NTNU
Det skapende universitet

User Privacy in Recommender Systems

Itishree Mohallick

Master i informatikk
Innlevert: juni 2017
Hovedveileder: Jon Atle Gulla, IDI

Norges teknisk-naturvitenskapelige universitet
Institutt for datateknologi og informatikk

NTNU – Trondheim
Norwegian University of
Science and Technology

User Privacy in Recommender Systems

Itishree Mohallick

June 29, 2017

Supervisor : Professor Jon Atle Gulla
Co-supervisor : Özlem Özgöbek

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

This page is intentionally left blank.

Abstract

With the increasing ubiquity of access to online information sources, the recommender systems have emerged as a powerful tool to reduce information overload and provide customized information access for the targeted audience. Recommender systems are prevalent in every aspect of the web starting from the e-commerce to the most dynamic environment of news. Despite the growing popularity, these recommender systems are not 100% trustworthy, as the personal information used in these systems give rise to serious privacy concerns. Users whose privacy is invaded at least once are skeptical of using such systems in later times. Therefore, this thesis considers the research concerning user privacy in the recommendation context as a problem worth addressing. This thesis includes the privacy risks and the existing technical approaches to combat the same while considering the current privacy regulations as a safety measure for the concerned users.

Unlike prior privacy work concerning domain agnostic recommendation, news domain has been chosen as an additional research context. Specifically, this thesis identifies the various privacy aspects prevailing in the news recommendation domain. News personalization has become crucial on the web as user shows more interest to stay updated with the current news trend within a limited time span. The quality and accuracy of such personalized news recommendation rely on leveraging user profiles of the news readers. For instance, many news aggregator sites such as Google News suggest its users to provide sign in to the system for getting user-specific (relevant) news articles. For more generic news recommendation, the system collects user click history and page access pattern implicitly. The need and association of user profiles give rise to privacy concerns in the news domain, whereas privacy of user identity, user behavior in terms of page access patterns contributes to the overall privacy risks in the news domain.

Finally, a user-based research has been conducted through a set of the survey questionnaire to accumulate the privacy-centered opinions of the online users. It is found that user's privacy preferences, awareness, and ownership (control) over their own data can highly influence online users privacy concerns. In addition, the analysis of the survey results reveals that the Norwegian users are less concerned about online privacy as compared to the non-Norwegian users.

This page is intentionally left blank.

Sammendrag

Med økningen i tilgangen til nettbaserte informasjonskilder så har bruken av anbefalingssystemer tredd frem som et kraftig verktøy for å redusere mengden overfloden av informasjon og samtidig tilby tilpasset innhold for den spesifikke målgruppen. anbefalingssystemer er mye brukt i forskjellig aspekter knyttet til nettet med alt fra netthandel til dynamisk nyhetsområder. Til tross for den økte populariteten så er ikke rekommendasjonssystemene nødvendigvis 100 % troverdige på grunn av at den personlige informasjonen som disse systemene samler inn kan utgjøre en personvernrisiko. For en bruker som opplever at privat informasjon blir misbrukt av et slikt system vil naturligvis være skeptiske til slike systemer senere. Derfor tar denne oppgaven utgangspunkt i undersøkelse av personvern i anbefalingssystemer som viser at dette kan være et problem som det er verdt å se nrmere på. Denne oppgaven inkluderer også personvernrisikoer og de de eksisterende tekniske løsningene brukt for å beskytte personlig informasjon, samt de nåværende lovene rundt personvern med tanke på bekymrede brukere.

I motsetning til tidligere forskning utført på personvern ved domeneuavhengige anbefalingssystemer så har nyhetsdomener i denne avhandlingen blitt valgt som et ekstra forskningspunkt. Mer konkret så vil denne avhandlingen identifisere de personlige opplysningene som inngår i anbefalingssystemer for nyhetsdomener. Personalisering av nyheter har blitt mer viktig da en bruker er mer interessert i å holde seg oppdatert på spesifikke nyheter innenfor en kort tidsperiode. Kvaliteten og nyaktigheten til slike persontilpassede nyheter er avhengig av å sanke informasjon om leserne. Som et eksempel så er ønsker nyhetssamlere slik som Google News at brukere skal kunne logge inn i systemet for å få persontilpassede nyheter. For mer generiske nyhetsforslag så samler systemet brukerens netthistorie og ser mønster i nettsidene brukeren har besøkt. Behovet for brukerprofiler øker risikoen for personvernet i nyhetsdomener, mens logging av en brukers netthistorie fører til en økt risiko for personvernet til en hvilken som helst bruker av nyhetsdomenet.

Til slutt så har det også blitt utført en brukerunderskelse gjennom en serie med spørreskjemaer for å kartlegge brukernes meninger om personvern på nettet. Det ble konkludert med at en brukers preferanser med tanke på personvern, hva brukeren visste om innsamling av persondata, samt det eventuelle eierskapet av den innsamlede dataen hadde en stor innvirkning på en brukers mening om personvern. En analyse av resultatet fra underskelsen viste også at norske brukere er mindre opptatt av personvern på nettet sammenlignet med brukere fra andre nasjoner.

This page is intentionally left blank.

Preface

This report is submitted to the Norwegian University of Science and Technology (NTNU), as partial fulfillment of the degree of Master of Science (Informatics), and as part of the course *IT3901-Informatics Postgraduate Thesis (Software)*. The work culminating in this report has been performed at the Department of Computer and Information Science (IDI). This research work is supported by the NTNU SmartMedia[1] program on News recommendation and accomplished under the supervision of Professor Jon Atle Gulla.

[1] https://www.ntnu.no/wiki/display/smartmedia/SmartMedia+Program

This page is intentionally left blank.

Acknowledgement

First and foremost, I would like to express my sincere gratitude towards my supervisor Prof. Jon Atle Gulla for his supervision and invaluable feedback. I would also like to thank my co-supervisor Özlem Özgöbek for her contribution and guidance throughout the year. She has been incredibly patient and a huge motivator for me during the project.

My gratitude also goes to the students, friends, and professionals who have participated in the survey.

Finally, I would like to express deep appreciation to my parents (Purna Chandra Mohallick and Santilata Nayak), my husband (Dr. Kumar Ranjan Rout) and my daughter Eleena for their unconditional love and support and dedicate my thesis to them. Last but not least, my gratitude goes to my in-laws as well for their support.

This page is intentionally left blank.

Table of Contents

This page is intentionally left blank.

List of Tables

This page is intentionally left blank.

List of Figures

Abbreviations

ACM	=	Association for Computing Machinery.
AIS	=	Adaptive Information Server.
ANN	=	Artificial Neural Networks.
AOL	=	American Online.
CBF	=	Content Based Filtering.
CF	=	Collaborative Filtering.
DPD	=	Data Protection Directive.
EEA	=	European Economic Area.
EU	=	European Union.
GIC	=	Group Insurance Commission.
HTML	=	Hypertext Markup Language.
IDI	=	Department of Computer Science (Norwegian abbreviation).
IITF	=	Information Infrastructure Task Force.
IMDB	=	Internet Movie Data Base
IP	=	Internet Protocol.
ISP	=	Internet Service Provider.
LDA	=	Latent Dirichlet Allocation.
LSI	=	Latent Semantic Indexing.
MAE	=	Mean Absolute Error.
MAS	=	Multi-Agent System.
MDL	=	Minimum Description Length.
NN	=	Nearest Neighbor.
NRS	=	News Recommender Systems.
NTNU	=	Norwegian University of Science and Technology.
OS	=	Operating System.
OECD	=	Organisation for Economic Co-operation and Development.
PDA	=	Personal Data Act.
PDA	=	Personal Digital Assistant
P3P	=	Platform for Privacy Preferences Project.
PSP	=	Privacy Service Provider.
PLSI	=	Probabilistic Latent Semantic Indexing.
RS	=	Recommender Systems.
SPiD	=	Schibsted Payment ID.
SVM	=	Support Vector Machines.
URL	=	Uniform Resource Locator
US	=	United States.
W3C	=	World Wide Web Consortium.

This page is intentionally left blank.

Chapter 1

Introduction

Over the last decades, the Internet has become a ubiquitous part of our daily lives. Several factors such as the development of Web 2.0 technologies, has increased the deployment of mobile networks and the access to the mobile devices for the users. Hence, an extensive amount of information is readily available on the palms of the users for consumption. With a seemingly never-ending flood of information streams and limited time to evaluate each piece of information, users have to rely on a personal system which can filter, prioritize, and suggest the relevant content according to the user interests and preferences. As a whole, this is the problem of information overload. Recommender Systems (RS) have emerged as a powerful tool to reduce information overload and provide customized information access for the targeted audience (Adomavicius and Tuzhilin, 2005).

Recommender systems (Jannach et al., 2010) are information filtering systems associated with various application domains or websites. They strive to satisfy the user's need by providing tailored services by taking their tastes and interest into account. In most cases, these systems use computational methods to analyze users past actions and decisions. In addition, user-related or item-related information are used for generating the useful personalized recommendation. Recommender systems are used in multiple application domains starting from social networking sites, e-commerce to online content streaming sites. They are designed to improve the user experience by automatically filtering the extensive data about user preferences, behaviors and providing the item of interest to respective users. Thus, recommender systems are able to reduce individual user's cognitive load, and simultaneously provides them with more valuable and relevant product and services.

The scope of such 'personalized' services is not limited to any domain or any specific information content. However, 'personalization' requires more detailed information related to the user attributes and preferences. The accuracy of recommendation depends on the detailed user information and serves as the basis for generating the recommendation. On contrary, the same amount of collected and consolidated user data induces threat to the user's privacy in the RS (Jeckmans et al., 2013; Friedman et al., 2015; Lam et al., 2006).

Due to the fact that privacy risks associated with user data and RS are multifaceted, research regarding the privacy risks in few application domains remains challenging. This introduces the need to study the privacy concerns in RS from user and application domain perspective. User-centered research is important in information systems because the various web-based systems (including recommender systems) are developed and designed to be used by the end users. Primarily, privacy in recommender systems is concerned with user information. Hence, finding what the users think about privacy in the recommender systems is a relevant research objective. This introduces the requirement to conduct a user-centered survey to ascertain the opinion of RS users on privacy.

While many research work has already been done to understand the privacy risks associated with RS (in general) and the possible privacy preserving techniques, this thesis focuses on the privacy risks associated with a specific application domain, i.e., the news domain. A more focused approach is adopted to research the privacy characteristics of news recommender systems and the possible privacy preserving techniques. The final contribution of this thesis includes a user survey. The user survey is designed to find out the interesting and unique features related to user opinion concerning privacy in RS.

1.1 Problem Statement

RS are an inherent part of the web. The majority of the internet users must have come across some kind of RS during internet usage. For example, while reading online news, Google News suggests the readers "Top Stories" section irrespective of the user preferences. However, for a regular reader, the same online news website provides the opportunity to customize the news reading experience by knowing the user interest. Hence, generating a "personal newspaper" for each signed in user. Facebook suggests new friends for adding into the existing friend list. LinkedIn suggests job offers, news, interesting companies, and new connections in the relevant fields based on the user's resume and existing connections. In most of the cases, users provide the related information explicitly and build their own user profiles. But in some systems (News Recommender Systems) where explicit user feedback is rare, the system collects user feedback implicitly by storing the browsing pattern and click behavior for generating recommendation (Doychev et al., 2014). RS try to collect as much user data as possible because a precise and rich user profile results in a more accurate recommendation.

However, revealing the content of the user profile for receiving personalized convenience goes against the user privacy. In both of the aforementioned cases, personal information related to the user might be violated or manipulated by the service provider, sold to or shared with a third party or leaked by an attacker (hacker). This phenomenon is known as the "privacy-personalization tradeoff" (Chellappa and Sin, 2005; Awad and Krishnan, 2006). The privacy risk increases with more advanced recommendation scenario. Therefore, the main challenges are to understand the various privacy risks which can later contribute to designing robust RS. As different domains posses different unique properties, a later research included the privacy aspects from news recommendation perspective.

Although there exists several efficient RS which can provide accurate recommendation, very few of them deal with privacy concerns or aim to deal with the privacy risks as addressed in (Ramakrishnan et al., 2001). Many privacy preserving techniques such as anonymization (Sweeney, 2002), by applying perturbation (adding random value) to user ratings (Polat and Du, 2005) and differential privacy (Dwork, 2006) have been suggested and evaluated in different recommendation domains. Considering these existing techniques, an evaluation is performed for the possible application of such methods in the news domain. To gain a better insight of user's opinion regarding privacy, a user-centered survey has been required. Hence a survey has been designed and conducted. The outcomes of the survey data are further analyzed to find out the different user opinions which can influence privacy aspects of the recommender systems.

1.2 Background & Motivation

RS are capable of identifying user's requirements. Modern RS deploy various sophisticated recommendation technologies for generating precise and accurate recommendation but at the same time falling out to provide the required privacy to users. In the past, different researchers have addressed the privacy breaches with the so-called robust RS. One such privacy violation was addressed with Netflix Prize data set. Netflix, an online movie rental, and service company have announced a million dollar prize for an improved movie prediction algorithm in 2006. To do so, Netflix published an 'anonymized' subset of its in-house customer's (more than 480,000 users) movie rating data. Although the prize was won by teams who came up with an improvised prediction algorithm with improved accuracy, later in 2008 the same data set led to a widespread privacy concern. In 2008, researchers were able to de-anonymize the users in the published Netflix dataset (Narayanan and Shmatikov, 2008). They were able to identify the customers by linking the existing Netflix dataset with the unanimous reviews of a popular online movie rating website (IMDB). Hence, revealing many potentially sensitive information (apparent political preferences, religion, beliefs, race, sexual orientation) of the customers.

This Netflix issue has ever since raised the privacy concerns because the privacy preserving techniques failed to acquire the desired privacy in its case. It also proved that the most prominent service providers are not taking enough measures to provide adequate privacy to the user's sensitive data as promised. This issue raises concern regarding the current state of privacy in the RS. Although many questions related to privacy are answered in the context of the RS in general, some application domains are quite untouched.

In this thesis, a thorough literature review is performed to understand the various aspects of privacy from both the user data and RS perspective. Later, the acquired knowledge is utilized to understand different privacy aspects of the news recommender systems. As seen in the above Netflix case, privacy is not achievable by providing only the technical solution. Therefore, the goal of this thesis is to broaden the view to look for solutions from a user, privacy policy, Laws and Regulations (Data Protection Laws) perspective. In addition, research is performed to find out the privacy risks in the news recommender systems where research regarding privacy concerns are found to be still young.

Invasion of information privacy in RS is the main focus of this thesis. Hence user-centered research is the appropriate way to understand users privacy preferences and privacy beliefs as its the user's information which is at risk. The most suitable way to do so is to conduct a survey for gathering the valuable feedback from the users. Hence, a user-centered survey is performed to find the answers for the final research objective.

1.3 Research Questions & Goals

The objective of the research is to explore and identify the various privacy characteristics associated with RS. The research consists of two phases, where the primary phase focuses on researching the current work concerning privacy risks in RS and the technological solutions for retaining the privacy features. This part of research aims at filling the gap between current literature study and most recent developments regarding the privacy issues prevailing in RS irrespective of any specific domain. To investigate the specific characteristics of the privacy risk and the possible solution for preserving privacy in news recommender systems is another criterion of this thesis. Another important contribution is the user survey which addresses the interesting features regarding the user's privacy concerns. This part of the research is performed in the second phase.

The following goals are identified within the research context towards understanding privacy risks and solutions associated with RS.

G1: Research on the state-of-the-art of the privacy risks and the possible solutions of the recommender systems. The goal is to learn about the research done on privacy and form a knowledge base to support an assessment of the domain.

G2: Based on the previous research outcome, the objective of this thesis is to identify the particular characteristics of privacy risks and domain aspects in the news domain.

G3: On the basis of the above two research goals we try to explore the feasibility of possible privacy preserving solution in the news recommender systems.

G4: Explore the privacy attitude of the users by conducting an user-centred survey. A survey has been carried out in order to investigate the user's attitude towards controlling individual data in recommender systems.

This thesis primarily seeks to answer the following research questions.

RQ1: What are the privacy risks in recommender systems?

RQ2: What are the particular characteristic privacy risks in news recommender systems?

RQ3: What are the techniques as a solution to the privacy risks of recommender systems?

RQ4: How people think about privacy issues in recommender systems?

1.4 Research Context

The area of privacy in RS is widespread as the application of such systems in multiple domains are boundless. The privacy in RS stretches further where the RS are deployed across domains. As privacy is multifaceted and every domain has its own set of characteristics, it is not possible to research everything in this context. Therefore this thesis focuses on identifying the privacy risks and possible solutions associated with the news recommender systems. The result of the aforesaid research objective is based on the preliminary research done through an extensive literature study concerning user privacy and the privacy solutions in recommender systems. Here, the identified privacy risks are evaluated with multiple perspectives. Furthermore, the evaluation of the research work has been carried out by conducting a user-centered survey.

1.5 Research Methodology

This section presents the research methods adopted in the thesis to investigate the research questions. Besides, the research challenges and limitations are also addressed regarding these methods.

The research process described in (Oates, 2006) consists of the following components: experiences and motivation, research question (s), literature review, conceptual framework, strategies, data generation methods and data analysis. Selection of right methodology is important in research for finding the most appropriate answers for the research questions 1.3.

The objective of this thesis is to study the privacy risks and existing privacy solutions concerning recommender systems to later identify the privacy risks associated with news recommender systems. In addition, this thesis seeks to study user's opinion regarding privacy risks in the recommender systems. In order to answer these research questions, the following methods have been applied: literature review, survey, data collection, and evaluation. Figure 1.1 highlights the research methods applied in this thesis.

1.5.1 Literature Review

Through the literature review, the first two research questions, privacy risks and existing privacy solutions in recommender systems are studied. Relevant research findings related to the aforementioned topic (from both the technical and non-technical perspective) are revised which is included in the later Chapter 3. Based on the findings of those literature reviews, possible privacy risks in the field of news recommender systems are identified and included in Chapter 4 which answers our third research question. This chapter concludes with a discussion to find out if the privacy preserving techniques stated in the previous chapter are suitable for the news recommender systems. Besides, the various evaluation processes, recommendation methods and characteristics of the news domain are studied

through the literature review which helps to understand the basic theories behind a recommendation process.

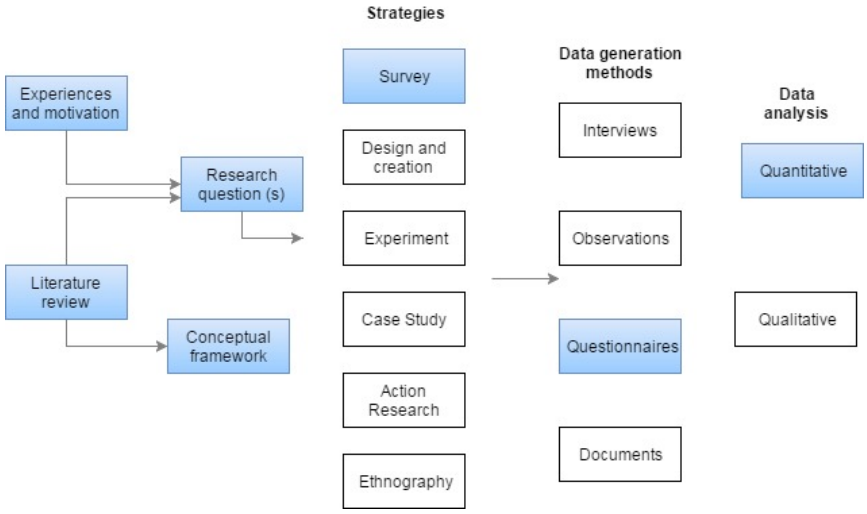


Figure 1.1: Model of the research process (Oates, 2006)

1.5.2 Survey

The primary objective of a survey strategy is to gather similar data from a group of people in an organized manner. Then the found statistical patterns are utilized to establish a general trend for a larger population (Oates, 2006).

The final research question of this thesis is the main driving force behind conducting an opinion-based user survey. The results of the survey are included in Chapter 5. The following paragraphs include the survey design and participant selection process.

Design of Survey

To be precise while designing the survey, a limited number of questions are selected. The survey is intended to focus more on the user's opinion regarding the privacy concepts in the recommender systems. In addition, the survey covers the dimensions such as user interest related to news recommendation, ownership and control over the data, and user's privacy behavior, among others.

The survey is designed using the Google Forms and comprised of 26 questions separated into the following categories;

- General Information (demographic information such as gender, age, and nationalities of the users).
- User's knowledge of privacy in RS.

– User's opinion concerning privacy in RS.

Finally, the respondents are informed about the motive and purpose behind the survey.

Participants

A selected group of professionals and students are targeted to participate in the survey ranging between 18 and 65+ years old. The purpose behind choosing a group of people from the academic background is to ensure that the subjects have the minimal knowledge related to the research topic. This type of participant selection process corresponds to *non-probabilistic* "purposive sampling" where the research motive is to explore a topic in depth rather than make generalization (Oates, 2006).

1.5.3 Data Collection

A pre-defined questionnaire is used for collecting data during the survey. The questionnaire is designed to collect exact feedback from users to understand privacy from the users perspective in recommender systems. This also aims at collecting expected privacy requirements from the users and the current trend of user interest in the various domains (including News). The set of questions includes 25 closed questions and only one open question. Most of the closed questions and responses are based on the "Likert scale"(1 to 10). The open question enables the user to state their exclusive opinion regarding privacy and recommender systems. The survey questions are listed in Appendix B.

1.5.4 Evaluation

A quantitative analysis of user data is performed to find out users opinion regarding privacy in recommender systems and included in Chapter 5.

