and interaction within the recommendation system were investigated. Users were categorised into active and inactive groups based on their history of interaction. Interestingly, no significant differences in channel relevance precision were found between these groups, indicating consistent system performance regardless of user activity status. However, the study had a limited response rate and the results were influenced by specific organisational characteristics and limitations. Therefore, these findings should be considered preliminary and subject to refinement as the system expands and diversifies its user base.

In the user feedback-based study of implicit recommendation algorithms, a survey was conducted to collect information on the use of sensitive information within a

8 Evaluation

privacy-preserving personalised recommendation system. The survey had a limited response rate and users exhibited varying levels of involvement with recommendations. The calculated average precision was 0.48, suggesting moderate recommendation accuracy. Users displayed various levels of satisfaction, with some achieving perfect precision scores while others were entirely dissatisfied.

Comparing the privacy-preserving system to traditional ones is challenging due to the small sample size and lack of a control group. To draw robust conclusions, a more extensive study with a diverse user population and benchmarking against established standards would be necessary. These results are subject to several challenges, including limited participation and interaction, departmental restrictions, and channel diversity. A more comprehensive evaluation will be possible as the notification system expands its user base and channel diversity.

9 Lessons Learned and Outcome

The first part of this chapter focusses on describing the lessons learned during this thesis and providing a retrospective view of each phase. The second part of this chapter focusses on providing detailed answers to the research questions stated in the beginning of this thesis, together with an overall outline of the contributions made.

9.1 Retrospection

The research in this thesis followed the Design Science-inspired approach complemented by waterfall methodology as the primary research methodology. This approach structured the research into distinct phases, each phase serving as a foundation for subsequent stages and contributing to specific sections of the thesis.

Conducting the literature survey proved to be a challenging work. The domain of privacy-aware recommendation systems contains a wide scope of topics, encompassing information generation, user profiling, profiling methodologies, and a variety of machine learning algorithms for recommendations. Additionally, the related work chapter offered an overview of existing solutions in multiple domains, including open data, recommendations for large organisations, machine learning, and privacy. This knowledge from various research areas formed the foundation for the creation of a privacy-preserving recommendation system.

The "Analysis of User Behavior" phase concentrated on studying user behaviour within the domain of information consumption. The user survey conducted used the insights obtained from the previous chapter to formulate research questions and steer the aim of the survey. It aimed to determine the critical components of a privacy-aware system, the nature and extent of information users consume, and the extent to which users are willing to compromise their privacy for tailored experiences.

The "Applicability of Social Media Elements in Notification Systems" phase explored various modalities of presenting information to users via notifications. It was facilitated by a user study that involved a custom tool capable of presenting notifications in various scenarios and about different user tasks. The findings of this chapter were highly important in investigating modifications to traditional notification-based recommendation systems, making notifications more useful and appealing to users.

Analysing a notification system still in active development and lacking comprehensive documentation posed unique challenges that were faced in the "Requirements

and System Design” phase. The task involved evaluating the system and proposing valid extensions while combining the knowledge from previous chapters. Balancing these considerations was essential to designing a recommendation system that could be efficiently implemented within the project’s timeframe, providing valuable support to the existing system.

Implementing the recommendation system was comparatively straightforward. Well-documented components and support from CERN and third-party providers streamlined the development process outlined in the “Development” phase.

The “Evaluation” phase encountered several challenges, mainly due to limited user participation stemming from the fact that the notification system was still in the developing stage. Despite these issues, it yielded two distinct types of evaluations. The first involved statistical assessments of various data sources (Mattermost and CERN Notifications) with similar structures, while the second contained the evaluation of multiple recommendation algorithms, including clustering algorithms and traditional methods. Additionally, two user studies assessed different recommendation algorithms, aspects of recommendation distribution, and user interaction with recommendations.

In summary, this retrospective overview underscores the structured progression of research phases within this thesis, culminating in the development and evaluation of a privacy-aware recommendation system tailored to the unique needs of large organisations, such as CERN. Despite the encountered challenges and complexities, the research outcomes contribute to the advancement of privacy-conscious information systems and offer valuable insights into user behavior and preferences.

9.2 Outcome

Within this section, the findings related to the research questions of the thesis will be discussed and the research questions will be answered.

Research Question 1

How do users behave and consume information in large, highly connected organisations?

The research performed in Chapter 3 investigates the patterns of user behavior and information consumption within large, highly connected organisations, with a particular focus on the context of CERN. This investigation included three distinct user studies: the CERN Newcomers Analysis, the CERN IT Department User Survey Analysis, and the CERN User Information Consumption Analysis. The overarching research question addressed by this chapter is, “How do users behave and consume information in large, highly connected organisations?” Here are the key findings and conclusions:

The CERN IT Department User Survey revealed that CERN’s computer community

utilises a diverse range of devices with varying operating systems. Notably, Microsoft Windows predominates as the desktop operating system, while laptops, predominantly running MacOS, are more commonly used. Importantly, over 20% of participants reported using non-CERN-supported operating systems. Email communication remains the preferred method, with 85% of users relying on their CERN email accounts. However, over 20% forward their CERN emails to other platforms. The survey demonstrated that CERN users exhibit strong preferences for specific hardware and software, making it challenging to propose one-size-fits-all solutions. These insights are invaluable for IT-CDA members in improving their services.

The newcomer behavior analysis revealed that within large organisations like CERN, users favor email and face-to-face meetings as their primary communication and knowledge-sharing methods. While these approaches offer confidentiality and protect sensitive information, they often hinder easy access to information. Email, especially for work-related purposes, is the preferred method of information discovery within CERN. However, the study highlighted that information retrieval via email can be cumbersome, involving multiple steps and waiting times for responses. In contrast, face-to-face meetings, while offering benefits like efficient information access through meeting minutes, can be challenging to organise and may lead to information loss.

The CERN User Information Consumption Analysis examined information consumption patterns, encompassing both personal and work-related aspects. Notably, it revealed that both male and female users exhibit similar preferred information consumption intervals. Mobile phones emerged as the preferred device for personal information consumption, followed by laptops and desktop PCs. For work-related information, desktop PCs were the most frequently used devices. Email-checking habits varied, with users commonly checking personal email "1-3 Times A Day" on both weekdays and weekends. However, work-related email was checked "More than 12 Times a Day" on weekdays, emphasizing its importance during the workweek. The content of information consumption differed between work and personal contexts. Work-related information centered on acquiring professional knowledge and skills, necessitating access to research, industry trends, and best practices. In contrast, personal information consumption encompassed a wider range of interests, including news, entertainment, hobbies, and personal development.

Research Question 2

How can sensitive organisational data be used for reproducible research and development?

Overall studies indicated that the wider the search for knowledge, the higher the organisation's innovation. This has led to conclusions that the development of information openness can stimulate innovative activities, the creation of innovative approaches, and greater performance (Cruz-González et al., 2015; Lopez-Vega et al., 2016a). Previous studies have been inclined to demonstrate the benefits of openness in organisations.

Recently, studies have begun to stress the downsides of openness, which would justify why not many organisations have adopted the concept of openness (Huizingh, 2011b). One of the main disadvantages of openness in an organisation is the risk of losing competitive advantage and leaking private information (Isfandyari-Moghaddam, 2015). A study of information technologies for the sharing of knowledge in large healthcare organisations has shown that large, well-funded organisations struggle to develop and maintain such complex systems (N. Ali et al., 2012).

Sensitive organisational data stands as a cornerstone in today's data-driven research and development endeavours. Its usage, while ensuring privacy and reproducibility, is a challenge expertly addressed by the DataLift framework.

The DataLift framework emphasises the transformative power of anonymisation to transform sensitive organisational data into open data. Anonymisation is not just about concealing identities but also ensuring that the data retain its inherent value for research purposes. Striking a balance between privacy and data utility is important. While over-anonymisation may hinder research by omitting critical data, under-anonymisation can compromise sensitive information. The aim is to find a middle ground where transformed data remain a valuable resource for research and development and protect privacy concurrently.