1.5.5 Limitations

The primary challenge of this project is the chosen topic which is highly relevant yet undermined in the information privacy scenario. Defining the scope of the project and the research objectives are the other aspects of the challenges. The research questions covered a wide range of privacy aspects starting from technical, non-technical and user-based approaches made the projects more demanding. To complete a wide range of research topics within a time span of 36 weeks is found to be difficult. The research methods adopted and the results found are partially dependent on the opinion and comprehension of the authors (from literature study) and the users (results from user survey). This is considered as a limitation of this project.

1.6 Operational Definition of Terms

User Privacy

Privacy inherits numerous definitions from the context point of view. In this thesis context, privacy revolves around information privacy of the users. According to the Information Infrastructure Task Force (IITF), privacy of information or "information privacy" is defined as below (Kang, 1998):

Information privacy is "an individual's claim to control the terms under which personal information – information identifiable to the individual – is acquired, disclosed or used."

— IITF

Users

Users in recommender systems are the individuals or the group of individuals using the recommendation service from any online service providers. It is the user's sensitive personal data which is at risk of exposure in recommender systems. Both the online and offline users avail the recommendation services from various service providers. Mostly, users are considered, to be an honest and law abiding citizen while using different online services. But in contrast, some users try to use the recommendation services for their personal gain while presumed to be following the given protocols. Even some of the users try to invade other user's privacy. These users are known as "malicious users" and create profound privacy threats in recommender systems (Jeckmans, 2014).

Personal Data

According to The Norwegian Data Protection Authority, "Personal data (*personopplysning*) is a piece of information or assessment that can be linked to any person as an individual. This information includes individual's name, address, phone, email address, IP address, car registration number, photographs, fingerprints, iris pattern, head shape (Face) and identification number (including both the date of birth and social security number)" (Datatilsynet, 2016b). However, the inclusions are not limited.

Personalization

Personalization is directly related to information privacy. This can be defined as the ability of a system to proactively tailor products and services based on tastes and personal preferences of an individual user. Therefore, personalization critically depends on both the user and the service provider. The willingness of the user to share personal information for receiving personalized services and the ability of the service provider to collect and process that user information are key factors for the success of a personalized system (Chellappa and Sin, 2005).

Recommender Systems

Recommender systems aim at generating a meaningful recommendation to a group of users that might interest them. For instance, a suggestion for items (books, clothes and electronics devices) on Amazon, movies on Netflix are the few pioneers in the real recom-

mendation world. RS differ from each other in the way they analyze collected user data and generate the recommendation (Melville and Sindhwani, 2010).

News Recommender Systems

News recommender systems work in the same way the other recommender systems work except the fact that it filters out the most 'relevant' and "well-timed" news articles to the reader. Google News (Das et al., 2007), Daily Learner (Billsus and Pazzani, 2000), NewsWeeder (Lang, 1995a), GroupLens (Resnick et al., 1994) are few examples of the current news recommender systems.

User Profile

The user profiles are generated from the user feedback related to various artifacts in the recommender systems. This relates the attributes of the various items to user interests (ratings) (Aggarwal, 2016b; Jannach et al., 2010). User profiles or user models are developed and maintained for generating the recommendation in RS. Personal data regarding the users in the form of user ratings or user action constitute the user profile. Exposure of user profile data may lead to privacy concerns in RS.

1.7 Documentation & Collaboration Tools Used

Google Drive

Google Drive is a free web-based application developed by Google. This allows users to create, store, organize, edit and share the files or documents with anyone. It consists of GoogleDocs, GoogleSheets, GoogleSlides and more. Everyone has access to it and it makes the real-time collaboration easier. This allows the user to edit the document from anywhere in the real time and share the documents with the supervisor for review (Google Drive, 2012).

BibTex

Bibtex is an online referencing tool used with LaTex document. This is used for managing and formatting the reference list while writing LaTex reports. It is easier for users to follow BibTeX citation style which is comparatively easy and allows users to cite different sources with consistency (BibTex, 2016).

ShareLaTex

ShareLaTex is a web-based collaborative editor for writing research reports in LaTex. This is free and user-friendly. ShareLaTex allows real-time collaboration and online compiling of projects to PDF format. The report is written in the left part of the editor and the preview is visible on the right side of the editor in pdf format. The program is compiled when the report is edited and displays the errors if any (ShareLaTex, 2016).

Zotero

Zotero is a research tool which is used to organize the articles and other sources of information referred during the research process. Different databases are searched for the identified keywords and the Zotero plugin is then used to organize all the screened articles. Zotero further extends the flexibility to identify and eliminate the duplicate information sources (Zotero, 2006).

1.8 Report Structure

To start with, this report states the executive summary in the form of an abstract which provides a brief overview of the research work done. The rest of the project report is organized as follows.

- Chapter 1 introduces the thesis and states the research goals.
- Chapter 2 presents the state-of-the-art for information privacy concept from personalization point of view. In addition, this chapter introduces the basic privacy threats in online systems and privacy related regulations (non-technical privacy solutions) in EU and Norway.
- Chapter 3 includes basic background theory for the various privacy concerns and the research works related to the privacy preserving techniques in recommender systems.
- Chapter 4 describes the news recommendation with domain specific characteristics. The identified privacy risks in the news recommender systems are included along with a concluding discussion regarding the possible privacy solutions.
- Chapter 5 presents the user-centered survey and the subsequent analysis of the obtained data.
- Chapter 6 concludes the thesis with the concluding results and future work.

Chapter 2

State of the Art

The purpose of the literature review is to understand the research domain and the related aspects. The acquired knowledge is later used to identify the gaps (if any) in the research work and establish new results from the collected research data (Oates, 2006). This thesis includes research on privacy in recommendation context which requires an extensive literature review. This chapter presents some of the selected literature which has been reviewed and found to be relevant with regard to overall research domain. The primary source of articles has been the Google search engine, Google Scholar and ACM digital library.

2.1 Introduction

The aim of this section is to provide the readers with insight regarding the privacy considerations concerning collected user information and its possible exploitation in the RS with a closer look for the privacy in news recommendation. This section considers the state-of-the-art for basic theories and legal solutions concerning privacy. Precise technical details concerning privacy risks in recommender systems and the possible solution measures are described in Chapter 3. Recommender systems are meant to provide personalized services to online users. To be more generic while describing privacy, personalized systems are used rather than RS in this chapter.

2.2 Personalized Systems

The pervasiveness and the growing availability of online products and services are the landmark of the current digital era. Online users entrust personalized systems by sharing

their personal information, such as name, age, address, profession or credit card numbers while availing different services. At the same time, sophisticated web technologies enable these personalized systems to track the user's movement on the web (Steinke, 2002). The personal data can be collected, stored, analyzed or shared easily with third parties without users knowledge by these systems. User data is treated as a valuable commodity and the internet makes it easier to access these unlimited data available on countless website irrespective of any geographic boundaries. The possibilities of misuse and manipulation of personal data increase with the lack of proper rules and regulations. This particular concern regarding the malicious use of user data is known as *"digital data privacy"* issue.

2.3 Data Collected in Personalized Systems

Information collected for the users are stored in user profiles in the web based systems. A user profile consists of data which tells everything about a user. A more detailed user profile leads to better-personalized services. User profiles include both the directly identifiable information about the person (name, age and email) and other information related to person's online behavior. The type of data collected and used on any standard web-based personalized system is given below (Rao et al., 2014) whereas a more comprehensive classification of information used in RS is included in Section 3.8.

- Demographic data consists of the background information regarding the user. This includes name, address, sex, age, marital status, zip code, education level, employment (type of industry) income, the number of family members in the household, the number of children, the age of children, ethnicity, religious affiliation and so on. This kind of data is obtained when the users sign up for receiving new services provides their personal details by themselves. Also, tracking technologies help the online systems in acquiring the demographic data regarding a user.
- Location data is retrieved through Wi-Fi, GPS and IP-address used by the user.
- Technical data consists of the details regarding user's digital devices such as the computer, smart phone, tablet and other devices used for establishing the connection between users and internet. For instance, IP address, operating system (like Windows 10), or browser (Google Chrome) is the technical data collected by any online system.
- Predictive data consists of the prediction of interest, behavior, and attitudes of a user which are derived from a large amount of aggregated data by the various online systems (including personalized systems).
- Psychographic data consists of user's interest and attitudes. For example, an online user might be interested in health and fitness related news or products.
- Behaviour data consists of user's lifestyle, activities, and personality.

– Life Event data consists of in a certain event in user's life which impacts users behavior and requirement. For example, a status update on Facebook, "On vacation at Madrid" may expose details of user's current location.

2.4 Online Tracking Technologies

Different types of tracking technologies are used for tracking online user's data. Various Internet Service Providers (ISP) use different tracking technologies to provide targeted advertising or personalized services for its online users. There is a common assumption prevailing among online users that the personalized services (recommendations) costs them nothing. On contrast, users receive those services at the cost of their personal information (Datatilsynet, 2015a; Ersdal and Skjrstad, 2016). Given below are some the tracking technologies used for collecting online user information and the preventive measures.

– Browser cookies are used widely to track online users (Datatilsynet, 2015a). A cookie is a small file that is stored in the user equipment when the user visits a website. Every time the user visits the site, the web browser sends information back to the site's server to notify the website about the user's activity on the page. Depending on the usage, cookies are distinguished as first-party cookies and third-party cookies. First-party cookies are placed on the domain website by the website owner whereas the later is placed on a domain website by a third party owner. First-party cookies are deleted when the web sessions end, but third-party cookies are not session dependent. Online service providers are able to track individual users over different websites and build exclusive user profiles due to the presence of third-party cookies over multiple websites. Nowadays, cookies are facing a lot of resistance from privacy inclined users.

– Web Widget is a small application placed on the websites (Wikipedia, 2016). These are used to interact with different websites by displaying contents from and redirecting users to other websites. The end users are able to place these small functional codes on their websites, blogs or personalized start page as standalone applications. A common example of web widgets is the Google advertisements. Typical widgets vary from pop-ups to social sharing buttons (Ersdal and Skjrstad, 2016). For instance, in an online news site, social sharing buttons are embedded for every news article. So that, the interesting news may be shared with friends in the social media. In this way, the news website tries to reach a larger audience. So, web widgets are useful for enhancing the websites.

– IP address is a unique identifier associated with any digital devices such as desktops, laptops, and tablets (Datatilsynet, 2015a). The information collected by IP-address includes the location information of a user and network information. Typically, most of the users use the same IP address for a longer period of time. For example, a user 'X' is using the internet through the same desktop for past 3 years from his home network. Therefore, it's easier to track user 'X' over time through the IP-address. The advantage of an IP-address is the ease of accessibility for the website owners.

- Web-beacons is a small, invisible graphic image file. Generally, they are placed within the HTML documents on a website (Sipior et al., 2011). They have used alone or combined with cookies for collecting additional information. The information collected by web beacons may include user interaction on the web page, mouse movement, typed entries, search queries, IP addresses, user's demographic data or clickstream data.
- Digital fingerprint, also known as *"Device fingerprint"* is an advanced tracking technology used for uniquely identifying and tracking users across the web. These are used by websites as an alternative to tracking cookies (or when the tracking cookies are turned off by the users). When a computer is connected to the Internet, it gets a unique electronic imprint (Datatilsynet, 2015a). This electronic imprints with added information like browser type, Operating System (OS) type, installed items (plugins, fonts etc), IP address, location and time zone settings can be aggregated to create digital fingerprint (Zawadzinski, 2016). Typically, a digital fingerprint is able to operate from a single browser for identifying users. The recent advances in digital fingerprinting have enabled to track users over multiple browsers on the same device.
- Unique ID is the tracking solutions proposed to trace both the online and mobile users by outperforming the flaws of the previously stated tracking technologies Datatilsynet (2015a). This unique ID is adopted by major online service providers for generating login solutions for users. One such service provider is Schibsted media group (http://www.schibsted.com/no/), which provides unique login solution 'SPiD' for different websites such as Finn.no, VG+. SPiD is used as single login and payment solution for the users in multiple websites. This provides more accurate data about users such as email address or mobile numbers. Unique ID provides service providers to have more control over their user data.

2.5 Preventive Measures for Online Tracking

Online tracking contributes to a great deal of user's privacy loss. Therefore, various preventive measures are provided to the users to avoid online tracking and limiting information collection.

Do Not Track (DNT) is a web browser setting which is used to disable online tracking if turned on by a user. DNT sends a special request to websites and other related web services, to stop tracking the concerned individual. There is no current standard concerning the use of Do Not Track in ISP. So, most ISP ignores the DNT requests and continue with their current practices (Future of Privacy Forum, 2016). Other options such as Opt-out cookies (allaboutcookies.org, 2016) and browser extensions (Ersdal and Skjrstad, 2016) are used for manually opting out for cookies or blocking the third-party tracking companies.

Despite the fact that, online tracking raises privacy concerns for online users, tracking is essential for the website owners. The aforementioned tracking technologies provide the

adequate knowledge of a user to the websites and hence enables them to produce user-specific services.

2.6 Privacy

Privacy is derived from the Latin word 'privatus' which means to withdraw from the public life and or to have seclusion from the public. The definition of privacy varies from situation to situation while the central concept remains the same. Privacy is associated with multiple subjects. Any system dealing with personal identity information is subjected to the potential privacy risk. Maintaining privacy by these systems involves various aspects: legal, organizational, behavioral and technical aspects. This section describes privacy of any personal data which is collected and exploited by the various web-based service providers. In addition, this chapter documents the various legal privacy requirements such as privacy laws and regulations within the context of personalization and recommendation.

Privacy is described as one of the many potential research challenges posed by the RSs by John Riedl (Riedl, 2001). The term privacy in RS is hard to describe from the research perspective without mentioning personalization, as both the terms are very closely associated. Typical online users consider these personalized services as a privilege and share their preferences, as long as the desired service is received. But, the users hardly know about the owner or usage of their web data once the online communication is over. There are possibilities that the user data might be sold or shared with third party systems afterward without the knowledge of the users. For example, people often wonder after receiving a marketing call while doing something important. They might think how did these people get their personal mobile number? This is a case of invasion of user privacy in return for the received personalized/non-personalized services. This user must have shared his/her personal mobile number during any online transaction in the past. Later, the advertisement agencies might have received the authorization of this user's data by some possible ways.

So, privacy is an important aspect of personalization based recommender systems. The Details of the privacy risks in personalization context (technical aspects) is included in Chapter 3. In principle, recommender systems are also subject to privacy rules and regulations, as they collect personal data which may be used to identify respective individuals

2.6.1 Platform for Privacy Preferences Project (P3P)

P3P is a part of the proposal adopted by the World Wide Web Consortium's (W3C) (World Wide Web Consortium, 2000). This is designed as an international standard for online privacy. This provides a computer-readable format for privacy policies and a protocol. The P3P protocol enables web browsers to read and process the privacy policies automatically. The main objective of the P3P project is to develop a variety of tools and services which

empower the users by giving them greater control over their personal information. Thus, P3P helps in increasing trust between online users and web-based systems.

2.6.2 Privacy Policy

A privacy policy is an appropriate tool for incorporating the various privacy laws, guidelines and privacy statements. This is a written statement which explains the collection and usage of personal data specific to any web based systems (Awad and Krishnan, 2006). Primarily, privacy policies express the right, permission, and obligation of individuals (a person or a system). These are articulated and stated in a variety of context in every sector (e-commerce, financial, health or government). Privacy policy associated with any website describes the basic rights of its end users and the permissions retained by the system itself. This also describes the obligation of the website towards its customers adhering to the laws and regulations. Privacy policies are presented to users during user registration process. For example, during a user registration process for Yahoo!, users are shown an option "I agree to the Yahoo Terms and Privacy". This states the terms and conditions including the privacy policy for Yahoo. Privacy policies make sure that the end users know about the privacy practices of the specific system. The users must agree to the privacy policy associated with a system before using its service.

2.6.3 Legal and Legislative Approach

This section provides a brief overview of the privacy and data protection laws for regulating privacy concerns in personalized systems from the European Union (EU) and Norwegian legislation point of view. The opinion and acceptance regarding the term privacy vary from people to people around the globe, so does the approaches to privacy regulation. The basis for this research helps in identifying the potentials and threats for privacy in various media industries within Norway and abroad in the later sections. This section concludes with the discussion of all the findings from the research.

Many initiatives are taken from the legal and legislative purpose to retain privacy in the various sections of the enterprises (including online systems). This includes the US privacy laws, EU data protection privacy laws and many more specific national privacy initiatives (Casassa Mont, 2004). Various guidelines, such as OECD guidelines (OECD, 2013), are established to ensure the protection of privacy and the transborder flows of personal data. The following sections briefly describe the privacy concerns and the adopted policies to fight the issue within EU and Norway.

EU Regulation

European Union (EU) has a very strong stand on the regulation of personal data and its movement on the web. The EU regulation provides the highest level of protection to personal data from rest of the world by providing "right to privacy" to individual user. The revision of EU data protection rules "Regulation (EU) 2016/679" and "Directive (EU) 2016/680" ensures a more stricter privacy guideline for the European consumers across Europe and outside as well (European Commission, 2017d). . **Data Protection Directive**

The *ePrivacy Directive* and *General Data Protection Regulation* constitutes the standardized EU legal framework for safeguarding digital privacy within Europe (European Commission, 2017a). The enactment of the above directive has a greater impact on protecting the personal data within EU and outside of EU as well. Therefore, the service providers like Amazon, eBay, America Online, and Yahoo! have set up their websites in EU countries to keep EU data separate from the rest of the world (Steinke, 2002). This act forbids the online tracking of user's movements by Doubleclick (through cookies) inside Europe. However, Doubleclick is allowed to track the online user's movement inside the *US*.

- *ePrivacy Directive* was first introduced in 1995 by the European Union as the Data Protection Directive (DPD) and took effect from 1998. This ensures the best possible protection to data while the data is accessed or exported abroad. After the revision in 2009 to the *ePrivacy Directive*, "informed consent for cookies" are made mandatory. In addition, this ensures that any kind of privacy violation with user data is reported by the respective service providers. The European Commission has adopted a new proposal for replacing the existing *ePrivacy Directive* (European Commission, 2017b) on January 10th, 2017. This includes the following change of rules.
  - New players like WhatsApp, Facebook Messenger, and Skype are included under electronic communication application must provide the similar level of data protection as traditional telecom operators.
  - Stronger rules for the protection of user data across EU.
  - Protection of communication content and metadata: location, content and time of a call
  - New business opportunities for telecom service providers once the user has provided the consent regarding collection of communication content and meta data.
  - Simpler rules on cookies makes the process of acceptance and rejection of cookies in the web browser more user-friendly.
  - Protection against spam.
  - More effective enforcement of the data protection regulations.

- *General Data Protection Regulation* is adopted by EU in 2016 to (European Commission, 2017a) ensure that the collection of personal data meets the required guidelines. A valid purpose is required for the collection of personal data by any of the service providers. Later, the respective service providers ensure privacy for users by protecting the misuse of user data. A revision of the Data Protection Regulation includes a set of updated rules to provide better control to an individual over their personal data. These rules are listed below and going to effective by May 2018 (European Commission, 2017d).
  - The right to be forgotten
  - Better control over who holds ones private data
  - The right to switch ones personal data to another service provider
  - The right to be informed in clear and plain language
  - The right to know if your data has been hacked
  - Clear limits on the use of profiling
  - Special protection for children

Norwegian Regulations

The Norwegian Data Protection Authority (Datatilsynet), established in 1980, is responsible for regulating both the national and international processing of personal data and the associated risks inside Norway. The "Personal Data Act" (*Personopplysningsloven*) (Datatilsynet, 2017a) and the "Personal Data Regulation" (Datatilsynet, 2017b) (*Personopplysningsforskriften*) act as the two main pillars of Norwegian data regulation. The transfer of personal data from Norway to other countries takes place under the strict supervision of these regulations. Norwegian regulations work in accordance with the EU Data Protection Directive as per the EEA agreement. Therefore, the same set of EU rules stated earlier is going to be effective for personal data inside Norway as well from May, 2018.

The PDA ensures "right to privacy" for every Norwegian citizen by securing the processing of their personal data. The transfer of personal data is only possible in the countries of EU or European Economic Area (EEA) which provides an adequate level of protection to personal data (Datatilsynet, 2017a).

Safe Harbor agreement

Enforcement of the Data Protection Directive provides safety for personal data within the countries of EU/EEA. The set of rules from the EU Directive along with PDA ensure safety for Norwegian data as well. However, the Data Protection Directive prevents the flow of personal data between EU and US, as the latter does not comply with the EU privacy standards.

The Safe Harbor agreement between EU and the US offers a convenient way of complying with the adequate level of safety requirements of the EU Directive. This allows the

personal data to be transferred to the US in a secure way (Steinke, 2002). This principle also regulate the transfer of Norwegian data to the US.

On October 6th, 2015, the Safe Harbor framework is declared as invalid by the European Commission (Datatilsynet, 2015b). However, a set of existing standard contracts between EU and the US is working on the legal basis for the transfer of personal data outside the EU territory. These set of contracts are also applicable to countries which do not satisfy the adequate level of protection as stated by EU.

EU-US Privacy Shield

On July 12th, 2016, the EU-US privacy shield is adopted by the European Commission to control transatlantic data transfer and transfer of personal data to US (European Commission, 2016). This enforces stronger obligations on US based enterprises to provide safety to personal data. It ensures greater transparency for transfer of personal data to the US. The new framework includes the following set of rules:

- strong data protection obligations on companies receiving personal data from the EU safeguards on;
- U.S. government access to data;
- effective protection and redress for individuals;
- annual joint review to monitor the implementation.

This section has documented the various legal aspects till date, as the legal approaches aim at providing protection for personal data irrespective of the application. Hence, the legal approaches are applicable to protect personal data in RS as well. The technical approaches for preserving privacy in RS is described in Section 3.10.

2.7 Privacy Policy in News Domain

In the earlier sections of this chapter, online privacy issue has been discussed with various examples. Privacy regulation from EU perspective has been included to provide the viewers an initial understanding of how the information privacy violation takes place and what are the non-technical measures (organizational or legislative measures) to deal with user's information privacy.

An additional research objective of this thesis is to investigate privacy aspects from online news recommendation perspective as well. While studying the various privacy-related regulations and solutions, the privacy policy is found to play an important role in online privacy. The privacy policy is a salient document which states the privacy statements applicable to the service providers and the user. This is the only document which is readily available for the users before they can avail any online services and states how and why websites collect, use and manage user information. User's awareness can be increased and trust can be built for the online service providers by understanding the given privacy policies. Hence, privacy policies from two online news website (Adresseavisen and Google

News) are studied. Among the two online news websites, Adresseavisen is a Norwegian online newspaper and Google News is the most popular news site across the globe. Both of the online news websites have deployed multiple recommender systems for providing personalized experience for their readers. The former news site provides privacy through PDA (including EU Regulation) whereas the latter follows the privacy regulations according to the geographic location.

This section deals with privacy in online news domain from the policy perspective and is dedicated to finding out how the privacy policies stated in the online news websites addresses users privacy concerns.

2.7.1 Adresseavisen

Adresseavisen[1] is the oldest newspaper in Norway which is currently owned by Polaris Media Group. It has started the internet version of the newspaper in the year 1996. When the online media is dominated by the service providers like Google and Facebook for their personalized services, Norwegian media is trying hard to make their own platform in the field of personalization. In the race between the "most data" and "best technology", user data and their interest are traded as a commodity by these service providers. Every website claims to protect the privacy of their users by taking the consent of the users before providing any of the personalized services. The user is asked to accept the "terms and conditions" for the service, it is going to avail. However, according to (Datatilsynet, 2016a), the information provided by the policy statement as for how the user data is protected, is quite vague and very generic. Different technologies are used to gather user data on online platforms. Cookies, IP-addresses, web beacons, and digital fingerprints are few techniques used to gather user data. Currently, Login solutions are introduced to overcome the shortfalls over these techniques which can track user's unique identity (name, address, phone number). Polaris Media Group is currently using the unique login solutions (Unique ID) provided by the Schibsted media group ('SPid') for collecting the more valuable user data than the cookies (Datatilsynet, 2016a). In the context of providing privacy to the online users, it would be worth noting the privacy information provided by Adresseavisen (Polaris Media Group) (Polaris Media, 2009). A survey conducted in (Datatilsynet, 2016a) shows, an online newspaper page of Adresseavisen contains 139 cookies, 37 third parties and 57 IP-addresses for tracking the online user's activity and interest for user profiling and segmenting. The claim for using anonymous user data being used for user profiling by the news website is difficult to verify by the authorities. Also, the policy includes how the user's digital identity can be defended by giving the user control over their own data. This can be achieved via ghostery[2] browser extension or by using privacy tools such as disconnect[3]. But online data retention is a hidden risk embedded with personal data. Despite the stated privacy statements in Adresseavisen, it is difficult to predict the extent of privacy protection to personal data.

[1] http://www.adressa.no/
[2] https://www.ghostery.com/
[3] https://www.disconnect.me/

2.7.2 Google News

Google News[4] is a news aggregation website, first introduced in September 2002 and operated by Google (biggest internet service provider). Aggregation techniques act as an unbiased human editor and enable Google News to generate the front page without any human input. It collects the news stories from multiple news providers. Hence, a wide variety of news stories is covered in Google News which is not possible in case of a single news provider. The "Top Stories" section of Google News is carefully chosen from the top ranked stories of prominent news providers. The precise details regarding the techniques behind the personalization and ranking (recommendation) algorithms are the proprietary of the Google system (Billsus and Pazzani, 2007). However, given policy statements from Google are referred to gain insight of the topic.

This paragraph aims at understanding the general concepts related to collection and usage of user data in Google News by researching the available documents (Privacy and Terms[5]) from Google. Google has emerged as the "big brother" in the digital world by acquiring a vast amount of user data and possess a greater privacy risk as compared to the collected user data. A previous work (Ersdal and Skjrstad, 2016) has detailed the policy from Google in the context of the social networking site Google+. Privacy policy from Google is applicable for all Google services, except Gmail and YouTube. Hence, Google's privacy policy and "Terms of Use" is applicable for Google News as well. User data is accessible to the users through Google's transparency services such as dashboards, account activity, and ad preferences. Users are allowed to choose the desired ad or opt-out from the advertisement sites through a given website *youronlinechoices*[6]. Furthermore, Google provides various browser plugins for enabling opt-out options for users.

The research work from (Ersdal and Skjrstad, 2016) found Google's privacy policy to be extensive. Besides, it is hard for the users to get precise information related to their data specific to Google News, as the privacy policy is applicable to different Google products including Google News. The authors of (Ersdal and Skjrstad, 2016) found the policy to be partial and vague. The same has been concluded for Adresseavisen as well. So, there is much scope for improving user-friendly privacy policies from the organizational perspective.

[4] https://news.google.com/
[5] https://www.google.com/policies/privacy/
[6] http://www.youronlinechoices.com/nor/dine-valg

This page is intentionally left blank.

Chapter 3

Background Theory

This chapter presents concepts concerning privacy in recommender systems. These theoretical and technical concepts capture reasons for privacy risks in recommender systems. Furthermore, this relevant background knowledge is used to understand and identify privacy aspects in news recommender systems.

3.1 Historical Background

Privacy in recommender systems holds a considerable amount of importance for the successful evaluation of such intelligent and adaptive systems. The roots of privacy as a concept can be traced back through the centuries. However, privacy in recommender systems came into the limelight after the invasion of Governor William Weld's Medical Information by a graduate student back in 1998 (Sweeney, 2002).

3.1.1 Re-identification of Governor's data

In an attempt to re-identify personal data by linking the publicly available dataset, a graduate student was able to identify the medical records of William Weld (the governor of Massachusetts of that time). This resulted in severe privacy loss of the concerned subject (Sweeney, 2002). In the process of re-identification, Latanya Sweeney, a graduate student, tried to identify the unique users by matching them against the available information in two databases. One of the two databases was the anonymized dataset released by the Massachusetts-based Group Insurance Commission (GIC) which was responsible for purchasing health insurance for state employees. The second database was the voter registration list for Cambridge Massachusetts. She purchased this voter registration list for 20 dollars which contained the details like the name, address, zip code, birth date, and

gender of each voter. The matching of the two databases against the ZIP code resulted in medical records of the governor who was a resident of Massachusetts. This kind of linkage attack draws further attention of researchers towards loss of privacy due to the risks of re-identification: i.e., the (*"ability to relate supposedly anonymous data with actual identities"*). A new concept of "quasi-identifiers" is introduced in this context. Such identifiers are used to collect a combination of information (such as name or social security numbers) in a private dataset which may help for identifying a person among a population (identified datasets). For example, as shown in Figure 3.1 the combination of ZIP code, birth date, and gender constitutes a quasi-identifier which was able to re-identify 87% of people in the US Census data.

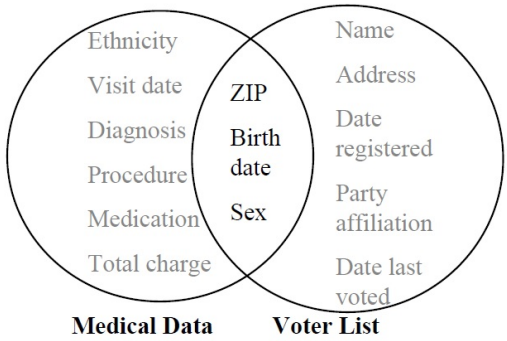


Figure 3.1: Linking to re-identify data (Sweeney, 2002)

To prevent such attacks, k-anonymity privacy model was developed. In this model, the published dataset (anonymized) is considered to be k-anonymous only if every tuple in the dataset looks like k–1 other tuples. For anonymization of the dataset, the personal identifiers such as name and social security numbers are deleted, and other related information is modified before publication. However, the anonymized database still possesses the risk of *re-identification* or *de-anonymization* by intruders or data terrorists.

3.1.2 Re-identification of AOL Searcher No. 4417749

America Online, better known as AOL is a web-based system. In 2006, AOL released poorly anonymized query logs for 20 million web search queries from 6,57,000 users (Barbaro, 2006). The goodwill behind the dataset release is to help the information retrieval research community. However, the initiative is proved to be risky for several of its website users whose web search queries are published publicly. Generally, a query log comes in the form of 5-tuple (Navarro-Arribas et al., 2012): (id, q, t, r, u), where id is the unique identifier assigned to each user by the service providers, q is the search query, t is the time stamp, r is the rank of the clicked url, and u is the url clicked by the user.

Figure 3.2 displays some real user queries from AOL which corresponds to the above given 5-tuple format. As the dataset is already anonymized, the original url is truncated and only the domain url is included in the log. In addition, the user identifiers are anonymized by the hash function. Despite all these privacy measures, user 4417749 (searcher no.) is identified by matching the various search queries and clicks of the users. The anonymized dataset from AOL failed to protect the anonymity of the user 4417749 who is identified as a woman from Georgia named "Thelma Arnold" (Barbaro, 2006).

```
24963762 myspace codes  2006-05-31 23:00:52  2  http://www.myspace-codes.com
24964082 bank of america  2006-05-31 19:41:07  1  http://www.bankofamerica.com
24967641 donut pillow  2006-05-31 14:08:53
24967641 dicontinued dishes  2006-05-31 14:29:38
24969374 orioles tickets  2006-05-31 12:31:57  2  http://www.greatseats.com
24969374 baltimore marinas  2006-05-31 12:43:40
```

Figure 3.2: Example of AOL search query log (Navarro-Arribas et al., 2012)

The unintended AOL data leakage issue created privacy uproar for the personalized web search engines such as Google and Yahoo! This also brings forth the privacy issues lying behind the personalization based web service providers and their inability to comply with stated privacy policies for protecting user's personal data. Hence, a mere anonymization technique is not enough to shield user privacy as promised by the various service providers. Additional research is required to improve existing privacy measures.

Another privacy breach in Netflix context is already discussed in Section 1.2. It is clear from the Netflix fallout that preserving privacy in the so-called robust systems are not always achievable. The proficient methodologies employed with goodwill in various online service providers dealing with public data does not always succeed to provide data privacy for its user. Therefore, the various systems which are dealing with sensitive personal data (with a probability to share or sell the data) need to re-consider and carefully evaluate the privacy risks before publishing the data.

3.2 Privacy and Personalization

Recommender systems proactively tailor the online products and services in accordance with individual user's preferences and needs. This process of tailoring product and services is known as personalization (Chellappa and Sin, 2005). Personalization-based systems enhance user experience in multiple directions on the web and at the same time raise the concern for user privacy (Riedl, 2001; Kobsa, 2007b). Most of the recommender systems aim at providing personalized service and hence comes under the personalization-based systems category. For instance, MovieLens is a personalized recommender system. This recommender system suggests movie for users based on their past seen movie and the opinions given for those movies. So, it is important for such system to know about the preferences of its users before predicting such personalized recommendation. Amazon.com, a

pioneer in the field of e-commerce, uses *automated collaborative filtering* (Adomavicius and Tuzhilin, 2005; Resnick et al., 1994) techniques for providing highly personalized experience to users based on user's past purchase history. User information in the form of purchase history or movie ratings lead to better personalization but also contributes in invading user privacy. The privacy concerns in automated collaborative filtering systems are high where the system tries to maximize the utilization of the user-given contents (in the form of ratings, tagging, or other means of behavioral preferences). Examples of privacy concerns in the context of personalization based services are (Kobsa, 2007b):

- Many personalization-based websites are able to disambiguate user's search queries and present them with relevant search results. But at the same time stores the keywords related to search. This disclosure of individual's genuine search interests might be privacy revealing for the users.
- Online users using personalized stores to purchases private stuff (such as medicines for some critical diseases or fitness equipment for some physical disability) might not want to disclose the purchase history to the system or anyone else. The revelation of such private purchase records may create a privacy threat for the related users.

3.3 Privacy and Recommendation

Recommender systems are widespread in every aspect of the web starting from the e-commerce to the most dynamic environment of news. The process in which these systems gather personal information for predicting personalized decision is known as *recommendation*. Despite the growing popularity, these recommender systems are not 100% trustworthy, as the personal information used in these systems gives rise to serious privacy concerns. Users whose privacy is invaded at least once are skeptical of using such systems in later times. One such privacy invasion scenario describes the issue in a collaborative filtering based recommender systems (Shani and Gunawardana, 2011). Here, the specific user bought a book "The Divorce Organizer and Planner" from a website who is also interested in the art of growing Bahamian orchids. The spouse of that user may get a recommendation for the above-said book under the recommendation section "people who bought this book also bought" while (s)he actually browses for the book "The Bahamian and Caribbean Species (Cattleyas and Their Relatives)" in the same website. So, the probability of revealing a sensitive information always remains with a recommender system.

Recommender systems are designed in many different ways, such as distributed, peer-to-peer, without a server, typical client-server or recommendation via an agent. For example, privacy threats in a traditional client-server architecture based system are discussed in (Lam et al., 2006) where the user trust is prioritized over the value of the information. The privacy concerns discussed in this paper mostly deals with user's trust and the privacy of the personal information which may include user preferences or any form of identifiable information, such as name, address, and gender. The privacy concerns in such a scenario generate due to the violation of user trust in three major forms (Lam et al., 2006):

- Exposure of user-contributed personal information.
- Bias introduced by some external entity to modify the user's recommendation with some malicious intent.
- Sabotage where the attackers deny for carrying out service attacks on the recommender systems.

Privacy concerns raised due to exposure of loosely coupled data is already discussed in section 3.1. Here, the exposure risk is dependent on the domain of recommendation up to a certain extent. For example, the recommendation used in the health care domain contains more sensitive personal information than the acquired data in a music domain. In a music domain, the users might share their preferences more willingly and without thinking much about their own privacy. But in the case of the sensitive domains, any given unusual user–preferences might lead to a privacy breach for users, where the given user-attributes acts as a *quasi-identifier* (Sweeney, 2002). Another kind of exposure risk is discussed in (Ramakrishnan et al., 2001) due to the presence of a specific user (with *eclectic* preferences) in the recommender systems. Here, the risk prevails due to the direct exposure of personal information or the inferred personal information from the presence of *straddlers* in the recommender systems.

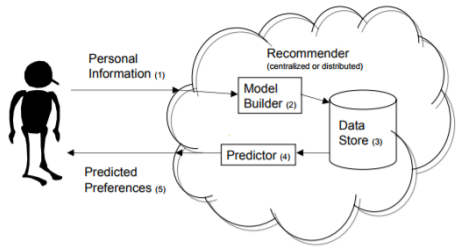


Figure 3.3: Model of Interaction between User and RS (Lam et al., 2006)

The above given Figure 3.3 shows a high level data flow in between a user and the recommender system (Lam et al., 2006). Also, it may represent the classic client-server architecture based recommender system, where privacy concern arises due to the centralized server.

Apart from the risk of exposure, the privacy concerns are induced by user bias (a group of human users or software agents) in the form of a "shilling attack" (Lam et al., 2006). The main potential benefit of such attacks is to manipulate the potential buyers' community. Such attacks are carried out by inducing special opinions (positive or negative) about any products or services with some vested interest in mind to bias the output prediction. One similar example attack in the form of a human user is discussed in the context of book recommendation in Amazon.com. In this case, the author of a particular book wrote multiple positive reviews about his book and published them online to increase the prediction output. Although recommendation in a distributed or peer-to-peer environment can overcome the existing challenges lying with centralized servers integrity and practical application

needs to be verified. These questions are further discussed while we seek the answer to preserve the privacy in case of these attacks in the Section 3.10.

3.4 Recommender Systems

In the everyday life, people rely on suggestions from other people regarding various artifacts (items, places, services, etc,) which are unknown to them. These suggestions come in many forms. For instance, this may come from a friend with similar taste (by *"word of mouth"*) who has an interest in reading books and suggests "The Fountainhead" for a weekend reading. The other kind of recommendation may come in the form of a recommendation letter which can play a crucial role for selection process in a job interview. On this era of the web, recommendation systems are the realization of such social processes (Resnick and Varian, 1997). Recommender systems based on collaborative filtering methods were first introduced in the mid-nineties to combat the information overload problem (Konstan and Riedl, 2012). The information overload is the result of the abundant source of information available on the web due to the increasing popularity and use of the internet. Recommender systems are defined as the information filtering systems that suggest the most relevant item to users from a larger set of items. Recommender systems serve as a solution for the information overload problem and extend the service to provide personalized recommendation. These systems traditionally rely on the user-system interaction history to build user profiles and represent relevant suggestion based on the user's interest. A typical recommender system could provide suggestions like:

- Which movie should 'X' watch in the weekend?
- Which country should 'Y' visit next?
- Which news article would be of interest to 'Z'?

Examples of such recommender systems are:

- Recommendation of movies at Netflix (Gomez-Uribe and Hunt, 2015)
- Recommendation of popular travel destinations by Triplehops TripMatcher (used by www.ski-europe.com, among others) and VacationCoachs expert advice platform, MePrint (used by travelocity.com) (Ricci, 2002)
- Recommendation of netnews to help people find articles at GroupLens (Resnick et al., 1994)

A common recommender system based on "items you own", deployed in the book section of Aamzon.com is shown in the below Figure 3.4. So, these RS not only provide value for the users by narrowing down their set of item choices but also add value to the service providers like Amazon.com by increasing sales and acquiring more knowledge about customers. In the case of Amazon.com, user ratings and user models (profiles) are used to tailor the services or products. Then, these artifacts are shown as recommended to the users on its web page. Therefore, user ratings are counted as the input for these auto-

mated RS, and the output recommendation is further generated and directed towards the appropriate recipients (Resnick and Varian, 1997).

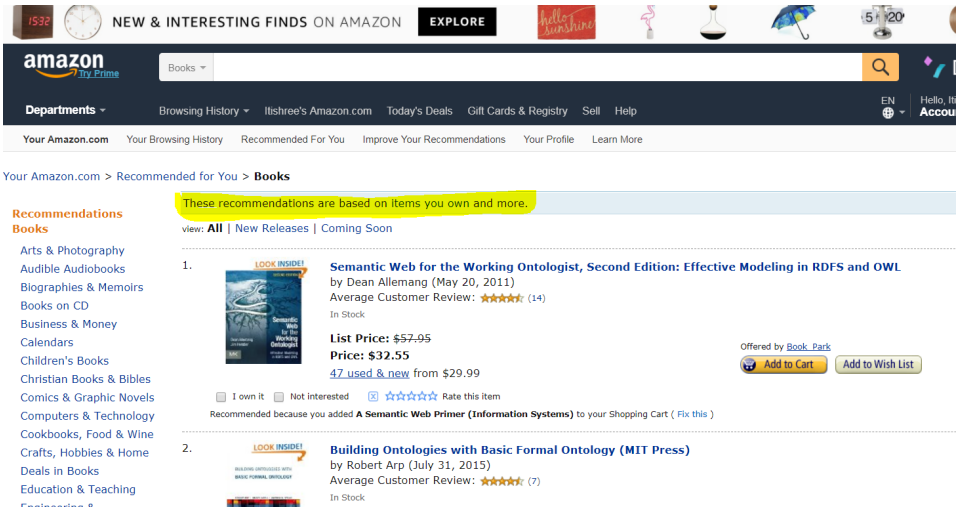


Figure 3.4: Recommender System (RS)

Recommender systems are viewed as functions which compute the relevance scores with a given user profile (e.g. user ratings, preferences, demographics, situational context, social context) and item set (with or without description). These relevance scores are further used for generating the ranking list for items. Based on the ranking list the most relevant items are recommended to respective users. These recommended (relevant) list of items are often dependent on the context which is a major drawback of this recommendation process (Jannach et al., 2010). So, a closer look at relevancy is given below in this context.

Relevancy of recommendation is essential for the success of a recommender system. The term relevance is defined as (Borlund, 2003): *" the utility or usefulness of the information in regard to the user's task and needs"*. For achieving the relevant recommendations, recommender systems predict the "relevance scores" for all the items that are unknown to the users. Items with the highest relevance score are recommended to the particular user. The recommendation process can be stated as follows: Let us consider a set of m-number of users and n-number of possible items are present in the system. The mapping of a user to the item depends on the estimated value of relevance of the item.

$$RelevanceEstimation : \hat{r}(U \times I) \rightarrow \mathbb{R} \cup \{null\} \tag{3.1}$$

$$\text{where } U = \{u_1, u_2, ..., u_m\}, \text{ and } I = \{i_1, i_2, ..., i_n\}$$

From the above Equation 3.1, the recommender system selects the top-N items based on their relevance score from set I. Here, I is the set of unknown yet interesting items for the selected users. Then the items are ranked based on their perceived relevance score before

suggesting them to the particular users. In the process, irrelevant items (all the items below the threshold t) are removed from the information source (Jeckmans et al., 2013; Barbosa, 2015). The below Figure 3.5 displays a typical recommender system paradigm where the recommendation list is generated by predicting the relevance score of the items.

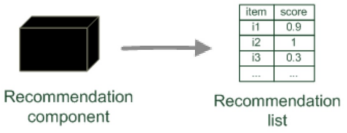


Figure 3.5: Recommender systems: A solution for information overload (Jannach et al., 2010)

RS explore two types of information for predicting the relevance of an item for a user: explicit feedback and implicit feedback. For instance, users generally provide explicit feedback in the form of ratings if they are satisfied with the product or services. These ratings are the exclusive opinions of the user which might come either in the form of a numerical rating (1 to 5), binary ratings (like, dislike) or rating on a Likert scale (strongly agree, neither agree nor disagree, agree and strongly agree). Moreover, implicit feedback is collected by the recommender systems without the knowledge of the user in the form of user's online behavior (browsing patterns and user clicks). This type of feedback is mostly gathered by news recommender systems (Google News) as the users hardly provide any explicit feedback for the news articles (Doychev et al., 2014; Ilievski and Roy, 2013).