Before even starting the process of data-driven research and development, it is crucial to define the objectives of the research clearly. A clear understanding of the purpose and scope of the data's usage ensures secure data processing while meeting explicit research or development goals. Data classification acts as the next layer of refinement, providing a framework for determining variable sensitivity levels between different data subsets. This categorisation leads to the subsequent stages of risk assessment and data transformation. A detailed risk assessment reveals the potential pitfalls and vulnerabilities associated with data publication, ranging from privacy concerns to strategic or economic challenges. Based on these insights, the data undergo transformation and anonymisation, while preserving sensitive elements.

The DataLift framework's benefits come from its defined structure, which offers accessibility to a broad spectrum of users, ranging from novices to experts. It ensures uniformity in data preparation, providing a basis for reproducibility. The framework, while comprehensive, could be resource intensive, posing challenges for smaller organisations with limited resources. Moreover, there is a grey area in defining 'sensitive' data, leading to potential inconsistencies in data preparation, which could, in turn, affect reproducibility.

Embracing a standardised framework such as DataLift can pave the way for reproducibility in research and development. Consistency in data preparation and processing becomes a norm rather than an exception. When processed appropriately, sensitive organisational data provide various insights, driving innovation without infringing on ethical and legal restrictions. The knowledge produced from structured approaches lay

the groundwork for privacy-preserving technologies, especially in domains such as personalised recommendation systems and advanced information retrieval.

Research Question 3

How important are privacy and privacy preserving concepts for employees in large organisations?

In the context of large organisations, the importance of privacy and the preservation of private concepts among employees cannot be understated. Research presented in chapters 3, 4, and 5 provides substantial information, revealing details of employee behaviour, preferences concerning their personal and work-related information, and the importance of privacy.

A significant indication of privacy preference is evident in the tendency of employees to use their personal email and other private communication channels to share information. Such a preference is not merely a reflexion of a desire for enhanced privacy, but it could also be rooted in a desire for greater control over communication's content and its frequency. Many employees' decisions to use personal email accounts for various purposes or even to forward their official emails to other systems can be seen as an assertion of personal autonomy and control in their communication. It is an indication of a subtle relationship between professional obligations and personal boundaries.

Securing privacy is not without challenges, especially in relation to efficient information sharing. Large organisations, such as CERN, provide an ideal case study of this dynamic. The tendency towards private methods of communication, including face-to-face meetings and email, might offer employees protection of confidentiality. However, they also inadvertently impose limitations on the broader dissemination and accessibility of information, potentially hindering collaborative efforts.

The combination of personal and professional spheres is further evident in the prevalent use of personal devices such as mobile phones and laptops for work-related tasks, with many employees handling sensitive and potentially dangerous information.

However, privacy extends beyond communication channels and personal devices. The chapters mentioned offered insights into the attitudes of users towards personalised recommendation systems. Although these systems offer customisation and personalization, they often come at the potential cost of privacy, demanding access to personal data. The cautious, if not apprehensive, attitude of many users towards such systems is indicative of a broader trend of valuing online privacy and a reluctance to share personal information, even if it is in exchange for the mentioned benefits.

Drawing from a broader perspective, it is common knowledge that in large organisations, the problems surrounding data privacy have a larger impact than on individuals. Employees are often guardians of proprietary, sensitive, or even classified information. Any unauthorised access or breach can not only jeopardise organisational

integrity, but can also have severe consequences for the individuals involved. This heightened responsibility is further emphasised by regulatory frameworks like the General Data Protection Regulation (GDPR) and the organisation's internal guidelines on data protection and privacy.

Furthermore, the challenges reported by participants in rediscovering personal-related information point toward another dimension of privacy—personal information management. The difficulty in retrieving such information is not just an indication of challenges in managing data effectively. It subtly underscores the lengths to which employees go to ensure their information's privacy, even if it occasionally poses challenges for them.

In conclusion, privacy in the context of large organisations presents a multifaceted landscape. While modern communication and information systems offer convenience and capabilities, they also bring forth complex challenges surrounding data privacy. Employees, in their search for personal autonomy and discretion, balance the conveniences of digital platforms with the imperatives of privacy. As digital platforms continue to evolve, understanding and addressing these dynamics will be crucial for both organisations and employees.

Research Question 4

How can sensitive information be used in a privacy-preserving way for the creation of a personalised recommendation and information retrieval system and what is the performance of such a system compared to traditional systems?

Sensitive information, especially in the era of data-driven decision making, has tremendous potential to enhance the abilities of personalised recommendation systems. However, escalating data privacy concerns and increasing regulatory pressure aim to ensure that the ethical use of such data becomes mandatory to advance and maintain trust in digital ecosystems.

Sensitive information can be used in a privacy-preserving manner for the development of a personalised recommendation and information retrieval system through the implementation of various strategies and technologies, which are covered in Chapters 2, 3, 4 and 5. Anonymisation and pseudonymization of personal data are effective measures. These techniques render it challenging, if not impossible, to associate specific information with individual users, thus protecting user identities and sensitive data. Differential privacy techniques can be integrated to introduce carefully calibrated noise into data, preventing the inference of sensitive user information while offering personalised recommendations. Federated learning is a decentralised approach that allows model training on distributed devices or servers. This ensures that sensitive data remain decentralised and never centralised, as models are collaboratively updated without exposing raw data. Homomorphic encryption presents another viable strategy; by encrypting data in a manner that enables computations on the encrypted data

itself, without requiring decryption, sensitive information remains secure throughout processing. This ensures that individual data contributions are concealed during recommendation generation.

User consent and transparency have been shown to be important, as outlined in Chapters 4 and 5. Systems must prioritise user consent, ensuring that users are informed about how the data is used. Users should retain control over their data, including the ability to opt-in or opt out of personalised recommendations. Data minimisation should be practised diligently. This involves collecting and retaining only the minimum amount of data necessary to generate recommendations while avoiding the collection of sensitive information whenever possible.

Chapters 6 and 7 illustrate that secure user authentication is crucial. Robust authentication mechanisms should be implemented to protect user accounts and ensure that recommendations are provided to the rightful user without exposing sensitive data. The use of secure, authenticated, and authorised applications to access and share data is vital.

By adopting these privacy-preserving techniques and technologies, it is possible to use sensitive information to create a personalised recommendation and information retrieval system. These approaches achieve a balance between personalization and privacy.

The evaluation in Chapter 8 included multiple aspects of recommendation systems in the context of large organisations, with a particular focus on implicit data. Enhanced metrics and data sparsity reduction techniques positively impacted recommendation results. The conversion of implicit data to explicit data, that is, explicit grouping, demonstrated substantial improvements in recommendations when combined with clustering approaches. Furthermore, it was established that the construction of a high-performing recommendation system was feasible without the need for sensitive user data or privacy intrusion. Comparison of various techniques, including BPR, LMF, ALS, and KNN, was the central element for the evaluation of traditional recommendation algorithms with implicit data sets. This analysis implied that personalised systems, often employing enhanced metrics and reducing data sparsity, were competitive with their traditional counterparts.

However, the persistent challenge of data sparsity remained, which restricted models from attaining the highest levels of recommendation accuracy, a limitation shared by both personalised and traditional systems.

Another evaluation involved cluster-based recommendation algorithms. These assessments revealed that specific community detection algorithms exhibited high effectiveness in identifying relevant communities. This insight indicated that the incorporation of community and group information could extend personalised recommendations.

User engagement studies offered valuable information. No significant differences in channel relevance precision were evident between active and inactive users. This

9 Lessons Learned and Outcome

implied that personalised recommendation systems could maintain consistent performance regardless of the status of user activity.

However, the precision of the recommendation, measured by an average precision value of 0.48 in one study, resulted in the attribution of cluster-based recommendation systems as "moderate". Users showed varying levels of satisfaction, suggesting prospects for enhancement.

It is essential to acknowledge the array of challenges and limitations, which include factors such as limited response rates, departmental constraints, and channel diversity. Therefore, they underscored the need for more research and system refinements to comprehensively outline the potential of personalised recommendation and information retrieval systems.

10 Conclusion and Future Work

In this chapter, ideas for future research projects and detected limitations are explained along with the summarization of the research questions and the goals of the thesis.