3.5 Classification of RSs

According to (Adomavicius and Tuzhilin, 2005), recommender systems are classified into three broad categories depending on the methods they adopt for getting the recommendation: collaborative filtering, content-based filtering, and hybrid systems. Collaborative filtering (CF) approach considers only user opinions regarding the item whereas the content-based filtering (CBF) considers the properties of items for making the recommendation. Hybrid approach ensembles any of the above two basic approaches to overcome the problem which might arise with the use of a single approach. However, the hybrid systems are further divided into 7 other categories such as knowledge-based recommender systems and demographic-based recommender systems (Burke, 2002). This section provides an overview of the various types of recommender systems.

- Collaborative Filtering (CF):

  Collaborative filtering is the most prominent approach used for recommendation starting from the good old days. One such application is 'Tapestry', which is the first

collaborative recommender system designed to retrieve relevant email messages for a particular user from 'Usenet' mailing list. The term of collaborative filtering is coined by the authors for the first time in the paper (Goldberg et al., 1992). The techniques used for Tapestry are still widely used in the current recommendation space: ratings. Every user is supposed to rate a given item. These ratings are further used for generating user similarity matrix or item similarity matrix for that specific user. Finally, the recommendations are made for the highest rated items by peers with similar interest or the most similar items preferred in the past. A common assumption in this system is *"Customers who had similar tastes in the past, will have similar tastes in the future"*. This popular recommendation approach is widely used in large, commercial ecommerce sites such as Amazon.com but not limited to any specific domain (Jannach et al., 2010). A major drawback of this system is the cold start problem, i.e.new items which are not rated by any person or new users who have not yet bought any items.

- Content-Based Filtering (CBF):

  Content-based (CBF) recommender systems require both item and user related information. This kind of recommender systems gathers knowledge about user preferences and attributes (feature) of an artifact (item). Content-based recommender systems assume that: The user will like the items similar to the ones (s)he preferred in the past. Therefore, item similarities are calculated for getting the recommendation. For instance, item similarities are computed for recommending a book (sports) to users who liked the same kind of book in the past. They have been applied in multiple application domains, such as movie, books and web pages. For instance, Newsweeder (Lang, 1995b), a net news-filtering system employs the content-based filtering as the core underlying techniques for generating the recommendation. However, most content-based recommendation techniques are used to recommend text documents (News Groups or Web sites). Limitation of such systems arise due to too short content, new items (cold start problem due to new items) and difficulties in deriving implicit feedback (Jannach et al., 2010). However, this RS does not require a user community to function like the collaborative filtering based recommender systems. The item similarity for the content-based filtering system is computed by item meta data (genre for movies, Top Stories for news). The most interesting and preferred items in the past get recommended to the user. Many similar items in an application domain may lead to a less accurate recommendation and considered as a potential threat in the content-based system. So, pure content-based recommender systems are rarely used in the commercial application domain.

  The following two categories of RS relies on the user and item properties. Hence, they have included under the content-based RS (Barbosa, 2015) category.

  - Demographic

    Analogous to knowledge-based recommender systems, the detailed user preferences are rare in the demographic recommender system. The demographic information such as age, gender, country of residence, job status, educational status, and marital status are used to generate a partially personalized rec-

ommendation. Grundy (Rich, 1998) is an example of a demographic recommender system which acts as a librarian to provide book recommendation to the readers. In a real life situation, a librarian has to build an initial user model by looking at a user before providing any suggestion. This user model completely depends on the librarian's perception on the related user's attributes such as age and nationality. Similar to the given scenario, Grundy matches the preliminary demographic information to the *stereotypes* and recommends the items which are associated with such *stereotypes*. *Stereotypes* are defined as: *"a collection of frequently occurring characteristics of users"*. The major limitation of such method is the generalization of stereotypes.

- Knowledge-Based

  The knowledge-based recommendation comes into picture when there exists multiple artifacts with a low number of available ratings in the system. In this scenario, the user's explicit requirements can help to form the knowledge model for the recommender system. These explicit user requirements are also known as the constraints. Generally, the user specifies all the initial preferences at once. But, the users may also state their preferences in different iterations while seeking for a recommendation. For example, a user specifies the color of the shirt to be 'navy blue' while trying to look (buy) for one on the web. So, a number of outputs get generated based on the explicit knowledge available in the system. The recommendation gets more precise with time as the user defines the specific preferences explicitly. The final output recommendation is delivered to the user after a few such iteration. The feedback of the user is utilized by the recommender systems to further enhance the knowledge base. For example, Entree (Burke, 2000) is an example of aforesaid knowledge-based recommender system which recommends suitable restaurants to the diners. The limitation of such system depends on the cost of knowledge acquisition and the accuracy of the preference models (Jannach et al., 2010).

- Hybrid Filtering

  Hybrid recommender systems unify the techniques from collaborative and content-based methods (Adomavicius and Tuzhilin, 2005; Burke, 2002). These hybrid systems aim at performing better than the rest of the recommender systems by outperforming certain drawbacks of content based and collaborative filtering approach. There are four different ways for obtaining hybrid recommender systems. The approaches are (Adomavicius and Tuzhilin, 2005): *change the enumeration style to roman*

  - Implementing separate RS (collaborative and content-based) with combined prediction.
  - Incorporating some content-based characteristics into a collaborative approach.
  - Incorporating some collaborative characteristics into a content-based approach.

  - Constructing a general unifying model that incorporates both content-based and collaborative characteristics.

  For instance, in a movie recommendation scenario, the hybrid content features are obtained by combining the social feature and content features as described in (Basu et al., 1998). Limitation of this approach lies within the dataset because most of the datasets don't allow to compare the different recommendation features. To be precise, a single dataset rarely contains all the properties like ratings, requirements, attributes of each artifact, domain knowledge or feedback (Jannach et al., 2010). Different types of hybrid recommender systems are described in (Burke, 2002).

- Context-aware

  In many recent applications, the two traditional entities (user and item) are not sufficient enough for generating a recommendation. For instance, while recommending a movie for the user, the system most likely looks for other contextual information such time, date, or the company one wants to be with, for an improved recommendation. Hence, unlike the traditional recommender systems, the contextual information is used to improve the recommendation. This kind of information is dynamic in nature. So, with variable context, the recommendation even varies for the same user. As in (Adomavicius and Tuzhilin, 2015), various ways are discussed for integrating contextual information in recommender systems.

- Ensemble

  Systems using the ensembles techniques come under the category of hybrid recommender systems as the system might combine the rating outputs of both the content-based and the collaborative filtering based recommender system to generate a single output (Aggarwal, 2016a). Contrary to the aforementioned approach, ensembles system combine multiple recommender systems which use similar recommendation approach as well. In this process, multiple opinions are gathered from associated recommender systems before making decisions. This system can generate effective recommendation and increase the performance. As a study of fact, all the ensemble systems come under the umbrella of hybrid systems, but the reverse does not hold true always.

- Social

  With the increased popularity of applications like Facebook, Twitter, Instagram on the web, the use of online social networking sites is rising. This has enhanced the availability of user's social information on the web. The recommender systems employed in social networking sites are exploring the information available on those sites for predicting recommendations. The friendship network is one such social information readily available on the web to be used by such systems. Facebook suggests new friends to a user based on her existing friend list and social interactions on its website. LastFM, an example of collaborative filtering based social recommender system is given in (Konstas et al., 2009) uses available social information about the user to improve the recommendation.

3.6 Similarity Measures in RS

The most important features of a recommender system is to find similarity between some sort of entities (users, items or contents) to provide the recommendation in both the neighborhood-based collaborative filtering and the content-based filtering. Following are some of the popular similarity functions that are used in the RS.

Pearson Correlation Co-efficient:

This is the most popular algorithm for user-based (memory-based) collaborative filtering algorithm (Jannach et al., 2010). This method is used by GroupLens (Resnick et al., 1994) to compute the statistical correlation between the common ratings of two different users to determine the similarity. The similarity corresponds to the cosine of the standard deviation from the average. Pearson correlation coefficient between two random users a and b is defined as:

$$Pearson(a,b) = \frac{\Sigma_{\{p \in P\}}\{(r_{a,p} - \bar{r_a}) \times (r_{b,p} - \bar{r_b})\}}{\sqrt{\Sigma_{\{p \in P\}}(r_{a,p} - \bar{r_a})^2 \Sigma_{\{p \in P\}}(r_{b,p} - \bar{r_b})^2}} \tag{3.2}$$

where $a, b = users$,

$r_{a,p} = ratings\ of\ user\ a\ for\ item\ P$,

$r_{b,p} = ratings\ of\ user\ b\ for\ item\ P$,

$P = set\ of\ items\ rated\ both\ by\ user\ a\ and\ b$,

$\bar{r_a}, \bar{r_b} = user\ a's\ and\ b's\ avaerage\ ratings$

Possible similarity values between -1 and 1

Cosine similarity:

This is also known as vectorial or L2–norm (Jannach et al., 2010). The Cosine similarity is used in both the user-based and item-based collaborative filtering approaches. This produces best results for item-item based filtering. Here, the similarity between the vectors of evaluation for user a and b is measured for a common set of items P (for which a and b has given their ratings)

$$cosine(a,b) = \frac{\Sigma(r_{a,p} \times r_{b,p})}{\sqrt{\Sigma(r_{a,p})^2 \Sigma(r_{b,p})^2}} \tag{3.3}$$

The Co-sine similarity is a much simpler approach than the Pearson correlation coefficient as it does not consider rating average of similar users. It can be used in content based filtering for finding the similarity of items as well.

$$sim(i,j) = \vec{v(i)}.\vec{v(j)} = \frac{\vec{V(i)}.\vec{V(j)}}{\|\vec{V(i)}\|.\|\vec{V(j)}\|} \tag{3.4}$$

$$\text{where } \vec{v(i)} = set\ of\ attributes\ from\ item\ i,$$
$$\text{and } \vec{v(j)} = set\ of\ attributes\ from\ item\ j$$

Jaccard Similarity:

Pearson and cosine similarity measures only consider the common set of attributes between two vectors. Thus two vectors may be completely similar even if they only share one rating on one attribute. For example, one user is interested in fiction novels and the other user is interested in comic novels. There exists a novel which is based on both science and fiction, then the novel is going to be liked by both the users according to their stated preferences. Here, the two different users are treated as completely similar users by the aforesaid similarity measures. Jaccard similarity measure is introduced to overcome the above-said limitations. This considers the difference between two set of items but ignores the difference of ratings associated with the items. So, this measure is suitable for binary and one-class rating (Borges and Lorena, 2010).

$$Jaccard(a,u) = \frac{|\{S_a \cup S_b\}|}{|\{S_a \cup S_b\}|} \tag{3.5}$$

Other measures such as *adjusted cosine* presented by (Jannach et al., 2010), distance based similarity measures such as Euclidean distance (L_2) and Manhattan distance (L_1) (Aggarwal, 2016b) are also used for finding the similarity between users and items.

3.7 Evaluation of RS

Recommender systems adopt several evaluation measures for estimating the performance. There are three primary types of evaluation metrics of RS which is discussed in (Aggarwal, 2016b): *user studies*, *offline* and *online* evaluation. Offline evaluation is mostly used in RS as an active user are not involved during the evaluation process. This kind of evaluation depends on the historical rating data for finding out the quality of recommendation. However, user studies and online evaluation require active user participation.

A common goal of evaluation measure is accuracy which is used to evaluate the performance of the recommendation methods. Because there is a common assumption in RS which states that the most accurate prediction given by a RS is preferred by the users. The typical accuracy measures are the *Mean Absolute Error (MAE)*, *Root Mean Squared Error (RMSE)* and *precision measures* as described in (Ricci et al., 2010). *MAE* and *RMSE* are meant to measure the accuracy of ratings and precision measure is used for finding usage prediction in RS.

Mean Absolute Error (MAE):

This method measures the difference between the original rating and the predicted ratings. In a given test set $\mathcal{T}$ of user u, item i, the true rating is denoted as r_{ui} and the predicted rating is stated as $\hat{r}_{ui}$. The *MAE* between the predicted and actual ratings is given by:

$$MAE = \frac{1}{\mathcal{T}} \sum |\hat{r}_{u,i} - r_{u,i}| \tag{3.6}$$

where $(u, i) \in \mathcal{T}$

This method measures the absolute difference between an actual and predicted ratings. RS predict user ratings more accurately with lower MAE value.

Root Mean Squared Error (RMSE:

This is the most popular measure for finding accuracy in rating predictions. This method considers large errors as the errors are squared before they are summed. The *RMSE* between the predicted and actual ratings is given by:

$$RMSE = \sqrt{\frac{1}{\mathcal{T}} \sum (\hat{r}_{u,i} - r_{u,i})^2} \tag{3.7}$$

Precision Measures:

There are many applications, where RS predict items instead of predicting user preferences. For example, in Netflix movie recommender system, movies (items) are recommended for the users instead of predicting the ratings (preferences). The usages preferences aim at measuring the use of items by the users. Precision and recall are the two best-known metrics used for measuring the usage prediction.

Categories	Selected	Not Selected
Relevant (Rated Good)	True Positive (tp)	False Positive (fp)
Irrelevant (Rated Bad)	False Negative (fn)	True Negative (tn)

Table 3.1: Classification of possible outcomes of a movie recommendation (Jannach et al., 2010)

$$Precision = \frac{|tp|}{|tp + fp|} = \frac{good\ movies\ recommended}{all\ recommendations} \tag{3.8}$$

$$Recall = \frac{|tp|}{|tp + fn|} = \frac{good\ movies\ recommended}{all\ good\ movies} \tag{3.9}$$

where $tp = movies\ recommended\ and\ used$,

$tn = movies\ not\ recommended\ and\ not\ used$,

$fp = movies\ recommended\ but\ not\ used$,

$fn = movies\ not\ recommended\ but\ used$

The F1 metric is used to produce evaluation results that are more universally comparable by combining both the precision and recall values into a single measure.

$$F_1 = 2.\frac{precision.recall}{precision + recall} \tag{3.10}$$

Accuracy is no doubt one of the measure evaluation criteria in RS. But other factors, such as robustness in the RS (in case of CF), computational complexity, scalability, novelty, and reliability should not be overlooked for obtaining good predictions (Borges and Lorena, 2010).

3.8 Information Collected by RS

As discussed and documented in the above sections, privacy is a concern due to two types of basic information: Personal information and Prediction output. Therefore, it is important to look at the different types of information used in the context of RS. Personal information comes in many forms, either explicitly expressed by the user or implicitly collected by RS. Given below is the list of information used in the RS in a recent work as given in (Jeckmans et al., 2013). This information is later used while addressing the fundamental research questions for this thesis, "privacy risks" in the next section.

- Behavioral information is collected implicitly by the respective systems as the user–system interaction log.
- User's purchase history is implicitly collected by the RS to generate future recommendations.
- Contextual information such as location, time, date is collected implicitly by the system. But sometimes, the user explicitly states such information for getting relevant recommendations as well.
- Domain specific knowledge changes over time and gathered by the recommender systems.
- Item meta data gives additional information regarding the content items by which the user query becomes more specific.
- Recommendations are the outputs of a recommender system which is used for identifying an individual in future. These are readily available to the involved recommender systems.
- Social information is readily available for users who are active on social networking sites. Such information is implicitly used by the recommender systems.
- User feedback is given explicitly by the user regarding a received service at any given point of time. These feedbacks come in the form of positive, negative or something specific.
- User preferences are collected in the form of ratings (numerical indicators) or text (tags and comments).

3.9 Privacy Risks in Recommender Systems

This section describes the particular privacy risks associated with recommender systems. In a typical collaborative filtering environment, users often share their preferences willingly over a service or products for receiving useful recommendations. However, users never prefer to disclose their sensitive personal information stored in the recommender systems. Most of the users prefer to remain private if their personal information carries any sensitive information. On the pretext of providing personalized recommendation, large amount information is consolidated and additional information is inferred from already acquired data against the existing OECD guidelines (OECD, 1980).

Given that a number of research activities have already been published concerning the existing privacy risks in the recommender systems, this section aims at consolidating those works in one place which provides us the foundation for identifying the privacy risks in news recommender systems in a dynamic environment. Privacy risks in the recommender systems are classified into two broad categories as in (Friedman et al., 2015). The first category of risk takes place due to the direct exposure of personal information in recommender systems whereas the other kind of risk is involved with the inference of new information from already existing personal data. Information used in the recommender systems are briefly described in subsection 3.8. The below section gives an overview of the inference techniques in the recommender system from the data mining perspective where the data is collected and processed using the advanced techniques.

Privacy risks are identified in many prior works as in (Lam et al., 2006; Jeckmans et al., 2013; Friedman et al., 2015; Ramakrishnan et al., 2001). We have explicitly adopted the classification used in (Jeckmans et al., 2013) and followed the same for identifying the privacy aspects in news domain.

Privacy concerns as classified in (Jeckmans et al., 2013) is given as follows:

- Collection of Data

  Nowadays, data is treated as a commodity and there is no limitation for the data collection. The Internet community is gathering as much user data as possible due to the availability of large storage space. Truly said, once the user has entered some kind of user data in the web, it is stored forever. Users mostly lack knowledge regarding the service providers ability to collect the user data and the further usage. Beyond the stated privacy policy, the users are not even aware of their basic rights regarding the personal data on the web. To be precise, as users are more habituated towards online activities, they have become more vulnerable towards the intended privacy invasion due to the lack of awareness and knowledge. Being in the age of mobiles, the users are using various apps for readily availing services. One such app is known as “Pandora Internet Radio” is able to access the contact list from the mobile devices. This is a perfect example of how the ubiquitous systems are adding to the privacy risks in the recommender systems.This kind of risk exists due to the direct involvement of user data in RS.

- Retention of Data

  Retrieving user data irrespective of the given condition is a case of direct exposure of personal information. The data is collected for a stated duration by the systems for a given service. But, the value of information associated with it may prevent the removal of data by the concerned service provider. Hence, the data might be accessible beyond the intended period of time. Any unauthorized access to the data can lead to the privacy breach. Hence, personal data once available in the web holds privacy risk for a lifetime.

- Sale of Data as a Commodity

  E-commerce is the biggest marketplace of the current society. In such a place, data entails the highest value as it carries the potential to increase or decrease marketing credentials of various online services. Hence, the user data poses a never diminishing value on the web and treated as the commodity. The different aspects of the selling of raw data are given in the below paragraphs.

  The systems dealing with data are scooping up an enormous amount of user data from the web. User data sold to any third party is economically beneficial for the system who owns the data. At the same time, the sold data can hold value for the third party as analysis of such data can boom their business. In both the cases, the user data is used beyond the promised time frame and without being acknowledged by the concerned users (Beckett, 2013). These unethical practices are the main cause of user's privacy concern in the current personalized recommender systems.

  The system holding the user data may share the anonymized version of data for the research community and collaboration projects. For example, AOL released anonymized search queries of its users (Barbaro, 2006), and the Netflix Prize organizers released the anonymized dataset for the competition (Bennett et al., 2007). Many prior research as in (Sweeney, 2002; Narayanan and Shmatikov, 2008) has already addressed the potential risk associated with such publicly published datasets by re-identifying the individuals. This proved the previous myth regarding the robust technologies behind the recommender systems is not enough to keep the user data safe.

  Another threat to privacy arises due to the outsourcing of user data from the parent company to trusted third parties. The third party eventually processes, analyzes and generates the recommendation for the parent system. Although such companies follow the guideline for removing the data after the completion of stated purpose, the copy of an anonymized dataset can still be found.

- Employee's Accessibility for User Information

  Systems dealing with user data has complete control over it and so does the human resources who are in charge of managing these systems. The employees working for the system might access the personal information against the intended guidelines. In the worst case scenarios, user privacy is breached by the unsolicited access of personal information by the trusted employees of the concerned system. This type

of risk involves the direct involvement of user data.

- Recommendations Breaching User Privacy

  Recommendations are the output of the recommender systems. Typically, recommendation inherits a small piece of user information which can be used to reveal personally identifiable information. These recommendations are based on the user preferences which comes in the form of user ratings (collaborative filtering) or the associated context in the context-based recommender systems. RS provide different types of recommendation such as item-to-item, user-to-user, user-to-item or item-to-user recommendation. Privacy risks due to item-to-item and user-to-item recommendation are discussed in (Calandrino et al., 2011). Generally, recommendations based on related items lists are considered to reveal only the relationship between the items. So, this may suggest that no potential privacy risk exists in such a scenario as no user-related information is revealed through recommendation. But opposing this assumption, this paper describes the potential privacy risks associated with the public recommendation (items-lists) generated by the CF approach. For example, $Amazon.com$[1] uses the item-to-item collaborative filtering approach for generating public review regarding the top-N items. One such feature of item similarity on Amazon.com is "Customers who bought the item also like..". In the case of user-to-item recommendation, websites like Amazon generates a personalized recommendation for the signed in users based on both the users and items. The recommendations generated from the above two approaches can be violated by the attackers. The passive form of attack can take place where the attacker has access to the public recommendation list, popularity lists and so on. This kind of inference attack takes a longer period of time as the attacker has to observe the changes in the recommendation lists. For example, a change in the items sales rank can be captured for the inference attacks. The other kind of risk discussed in the paper is due to active attacks where the attacker creates fake user profiles known as *sybil user"*. The paper (Calandrino et al., 2011) also discusses various item related auxiliary information available in the output recommendation to aggregate all the information for the attackers. The attack algorithms discussed here are able to infer the personal (non-public)-information from the publicly displayed outputs of the recommender systems. Hence, breaching the user privacy by any of the system users who are capable of carrying out a passive attack.

  The research work as in (Ramakrishnan et al., 2001) shows the privacy breach due to the presence of a special user who rates products across different types or domains in the system. The presence of such user with *eclectic* tastes enables the RS to generate the serendipitous recommendation. On the other hand, these recommendations can be used for revealing the personal information and to identify individuals in the system. These special users know as 'straddlers' poses the highest risk than the other users. Many such inference attacks are given (Friedman et al., 2015) in details.

- Private Information revealed through Shared Devices

  This kind of risk might be generated due to the shared services or shared devices. In the case of group recommender system, the system recommends information or

items that are relevant to a group of users rather than to an individual. Hence, such recommendation is capable of revealing other user's activities (Masthoff, 2015) to the group of users. A more common scenario is found in individual houses where multiple members of the family might be using the same shared device such as a desktop. As browser cookies are used to identify users (between consecutive sessions) in the computers, the risk of revealing a private purchase of a family member is high unless that person has a personal online account (Friedman et al., 2015). Because of the tracking cookies, different users from the same family are able to see each other's ads (advertisements) while using the same computer and the same web browser. The risk of revealing private information (in the form of any purchase or browsing behavior) is more if the ads are displayed in the shared devices which are accessed by the co-workers from the same organization.

- Risks due to External Entities

  Lawfully, data can be accessed by the Government or Judiciary without the consent of the user or the system. The attacks carried out by third party entities try to access the user data with malicious intentions, resulting in data theft. These external entities are capable of collecting user data in both the raw form and anonymized form. As discussed in (Narayanan and Shmatikov, 2008), the anonymized dataset in the most robust and sparse environment is not free from the re–identification risk. Therefore, it is difficult to guarantee anonymity in common transactions and preference records in the database. In addition, any third party (intruder) with the knowledge of user attributes is capable of re–identifying individuals in the dataset.

Privacy risks associated with user interest is not always linked with user's direct disclosure of sensitive personal information. Privacy-aware users may hide some of the sensitive information from the RS, but the inference capability of the RS collects much more related data about the user. Similar users interest who does not carry a direct connection with the aforesaid user, also helps the RS in the inference process. Hence, a user's privacy preferences are not enough to guarantee his/her own privacy. Therefore, privacy must be ensured in the design of RS. Privacy in RS can be achieved by multiple ways. The following section describes the research work concerning the privacy solutions in RS.

3.10 Privacy Preserving Techniques

Privacy in RS are multifaceted and not limited to a single category. So, there is no single solution present which can achieve complete privacy solution in a RS. This section introduces research works concerning privacy solutions from both technical and nontechnical aspects. A current research (Friedman et al., 2015) has classified the solutions as given below:

- Design of RS Architecture
- Algorithmic Solution
- Laws and Regulation

– User Contribution

In the original literature, user's perception of privacy is discussed separately. However, user contribution is emphasized in this thesis. So, this approach is discussed in a separate category as it carries more importance from the news recommendation perspective. The former two approaches related to RS design and various privacy preserving algorithms intend to provide technical solutions to retain privacy in RS. Whereas the later two corresponds to the nontechnical solutions which contribute to protect privacy in RS. Privacy can be achieved by combining multiple strategies instead of just one due to the limitations

3.10.1 Design of RS Architecture

The design and underlying architecture of the recommender systems can regulate the exposure of user data as well as limit the propagation and linkability of user profile data (Pfitzmann and Hansen, 2009; Friedman et al., 2015).

A larger possibility of privacy breach remains within cross-domain recommendation where the user profile data can easily be accessed across applications. The user profiles can be accessed after the recommendation process gets over. Also, the user profiles can be shared with third parties. In this scenario, user privacy remains at stake. Therefore, various service providers must seek for user's explicit consent before any kind of disclosure of profile data. In addition, the linkability between different user sessions or user profiles must be restricted by the service providers. There should be a limit for the service providers for storing temporary profile data. The aforesaid privacy requirements are implemented through three ways such as reputation, certification and trusted computing (Cissée and Albayrak, 2007). Service providers must comply with these methods for gaining the trust of users. A multi-agent system (MAS) as described in (Cissée and Albayrak, 2007) is developed as a privacy preserving event planner where only trusted third parties can cross-link user profile data.

In a different scenario, (Friedman et al., 2015) privacy risks originating from social networking websites are discussed where the profiles are managed by the users. Different architectures are discussed to protect user data in user-manged profile perspective.

Most of the service providers are based on the client-server architecture where the privacy arises from the server side. To eliminate the server side privacy risk (Friedman et al., 2015), recommendations are generated on the client instead of the server. As an alternate approach to the traditional client-server based architecture peer-to-peer system is proposed by (Lathia et al., 2007; Berkovsky et al., 2006). This system directly interacts with user data for generating recommendations. Hence, the exposure risk still remains with the peer-to-peer system. A hybrid approach is proposed by (Shokri et al., 2009) for preserving privacy in collaborative filtering based systems. In this approach, each client can interact with a centralized server for generating recommendation but at the same time also communicates with other system users. Privacy is maintained in this hybrid system by maintaining two aggregated user profiles. Each user possesses an offline profile which

is stored at the client whereas an online profile is stored at the server. In such distributed architecture, the computational difficulty is considered as a limitation.

Client-side architecture is particularly helpful in addressing user's privacy risk raised due to the centralized server. But, the exposure risk can not be alleviated completely by the client-side architecture as the user data can still be exposed during interaction with other system users and server. The limitation of privacy preserving architecture can be addressed by the cryptographic procedures (Friedman et al., 2015).

3.10.2 Algorithmic Solution

Privacy risks concerned with inference user data can be reduced by the use of the various algorithmic solution. The value of the data decreases with the added uncertainty of the raw user data. This section includes the primary algorithms such as anonymization, perturbation, differential privacy and cryptographic procedures (Friedman et al., 2015; Jeckmans et al., 2013).

- Anonymization

  This method is used to protect individual privacy in large databases by removing the personally identifiable information like name, phone numbers or social security numbers. This technique ensures the safety of data when the data is published (Aggarwal, 2016b). Privacy of user data is assured by removing any possible link between users and data sold (Sweeney, 2002).

  One such privacy model is k-anonymity where anonymization is obtained through *data generalization* and *cell value suppression* processes. The k-anonymity model provides protection against the published dataset from linkage attack and safeguard individuals from re-identification. However, the model fails to secure the data against re-identification risk in case of homogeneity attacks where enough diversity is not found inside the group. Another limitation of k-anonymity model lies with the attackers who possess enough background knowledge (Aggarwal and Yu, 2008). To overcome the *homogeneity attacks* and *background knowledge attacks*, *l diversity* model was proposed which is used to maintain the required diversity among the sensitive attributes of a group of k-tuple.

  In the NetFlix fallout, the user data was anonymized using random numbers which are later re-identified by a group researchers (Narayanan and Shmatikov, 2008). Sparsity and the large volume of data are two limitations of anonymization technique which lead to the NetFlix fallout.

  An alternate privacy model is known as *Agent-based Approach* (Cissée and Albayrak, 2007). This works in the same way as anonymization except that the users must rely on and trust the agent rather than the recommender systems. The users remain anonymous as the agent (either hardware or software) acts as an intermediary between the users and the recommender systems. So, the user can easily hide their personal details and the ratings from RS. Information filtering approach is adopted

to ensure *user privacy*, *provider privacy* and *filter privacy* in this agent-based approach.

Analogous to anonymization algorithms, pseudonyms are also used to hide user's actual identity. Pseudonymity framework in recommender systems are achievable through user anonymization, user data encryption, role-based access, and selective access permission (Friedman et al., 2015; Kobsa and Schreck, 2003).

- Perturbation

  This is also known as data *obfuscation* or data *randomization* technique. This technique ensures the safety of data when the data is collected either in *a perturbed way or in the form of an aggregate* (Aggarwal, 2016b). Perturbation approach adds a degree of uncertainty to the original data. In collaborative filtering approach, the original rating gets replaced by different values before the rating are submitted to a central server (Friedman et al., 2015; Jeckmans et al., 2013). The real data still remains safe from misuse or manipulation, in case the disguised user profiles become accessible to any of the untrusted third parties. This altered data can offer "plausible deniability" to users where they can deny the accuracy of the data if they suspect that the data has been compromised (Walton, 1996). The privacy of the user is enhanced while perturbation techniques are used but users have to rely on the centralized, domain specific server for receiving the recommendation (Polat and Du, 2003).

  An alternate approach to perturbation is aggregation where data from multiple users is aggregated into the profile of a single user. A degree of uncertainty is added to users actual information so that it becomes difficult for the recommender system to identify and link the aggregated and actual user data (Shokri et al., 2009).

  The aforesaid approaches have a negative effect on the accuracy of RS. More user data need to be aggregated and a larger amount of noise need to be added to achieve the desired privacy in RS. Added uncertainty to user data during perturbation decreases accuracy in the RS. In addition, only the service provider adds noise to the user data to preserve accuracy. Hence, privacy against the service provider is not achievable. Whereas aggregation leads to loss of user privacy as actual user data is required while creating aggregates (Jeckmans, 2014). So, perturbation is not the most suitable technique for preserving privacy in the RS.

- Differential Privacy

  This is a reliable trend for preserving privacy in the recommender systems. This method is adopted as a solution for the anonymization techniques where the adversaries with enough background knowledge pose risk for the user. Because, it is difficult to distinguish between a legitimate user and an adversary except in the cryptographical algorithms.

  Differential privacy tries to remove the possible link between a user's input preferences and the recommendation output by making the users computationally indistinguishable in the published dataset (Dwork, 2008, 2006). This is achieved by adding

an adequate amount of noise to input or output of the recommender systems. The amount of noise determines the level of accuracy of output recommendation and privacy of the input user information. The level of noise depends on the sensitivity of the user data (user query data). In the case of statistical queries (for computing the maximum or mean), differential privacy can be implemented by simply adding Laplace noise (Dwork, 2006). This framework was first applied to collaborative filtering setting in the recommender systems (McSherry and Mironov, 2009). In this CF setting, noise is added to the input ratings and then a differentially private item co-variant matrix is computed.

Differential privacy is based on the principle that inference of any personal record in the input user data is not possible from the output recommendation. There are multiple definitions for achieving differential privacy as given in many researches. According to (Dwork, 2006, 2008), differential privacy can be defined as:

Definition: $A\ randomized\ function\ \mathcal{K}\ gives$ ϵ-differential privacy $if\ \ for\ all$ $data\ sets\ \mathcal{D}_1\ and\ \mathcal{D}_2\ differing\ on\ at\ most\ one\ row,\ and\ all\ S \subseteq Range(K),$

$$\Pr\{K(D_1) \in S\} \leq exp(\epsilon) \times \Pr\{K(D_2) \in S\} \tag{3.11}$$

$where\ the\ probability\ space\ in\ each\ case\ is\ over\ the\ coin\ flips\ on\ \mathcal{K},\ and$ $smaller\ \epsilon\ yields\ a\ stronger\ privacy\ guarantee.$

In the above Equation 3.11, the randomized function K is the algorithm applied by the service provider before releasing any user information. The given database D_1 and D_2 consists of a set of rows where each row holds data for an individual. These two databases differ from each other with exactly one extra row and one database is the subset of another database. Therefore, function K satisfying the above-given differential privacy definition is able to prohibit the leakage of user data. For instance, if a user Terry Gross has removed his personal data from one of the databases in fear of privacy leakage, it does not impact her chance of receiving coverage from an Insurance company who has consulted the related databases (Dwork, 2006) before providing coverage to the beneficiary.

Differential privacy offers the privacy guarantee to users. This makes the identification process difficult to differentiate if a record (user data) is contained in a database or not. The major drawback of this method is adding the right level of noise, as too much of noise can have an adverse effect on the output recommendation and less noise fails to hide the contribution of the user (Friedman et al., 2015; Jeckmans et al., 2013).

- Cryptographic Procedures

  Cryptographic procedures are helpful in addressing the privacy risks when the data is exposed to or shared by third parties either by purpose or by force. Different tools are used for providing cryptographic privacy solutions such as secure multi-party computation, secret sharing, homomorphic encryption, and zero-knowledge

proofs (Jeckmans et al., 2013; Goldreich, 2005). However, secure multiparty computation and homomorphic encryption are the most used cryptographic techniques. Therefore, a brief overview of the two techniques is included in the last part of this section.

Cryptographic procedures are used with or without a centralized server for providing a different level of privacy in recommender systems. In a distributed setting, a combination of secure multiparty computation, homomorphic encryption, and zero-knowledge proofs is used for preserving privacy in a collaborative recommender system (Canny, 2002). Here, the trust associated with a service provider is removed as the users collaborate to compute the intermediate values without the presence of a central server. However, this decentralized structure is less preferred by the recommender systems as it does not strengthen the business model. Also, the need of more user involvement for generating recommendation can lessen the accuracy of output in the recommender systems (Jeckmans et al., 2013).

On contrary, cryptographic procedures used in a centralized setting ensures user privacy through secure multiparty computation and homomorphic encryption while utilizing the centralization offered by the service provider. A similar scenario is described in (Aïmeur et al., 2008) where user information is stored separately by two parties: an agent and a service provider. The agent and the service provider has access to user ratings and items respectively. Both the parties are responsible for generating a recommendation, but none of them can link the ratings with the item. The centralized structure is still stored while user privacy is acquired. However, the efficiency of the recommender system decreases in the centralized setting.

Definition: A secure multiparty computation for function f can be stated as a joint protocol between different parties $P_1 ... P_n$ for securely computing $y = f(x_1, ..., x_n)$ where $x_1, ..., x_n$ are the input values of the respective parties. This technique yields the correct output value y for the involved parties even when certain parties are corrupted. In addition, no information breach takes place on the input information of the concerned parties (Tilborg and Jajodia, 2011).

Definition: An encryption technique E is known to be homomorphic when the plaintext space and ciphertext space maintain either additive or multiplicative structure keeping the public key fixed. For example, the public key cryptosystem E is homomorphic when the product of two ciphertexts $E(m_1)$ and $E(m_2)$ results in a ciphertext $E(m_1 + m_2)$, containing the sum of the values m_1 and m_2 (Tilborg and Jajodia, 2011), i.e.

$$E(m_1) * E(m_1) = E(m_1 + m_2) \tag{3.12}$$

In the case of homomorphic encryption, one operation is allowed on the encrypted value followed by another on the ciphertext. A basic function on the encrypted value is calculated without the prior knowledge of the actual data. The result of the function is then obtained by decryption (Jeckmans et al., 2013). There are different homomorphic cryptosystems which are used to enhance user privacy. Examples of

the most used cryptosystems such as RSA, Paillier and ElGamal are included below with definition (Tilborg and Jajodia, 2011).

A multiplicatively homomorphic RSA encryption can be stated as:

$$({m_1}^e \ mod\ n) * ({m_2}^e \ mod\ n) = (m_1 m_2)^e \ mod\ n \tag{3.13}$$

where $the\ ciphertexts\ m_1, m_2 \in {\mathbb{Z}_n}^*$

A multiplicatively homomorphic Paillier encryption can be stated as:

$$(g^{m_1}\ {r_1}^n \ mod\ n^2) * (g^{m_2}\ {r_2}^n \ mod\ n^2) = (g^{m_1+m_2}(r_1 r_2)^n \ mod\ n^2) \tag{3.14}$$

where $the\ ciphertexts\ m_1, m_2 \in \mathbb{Z}_n\ and\ r_1, r_2 \in_R {\mathbb{Z}_n}^*$

A multiplicatively homomorphic ElGamal encryption can be stated as:

$$(g^{r_1},\ y^{r_1} m_1) * (g^{r_2},\ y^{r_2} m_2) = (g^{r_1+r_2},\ y^{r_1+r_2} m_1 m_2) \tag{3.15}$$

where $the\ ciphertexts\ m_1, m_2 \in \langle g \rangle\ and\ r_1, r_2 \in_R \mathbb{Z}_q$

with $q = ord(g)$

Online recommender systems do not find the cryptographic methods feasible (always) due to the need of more computational resources, time, storage and communication overhead (Friedman et al., 2015).

3.10.3 Laws and Regulations

Privacy in the context of personalization is covered in the literature stating the different regulations and guidelines available for protecting consumer privacy (Kobsa, 2007a). Laws and regulations are adopted by many countries to regulate the user privacy and various service providers dealing with user data. As recommender systems are deployed with many online service providers and deals with user data, the same rule and regulations are applicable to the RS as well. For instance, the revision of EU data protection rules in 2017 ensures stricter privacy guidelines for the European consumers (European Commission, 2017b). A more detailed discussion regarding the various privacy rules and regulation within EU and Norway is covered in Subsection 2.6.3. Privacy standards like P3P and privacy policies are described in Chapter 2. The OECD guidelines (OECD, 2013) are established for protecting personal data and the transborder flows of personal data. The EU data protection laws ensure that users are given required freedom to provide their consent before any of the service providers can process respective user data. This provides the user participation rights so that the user can access, modify or delete the data. Thus, the EU data protection law is empowering the users by allowing them control over their own data.

The legal approach is helpful for prevention of any problem raised after the violation of personal information takes place whereas the technical solutions prevent the violation itself.

3.10.4 User Contribution

Most of the privacy solutions are concerned to provide technical solutions. However, privacy in the recommender systems is inseparable from the users as the privacy threat arises from users to users. To be precise, privacy issues are caused by humans (through s/w or h/w) in the form of data theft or attacks. Also, the privacy of users remains in stake due to their own privacy behavior and attitude. Therefore, user contribution is an inherently important solution for dealing with privacy in recommender systems. This part of the thesis addresses some of the user-centered privacy solution.

- Awareness

  Users are not often aware regarding their own rights (European Commission, 2017c). Privacy awareness of a user is based on his/her attention, perception and cognition of the subject privacy (Pötzsch, 2009). A privacy-aware user might be concerned about the below facts:

  – If the user's personal information or activities are exposed by any means

  – Which information is revealed and to whom

  – How the information is going to be used or for how long

  – If the user can take some action regarding his/her own privacy

  Privacy on the web can be successfully addressed by enhancing the user awareness regarding the issue. Awareness can help in user's privacy considerations. One such privacy initiative is the Platform for Privacy Preferences (P3P) (World Wide Web Consortium, 2000). This provides a standardized format for websites where they can define their privacy policies (Jeckmans et al., 2013). The P3P enabled websites to send an alert if a conflict of a privacy interest is found between the user's stated privacy preferences and the stated privacy policies of the website. Thus, P3P helps to enhance user's privacy awareness. A study (Tsai et al., 2011) shows how user's privacy awareness depends on the privacy-related information available with the product in an online shopping experiment. User awareness regarding the prevention of online tracking technologies (opt-in and opt-out options) such as DoNotTrack (Future of Privacy Forum, 2016) can help to retain user privacy (See more in Chapter 2).

- Control

  User control exhibits higher importance in information privacy (Malhotra et al., 2004). It is observed from many studies that users want to have control of their personal information because they take a high risk by sharing their personal information while receiving various services. Users can gain control over their data by gaining access to modify the data, delete the data or approve the data usage. User control is also achieved when they are able to consent before the data is collected from them through opt-in opt-out options. Hence, the user control can aid and manage user privacy in the recommendation context as well.

Chapter 4

News Recommender Systems & Privacy

Recommender systems (RS) are an inherent part of the many advanced websites and mobile applications. RS work towards providing an additional item of interest to users by ranking or filtering the items. Recommendations are generated according to the past user preferences and the current contextual situations of the respective users. Nowadays, RS are not confined to any specific domain, rather distributed over a large space of application domains covering from e-commerce to news.

Mainstream research in the field of privacy in recommender systems are either domain independent or focused on relatively stable domains such as movie, music or e-commerce. News recommendation is a comparatively younger and less researched branch of RS. This poses many unique challenges starting from real-time requirements to the lack of explicit user due to the unique characteristics of the domain. A large amount of research has been devoted to the generation of personalized news experience in real time with utmost accuracy, whereas very less attention has been paid to addressing the privacy concerns in the highly-dynamic news domain.

This chapter explores the various characteristics of the news domain from the context of news recommender systems. In addition, various privacy aspects concerning news recommender systems are addressed. The findings heavily rely on the literature study done in Chapter 3 regarding the privacy risks in the RS in general. An extensive literature review regarding the existing news recommender systems is included from the user privacy perspective in the later section. This chapter concludes with a discussion regarding the possible privacy preserving solution in news recommender systems.

4.1 News as a Recommendation Domain

A recommendation domain is defined as *"the set of items that the recommender will operate over, but may also include the set of aims or purposes that the recommender is intended to support (Burke and Ramezani, 2011)"*. In this context, the news domain explicitly deploys the news recommender systems for identifying interesting stories for its online readers and operates over a set of unlimited news sources.

The news consumption pattern among the readers has evolved over time with the rapid growth of World-Wide Web. The news readers have shifted their focus from the traditional model of news (in terms of subscribing the physical newspaper) to accessing online news sources (Liu et al., 2010). A large amount of news content available on the web causes the information overload problem for the online news readers. The news readers find it difficult to chose relevant news articles from the ample amount of news sources available on the news websites. News recommender systems help these online readers in alleviating their effort in terms of time and choice by providing personalized list of news articles. For achieving personalization in news domain, these tools consider, store and analyze the online reader's past usage patterns. However, the news recommender systems may vary from each other in accordance with the different implementation consideration within the application domain. For instance, the Google News operates as a generic news recommender where as a specialized news recommender system identifies the specific stories which interest the government intelligent analysts (Burke and Ramezani, 2011). The various application domains pose different characteristics as they hold different knowledge sources in the context of recommendation problem. News domain is different from the rest of the application domains due to the unique features it carries. Online news sources are large in number and undergo constant changes. This leads to a large volume of a sparse dataset and dynamic behavior of the news domain. The size of the application domain matters the most in terms of collecting user feedback. For instance, when recommender systems are embedded with a larger application domain (e-commerce sites), they might have to deal with implicit user-system interaction. But, if the recommender systems are integrated into a smaller application domain (basically used as the primary usable source for the service provider), then the recommender system can gather more explicit user feedback. Thus, the size of the application domain matters for the embedded recommender systems. The large volume of the application domain results in infrequent explicit feedback for the recommender systems.