10.1 Conclusion

This doctoral research was designed to address a significant challenge faced by large organisations, which is the use of sensitive information for advanced data-based tools, such as recommendations. This research focused on CERN, namely enhancing the CERN notification system in order to make it proactive and provide a highly flexible information retrieval and information navigation through a recommendation system. Through the development of a new recommendation system integrated within the CERN notification system, this thesis aimed to investigate how privacy-aware information retrieval, user navigation, and information visualisation methodologies could improve and complement the existing system.

To achieve this objective, a series of research questions were formulated:

1. How do users behave and consume information in large, highly connected organisations?
2. How can sensitive organisational data be used for reproducible research and development?
3. How important are privacy and privacy-preserving concepts for employees in large organisations?
4. How can sensitive information be used in a privacy-preserving way for the creation of a personalised recommendation and information retrieval system, and what is the performance of such a system compared to traditional systems?

A comprehensive investigation of these research questions resulted in several significant contributions.

Understanding User Behaviour Our research investigated the difficulties of user behaviour and information consumption in large organisations. By analysing user interactions within the CERN, we gained valuable insights into how users navigate, consume and seek information in large organisations.

Leveraging Sensitive Data for Research: We introduced the DataLift framework, which focusses on opening up organisational data for reproducible research while

addressing privacy and compliance concerns. This framework offers a viable solution for organisations looking to balance data accessibility with security and ethical considerations.

Privacy Awareness: The importance of privacy cannot be overstated. Our investigations highlighted the significance of privacy concerns among individuals in large organisations (employees, contractors, users, and others), emphasising the need for privacy-preserving solutions in various systems.

Privacy-Aware Recommendation System: The end result of this research is the development of a privacy-aware recommendation system. This system uses sensitive organisational data in a secure and privacy-conscious manner to generate personalised recommendations and improve information retrieval. Our evaluation demonstrated that this system offers a viable alternative to traditional recommendation systems while ensuring data privacy.

In conclusion, this research underscores the importance of addressing privacy concerns while improving information retrieval and recommendation systems in large organisations. The findings and solutions presented here provide valuable guidance for organisations like CERN and beyond, aiming to harness the power of data for enhanced productivity and user satisfaction, all while safeguarding individual privacy. As the digital landscape evolves, the lessons learnt from this research will remain relevant and influential in shaping the future of information systems in large, dynamic and privacy-conscious environments.

10.2 Limitations

In this section, the limitations of this thesis are discussed.

In Chapter 3, a comprehensive review of different user studies has been conducted with the main goal of understanding user information consumption behaviours. These studies aimed to determine user preferences regarding information consumption mediums, devices, temporal patterns, and attitudes toward privacy and data utilisation within recommendation systems. The findings of these surveys have revealed that, over time, there have been minimal changes in the devices and channels through which users seek information. These relatively static patterns can be largely attributed to organisational constraints and limitations that hinder rapid adaptation to emerging technologies. Conversely, it has become evident that user information needs change significantly. This observation highlights the challenges and limitations organisations face in aligning their information dissemination strategies with the evolving requirements of their user base. Furthermore, in the context of large organisations characterised by diverse user backgrounds and occasionally conflicting user group beliefs, accommodating heterogeneous information preferences presents a difficult endeavour.

The DataLift framework was introduced in Chapter 4 as a mechanism for the open-

sourcing of organisational data, specifically designed to facilitate its use in the context of reproducible research. The DataLift framework aggregates best practises from diverse open data frameworks, resulting in a generic approach to creating open data. However, this generically adaptable approach has certain inherent limitations. These limitations manifest themselves in the form of ambiguities in defining risk levels and articulating the data lifecycle. As a consequence, adhering to such a framework may present challenges, particularly for larger organisations. These challenges are primarily rooted in the imperative nature of compliance requirements that span organisational, governmental, ethical, and other regulatory dimensions.

The use of social media elements and additional information elements for the display of notifications has been covered in Chapter 5. Although the study was well designed, its capacity to continuously monitor notification usage was limited, restricting a comprehensive assessment of all social media elements within notifications. Additionally, the study was unable to discern the evolution of user preferences over time, which could introduce bias into the conclusion. Furthermore, the study's homogeneous user group may have skewed the insights obtained, primarily toward technical audiences.

The recommendation system developed for this thesis serves as a demonstration solution, showcasing the potential of a privacy-aware approach to recommendation system design. However, it is important to acknowledge that this system's integration was limited to the organisation's notification system, which has a relatively modest user base of approximately 700 users, all from the same department within the organisation. Consequently, this imposed certain limitations on the completeness of the evaluation of the system, preventing a more exhaustive assessment of its functionalities and potential impact. Due to these factors, it was also not possible to evaluate multiple recommendation algorithms or other machine learning technologies.

10.3 Future Work

Many of the topics discussed in the thesis can be used as a basis for further development and research projects.

In Chapter 5 not all social media element use cases were investigated due to time restrictions. Additional social media elements, other media and devices for notification display could be considered for review in future work. Future research may also examine how users respond over time to messages that provide additional information. This could allow us to evaluate other social media elements in notifications more effectively. It may allow for a better evaluation of the analysed social media elements. Tracking user reactions for longer periods to different combinations of social media elements in notifications could lead to a novel approach to their use within notification systems.

In order to promote the utilisation of open data for the purpose of facilitating

10 Conclusion and Future Work

reproducible research, it becomes crucial to enhance the DataLift framework. This enhancement should aim at mitigating the existing limitations mentioned Section 10.2 in within the framework and rendering it more comprehensible and user-friendly.

In the context of the recommendation system, as previously discussed in Section 10.2, the scope of this study was limited to the evaluation of a limited set of recommendation algorithms and conventional machine learning technologies. In particular, the examination was centred around well-established algorithms. However, prospective research endeavours have the potential for expanded exploration that includes contemporary approaches such as neural networks and large language models, thus enriching the repertoire of recommendation techniques.

Furthermore, to enhance the transparency and interpretability of the models developed, the implementation or integration of a monitoring console capable of generating comprehensive evaluation metrics is deemed imperative. Such an addition would provide valuable insights into the system's performance and contribute to the refinement of recommendation processes.

References

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In (p. 308–318). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2976749.2978318> doi: 10.1145/2976749.2978318
- Abel, F., Henze, N., Herder, E., & Krause, D. (2010, 09). Linkage, aggregation, alignment and enrichment of public user profiles with mypes.. doi: 10.1145/1839707.1839721
- Abrams, L. (n.d.). *533 million facebook users' phone numbers leaked on hacker forum*. Retrieved 2021-10-26, from <https://www.bleepingcomputer.com/news/security/533-million-facebook-users-phone-numbers-leaked-on-hacker-forum/>
- Adomavicius, G., & Tuzhilin, A. (2005, 07). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17, 734-749. doi: 10.1109/TKDE.2005.99
- Aftab, S., & Ramampiaro, H. (2022). Evaluating top-n recommendations using ranked error approach: An empirical analysis. *IEEE Access*, 10, 30832-30845. doi: 10.1109/ACCESS.2022.3159646
- Aggarwal, C. C., & Yu, P. S. (2008). A general survey of privacy-preserving data mining models and algorithms. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-preserving data mining: Models and algorithms* (pp. 11–52). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-70992-5_2 doi: 10.1007/978-0-387-70992-5_2
- Akman, I., & Mishra, A. (2010). Gender, age and income differences in internet usage among employees in organizations. *Computers in Human Behavior*, 26(3), 482-490. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0747563209001988> doi: <https://doi.org/10.1016/j.chb.2009.12.007>
- Alarcón-del Amo, M.-d.-C., Lorenzo-Romero, C., & Gómez-Borja, M.-Á. (2011). Classifying and profiling social networking site users: A latent segmentation approach. *Cyberpsychology, behavior, and social networking*, 14(9), 547–553.

References

- Ali, M., Shaikh, Z., Khan, M., & Tariq, T. (2015). User profiling through browser finger printing. In *International conference on recent advances in computer systems* (pp. 135–140).
- Ali, N., Whiddett, D., Tretiakov, A., & Hunter, I. (2012, 03). The use of information technologies for knowledge sharing by secondary healthcare organisations in new zealand. *International journal of medical informatics*, 81, 500-6. doi: 10.1016/j.ijmedinf.2012.02.011
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, 3–21.
- Amati, G. (2009). Information retrieval models. In *Encyclopedia of database systems* (pp. 1523–1528). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-39940-9_916 doi: 10.1007/978-0-387-39940-9_916
- Antony, S., & Salian, D. (2021, 10). Usability of open data datasets. In (p. 410-422). doi: 10.1007/978-3-030-89022-3_32
- Baeza-Yates, R. A., & Ribeiro-Neto, B. A. (1999). *Modern information retrieval*. ACM Press / Addison-Wesley. Retrieved from <http://www.ischool.berkeley.edu/~hearst/irbook/glossary.html>
- Bangor, A., Kortum, P., & Miller, J. (2009, May). Determining what individual sus scores mean: Adding an adjective rating scale. In (Vol. 4, p. 114–123). Bloomingdale, IL: Usability Professionals' Association.
- Barbaro, M., & Jr., T. Z. (2006). *A face is exposed for aol searcher no. 4417749*. Retrieved 2006-08-09, from <https://www.nytimes.com/2006/08/09/technology/09aol.html#:~:text=By%20Michael%20Barbaro%20and%20Tom%20Zeller%20Jr.&text=Buried%20in%20a%20list%20of,not%20much%20of%20a%20shield.>
- Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1, 2017.
- Baroudi, J. J., Olson, M. H., & Ives, B. (1986, March). An empirical study of the impact of user involvement on system usage and information satisfaction. *Commun. ACM*, 29(3), 232–238. Retrieved from <https://doi.org/10.1145/5666.5669> doi: 10.1145/5666.5669
- Bieniasz, J., & Szczypiorski, K. (2019). Methods for information hiding in open social networks. *JUCS - Journal of Universal Computer Science*, 25(2), 74-97. doi: 10.3217/jucs-025-02-0074

- Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Berlin, Heidelberg: Springer-Verlag.
- Bishop, C. M. (2007). *Pattern recognition and machine learning (information science and statistics)* (1st ed.). Springer.
- Bobic, A., Jakovljevic, I., Gütl, C., Le Goff, J., & Wagner, A. (2022, April 1). Implicit user network analysis of communication platform open data for channel recommendation. In *9th international conference on social networks analysis, management and security - snams 2022*. (9th International Conference on Social Networks Analysis, Management and Security : SNAMS 2022, SNAMS 2022 ; Conference date: 29-11-2022 Through 01-12-2022)
- Boland, M. R., Trembowelski, S., Bakken, S., & Weng, C. (2012). An initial log analysis of usage patterns on a research networking system. *Clinical and Translational Science*, 5(4), 340-347. Retrieved from <https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-8062.2012.00418.x> doi: <https://doi.org/10.1111/j.1752-8062.2012.00418.x>
- Bruns, A. (2021, 08). Echo chambers? filter bubbles? the misleading metaphors that obscure the real problem. In (p. 33-48). doi: 10.4324/9781003109891-4
- Brusilovsky, P., & Tasso, C. (2004, 06). Preface to special issue on user modeling for web information retrieval. *User Model. User-Adapt. Interact.*, 14, 147-157. doi: 10.1023/B:USER.0000029016.80122.dd
- Canales, K. (n.d.). *Hackers scraped data from 500 million linkedin users – about two-thirds of the platform’s userbase – and have posted it for sale online*. Retrieved 2021-10-26, from <https://www.businessinsider.com.au/linkedin-data-scraped-500-million-users-for-sale-online-2021-4>
- Capdevila, J., Arias, M., & Arratia, A. (2016). Geosrs: A hybrid social recommender system for geolocated data. *Information Systems*, 57, 111-128. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0306437915001842> doi: <https://doi.org/10.1016/j.is.2015.10.003>
- Carberry, S. (2001, 03). Techniques for plan recognition. *User Model. User-Adapt. Interact.*, 11, 31-48. doi: 10.1023/A:1011118925938
- Center, I. I. (2012). *Big data analytics intel’s it manager survey on how organizations are using big data*. Retrieved 2012, from <https://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf>

References

- Chai, T., & Draxler, R. (2014, 01). Root mean square error (rmse) or mean absolute error (mae)? *Geosci. Model Dev.*, 7. doi: 10.5194/gmdd-7-1525-2014
- Chatterjee, A. (2017). Chapter h - selective dissemination of information. In A. Chatterjee (Ed.), *Elements of information organization and dissemination* (p. 117-123). Chandos Publishing. Retrieved from <https://www.sciencedirect.com/science/article/pii/B9780081020258000089> doi: <https://doi.org/10.1016/B978-0-08-102025-8.00008-9>
- Chen, H., Chen, J., & Ding, J. (2021). Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability*, 70(2), 831-847. doi: 10.1109/TR.2021.3070863
- Correa, D., & Sureka, A. (2011). Mining tweets for tag recommendation on social media. In *Proceedings of the 3rd international workshop on search and mining user-generated contents* (p. 69-76). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2065023.2065040
- Costa, P., Cordeiro, A., & Oliveira Jr, E. (2021, 12). Comparing open data repositories. In (p. 60-69). doi: 10.5753/eres.2021.18451
- Cruz-González, J., López-Sáez, P., Navas-López, J. E., & Delgado-Verde, M. (2015). Open search strategies and firm performance: The different moderating role of technological environmental dynamism. *Technovation*, 35, 32-45. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0166497214001230> doi: <https://doi.org/10.1016/j.technovation.2014.09.001>
- Cruz-González, J., Sáez, P., Navas-López, J., & Verde, M. (2014, 01). Open search strategies and firm performance: The different moderating role of technological environmental dynamism. *Technovation*, 35. doi: 10.1016/j.technovation.2014.09.001
- Cunningham, P., Cord, M., & Delany, S. (2008, 01). Supervised learning. In (p. 21-49). doi: 10.1007/978-3-540-75171-7_2
- Dabbagh, M., Peck Lee, S., & Parizi, R. (2015, 07). Functional and non-functional requirements prioritization: empirical evaluation of ipa, ahp-based, and ham-based approaches. *Soft Computing*. doi: 10.1007/s00500-015-1760-z
- De Cristofaro, E. (2021). A critical overview of privacy in machine learning. *IEEE Security Privacy*, 19(4), 19-27. doi: 10.1109/MSEC.2021.3076443
- Delgado, J., Davidson, R., & Triplehop. (2002, 01). Knowledge bases and user profiling in travel and hospitality recommender systems. doi: 10.1007/978-3-7091-6132-6_1

- De Pierrefeu, A., Fovet, T., Hadj-Selem, F., Löfstedt, T., Ciuciu, P., Lefebvre, S., ... Duchesnay, E. (2018). Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity. *Human brain mapping*, 39(4), 1777–1788.
- Dickinson, I., Reynolds, D., Banks, D., Cayzer, S., & Vora, P. (2003, 03). User profiling with privacy: A framework for adaptive information agents. In (Vol. 2586, p. 123-151). doi: 10.1007/3-540-36561-3_6
- Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14, 241–258.
- Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international acm sigir conference on research and development in information retrieval* (p. 787–788). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1835449.1835616
- Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2019). A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access*, 7, 144907–144924.
- Ergüner Özkoç, E. (2021, 07). Privacy preserving data mining.. doi: 10.5772/intechopen.99224
- European Commission and Directorate-General for the Information Society and Media. (2002). *Commercial exploitation of europe's public sector information : executive summary*. Publications Office.
- Felden, C., & Linden, M. (2007). Ontology-based user profiling. In W. Abramowicz (Ed.), *Business information systems* (pp. 314–327). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Ferguson, S., & Hebels, R. (2003). Chapter 2 - information sources and services. In S. Ferguson & R. Hebels (Eds.), *Computers for librarians (third edition)* (Third Edition ed., p. 41-79). Chandos Publishing. Retrieved from <https://www.sciencedirect.com/science/article/pii/B9781876938604500082> doi: <https://doi.org/10.1016/B978-1-876938-60-4.50008-2>
- flexfibre. (2019). *How much data does your organisation generate? and why is it important to know?* Retrieved 2019-04-23, from <https://flexfibre.com/2019/04/23/how-much-data-does-your-organisation-generate-and-why-is-it-important-to-know/>
- Folorunso, O. O. (2015). Information-seeking behavior of social sciences scholars: a nigerian case study. *Brazilian Journal of Information Science: research trends*, 9(1). Retrieved from <https://revistas.marilia.unesp.br/index.php/bjis/article/view/5218> doi: 10.36311/1981-1640.2015.v9n1.07.p107