4.1.1 Characteristics of News Domain

In order to build a successful recommender system, one has to gain insight of the domain. The characteristics of a domain have the potential to affect the availability and utility of different knowledge sources (Burke and Ramezani, 2011). News domain, being dynamic in nature undergoes constant changes. Hence, it is not possible for the online news readers to rate or experience each of the news sources or articles. Ratings are considered as an important source of knowledge in recommender systems. But in news recommender systems, ratings are not considered as an important characteristic because they are rarely

available (Ilievski and Roy, 2013; Doychev et al., 2014). The following section gives a precise overview of the characteristics of the news domain.

- *Heterogeneous* nature of information sources present in the news domain makes the recommender systems to satisfy different goals while recommending news. The news domain consists of different items (news articles) spaces. The different news stories (from sports, science, health, technology, etc.,) represent unique characteristics and satisfy different user's preferences. For example, Google News classifies the various heterogeneous news articles into different topic categories such as 'India', 'Business', 'World', etc (Ilievski and Roy, 2013). Although the different news articles come under a common category 'News', the disparate categories like 'Entertainment', 'Science' etc., which are further sub-classified and co-exists under the same item space, i.e. News. In this case, only content knowledge specific to the news articles are not enough as a camera-only site, which recommends only cameras.

- *Unstructured format* of the news stories makes the recommendation process difficult to analyze and might result in an unreliable recommendation. News recommender systems are mostly text-centered as the news domain is rich in text and unstructured in nature. A significant amount of item attributes (subjective content or text description) is available in every news article from which the text attributes are extracted. These text attributes known as the keywords, are later utilized to identify the specific features of the news articles. This feature extraction is used to provide the content-based recommendation in news recommender systems (Aggarwal, 2016b).

- *Large volume* is an important property of the domain as multiple news articles overload the web within a limited time span. This requires more computation for generating news recommendation.

- *Greater item churn* (Das et al., 2007; Ilievski and Roy, 2013) is a key characteristic in the news domain where the items (news stories) enter and leave the system rapidly (Ricci et al., 2010). For example, Google News has a higher item churn than most of the other recommender systems (Das et al., 2007). The underlying item-set on Google News continually goes through churn i.e., insertion and deletion in every few minutes. By undergoing churn, Google News keeps track of the most interesting stories which appeared in the last couple of hours, in any given period. Although memory-based methods are adopted to mitigate the item-churn issue, this method fails to handle when the system deals with many users and items. In this scenario, model-based methods used, but no model older than a few hours can mitigate the item churn as news articles are inserted or deleted at a high frequency in Google News.

- *Named entities/entity preference* is important in news domain as most of the news articles describe the occurrence of a specific event. The description of the events includes the time of the event, the place of the event, the entities involved and the information of the event. News readers might have special preferences for news articles with some named entities. So, preferences for certain entities are important while recommending news to individual readers.

- *Context* can be the time of the day or the day of a week when the user is reading the newspaper. More specifically, one typical news reader will access the news headlines in the morning while looking for some entertainment in the evening. This kind of contextual information is used to alleviate the cold start and data sparsity issues in the existing systems. A detailed study regarding the context aware news recommendation can be found in (Ingvaldsen et al., 2015b). Location of the user can be treated as spatial context and can be used to discover the users changing preferences (Noh et al., 2014).

- *Recency* is a crucial feature in news domain as the news articles have very short life span (Li et al., 2011). Most of the stories are consumed within few hours of their arrival. For instance, a story regarding a rugby match remains popular for the day of publication of that article. After 2 days, the popularity of the same story changes. The same scenario differs in case of a popular movie which might remain popular for weeks, months or even for years. As recency of the news articles depends directly on the time, time is considered as an important characteristic of the news domain. The popularity of the news articles and the user preference both changes over time. The approach adopted in (Wen et al., 2012) has considered time factor along with user interests and preference models to recommend a news item.

- *Filter Bubble* (Pariser, 2012), term coined by Eli Pariser can be well fitted in the context of news domain. Filter bubble is a special characteristic of the various personalization-based service providers like Google News, Yahoo! News, AOL, Facebook and ABC News. This has changed the way users consume information. Filter bubbles are used in the service providers to present the most pleasant and familiar piece of information to the user community. For instance, users risk to get only news articles that match their previous reading behavior due to the presence of this invisible filters. Therefore, users risk to miss important news stories due to a personalized filter.

- *Interaction Style* in the news recommender systems is unique in the way the news readers access the system. Collecting user data is challenging in personalized news recommendation as the recommender must deal with the unavailability of explicit user ratings (Ilievski and Roy, 2013). Explicit feedback data (e.g., ratings and votes) are rare or mostly absent in the news domain as the user unanimously interacts with the system (Doychev et al., 2014). So, collecting implicit feedback data is important for creating the user–item matrix for news recommendation. For example, the click events are considered as a vote for the news and are used to create the binary matrix for predicting the users news preferences. Implicit feedback data (e.g., news click history) which is also known as implicit user ratings are collected for the news articles in the form of clicks (Click Through Rates-CTR). The ratings are considered binary in nature: the read articles are considered as the '1' rating and the unread articles are treated as the '0' rating (Das et al., 2007). However, the unread articles do not always express the unlikeliness of the reader for that news article (e.g. in the case where the article is not noticed by the reader). Also, the time spent on the news stories not necessarily express the preferences of the user for that specific article. Examples of news recommender systems with explicit feedback are given in (Liu

et al., 2010). News Dude, a news recommender system, provides a list of options to the user such as "interesting", "not interesting" while reading the news stories to collect explicit user opinions about the read article.

- *Changing user preferences* is an important characteristic in the news domain. News domain has a dynamic environment. The latest news articles continuously get arriving while the older news article gets outdated. Users preferences changes over time is driven by the changing environment of the news cycle. For example, a reader may prefer the news related to politics during the "US Presidential Election" or sports related news during the "FIFA World Cup" which will change once the events are over. This can be referred as short-term interest. Contrary to this, long-term interest reflects a user's actual steady preference for a specific news section (Liu et al., 2010). Also, preferences for some type of news will increase or decrease as the user naturally tends towards something specific. Stable preferences are rare in the news domain. Bias can be another factor which may influence the user preferences in the news domain.

- *Explanation* is mostly associated with elevated risk applications where the recommendation is explained with a valid reason. The explanation (scrutability) helps the recommendation to get accepted. Explanation for the news articles can help readers to better understand the reason behind the recommendation (Blanco et al., 2012), i.e., why the recommended news articles might be interesting for that reader? Explanation for the recommended news comes as explanatory statements in the news recommender systems. These explanations are generated on the basis of text, entity or usage. These explanations are ranked by using a Markov Logic Networks based on their effectiveness.

- *Novelty* of a news story remains within the information which is unknown and yet interesting for the users (Billsus and Pazzani, 2007). In most of the recommendation domain finding similar items or user (" more of the same") are useful for the selection process. But, novelty of news access differentiates the news recommendation from other domains. To deal with this specific characteristic, the news recommendation techniques try to find similar contents which are previously accessed by the news readers but at the same time not completely identical to the already read news articles. A novelty based news personalization is discussed in (Gabrilovich et al., 2004).

- *The privacy risk* in news domain may be lower as compared to the risk involved in the health domain (medical diagnosis). But when we consider the diversity of topics mentioned in news articles, learning a users preferences on news domain can reveal much more sensitive information than expected. The collection of user data, the management of user profiles and the generation of personalized recommendations raise several privacy issues. User's tolerance for false positive recommendation is determined by the risk factor. The tolerance for false positive is going to be high in the low-risk items (news articles) in the highly voluminous and ephemeral news domain (Ricci et al., 2010).

4.2 News Recommendation

The benefits of recommendation are most salient in the news domain due to the large volume and relatively short lifespan of the news articles. A news recommender system is defined as a tool for filtering out the incoming news and presenting a ranked list of relevant news articles for an individual user based on his past behavior, preferences, context and so on. The problem involved in news recommender systems is to evaluate the news articles which are not yet read by the users or known to users. A ranked list of news articles is presented to users based on this evaluation. The task of news recommendation is stated as a utility function which automatically evaluates the news articles for a user (Gulla et al., 2014). In a given set of M users and N news articles, the utility function (u) evaluates (v) the usefulness of news article n for the user m:

$$u : (M \times N) \rightarrow V \tag{4.1}$$

In the above equation, V is a completely ordered set formed by non-negative values within an interval. For example, the interval can vary from 0 to 10 or from 1 to 100. The elements in the set of users M are defined by the distinct characteristics of the user profiles whereas the elements of the news articles set N is defined by the domain specific characteristics, i.e. characteristics of the news domain such as topic category, editor, author, date and so on. Usually, utility of an item is represented by the given rating but this can also be represented by the function. In the given condition, a news article n' is recommended by the system which maximizes the utility function for a specific user m (Borges and Lorena, 2010).

$$\forall m \in M, n'_m = argmax(u(m, n))n \in N \tag{4.2}$$

The potential problem with this recommendation approach lies with the utility function which is not always defined within the domain of users and news articles. Therefore, the utility function is often *extrapolated* and news articles are presented to the users which might be interesting for them. In this case, the extrapolation of the utility function depends on the evaluation of other similar news articles which are read by the user. The news articles which are not yet evaluated by the users are then rated based on this evaluation. There are different techniques which are used for the estimation process such as Bayesian classifiers, Support Vector Machines (SVM), decision trees, Artificial Neural Networks(ANNs) and clustering (Adomavicius and Tuzhilin, 2005).

4.3 News Recommendation Approach

Chapter 3 presents a detailed discussion regarding the different types of recommender systems based on the three fundamental recommendation techniques, i.e., Collaborative Filtering, Content-Based Filtering and Hybrid Filtering. This section describes exclusively the application of those techniques employed for the recommendation of news with examples.

4.3.1 Collaboartive Filtering Approach

Collaborative filtering approach tries to predict the utility of news article for a particular user based on the news articles rated by other users in the past. So, the utility $u(m, n)$ of article n for user m is calculated on the basis of utilities $u(m_i, n)$ of other "similar users" ($m_i \in M$) as user m for the same article n (Adomavicius and Tuzhilin, 2005). This approach collects user preferences (ratings) and community data for recommending news articles to the target users or a group of users (Das et al., 2007). The same method can be applied to the items by calculating the item similarity and then the new item gets recommended to the specified set of items. Collaborative Filtering approach is further divided into *memory-based (heuristic-based)* and *model-based* (Adomavicius and Tuzhilin, 2005) techniques.

- Memory-based algorithms: Memory-based algorithms utilize all the available dataset for computing similarity (Ricci et al., 2010). The similarity measures can be calculated by using Pearson correlation coefficient, cosine similarity or vector similarity. Both the item and user similarities are taken into account while calculating the weighted average of ratings. In the process to compute prediction and recommendation, the memory-based algorithm keeps on tracing the history dataset all the time which is a drawback of this method. The other disadvantage of such method is due to scalability. Due to which this method alone is not sufficient to be used in the dynamic news domain such as Google News where the item churn is very high. However, the simplicity of this techniques makes these algorithms popular (Das et al., 2007)

- Model-based algorithms: On the contrary, these algorithms are used to derive user models based on users' past ratings which are further used for the recommendation of new (unseen) items (Ricci et al., 2010). Preferences of the users differ from topic to topic while reading the online news. Hence, the different preferences of the users are classified into different clusters or classes as described in (Das et al., 2007). Examples of model-based algorithms include latent semantic indexing (LSI), Bayesian clustering, probabilistic latent semantic indexing (PLSI), Markov Decision Process, Latent Dirichlet Allocation (LDA) and multiple multiplicative Factor Model. As compared to memory-based algorithms, these are more complex and computationally expensive algorithms.

The most popular application of collaborative filtering in news domain is *Google News* which is highly dynamic in nature as the news articles are continuously changing. The presence of millions of users and news articles makes *Google News* very large in volume. *Google News* is a news aggregation system which accumulates the news headlines from more than 4,500 news sources. Generating recommendation is highly challenging in the *content-agnostic Google News* due to (Das et al., 2007) the implicit feedback and time constraint. It provides personalized news services for the signed-in users based on implicit feedback, i.e., user click history and click history of the community (visitor or reader). Every click of the user is considered as a positive vote for that news article which results in more noisy data while user's dislikes (negative votes) are not known to the systems. The response time in case of *Google News* is limited to few hundred milliseconds for

generating a real-time recommendation. For achieving the scalable news recommendation, a combination of memory-based (counting co-visitation) and model-based (*Min Hashing*, PLSI) algorithms are used in *Google News*. The PLSI and MinHash techniques are used for clustering news items, and item co-visitation is used for the recommendation in *Google News*.

NRS	Domain	Relevance Feedback	References
Google News	Online News Aggregator	Implicit (clicks, page visits)	(Das et al., 2007)
GroupLens	Net News Recommender	Explicit and Implicit	(Resnick et al., 1994)

Table 4.1: CF based News Recommender Systems (Borges and Lorena, 2010)

Another well-known application of collaborative filtering on news recommender systems is *GroupLens* which is used for filtering Usenet news (Resnick et al., 1994). Both implicit (time spent on each news article) and explicit feedback (ratings, text comments) are available in the GroupLens system. A client-server based open architecture is proposed for obtaining a compatible, easy to use, scalable and privacy-preserving system. The *GroupLens* client library records the user ratings while the users read a news. The server is responsible for collecting the user ratings and scoring the stories before sending them back to the clients. This system is based on the opinion of other people who have already rated the news articles. The scoring methods (such as, *reinforcement learning, multivariate regression*, and *pairwise correlation coefficients*) used in the system assumes that opinions of a user remain same who has shown a similar opinion in the past on articles in the same newsgroup. The explicit user ratings are used by the *GroupLens* recommendation functions as an input. Then, Pearson correlation coefficient is computed for finding out the relation between two different user's rating behavior. The possibilities of improvement (by considering the time spent on news articles and the content of the articles) over the current evaluation techniques are there for the *GroupLens* system.

4.3.2 Content-Based Filtering Approach

Content-based filtering approach tries to predict the utility of news article for a particular user based on the content of the news articles read in the past. So, the utility $u(m, n)$ of article n for user m is calculated on the basis of utilities $u(m, n_j)$ of user m for the "similar article" n_j where $(n_j \in N)$ (Adomavicius and Tuzhilin, 2005). This approach is based on the assumption that the users read only topic-based interesting articles and the user's interest remains consistent over a period of time. For instance, if a user is interested in sports related news one day, s(he) might read the similar news (sports) the other day as well. In addition, a user's page visit or click history pattern states about the degree of interest regarding the topic or category of the given news articles (Gulla et al., 2014).

The utility function in a content-based news recommender system can be stated as (Adomavicius and Tuzhilin, 2005):

$$u(m,n) = score(ContentBasedUserProfiles(m), Content(n)) \tag{4.3}$$

In the above equation 4.3, ContentBasedUserProfiles(m), Content(n)can be represented as the*Term Frequency* and *Inverse Document Frequency*(TF-IDF) vectors. The TF-IDF weighting is further used to calculate the cosine similarity which is the utility function for the news article content.

Typical examples of heuristic based methods are TF-IDF, clustering whereas the model-based techniques include Bayesian classifiers, Decision Trees, Artificial Neural Networks (ANN) in the content-based filtering approach. Content-based approaches have their benefit over the CF methods, however, a problem remains where the similarity measures identify the news articles from different 'Topic' (Doychev et al., 2014).

The below-given table 4.2 includes the examples of content-based news recommender systems and available feedback. *DailyLearner* is a web-based adaptive news service aims at overcoming the classic assumption of the content-based filtering techniques, i.e., static user interest (user interest does not change over a certain period of time). This system, however, works towards considering the user interests as dynamic and prone to change over time. The *Adaptive Information Server (AIS)* is the core underlying technique of this client-server based *DailyLearner* framework. Two different versions of *DailyLearner* newsagent is discussed in (Billsus and Pazzani, 2000). One of the personalized newsagent provides the news access through a user interface while second newsagent is dedicated for the Wireless Information Devices (e.g., PDA and cell phone). The user profiles are based on the explicit feedback available in the system. The system allows the user to choose the 9 topic category and provide rating or feedback in the form of 'interesting' or "not interesting". The users are also allowed to skip the rating, request for more information on the news articles (*tell me more*) or let the system know about their prior page-visits (*I already know this*). The short and long term interests of the users are addressed by using the different machine learning approaches such as nearest neighbor (NN) algorithm and Naive Bayesian Classifier. The nearest neighbor algorithm is capable of addressing multiple user-interest and quickly adapting the changing user interest. A major advantage of using the NN algorithm in this system is the limited (single) requirement of training data for providing similar user story. The web client involved with *DailyLearner* (web-based version) performed well as compared to the mobile based client (mobile version) due to the associated explicit ratings.

NRS	Domain	Relevance Feedback	References
ACR News	Net News Filtering	Implicit	(Mobasher et al., 2000)
Daily Learner	Net News Recommender	Explicit	(Billsus and Pazzani, 2000)

Table 4.2: CBF based News Recommender Systems (Montaner et al., 2003)

ACR News is a usage-based personalized news service. Different web usage mining techniques, i.e., transaction clustering, usage clustering, and association rule discovery are used to extract usage knowledge regarding the web personalization. These techniques makes the personalization process automatic and dynamic by learning the user preferences from the web usage data which makes the news personalization effective. In case of ACR news, ACR logs are used as web cluster logs and transaction clustering is used to extract the URL clusters. Implicit feedback in terms of users access pattern is used for providing recommendation in the ACR news. A detailed recommendation process regarding the content based filtering in ACR news is given in (Mobasher et al., 2000).

4.3.3 Hybrid Filtering Approach

A RS unifies the aforementioned filtering approaches (i.e., CF and CBF) to get a hybrid system. Despite the strengths of pure CF method, there lie several short-comings like sparsity, early rater (cold-start) and grey-sheep problems which make this approach ineffective if used alone in the recommendation process (Claypool et al., 1999). Similarly, the pure CBF techniques are less effective when the number of items increases in the given space of possible items. So, the hybrid systems aim at aggregating the benefits and alleviating the problems associated with these techniques. There are different ways to combine this two techniques (Adomavicius and Tuzhilin, 2005) which is discussed in the section 3.5. An alternate approach for achieving the hybridization is discussed in (Burke and Ramezani, 2011). There are seven different techniques like *weighted*, *switching*, *mixed* and so on which are used to obtain the benefits of hybridization. The following paragraphs discuss some of the above-mentioned techniques as used with different news recommender systems.

P-Tango (Claypool et al., 1999) is a personalized hybrid news recommender system. It provides a customizable and web-based interface for *Worcester Telegram and Gazette Online* newspaper. This uses a weighted hybrid approach by combining the CF and CBF recommendation scores. The rating of each news article is based on the weighted average of the CF prediction and CBF prediction. The pure CF approach uses *Pearson correlation coefficient* for calculating the similarity measures between the news reader (users), whereas the pure CBF methods utilize the keywords (article keywords and the keywords given in the user profiles) (Borges and Lorena, 2010). Feedback in the form of explicit numerical ratings, explicit keywords or the implicit keywords is used for calculating the similarity measures and the weighted average of the predictions. For evaluating the performance of *P-Tango* system, *mean absolute error (MAE)* is computed between predictions generated by the system and the numerical ratings given by the user for the news articles.

NRS	Domain	Relevance Feedback	References
P-Tango	Net News Filter	Implicit and Explicit	(Claypool et al., 1999)
News Weeder	Net News Recommender	Explicit (Ratings)	(Lang, 1995a)

Table 4.3: Hybrid News Recommender Systems (Borges and Lorena, 2010; Montaner et al., 2003)

NewsWeeder (Lang, 1995a) is a net-news filtering system. This system works towards removing the user dependency for creating and maintaining the user profiles as in most of the information filtering systems. Instead, it encourages its users to rate every read article with a numeric value between 1 to 5. The user profile is created by the system on the basis of user's explicit feedback (rating value) regarding the news stories. The CF approach used in *NewsWeeder* helps the system to learn about the user interest. The CBF approach is used to find the final weighted average between the predictions made by the user and the content-based predictions made by other users. So, *NewsWeeder* is a hybrid system due to the utilization of both the CF and CBF approaches. Machine learning techniques such as term-based *TF-IDF* weighting, cosine similarity and Minimum Description Length (MDL) are used. Precision technique is used for evaluating the ratio of relevant and retrieved news articles from the set of all retrieved news articles.

4.4 News Personalization

The primary goal of news personalization is to provide relevant news specific to individual interests. Different news personalization techniques are discussed in (Billsus and Pazzani, 2007). This section provides brief overview of these techniques.

- Personalizing News Content: This technique assists the user in finding the relevant news articles based on the user model. The user models are either explicitly defined by the user or implicitly created by the system and hold the user interests. The news articles are ranked automatically and then the relevant content is recommended for the specific user. However, content personalization in the news domain is different due to the special news characteristics discussed in Section 4.1.1.

- Personalizing News Context: This system is quite similar to the content-based personalized system except the fact that the news contents are suggested on the basis of the most currently viewed (*just-in-time*) news articles. For instance, a recent story regarding a local football match is recommended to a user after the user receives the e-mail message stating one of the player's name.