References

- Forbes. (2018). *How much data do we create every day? the mind-blowing stats everyone should read*. Retrieved 2018-04-23, from <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/>
- Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4), 1–53.
- Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In (Vol. 4321 LNCS). doi: 10.1007/978-3-540-72079-9{\-}2
- Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1), 129–149. Retrieved from <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bjso.12286> doi: <https://doi.org/10.1111/bjso.12286>
- Ghorab, M. R., Zhou, D., O'Connor, A., & Wade, V. (2013, 08). Personalised information retrieval: Survey and classification. *User Modeling and User-Adapted Interaction*, 23. doi: 10.1007/s11257-012-9124-1
- Ghosh, R., & Dekhil, M. (2009, 01). Discovering user profiles. In (p. 1233-1234). doi: 10.1145/1526709.1526944
- Gong, N. Z., & Liu, B. (n.d.). You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. Retrieved 2021-10-22, from <http://arxiv.org/abs/1606.05893>
- Grimmelmann, J. (2011, 01). Some skepticism about search neutrality. *The Next Digital Decade: Essays on the Future of the Internet*.
- Gunaratne, C., Baral, N., Rand, W., Garibay, I., Jayalath, C., & Senevirathna, C. (2020, 06). The effects of information overload on online conversation dynamics. *Computational and Mathematical Organization Theory*, 26, 1-22. doi: 10.1007/s10588-020-09314-9
- Guo, G., Qiu, H., Tan, Z., Liu, Y., Ma, J., & Wang, X. (2017). Resolving data sparsity by multi-type auxiliary implicit feedback for recommender systems. *Knowledge-Based Systems*, 138, 202–207.
- Gupte, M., & Eliassi-Rad, T. (2012). Measuring tie strength in implicit social networks. In *Proceedings of the 4th annual acm web science conference* (p. 109–118). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2380718.2380734> doi: 10.1145/2380718.2380734

- Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media.
- Hafeez, F., Nasirifard, P., & Jacobsen, H.-a. (2018, 12). A serverless approach to publish/subscribe systems. In (p. 9-10). doi: 10.1145/3284014.3284019
- Han, L., Chen, T., Demartini, G., Indulska, M., & Sadiq, S. (2020). On understanding data worker interaction behaviors. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (p. 269–278). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3397271.3401059> doi: 10.1145/3397271.3401059
- Haselton, M., Nettle, D., & Andrews, P. (2015, 09). The evolution of cognitive bias. In (p. 724-746). doi: 10.1002/9780470939376.ch25
- Helfat, C. (2006, 05). Open innovation: The new imperative for creating and profiting from technology. *Academy of Management Perspectives*, 20, 86-88. doi: 10.5465/AMP.2006.20591014
- Hensley, C. B. (1963). Selective dissemination of information (sdi): State of the art in may, 1963. In *Proceedings of the may 21-23, 1963, spring joint computer conference* (p. 257–262). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/1461551.1461584> doi: 10.1145/1461551.1461584
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004, jan). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1), 5–53. Retrieved from <https://doi.org/10.1145/963770.963772> doi: 10.1145/963770.963772
- Hernández-Álvarez, L., de Fuentes, J. M., González-Manzano, L., & Hernández Encinas, L. (2020). Privacy-preserving sensor-based continuous authentication and user profiling: a review. *Sensors*, 21(1), 92.
- Himeur, Y., Sohail, S. S., Bensaali, F., Amira, A., & Alazab, M. (2022). Latest trends of security and privacy in recommender systems: A comprehensive review and future perspectives. *Computers Security*, 118, 102746. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0167404822001419> doi: <https://doi.org/10.1016/j.cose.2022.102746>
- Horvath, I. (2007, 01). Comparison of three methodological approaches of design research..
- Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 eighth ieee international conference on data mining* (p. 263-272). doi: 10.1109/ICDM.2008.22

References

- Huizingh, E. (2011a, 01). Open innovation: State of the art and future perspectives. *Technovation*, 31, 2-9. doi: 10.1016/j.technovation.2010.10.002
- Huizingh, E. (2011b, 01). Open innovation: State of the art and future perspectives. *Technovation*, 31, 2-9. doi: 10.1016/j.technovation.2010.10.002
- Huston, P., Edge, V., & Bernier, E. (2019, 10). Reaping the benefits of open data in public health. *Canada Communicable Disease Report*, 45, 252-256. doi: 10.14745/ccdr.v45i10a01
- Igor, J., Martin, P., Christian, G., & Andreas, W. (2023, 7). Privacy-preserving user clustering: The application of anonymized data to community detection in large organizations. *IARIA JOURNALS*.
- Ingwersen, P. (1992). *Information retrieval interaction*. GBR: Taylor Graham Publishing.
- Institutional knowledge: What it is & how to use it*. (2022, May). Retrieved from <https://tettra.com/article/institutional-knowledge-what-it-is-how-to-use-it/>
- Isfandyari-Moghaddam, A. (2015, 04). Managing information in organizations: A practical guide to implementing management strategy, s.a. cox palgrave macmillan, hampshire, uk (2014), 436 pp., price: £44.99, paperback, isbn: 978-0-23029-884-2. *International Journal of Information Management*, 35. doi: 10.1016/j.ijinfomgt.2014.11.002
- Ives, B., Olson, M. H., & Baroudi, J. J. (1983, October). The measurement of user information satisfaction. *Commun. ACM*, 26(10), 785-793. Retrieved from <https://doi.org/10.1145/358413.358430> doi: 10.1145/358413.358430
- Jacobsen, H.-A. (2009). Publish/subscribe. In *Encyclopedia of database systems* (pp. 2208-2211). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-39940-9_1181 doi: 10.1007/978-0-387-39940-9_1181
- Jain, P., Gyanchandani, M., & Khare, N. (2018, 04). Differential privacy: its technological prescriptive using big data. *Journal of Big Data*, 5. doi: 10.1186/s40537-018-0124-9
- Jakovljevic, I. (n.d.). Codis survey tool. doi: 10.5281/zenodo.5345121
- Jakovljevic, I., Gütl, C., & Wagner, A. (2022, October 1). Towards a privacy-aware reproducible machine learning pipeline for open data. In *4th international open search symposium*. (4th International Open Search Symposium : OSSYM 2022, OSSYM 2022 ; Conference date: 10-10-2022 Through 12-10-2022)

- Jakovljevic, I., Gütl, C., & Wagner, A. (2023, 1). Privacy-preserving collaborative filtering: Evaluating a machine learning recommender system in a large interconnected organization. In *Proceedings of the 5th international open search symposium*.
- Jakovljevic, I., Gütl, C., & Wagner, A. (2022, 12). Analyzing the effects and applicability of social media elements in notification systems in large interconnected organisations. *IARIA JOURNALS*.
- Jakovljevic, I., Gütl, C., Wagner, A., & Nussbaumer, A. (2022). Compiling open datasets in context of large organizations while protecting user privacy and guaranteeing plausible deniability. In *Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*.
- Jakovljevic, I., Pobaschnig, M., Gütl, C., & Wagner, A. (2022). Privacy aware identification of user clusters in large organisations based on anonymized mattermost user and channel information. In *Proceedings of the 11th international conference on data science, technology and applications-iaria data analytics*.
- Jakovljevic, I., Russmann, S., Wagner, A., & Gütl, C. (2022). A proposal for client based user profiles for open search in large and highly connected organisations. In A. Wagner, M. Granitzer, C. Gütl, C. Plöte, & S. Voigt (Eds.), *Proceedings of the 3rd international symposium on open search technology* (pp. 27–32). (3rd International Open Search Symposium : OSSYM 2021, OSSYM 2021 ; Conference date: 11-10-2021 Through 13-10-2021) doi: 10.5281/zenodo.6840911
- Jakovljevic, I., Wagner, A., & Gütl, C. (2020). Open search use cases for improving information discovery and information retrieval in large anhighly connected organizations. In *Proceedings of ossym 2020*. CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH.
- Jakovljevic, I., Wagner, A., & Gütl, C. (2022). *Towards an open data based privacy-aware reproducible machine learning pipeline*.
- Jakovljevic, I., Wagner, A., Gütl, C., Pobaschnig, M., & Mönnich, A. (2022, March). *Cern anonymized mattermost data*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.6319684> doi: 10.5281/zenodo.6319684
- Javed, M. A., Younis, M. S., Latif, S., Qadir, J., & Baig, A. (n.d.). Community detection in networks: A multidisciplinary review. , 108, 87–111. Retrieved 2021-10-20, from <https://www.sciencedirect.com/science/article/pii/S1084804518300560> doi: 10.1016/j.jnca.2018.02.011
- Jeckmans, A. J. P., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L., & Tang, Q. (2013). Privacy in recommender systems. In N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver,

References

- & X.-S. Hua (Eds.), *Social media retrieval* (pp. 263–281). London: Springer London. Retrieved from https://doi.org/10.1007/978-1-4471-4555-4_12 doi: 10.1007/978-1-4471-4555-4_12
- Johnson, C. C. (2014). Logistic matrix factorization for implicit feedback data..
- Jokela, T., Ojala, J., & Olsson, T. (2015, 04). A diary study on combining multiple information devices in everyday activities and tasks. In (p. 3903-3912). doi: 10.1145/2702123.2702211
- Jones, P. (2017). It department user survey report. CERN.
- Jones, P. (2020). Know the new user communities. CERN.
- Joshi, K., Perkins, W. C., & Bostrom, R. P. (1986). Some new factors influencing user information satisfaction: Implications for systems professionals. In *Proceedings of the twenty-second annual computer personnel research conference on computer personnel research conference* (p. 27–42). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi-org.ezproxy.cern.ch/10.1145/317210.317220> doi: 10.1145/317210.317220
- Kay, R., & Loverock, S. (2008, 07). Assessing emotions related to learning new software: The computer emotion scale. *Computers in Human Behavior*, 24, 1605-1623. doi: 10.1016/j.chb.2007.06.002
- Kayacik, H. G., Just, M., Baillie, L., Aspinall, D., & Micallef, N. (2014). Data driven authentication: On the effectiveness of user behaviour modelling with mobile device sensors. *arXiv preprint arXiv:1410.7743*.
- Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241-251. doi: <https://doi.org/10.1016/j.bushor.2011.01.005>
- Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology*, 26(2), 195–197.
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (p. 591–600). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1772690.1772751
- Lattermann, M., Nauerz, A., & Kriha, W. (2009, 01). Implicit social network construction and expert user determination for context-aware web portal environments. In (p. 119-122).

- Lawrence, J. (1993). *Introduction to neural networks*. California Scientific Software.
- Lee, D., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *IEEE Software*, 14(2), 67-75. doi: 10.1109/52.582976
- Lee, M.-C. (2016, 01). Knowledge management and innovation management: Best practices in knowledge sharing and knowledge value chain. *International Journal of Innovation and Learning*, 19, 206. doi: 10.1504/IJIL.2016.074475
- Li, F. F., Larimo, J., & Leonidou, L. (2020, 06). Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda. *Journal of the Academy of Marketing Science*, 49, 51-70. doi: 10.1007/s11747-020-00733-3
- Li, J., Qiu, L., Tang, B., Chen, D., Zhao, D., & Yan, R. (2019, Jul.). Insufficient data can also rock! learning to converse using smaller data with augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 6698-6705. Retrieved from <https://ojs.aaai.org/index.php/AAAI/article/view/4641> doi: 10.1609/aaai.v33i01.33016698
- Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering* (p. 106-115). doi: 10.1109/ICDE.2007.367856
- Li, T., Li, Y., Zhang, M., Tarkoma, S., & Hui, P. (2023). You are how you use apps: User profiling based on spatiotemporal app usage behavior. *ACM Transactions on Intelligent Systems and Technology*.
- Li, Y., Zhang, Z., Peng, Y., Yin, H., & Xu, Q. (2018). Matching user accounts based on user generated content across social networks. *Future Generation Computer Systems*, 83, 104-115.
- Liang, S. (2019, Jul.). Collaborative, dynamic and diversified user profiling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 4269-4276. Retrieved from <https://ojs.aaai.org/index.php/AAAI/article/view/4335> doi: 10.1609/aaai.v33i01.33014269
- Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41(4, Part 2), 2065-2073. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0957417413007240> doi: <https://doi.org/10.1016/j.eswa.2013.09.005>
- Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021, mar). When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv.*, 54(2). Retrieved from <https://doi.org/10.1145/3436755> doi: 10.1145/3436755

References

- Lopez-Vega, H., Tell, F., & Vanhaverbeke, W. (2016a, 02). Where and how to search? search paths in open innovation. *Research Policy*, 45, 125-136. doi: 10.1016/j.respol.2015.08.003
- Lopez-Vega, H., Tell, F., & Vanhaverbeke, W. (2016b, 02). Where and how to search? search paths in open innovation. *Research Policy*, 45, 125-136. doi: 10.1016/j.respol.2015.08.003
- Lorenz-Spreen, P., Wolf, F., Braun, J., Ghoshal, G., Djurdjevac-Conrad, N., & Hövel, P. (2018). Tracking online topics over time: understanding dynamic hashtag communities. *Computational Social Networks*, 5, 5-9. doi: 10.1186/s40649-018-0058-6
- Lu, J., Jin, S., Liang, J., & Zhang, C. (2021). Robust few-shot learning for user-provided data. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4), 1433-1447. doi: 10.1109/TNNLS.2020.2984710
- Luo, G. (2016, 06). Predict-ml: A tool for automating machine learning model building with big clinical data. *Health Information Science and Systems*, 4. doi: 10.1186/s13755-016-0018-1
- Maher, N. A., Senders, J. T., Hulsbergen, A. F., Lamba, N., Parker, M., Onnela, J.-P., ... Broekman, M. L. (2019). Passive data collection and use in healthcare: A systematic review of ethical issues. *International Journal of Medical Informatics*, 129, 242-247. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1386505619302527> doi: <https://doi.org/10.1016/j.ijmedinf.2019.06.015>
- Mahesh, B. (2019, 01). *Machine learning algorithms -a review*. doi: 10.21275/ART20203995
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press. Retrieved from <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>
- March, J. G. (2008). *Explorations in organizations*. Stanford: Stanford University Press.
- McCartney, S. (1999). *Eniac: The triumphs and tragedies of the world's first computer*. Walker & Company.
- McCrickard, D., Chewar, C., Somervell, J., & Ndiwalana, A. (2003, 12). A model for notification systems evaluation—assessing user goals for multitasking activity. *ACM Trans. Comput.-Hum. Interact.*, 10, 312-338. doi: 10.1145/966930.966933
- McCrickard, D. S., Czerwinski, M., & Bartram, L. (2003). Introduction: design and evaluation of notification user interfaces. *International Journal of Human-Computer Studies*, 509-514. doi: [https://doi.org/10.1016/S1071-5819\(03\)00025-9](https://doi.org/10.1016/S1071-5819(03)00025-9)