- Personalizing News through Navigation: This system helps the news readers to navigate to the most frequently accessed news sections. The news navigation techniques analyze the access patterns of the news reader so that a relevant news section can be selected and appropriately placed within the menu hierarchy of the personalized news system. For instance, a user interested in the 'Sports' related articles would prefer to see the 'Sports' section in the top of the menu hierarchy for any personalized news systems. This kind of approach is suitable for the devices with a small screen such as mobiles and cell phones due to the ease of access.

- Personalizing News through Aggregation: The news aggregator sites such as *Google News*, automatically accumulates and classifies news articles from many different news sources. This helps the user to locate the most recent and popular news topics. News aggregation techniques adapt to the current news landscape and provide the

most emerging news trends. The emergence of Really Simple Syndication (RSS) news feeds helps in aggregating various news sources (Billsus and Pazzani, 2007).

4.5 User Privacy in News Recommender Systems

Personalization and recommendation are inseparable from the issues like privacy and trust irrespective of the domains (Riedl, 2001). News personalization has become crucial on the web as user shows more interest to stay updated with the current news within a limited time span. The quality and accuracy of such personalized news services rely on leveraging user profiles of the news readers. The need and association of user profiles give rise to privacy concerns in the news domain while recommending personalized news articles.

The pattern of news consumption among online news readers has evolved a lot in the last decades. With the emergence of several news aggregator sites, consolidated news is presented to the readers. Often the readers are either suggested or insisted upon to sign in by the system to avail relevant recommendations. But, mostly the readers are not obliged to such requests. Hence, recommender systems cannot have a persistent reader profiles or any identifiers for the future news prediction. In such scenario, the users log data is the only mean of generating the recommendation. Some recommender systems track the browsing pattern of readers by setting cookies on their devices. In this process, the reader's reading history or the list of the visited websites are easily stored by the system for further use. In most of the cases, the users are least aware of the accessed information and their future usage. In a news recommendation scenario, the user fears for the privacy of his/her identity and do not want to disclose his/her page access pattern among others. This "personally identifiable information" and other related information which can be linked together to identify the readers (at a later time) are coined as the primary privacy threats in the news recommender systems.

News context may play a vital role in revealing privacy of users. Users often provide their location details while using the online news if they are interested in local news. It is easier for the service providers to collect this contextual detail through the users mobile devices (through GPS or Wi-Fi) to provide location specific news. In this way, the current location or the neighborhood of the users can be revealed if these user clicks are disclosed.

As no explicit feedback is expected from the readers, the only relevance data available is the clickstream data. The readers can access the news sites from different devices or multiple users can use the same shared device. Although this complicates the ability of recommender systems to track the users browsing patterns, these shared devices can further lead to privacy breach for the readers. For example, in the case of a mobile news recommender, one members browsing pattern or recommended news can cause a privacy breach if accessed by another member of the family (in the same shared device: i.e. mobile). This can be a risk when the members of the family have a different political inclination or news preference (news related to crime, sexual orientation, religious opinion etc.,). The same holds true when the news is accessed from different news sources from different devices and exposed to strangers. The news recommender systems in the context of social

networking can also lead to privacy concerns.

News recommender systems active in the social networking sites such as Facebook, Twitter, Myspace, Google+ and LinkedIn also possess privacy risks for their users. News recommendation on a users profile page, news tagging, like or dislike of a certain category of news, and textual comments can reveal user preferences which may add to the violation of user interest. So, the user is more prone to losing privacy as the users personal history, friend list, and interest are readily available to the recommender systems. Although this type of privacy concern raised due to the social networking environment is not within the scope of this thesis, several privacy aspects lie in the social news domain.

An alternate privacy risk to the user lies within the service provider itself. Most of the news recommender systems clearly state their privacy policy regarding the usage of personal information. But in case, the system expands or shuts down due to some unforeseen circumstances, the future of the personal information gathered by the system remains in doubt. The data can either be sold or used for another purpose without any knowledge of the user. Even though the system claims to sale the anonymized data, there exists the risk of re-identification for the personal data. In addition, the authorized employees of any service provider with recommendation application may cause privacy concerns for the users. The employees supporting and managing various tasks in the recommendation engine might have access to user's personal information. Employees with malicious intention may take advantage of this situation and try to misuse user's personal data. However, this is against the work ethics and the privacy policy provided by the system.

The other kind of privacy risk exists due to the possibility of online data retrieval. Although the system claims to erase the data once the user is no longer registered within the system or due to "forget me" right of the user, the data can still be available from somewhere (e.g., backup) in the system. This kind of user data works as a potential threat to user privacy if available to any malicious user.

4.6 Conclusions

News recommendation is different from rest of the recommender systems due to its unique characteristics which are discussed in Section 4.1.1. Moreover, there are many prevailing challenges in the news domain which include cold-start, data sparsity, recency, scalability, serendipity, unstructured content, and privacy. Privacy is considered to be a major concern for many privacy inclined users. The various privacy preserving techniques addressed in Section 3.10 is relevant for recommender systems in general. This section provides a comprehensive discussion on possible privacy preserving solutions in news recommender systems.

To overcome the shortcoming of one technique, multiple approaches are combined while generating a recommendation for a personalized system. In Google news, the recorded click histories are kept secure by using anonymization techniques (Liu et al., 2010). As discussed in (Desarkar and Shinde, 2014), news recommendation has been generated for

the users without revealing their identities to recommender systems through diversification. In this process, the user must select his preferred publisher and no other user history is considered while building a profile.

Data perturbation techniques can help the users to secure their privacy with the received news recommendation. A similar scenario has been proposed in (Jeckmans et al., 2013) where random perturbation is combined with a peer-to-peer structure. In this dynamic random perturbation scenario, the user can control the data for each request.

Cryptographic procedures are not very suitable for news recommender systems as the computational difficulty can create a delay in generating recommendation and the cost to maintain the framework can be high. Contrary to this concept, the research work as in (Erkin et al., 2013) has proposed cryptographic methods to generate recommendations without revealing any sensitive user data (preference or ratings) in a highly dynamic environment where the number of user keeps on changing. To overcome the computational difficulties, a two-server model has been proposed where one server acts as Service Provider and the other server acts as the Privacy Service Provider (PSP). The feasibility of this system needs to be tested in the news domain to realize a privacy concerned news recommender system.

User control is a useful tool for dealing with privacy in the news domain. Transparency tools and user control are capable of yielding more satisfied users who can control their individual privacy (Hansen, 2008). A recent work (Ingvaldsen et al., 2015a) has considered these two concepts while engineering the mobile news recommender systems where the users are in control of their news stream recommendation via a user interface. Hence, retaining their own privacy while receiving the news service.

As a rule of thumb, awareness of the issue and more clarity (understanding) of how the news recommender systems deal with readers personal data are ideal for dealing with the privacy concern. Primarily, individual news aggregator sites, e.g. Google News, Adresseavisen, and Schibsted, should clearly state about the policies and methodologies they apply to the recommender systems instead of providing some vague description. The way personal data and personalized recommendation data is handled within or outside (trusted or questionable third party) the framework should be clearly stated by the news websites.

Privacy in news recommender systems holds a prominent place for the successful evaluation of such intelligent and adaptive systems. The scope of this chapter remains limited as very few literature address the various privacy concerns related to the news recommender systems directly. The privacy protection techniques can be combined to protect privacy and at the same time to maintain the level of accuracy and efficiency in news recommender systems. A more detailed research can help to build a robust news recommender system which complies with policy, user aspect, and technical perspective while considering privacy.

Chapter 5

User Perspective on Privacy in Recommender Systems

This chapter presents the results of the user survey as well as the subsequent analysis and interpretation of the obtained data. Researching user's opinion concerning privacy in RS is a salient feature of this thesis. The user survey is created on the basis of the literature review and background study performed in the earlier chapters. The outcomes of the survey are documented along with a discussion stating the interpretation and the possible constraints which might have influenced the data collection.

To understand the behavioral approaches and privacy concerns among online users, an online survey is conducted. The outputs of the survey are beneficial for providing user-centered guidelines or solutions in the said problem domain. This also supports the various theories related to user's privacy perspective studied in the earlier chapters and provides some detailed insight towards the news domain.

The online survey is conducted for a duration of 10 days and 52 responses are recorded. The aim of the user experience research is to gain adequate knowledge from a group of people who have the preliminary understanding of the personalized services. Hence, most of the respondents belong to the student and professional networks. But a common diversity designed during the survey is to find the opinion from different age groups and different nationalities.

5.1 Survey Outcomes

The first 3 questions (Q1-Q3) of the survey is designed to receive the basic demographic information of the users such as gender, age group, and nationalities. An overview of the

distribution of respondents who belong to 16 different nations given by gender and age is listed below:

- Gender
  - Male : 34
  - Female : 18
- Age groups
  - 18 - 24 : 7
  - 25 - 34 : 31
  - 35 - 44 : 11
  - 45 - 54 : 3

The overview of the survey results are given below:

Q4. Users availing recommendation service: 73.1% users avail the various recommendation services on a daily basis whereas 100% users avail the service at some point in time.

Q5. Users curious about their personal information collected by various service providers: A major user community lack in interest regarding privacy. Only 40.4% users tried to know about the user profile data.

Q6. User awareness regarding privacy regulations in recommender systems: Majority (53.8%) of users lack in knowledge regarding the problem statement whereas 33.3% users think that recommender systems do not adhere to the existing regulations.

Q7 User concern regarding privacy violation: Majority of user believes that recommender systems violate their privacy through collecting more data than approved and by sharing the data with third parties.

Q8. Online user's perceived knowledge regarding privacy: Most of the users (72% approx.) believes that their personal information is being exploited.

Q9-14. Usability study for recommendation domains: Users have shown less interest in news recommendation as compared to other domains such as movie, music, books, shopping, and tourism. However, movie, music, and books are found to be most preferred by the users. Interestingly, news and tourism are two domains where the users are less interested in receiving any recommendation.

Q15. Online users who think that sites who share personal information with other sites invade privacy: Majority (44.4%) of participants not at all prefer to share their user profiles across domains, although a common profile for multiple domains has its advantages (less time consuming while getting personalized services). Whereas certain users prefer to have a common profile with adequate user control.

Q16-17. Trust and online user's privacy: Trust is seen as important characteristic which can influence users privacy attitude. Users prefer to share their information with a trusted service provider. With trust, more users are ready to share their personal information across domain. User control and user consent regarding the data usage are the two key factors which can build trust for the service providers although the other options are equally relevant. User control increases trust and reduce privacy concerns for user

Q18-22 Behavioral preferences (user perception): Difficult to predict user preferences under different environment. However, a common trend for positive preferences is observed for book recommendation from the user behavior whereas news recommendation is found to be not desired.

Q23-25. Ownership and user control over user data: Ownership of the data can provide more user control over their online data. Ownership of the user data can reduce the privacy concerns of users. Most of the users (77%) believe to gain complete control (modification, deletion and usage control) over their data can provide them with actual ownership.

5.2 Additional Findings

The initial analysis of the survey results demonstrated that user's privacy concerns are a major and it directly impacts the recommender systems. Based on the results, the findings are divided into the following categories which can in return help to provide privacy solutions in the recommender systems. As found from the literature review, user awareness, user control and privacy regulations (nontechnical privacy solutions) contribute to a great extent in retaining privacy in recommender systems. This theoretical concept can be proved from the above survey results. The outcomes of the survey are further analyzed for the Norwegian users against the non-Norwegian users to find out any similarity.

5.2.1 Behavioral Preferences & Privacy

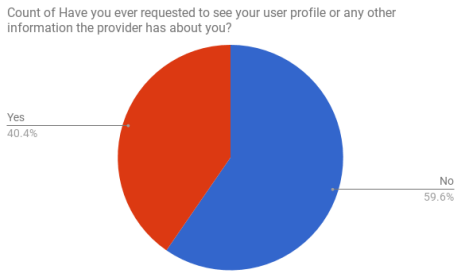
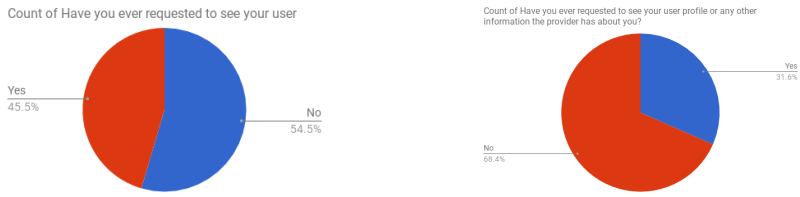


Figure 5.1: Privacy concerned users



(a) Non-Norwegian **(b)** Norwegian

Figure 5.2: Privacy concerned (a) Non-Norwegian (b) Norwegian users

On the topic concerning user's behavioral preferences and privacy, users are found to be more active in using the recommendation service on a daily basis. However, the preferences for using the recommendation service do not influence directly the privacy behavior of every user. Although all the participants have used the recommendation service at some point in time, respondents who have asked for the service providers to view their own user profiles or other information are found to be 40.4% (see the above Figure 5.1). User's perception of recommender systems following laws and regulations are found to be undermined.

To understand user's behavior from the demographic point of view, a further analysis has been done for Norwegian users versus non-Norwegian users. An interesting result has been found though. It was found that the non-Norwegian users are more privacy concerned than the Norwegian users whereas the Norwegian users most frequently use the recommendation services than the non-Norwegian users. Please refer to the above given Figure 5.2 for finding the privacy concern among Norwegian users and non-Norwegian users. However, Norwegian users are less interested in news recommendation as compared to the non-Norwegian users. Both the Norwegian and non-Norwegian users are found to be uncertain if the recommender systems are following the existing privacy laws and regulations.

5.2.2 Trust and Privacy

Trust and privacy have been interlinked in recommender systems. User's trust can be violated in many ways such as exposure, sabotage, and bias. Part of the questionnaire demonstrates the link between user trust and privacy in recommender systems.

For instance, when users are asked if they would prefer to have single user profile instead of having multiple user profiles across different domains starting from movie to news, a majority (46% approx) of users opted out by saying "'Not at all". Exposure of personal data through sharing user profiles is found to be a concern for both Norwegian as well as non-Norwegian community. In a follow-up question, it is found that added trust reduces the privacy concern among users and more users are willing to share their user profile across applications with trusted service provider. When the service provider is trusted, only 30% users refused to share their user profiles. It has been observed from the above

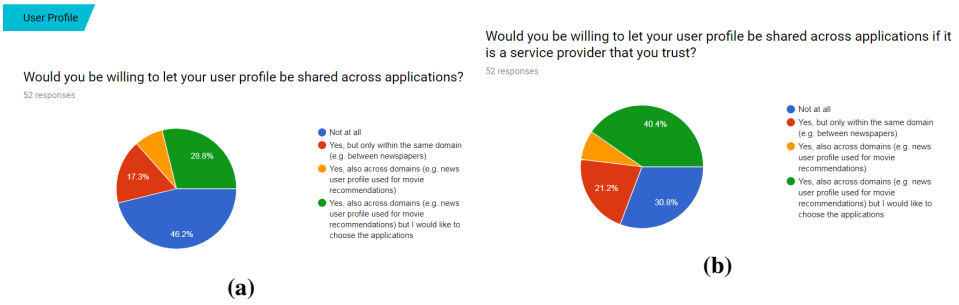


Figure 5.3: Sharing user profiles across applications with (a) any service provider and (b) trusted service provider

trend that added trust with the service provider increased the willingness of 16% users to avail the service by allowing their profiles to be shared across applications.

The link between user trust and exposure risk is clearly visible from the above charts (Figure 5.3). User trust for a service provider can be established by allowing the users to control their personal data, through privacy policies, and by followings privacy guidelines as found from the user opinion. Most of the user expressed their concern regarding service providers seeking the permission before using their data can build user trust for the concerned service provider.

Hence, user trust is found to be a primary factor from the survey results for reducing the privacy concerns of any user. Also, user trust can motivate the user for using the services of a trusted provider and increase the user's willingness to share their personal information (user profiles) with the service providers.

5.2.3 Ownership & Privacy

Ownership or control of user data plays a very crucial role in information privacy. One of the basic privacy requirement for any user is to have a minimal level of user control over their own data.

The survey results convey the concept of user's ownership over their data from the privacy perspective. By ownership of the data, the users are supposed to gain control over their data by being able to modify, access or delete their personal data (stored in the user profile) as and when they wish. Ownership over personal data makes the respondent less concerned regarding privacy in recommender systems, increase the trust for the service provider, and also encourages the disclosure of profile data across applications and system usage. The results concerning the above is shown in the below Figure 5.4. While inquiring about user's opinion regarding ownership, we found out an equal response (50/50 response) where the users would be allowed to modify and delete their data as well as require consenting before the data is shared. The results regarding "what the users actually mean by owning their online data" can be seen in the below Figure 5.5

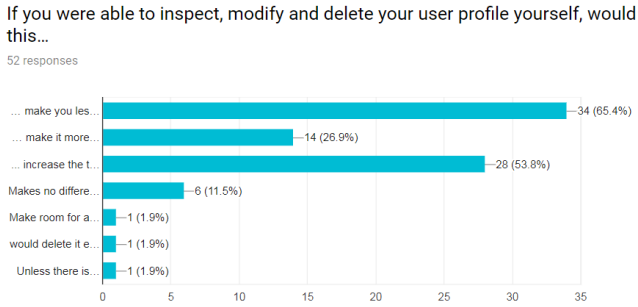


Figure 5.4: Impact of User Control

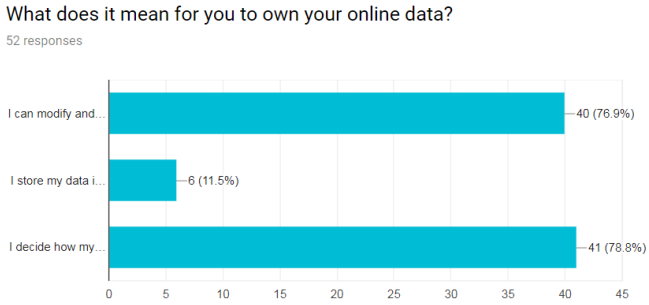


Figure 5.5: User Opinion on Ownership of Data

According to 85 % of respondents, ownership is important for the user community when it comes to online data and the perceived privacy risk.

5.3 Conclusion

Users' detailed comments revealed some explanations for the outcome of the study which is included in detail in Appendix C. Most of the user opinion in the very last open question indicated to the user's information privacy concern from three basic angles, data collection, user control, and awareness. However, one important comment we would want to state is:

" Recommender systems are very useful, but they can also isolate me in a bubble of similar choices, never allow me to see something completely different. This is contradictory: it is limiting my variety of different items, while showing me a variety of similar items from different sources. Trapped in an information bubble means being controlled by the recommender system."

which is referring to the bubble filter concept introduced by Eli Pariser. It can be easily understood how recommender systems or various personalized system are limiting online users freedom or convenience. However, it would be more interesting to study how the bubble filter can affect user privacy in recommender systems. User control can contribute to preserve privacy, but in the above scenario, users felt like being controlled by the recommender system.

The received opinion from users clearly states that how data collection and the control over individual data is undervalued in the user privacy scenario. Users are found to be concerned about the received benefit versus risk while receiving the recommendation services. Another concern reveals that online service provider's profits outweigh user privacy in practice. These user opinions if taken into account can certainly contribute to get a robust recommendation while preserving the privacy of users.

This page is intentionally left blank.

Chapter 6

Conclusion & Future Work

This chapter concludes this thesis with a summary of the research objectives which were achieved during the course of the project. Next, this chapter discusses the limitations of the research work and possible scope for improvements.

6.1 Discussion of Research Questions

This section includes the discussion regarding the stated research questions and the found results.

RQ1: What are the privacy risks in recommender systems?

In order to answer the first research question, a complete literature review has been performed to acquire knowledge regarding existing privacy risks in the recommender systems. The identified privacy risks such as privacy risks associated with recommender systems, privacy threats from the attacker, privacy threats from third parties are included in Section 3.9. Service providers collect a huge amount of personal data for providing recommendation and potentially exposes the data of many users. The consolidated and inferred user data leads to user privacy in the recommender systems.

RQ2: What are the particular characteristic privacy risks in news recommender systems?

In the news recommendation scenario, limited resources are available which deals with user privacy as a whole. Most of the research done in the news recommendation deals with real-time recommendation (with utmost accuracy). The limited resources regarding privacy in the context of news recommendation made the task challenging. The results found for this question is derived from the privacy related literature

available for domain independent recommender systems. The domain specific properties of news recommendation are considered while identifying the privacy risks of news recommender systems and presented in Section 4.5.

RQ3: What are the techniques as a solution to the privacy risks of recommender systems?

The potential privacy risks associated with RS further leads the need to protect the user data. As a prerequisite to investigate the existing privacy-preserving techniques, related works are studied through literature review. The various domain-independent approaches such as technical and non-technical are summarized in Section 3.10.

RQ4: How people think about privacy issues in recommender systems?

Based on the above results, a user-based survey is conducted to find out the explicit user opinion concerning overall privacy which can later be utilized for improving user privacy in recommender systems. The results are then specified into different categories such as privacy preferences, trust, and ownership. The analyzed results are presented in Chapter 5.

6.2 Future Work

Privacy risks included in this thesis is limited whereas the actual user privacy is more diverse. Addressing user privacy in an evolving recommendation domain is found to be challenging. Despite the obvious challenges associated with user privacy, the evaluation of the user data indicates the room of improvement for privacy in recommender systems. It is also observed from the survey that users have their own perception regarding privacy than what the service providers assure to provide. At the same time, the key factors like efficiency and accuracy may create additional conflict with respect to the user privacy. Considering all these facts, the possible scope of improvements for this thesis is listed below:

- In an attempt to identify the potential privacy risks and the user's opinion regarding cross-domain recommendation, it was found that users are concerned regarding sharing of user profiles and requires control over their data in order to avail these services. The cross-domain recommendation includes an inherent threat to privacy as the user profiles are directly linked to multiple applications. So, a further research can include user privacy from cross-domain recommendation perspective,
- An alternate improvement over this thesis can be achieved by implementing some of the suitable privacy preserving techniques such as anonymization or perturbation techniques in the context of news recommendation. This can realize the original problem statement regarding user privacy in news domain and can bring forth the results which are still considered to be an under-researched area.