- Mehrotra, A., & Musolesi, M. (2017). Intelligent notification systems: A survey of the state of the art and research challenges. *ArXiv*, *abs/1711.10171*.
- Milano, S., Taddeo, M., & Floridi, L. (2020, 12). Recommender systems and their ethical challenges. *AI SOCIETY*, *35*. doi: 10.1007/s00146-020-00950-y
- Mooers, C. N. (1960). The next twenty years in information retrieval; some goals and predictions. *American Documentation*, *11*(3), 229-236. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090110306> doi: <https://doi.org/10.1002/asi.5090110306>
- Navarro-Arribas, G., Torra, V., Erola, A., & Castellà-Roca, J. (2012). User k-anonymity for privacy preserving data mining of query logs. *Inf. Process. Manag.*, *48*(3), 476–487. Retrieved from <https://doi.org/10.1016/j.ipm.2011.01.004> doi: 10.1016/j.ipm.2011.01.004
- Newman, M. E. J., & Girvan, M. (2004, February). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. Retrieved 2021-07-23, from <http://arxiv.org/abs/cond-mat/0308217> doi: 10.1103/PhysRevE.69.026113
- Obuchowski, N. (2003, 11). Receiver operating characteristic curves and their use in radiology¹. *Radiology*, *229*, 3-8. doi: 10.1148/radiol.2291010898
- Oh, J., Hessel, M., Czarnecki, W. M., Xu, Z., van Hasselt, H. P., Singh, S., & Silver, D. (2020). Discovering reinforcement learning algorithms. *Advances in Neural Information Processing Systems*, *33*, 1060–1070.
- Ormancey, E., Wagner, A., Jakovljevic, I., Antunes, C., & Carpentier, C. (2022, Feb). CERN. Retrieved from <https://gitlab.cern.ch/push-notifications/notifications-docs/-/tree/master/docs>
- Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, *12*(3), 801–823.
- Panteli, A., & Boutsinas, B. (2023). Addressing the cold-start problem in recommender systems based on frequent patterns. *Algorithms*, *16*(4). Retrieved from <https://www.mdpi.com/1999-4893/16/4/182> doi: 10.3390/a16040182
- Park, S., Goo, J. M., & Jo, C.-H. (2004, 03). Receiver operating characteristic (roc) curve: Practical review for radiologists. *Korean journal of radiology : official journal of the Korean Radiological Society*, *5*, 11-8. doi: 10.3348/kjr.2004.5.1.11

References

- Patel, B., Desai, P., & Panchal, U. (2017). Methods of recommender system: A review. In *2017 international conference on innovations in information, embedded and communication systems (iciiecs)* (pp. 1–4).
- Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*.
- Personal Data Protection Commission Singapore. (2018). *Guide to basic data anonymisation techniques*. Retrieved 2022-01-26, from [https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation.v1-\(250118\).pdf](https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation.v1-(250118).pdf)
- Pielot, M., Church, K., & de Oliveira, R. (2014). An in-situ study of mobile phone notifications. In *Proceedings of the 16th international conference on human-computer interaction with mobile devices services* (p. 233–242). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2628363.2628364
- Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d’Alché-Buc, F., ... Larochelle, H. (2020). Improving reproducibility in machine learning research (A report from the neurips 2019 reproducibility program). *CoRR*, *abs/2003.12206*. Retrieved from <https://arxiv.org/abs/2003.12206>
- Pobaschnig, M., Gütl, C., Jakovljevic, I., & Wagner, A. (2023). *Community detection and community behavior in notification systems*.
- Poo, D., Chng, B., & Goh, J.-M. (2003). A hybrid approach for user profiling. In *36th annual hawaii international conference on system sciences, 2003. proceedings of the* (p. 9 pp.-). doi: 10.1109/HICSS.2003.1174242
- Pradhan, S., Qiu, L., Parate, A., & Kim, K. (2017). Understanding and managing notifications. In *“ieee infocom 2017 - ieee conference on computer communications* (p. 1-9). doi: 10.1109/INFOCOM.2017.8057231
- Pramanik, I., Lau, R., Hossain, M., Rahoman, M., Debnath, S., Rashed, M. G., & Uddin, M. (2021, 01). Privacy preserving big data analytics: A critical analysis of state-of-the-art. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*. doi: 10.1002/widm.1387
- Qiu, F., & Cho, J. (2006). Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on world wide web* (p. 727–736). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/1135777.1135883> doi: 10.1145/1135777.1135883

- Rafter, R., & Smyth, B. (2001). Passive profiling from server logs in an online recruitment environment. In *Proceedings of the ijcai workshop on intelligent techniques for web personalization (itwp 2001)* (pp. 35–41).
- Ramachandran, R., Bugbee, K., & Murphy, K. (2020, 11). From open data to open science. doi: 10.1002/essoar.10505011.1
- Reddy, M. C., & Jansen, B. J. (2008). A model for understanding collaborative information behavior in context: A study of two healthcare teams. *Information Processing Management*, 44(1), 256–273. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0306457307000313> (Evaluation of Interactive Information Retrieval Systems) doi: <https://doi.org/10.1016/j.ipm.2006.12.010>
- Reddy B, S., Krishnamurthy, M., & Asundi, A. (2018, 03). Information use, user, user needs and seeking behaviour: A review. *DESIDOC Journal of Library and Information Technology*, 38, 82–87. doi: 10.14429/djlit.38.2.12098
- Renaud-Deputter, S., Xiong, T., & Wang, S. (2013). Combining collaborative filtering and clustering for implicit recommender system. In *2013 ieee 27th international conference on advanced information networking and applications (aina)* (p. 748–755). doi: 10.1109/AINA.2013.65
- Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: bayesian personalized ranking from implicit feedback. *CoRR*, abs/1205.2618. Retrieved from <http://arxiv.org/abs/1205.2618>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).
- Rist, T. (2009). Information navigation. In *Encyclopedia of database systems* (pp. 1501–1502). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-39940-9_815 doi: 10.1007/978-0-387-39940-9_815
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65–386.
- Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., ... Merom, R. (2010). Suggesting friends using the implicit social graph. In (p. 233–242). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/1835804.1835836> doi: 10.1145/1835804.1835836

References

- Rouibah, K., Hamdy, H., & Al-Enezi, M. (2009, 03). Effect of management support, training, and user involvement on system usage and satisfaction in kuwait. *Industrial Management and Data Systems*, 109, 338-356. doi: 10.1108/02635570910939371
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Runeson, P., Olsson, T., & Linåker, J. (2021). Open data ecosystems — an empirical investigation into an emerging industry collaboration concept. *Journal of Systems and Software*, 182, 111088. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0164121221001850> doi: <https://doi.org/10.1016/j.jss.2021.111088>
- Saeed, F. (2022, 01). Filter bubble and fake news: Facebook and journalist ethics. *International Journal of Research and Innovation in Social Science*, 06, 59-65. doi: 10.47772/IJRISS.2022.61004
- Sahami Shirazi, A., Henze, N., Dingler, T., Pielot, M., Weber, D., & Schmidt, A. (2014). Large-scale assessment of mobile notifications. In (p. 3055–3064). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2556288.2557189> doi: 10.1145/2556288.2557189
- Salamatian, S., Zhang, A., Calmon, F. d. P., Bhamidipati, S., Fawaz, N., Kveton, B., ... Taft, N. (n.d.). Managing your private and public data: Bringing down inference attacks against your privacy. , 9(7), 1240–1255. Retrieved 2021-10-22, from <http://arxiv.org/abs/1408.3698> doi: 10.1109/JSTSP.2015.2442227
- Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression..
- Schafer, B., Konstan, J., & Riedl, J. (1999, 10). Recommender systems in e-commerce. *1st ACM Conference on Electronic Commerce, Denver, Colorado, United States*. doi: 10.1145/336992.337035
- Schiaffino, S., & Amandi, A. (2009, 01). Intelligent user profiling. *Artificial Intelligence*, 5640, 193-216. doi: 10.1007/978-3-642-03226-4_11
- Sedhain, S., Sanner, S., Braziunas, D., Xie, L., & Christensen, J. (2014). Social collaborative filtering for cold-start recommendations. In *Proceedings of the 8th acm conference on recommender systems* (p. 345–348). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2645710.2645772> doi: 10.1145/2645710.2645772
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. New York, NY, USA: Cambridge University Press.

- Shawi, R. E., Maher, M., & Sakr, S. (2019). Automated machine learning: State-of-the-art and open challenges. *CoRR*, *abs/1906.02287*. Retrieved from <http://arxiv.org/abs/1906.02287>
- Sherman, R. (2015). Chapter 18 - project management. In R. Sherman (Ed.), *Business intelligence guidebook* (p. 449-492). Boston: Morgan Kaufmann. Retrieved from <https://www.sciencedirect.com/science/article/pii/B9780124114616000186> doi: <https://doi.org/10.1016/B978-0-12-411461-6.00018-6>
- Shi, Y., Larson, M., & Hanjalic, A. (2014, may). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, *47*(1). Retrieved from <https://doi.org/10.1145/2556270> doi: [10.1145/2556270](https://doi.org/10.1145/2556270)
- Shirani, A., Aiken, M., & Reithel, B. (1994, November). A model of user information satisfaction. *SIGMIS Database*, *25*(4), 17-23. Retrieved from <https://doi-org.ezproxy.cern.ch/10.1145/192561.192570> doi: [10.1145/192561.192570](https://doi.org/10.1145/192561.192570)
- Sieg, A., Mobasher, B., & Burke, R. (2007, 01). Learning ontology-based user profiles: A semantic approach to personalized web search. *IEEE Intelligent Informatics Bulletin*, *8*, 7-18.
- Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.
- Singh, R., & Hsu, Y.-W. (2007). Analysis of usage patterns in experiential multiple perspective web search. In *Proceedings of the 15th acm international conference on multimedia* (p. 569-572). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/1291233.1291373> doi: [10.1145/1291233.1291373](https://doi.org/10.1145/1291233.1291373)
- Smith, T., & Wagner, A. (2020). *Open data, open science and open search*. Retrieved 2023-01-25, from <https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Findico.cern.ch%2Fevent%2F913056%2Fcontributions%2F3839957%2Fattachments%2F2196337%2F3713543%2F0penData-OpenScience-OpenSearch-OSSYM-2020-WM--v01.pptx&wdOrigin=BROWSELINK>
- Sousa, S., Guetl, C., & Kern, R. (2021). *Privacy in open search: A review of challenges and solutions*.
- Staunton, C., Barragán, C., Canali, S., Ho, C., Leonelli, S., Mayernik, M., ... Wonkham, A. (2021, 12). Open science, data sharing and solidarity: who benefits? *History and Philosophy of the Life Sciences*, *43*. doi: [10.1007/s40656-021-00468-6](https://doi.org/10.1007/s40656-021-00468-6)
- Sugimura, P., & Hartl, F. (2018). *Building a reproducible machine learning pipeline*. arXiv. Retrieved from <https://arxiv.org/abs/1810.04570> doi: [10.48550/ARXIV.1810.04570](https://doi.org/10.48550/ARXIV.1810.04570)

References

- Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. *IEEE Trans. Neural Networks*, 9(5), 1054–1054.
- Takács, G., Pilászy, I., & Tikk, D. (2011). Applications of the conjugate gradient method for implicit feedback collaborative filtering. In *Proceedings of the fifth acm conference on recommender systems* (p. 297–300). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2043932.2043987> doi: 10.1145/2043932.2043987
- The Radicati Group, I. (2019). *Email statistics report, 2015-2019*. Retrieved 2019, from <https://www.radicati.com/?download=email-statistics-report-2015-2019>
- Tkáč, M., & Verner, R. (2016). Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, 38, 788–804.
- Tsipenyuk, G., & Crowcroft, J. (2017, oct). An email attachment is worth a thousand words, or is it? In *Proceedings of the 1st international conference on internet of things and machine learning*. ACM. doi: 10.1145/3109761.3109765
- Uda, N., Mizoue, C., Donkai, S., & Ishimura, S. (2018, 11). Information seeking behavior of older adults in a public library in japan. *Libres*, 28, 1-12.
- Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, 109(2), 373–440.
- Van Schalkwyk, F., & Verhulst, S. (2017, 12). The state of open data and open data research..
- Verstrepen, K., Bhaduriy, K., Cule, B., & Goethals, B. (2017, sep). Collaborative filtering for binary, positiveonly data. *SIGKDD Explor. Newsl.*, 19(1), 1–21. Retrieved from <https://doi.org/10.1145/3137597.3137599> doi: 10.1145/3137597.3137599
- Villanueva, M. (2019, 01). Using gestures to interact with home automation systems. *Advanced Materials Proceedings*, 4, 18-25. doi: 10.5185/amp.2019.1443
- Vrouwenvelder, K., & Stall, S. (2023, Feb). *Community building for data sharing and open science within the earth, space, and environmental sciences*. Copernicus Meetings. Retrieved from <https://doi.org/10.5194/egusphere-egu23-2780>
- Weber, D., Shirazi, A. S., & Henze, N. (2015). Towards smart notifications using research in the large. In *Proceedings of the 17th international conference on human-computer interaction with mobile devices and services adjunct* (p. 1117–1122). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2786567.2794334

- West, J., Salter, A., Vanhaverbeke, W., & Chesbrough, H. (2014, 06). Open innovation: The next decade. *Research Policy*, 43, 805-811. doi: 10.1016/j.respol.2014.03.001
- Whissell, J., & Clarke, C. (2011, 10). Improving document clustering using okapi bm25 feature weighting. *Inf. Retr.*, 14, 466-487. doi: 10.1007/s10791-011-9163-y
- Wiederhold, G., & McCarthy, J. (1992, 06). Arthur samuel: Pioneer in machine learning. *IBM Journal of Research and Development*, 36, 329 - 331. doi: 10.1147/rd.363.0329
- Williams, O. (2019, Jan). *What are communication channels within an organization?* Retrieved from <https://smallbusiness.chron.com/communication-channels-within-organization-61447.html>
- Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), 2839-2846. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0031320315000989> doi: <https://doi.org/10.1016/j.patcog.2015.03.009>
- Yagisawa, M. (2015). Fully homomorphic encryption without bootstrapping. *Cryptology ePrint Archive*.
- Yao, X., Huang, T., Wu, C., Zhang, R., & Sun, L. (2019). Towards faster and better federated learning: A feature fusion approach. *2019 IEEE International Conference on Image Processing (ICIP)*, 175-179.
- Ye, N., Chai, K. M. A., Lee, W. S., & Chieu, H. L. (2012). Optimizing f-measure: A tale of two approaches. In *Icml*. icml.cc / Omnipress. Retrieved from <http://dblp.uni-trier.de/db/conf/icml/icml2012.html#NanCLC12>
- Yeung, C. M. A., Gibbins, N., & Shadbolt, N. (2008). Collective user behaviour and tag contextualisation in folksonomies. In *2008 ieee/wic/acm international conference on web intelligence and intelligent agent technology* (Vol. 3, p. 659-662). doi: 10.1109/WIIAT.2008.265
- Zafar, M. B., Valera, I., Ródriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics* (pp. 962-970).
- Zamberi, F., Adli, N., Hussin, N., & Ahmad, M. (2018, 01). Information retrieval via social media. *International Journal of Academic Research in Business and Social Sciences*, 8, 1375-1381. doi: 10.6007/IJARBS/v8-i12/5239
- Zhang, D., Zambrowicz, C., Zhou, H., & Roderer, N. K. (2004). User information-seeking behavior in a medical web portal environment: A preliminary study. *Journal*

References

- of the American Society for Information Science and Technology*, 55(8), 670-684. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20001> doi: <https://doi.org/10.1002/asi.20001>
- Zhang, J., Wang, Y., Yuan, Z., & Jin, Q. (2020, 04). Personalized real-time movie recommendation system: Practical prototype and evaluation. *Tsinghua Science and Technology*, 25, 180-191. doi: 10.26599/TST.2018.9010118
- Zhang, Q., Tan, S., Li, L., Zhao, Y., Yin, D., & Yin, S. (2021). Morse-stf: Improved protocols for privacy-preserving machine learning..
- Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, 5(1), 30-43.
- Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists* (1st ed.). O'Reilly Media, Inc.
- Zheng, H., Hu, H., & Han, Z. (2020). Preserving user privacy for machine learning: Local differential privacy or federated machine learning? *IEEE Intelligent Systems*, 35(4), 5-14. doi: 10.1109/MIS.2020.3010335
- Zheng, Y., Gao, C., Chang, J., Niu, Y., Song, Y., Jin, D., & Li, Y. (2022, 02). *Disentangling long and short-term interests for recommendation*.
- Zukerman, I., & Albrecht, D. (2002, 10). Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11. doi: 10.1023/A:1011175525451