Bibliography

Adomavicius, G., Tuzhilin, A., June 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17 (6), 734–749.

Adomavicius, G., Tuzhilin, A., 2015. Context-Aware Recommender Systems. Springer US, Boston, MA, pp. 191–226.
URL http://dx.doi.org/10.1007/978-1-4899-7637-6_6

Aggarwal, C. C., 2016a. Ensemble-Based and Hybrid Recommender Systems. Springer International Publishing, Cham, pp. 199–224.
URL http://dx.doi.org/10.1007/978-3-319-29659-3_6

Aggarwal, C. C., 2016b. Recommender Systems: The Textbook, 1st Edition. Springer Publishing Company, Incorporated.

Aggarwal, C. C., Yu, P. S., 2008. Privacy-Preserving Data Mining: A Survey. Springer US, Boston, MA, pp. 431–460.
URL http://dx.doi.org/10.1007/978-0-387-48533-1_18

Aïmeur, E., Brassard, G., Fernandez, J. M., Mani Onana, F. S., 2008. Alambic: a privacy-preserving recommender system for electronic commerce. International Journal of Information Security 7 (5), 307–334.
URL http://dx.doi.org/10.1007/s10207-007-0049-3

allaboutcookies.org, 2016. What is an opt-out cookie? Accessed: 2017-05-30.
URL http://www.allaboutcookies.org/manage-cookies/opt-out-cookies.html

Awad, N. F., Krishnan, M. S., Mar. 2006. The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. MIS Q. 30 (1), 13–28.
URL http://dl.acm.org/citation.cfm?id=2017284.2017287

Barbaro, Michael, Z. J. T., aug 2006. A face is exposed for aol searcher no. 4417749.

Accessed: 2017-06-01.
URL http://www.nytimes.com/2006/08/09/technology/09aol.html

Barbosa, A. D. M., 2015. Privacy-enabled scalable recommender systems. Ph.D. thesis, Universite Nice Sophia Antipolis.

Basu, C., Hirsh, H., Cohen, W., 1998. Recommendation as classification: Using social and content-based information in recommendation. In: Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence. AAAI '98/IAAI '98. American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 714–720.
URL http://dl.acm.org/citation.cfm?id=295240.295795

Beckett, L., 2013. Big data brokers: They know everything about you and sell it to the highest bidder.
URL http://gizmodo.com/5991070/big-data-brokers-they-know-\everything-about-you-and-sell-it-to-the-highest-bidder

Bennett, J., Lanning, S., Netflix, N., 2007. The netflix prize. In: In KDD Cup and Workshop in conjunction with KDD.

Berkovsky, S., Kuflik, T., Eytani, Y., Ricci, F., 2006. Hierarchical neighborhood topology for privacy enhanced collaborative filtering.

BibTex, 2016. Home. Accessed: 2016-12-02.
URL http://www.bibtex.org/

Billsus, D., Pazzani, M. J., Feb. 2000. User modeling for adaptive news access. User Modeling and User-Adapted Interaction 10 (2-3), 147–180.
URL http://dx.doi.org/10.1023/A:1026501525781

Billsus, D., Pazzani, M. J., 2007. Adaptive News Access. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 550–570.
URL http://dx.doi.org/10.1007/978-3-540-72079-9_18

Blanco, R., Ceccarelli, D., Lucchese, C., Perego, R., Silvestri, F., 2012. You should read this! let me explain you why: Explaining news recommendations to users. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. CIKM '12. ACM, New York, NY, USA, pp. 1995–1999.
URL http://doi.acm.org/10.1145/2396761.2398559

Borges, H. L., Lorena, A. C., 2010. A Survey on Recommender Systems for News Data. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 129–151.
URL http://dx.doi.org/10.1007/978-3-642-04584-4_6

Borlund, P., Aug. 2003. The concept of relevance in ir. J. Am. Soc. Inf. Sci. Technol. 54 (10), 913–925.
URL http://dx.doi.org/10.1002/asi.10286

Burke, R., 2000. Knowledge-based recommender systems. In: ENCYCLOPEDIA OF LIBRARY AND INFORMATION SYSTEMS. Marcel Dekker, p. 2000.

Burke, R., Nov. 2002. Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction 12 (4), 331–370.
URL http://dx.doi.org/10.1023/A:1021240730564

Burke, R., Ramezani, M., 2011. Matching Recommendation Technologies and Domains. Springer US, Boston, MA, pp. 367–386.
URL http://dx.doi.org/10.1007/978-0-387-85820-3_11

Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., Shmatikov, V., 2011. "you might also like: " privacy risks of collaborative filtering. In: Proceedings of the 2011 IEEE Symposium on Security and Privacy. SP '11. IEEE Computer Society, Washington, DC, USA, pp. 231–246.
URL http://dx.doi.org/10.1109/SP.2011.40

Canny, J., 2002. Collaborative filtering with privacy. In: Proceedings of the 2002 IEEE Symposium on Security and Privacy. SP '02. IEEE Computer Society, Washington, DC, USA, pp. 45–.
URL http://dl.acm.org/citation.cfm?id=829514.830525

Casassa Mont, M., 2004. Dealing with Privacy Obligations: Important Aspects and Technical Approaches. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 120–131.
URL http://dx.doi.org/10.1007/978-3-540-30079-3_13

Chellappa, R. K., Sin, R. G., 2005. Personalization versus privacy: An empirical examination of the online consumer's dilemma. Information Technology and Management 6 (2), 181–202.
URL http://dx.doi.org/10.1007/s10799-005-5879-y

Cissée, R., Albayrak, S., 2007. An agent-based approach for privacy-preserving recommender systems. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. AAMAS '07. ACM, New York, NY, USA, pp. 182:1–182:8.
URL http://doi.acm.org/10.1145/1329125.1329345

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M., 1999. Combining content-based and collaborative filters in an online newspaper.

Das, A. S., Datar, M., Garg, A., Rajaram, S., 2007. Google news personalization: Scalable online collaborative filtering. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07. ACM, New York, NY, USA, pp. 271–280.
URL http://doi.acm.org/10.1145/1242572.1242610

Datatilsynet, 2015a. Det store datakapplpet. Accessed: 2017-06-05.
URL https://www.datatilsynet.no/globalassets/global/04_planer_rapporter/kommersialiseringsrapport.pdf.

Datatilsynet, 2015b. Safe harbor-beslutningen kjent ugyldig. Accessed: 2017-06-08.

URL https://www.datatilsynet.no/Nyheter/2015/Safe-Harbor-beslutningen-kjent-ugyldig/

Datatilsynet, 2016a. Det store datakapplpet. Accessed: 2017-06-16.
URL https://www.datatilsynet.no/globalassets/global/04_planer_rapporter/kommersialisering-norsk-endelig.pdf.

Datatilsynet, 2016b. Hva er en personopplysning? Accessed: 2017-06-03.
URL https://www.datatilsynet.no/personvern/Personopplysninger/

Datatilsynet, 2017a. Personal-data-act-. Accessed: 2017-06-07.
URL https://www.datatilsynet.no/English/Regulations/Personal-Data-Act-/

Datatilsynet, 2017b. Personal-data-regulations. Accessed: 2017-06-08.
URL https://www.datatilsynet.no/English/Regulations/Personal-Data-Regulations/

Desarkar, M. S., Shinde, N., Oct 2014. Diversification in news recommendation for privacy concerned users. In: 2014 International Conference on Data Science and Advanced Analytics (DSAA). pp. 135–141.

Doychev, D., Lawlor, A., Rafter, R., Smyth, B., 2014. An analysis of recommender algorithms for online news. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 825–836.

Dwork, C., 2006. Differential privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II. ICALP'06. Springer-Verlag, Berlin, Heidelberg, pp. 1–12.
URL http://dx.doi.org/10.1007/11787006_1

Dwork, C., 2008. Differential Privacy: A Survey of Results. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–19.
URL http://dx.doi.org/10.1007/978-3-540-79228-4_1

Erkin, Z., Veugen, T., Lagendijk, R. L., Nov 2013. Privacy-preserving recommender systems in dynamic environments. In: 2013 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 61–66.

Ersdal, H., Skjrstad, S. S., 2016. Privacy and social media: Do users really care? Accessed: 2017-05-30.
URL https://brage.bibsys.no/xmlui/handle/11250/2403229

European Commission, 2016. The eu-u.s. privacy shield. Accessed: 2017-06-07.
URL http://ec.europa.eu/justice/data-protection/international-transfers/eu-us-privacy-shield/index_en.htm

European Commission, 2017a. Digital privacy. Accessed: 2017-06-06.
URL https://ec.europa.eu/digital-single-market/en/online-privacy

European Commission, 2017b. Proposal for an eprivacy regulation. Accessed: 2017-06-07.
URL https://ec.europa.eu/digital-single-market/en/proposal-eprivacy-regulation

European Commission, 2017c. Protecting your data: your rights. Accessed: 2017-06-15.
URL http://ec.europa.eu/justice/data-protection/individuals/rights/index_en.htm

European Commission, 2017d. Reform of eu data protection rules. Accessed: 2017-06-08.
URL http://ec.europa.eu/justice/data-protection/reform/index_en.htm

Friedman, A., Knijnenburg, B. P., Vanhecke, K., Martens, L., Berkovsky, S., 2015. Privacy Aspects of Recommender Systems. Springer US, Boston, MA, pp. 649–688.
URL http://dx.doi.org/10.1007/978-1-4899-7637-6_19

Future of Privacy Forum, 2016. All about do not track. Accessed: 2017-05-30.
URL https://allaboutdnt.com/

Gabrilovich, E., Dumais, S., Horvitz, E., 2004. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In: Proceedings of the 13th International Conference on World Wide Web. WWW '04. ACM, New York, NY, USA, pp. 482–490.
URL http://doi.acm.org/10.1145/988672.988738

Goldberg, D., Nichols, D., Oki, B. M., Terry, D., Dec. 1992. Using collaborative filtering to weave an information tapestry. Commun. ACM 35 (12), 61–70.
URL http://doi.acm.org/10.1145/138859.138867

Goldreich, O., 2005. Foundations of cryptography a primer. Foundations and Trends in Theoretical Computer Science 1 (1), 1–116.
URL http://dx.doi.org/10.1561/0400000001

Gomez-Uribe, C. A., Hunt, N., Dec. 2015. The netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6 (4), 13:1–13:19.
URL http://doi.acm.org/10.1145/2843948

Google Drive, apr 2012. Google drive - cloud storage & file backup for photos, docs & more. Accessed: 2016-12-02.
URL http://www.google.com/drive/

Gulla, J. A., Fidjestl, A. D., Su, X., Castejon, H., 2014. Implicit user profiling in news recommender systems. In: Proceedings of the 10th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,. INSTICC, ScitePress, pp. 185–192.

Hansen, M., 2008. Marrying Transparency Tools with User-Controlled Identity Management. Springer US, Boston, MA, pp. 199–220.
URL http://dx.doi.org/10.1007/978-0-387-79026-8_14

Ilievski, I., Roy, S., 2013. Personalized news recommendation based on implicit feedback. In: Proceedings of the 2013 International News Recommender Systems Workshop and Challenge. NRS '13. ACM, New York, NY, USA, pp. 10–15.
URL http://doi.acm.org/10.1145/2516641.2516644

Ingvaldsen, J. E., Gulla, J. A., Özgöbek, Ö., 2015a. User controlled news recommendations. In: IntRS@RecSys.

Ingvaldsen, J. E., Özgöbek, Ö., Gulla, J. A., 2015b. Context-aware user-driven news recommendation. In: Proceedings of the 3rd International Workshop on News Recommendation and Analytics (INRA 2015) co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 20, 2015. pp. 33–36.
URL http://ceur-ws.org/Vol-1542/paper5.pdf

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G., 2010. Recommender Systems: An Introduction, 1st Edition. Cambridge University Press, New York, NY, USA.

Jeckmans, A., feb 2014. Cryptographically-enhanced privacy for recommender systems. Ph.D. thesis, accessed: 2017-03-25.
URL http://eprints.eemcs.utwente.nl/24409/01/thesis.pdf

Jeckmans, A. J. P., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L., Tang, Q., 2013. Privacy in Recommender Systems. Springer London, London, pp. 263–281.
URL http://dx.doi.org/10.1007/978-1-4471-4555-4_12

Kang, J., 1998. Information privacy in cyberspace transactions. Stanford Law Review 50, 1193–1294.

Kobsa, A., 2007a. The adaptive web. Springer-Verlag, Berlin, Heidelberg, Ch. Privacy-enhanced Web Personalization, pp. 628–670.
URL http://dl.acm.org/citation.cfm?id=1768197.1768222

Kobsa, A., Aug. 2007b. Privacy-enhanced personalization. Commun. ACM 50 (8), 24–33.
URL http://doi.acm.org/10.1145/1278201.1278202

Kobsa, A., Schreck, J., May 2003. Privacy through pseudonymity in user-adaptive systems. ACM Trans. Internet Technol. 3 (2), 149–183.
URL http://doi.acm.org/10.1145/767193.767196

Konstan, J. A., Riedl, J., Apr. 2012. Recommender systems: From algorithms to user experience. User Modeling and User-Adapted Interaction 22 (1-2), 101–123.
URL http://dx.doi.org/10.1007/s11257-011-9112-x

Konstas, I., Stathopoulos, V., Jose, J. M., 2009. On social networks and collaborative recommendation. In: Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09. ACM, New York, NY, USA, pp. 195–202.
URL http://doi.acm.org/10.1145/1571941.1571977

Lam, S. K. T., Frankowski, D., Riedl, J., 2006. Do you trust your recommendations? an exploration of security and privacy issues in recommender systems. In: Proceedings of

the 2006 International Conference on Emerging Trends in Information and Communication Security. ETRICS'06. Springer-Verlag, Berlin, Heidelberg, pp. 14–29.
URL http://dx.doi.org/10.1007/11766155_2

Lang, K., 1995a. Newsweeder: Learning to filter netnews. In: in Proceedings of the 12th International Machine Learning Conference (ML95.

Lang, K., 1995b. Newsweeder: Learning to filter netnews. In: in Proceedings of the 12th International Machine Learning Conference (ML95.

Lathia, N., Hailes, S., Capra, L., 2007. Private distributed collaborative filtering using estimated concordance measures. In: Proceedings of the 2007 ACM Conference on Recommender Systems. RecSys '07. ACM, New York, NY, USA, pp. 1–8.
URL http://doi.acm.org/10.1145/1297231.1297233

Li, L., Wang, D.-D., Zhu, S.-Z., Li, T., Sep. 2011. Personalized news recommendation: A review and an experimental investigation. J. Comput. Sci. Technol. 26 (5), 754–766.
URL http://dx.doi.org/10.1007/s11390-011-0175-2

Liu, J., Dolan, P., Pedersen, E. R., 2010. Personalized news recommendation based on click behavior. In: Proceedings of the 15th International Conference on Intelligent User Interfaces. IUI '10. ACM, New York, NY, USA, pp. 31–40.
URL http://doi.acm.org/10.1145/1719970.1719976

Malhotra, N. K., Kim, S. S., Agarwal, J., 2004. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. Information Systems Research 15 (4), 336–355.
URL http://pubsonline.informs.org/doi/abs/10.1287/isre.1040.0032

Masthoff, J., 2015. Group Recommender Systems: Aggregation, Satisfaction and Group Attributes. Springer US, Boston, MA, pp. 743–776.
URL http://dx.doi.org/10.1007/978-1-4899-7637-6_22

McSherry, F., Mironov, I., 2009. Differentially private recommender systems: Building privacy into the net. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '09. ACM, New York, NY, USA, pp. 627–636.
URL http://doi.acm.org/10.1145/1557019.1557090

Melville, P., Sindhwani, V., 2010. Recommender Systems. Springer US, Boston, MA, pp. 829–838.
URL http://dx.doi.org/10.1007/978-0-387-30164-8_705

Mobasher, B., Cooley, R., Srivastava, J., Aug. 2000. Automatic personalization based on web usage mining. Commun. ACM 43 (8), 142–151.
URL http://doi.acm.org/10.1145/345124.345169

Montaner, M., López, B., de la Rosa, J. L., 2003. A taxonomy of recommender agents on the internet. Artificial Intelligence Review 19 (4), 285–330.
URL http://dx.doi.org/10.1023/A:1022850703159

Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy. SP '08. IEEE Computer Society, Washington, DC, USA, pp. 111–125.
URL http://dx.doi.org/10.1109/SP.2008.33

Navarro-Arribas, G., Torra, V., Erola, A., Castell-Roca, J., 2012. User k-anonymity for privacy preserving data mining of query logs. Information Processing & Management 48 (3), 476 – 487, soft Approaches to IA on the Web.
URL http://www.sciencedirect.com/science/article/pii/S0306457311000057

Noh, Y., Oh, Y.-H., Park, S.-B., Jan 2014. A location-based personalized news recommendation. In: 2014 International Conference on Big Data and Smart Computing (BIGCOMP). pp. 99–104.

Oates, B. J., 2006. Researching Information Systems and Computing. Sage Publications Ltd.

OECD, 1980. Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data. Accessed: 2017-06-01.
URL http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=114&Lang=en&Book=False

OECD, 2013. Oecd guidelines on the protection of privacy and transborder flows of personal data. Accessed: 2017-06-06.
URL http://oe.cd/privacy

Pariser, E., 2012. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin Books, New York, NY, USA.

Pfitzmann, A., Hansen, M., 2009. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management.

Polaris Media, 2009. Informasjon om bruk av cookies og personvernpolicy. Accessed: 2017-12-16.
URL http://www.polarismedia.no/om-polaris-media/datapolicy/

Polat, H., Du, W., Nov 2003. Privacy-preserving collaborative filtering using randomized perturbation techniques. In: Third IEEE International Conference on Data Mining. pp. 625–628.

Polat, H., Du, W., 2005. Svd-based collaborative filtering with privacy. In: Proceedings of the 2005 ACM Symposium on Applied Computing. SAC '05. ACM, New York, NY, USA, pp. 791–795.
URL http://doi.acm.org/10.1145/1066677.1066860

Pötzsch, S., 2009. Privacy Awareness: A Means to Solve the Privacy Paradox? Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 226–236.
URL http://dx.doi.org/10.1007/978-3-642-03315-5_17

Ramakrishnan, N., Keller, B. J., Mirza, B. J., Grama, A. Y., Karypis, G., Nov. 2001. Privacy risks in recommender systems. IEEE Internet Computing 5 (6), 54–62.
URL http://dx.doi.org/10.1109/4236.968832

Rao, A., Schaub, F., Sadeh, N. M., 2014. What do they know about me? contents and concerns of online behavioral profiles. CoRR abs/1506.01675.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. Grouplens: An open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. CSCW '94. ACM, New York, NY, USA, pp. 175–186.
URL http://doi.acm.org/10.1145/192844.192905

Resnick, P., Varian, H. R., Mar. 1997. Recommender systems. Commun. ACM 40 (3), 56–58.
URL http://doi.acm.org/10.1145/245108.245121

Ricci, F., 2002. Travel recommender systems. IEEE Intelligent Systems, 55–57.

Ricci, F., Rokach, L., Shapira, B., Kantor, P. B., 2010. Recommender Systems Handbook, 1st Edition. Springer-Verlag New York, Inc., New York, NY, USA.

Rich, E., 1998. Readings in intelligent user interfaces. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Ch. User Modeling via Stereotypes, pp. 329–342.
URL http://dl.acm.org/citation.cfm?id=286013.286035

Riedl, J., 2001. Personalization and privacy.

Shani, G., Gunawardana, A., 2011. Evaluating Recommendation Systems. Springer US, Boston, MA, pp. 257–297.
URL http://dx.doi.org/10.1007/978-0-387-85820-3_8

ShareLaTex, 2016. Sharelatex,evolved: The easy to use, online, collaborative latex editor. Accessed: 2016-12-02.
URL https://www.sharelatex.com/

Shokri, R., Pedarsani, P., Theodorakopoulos, G., Hubaux, J.-P., 2009. Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. In: Proceedings of the Third ACM Conference on Recommender Systems. RecSys '09. ACM, New York, NY, USA, pp. 157–164.
URL http://doi.acm.org/10.1145/1639714.1639741

Sipior, J. C., Ward, B. T., Mendoza, R. A., 2011. Online privacy concerns associated with cookies, flash cookies, and web beacons. Journal of Internet Commerce 10 (1), 1–16.
URL http://dx.doi.org/10.1080/15332861.2011.558454

Steinke, G., May 2002. Data privacy approaches from us and eu perspectives. Telemat. Inf. 19 (2), 193–200.
URL http://dx.doi.org/10.1016/S0736-5853(01)00013-2

Sweeney, L., oct 2002. K-anonymity: A model for protecting privacy. Int. J. Uncertain.

Fuzziness Knowl.-Based Syst. 10 (5), 557–570.
URL http://dx.doi.org/10.1142/S0218488502001648

Tilborg, H. C. A., Jajodia, S., 2011. Encyclopedia of Cryptography and Security, 2nd Edition. Springer Publishing Company, Incorporated.

Tsai, J. Y., Egelman, S., Cranor, L., Acquisti, A., 2011. The effect of online privacy information on purchasing behavior: An experimental study. Information Systems Research 22 (2), 254–268.
URL http://pubsonline.informs.org/doi/abs/10.1287/isre.1090.0260

Walton, D., 1996. Plausible deniability and evasion burden of proof. argumentation. Argumentation 10, 10–47.

Wen, H., Fang, L., Guan, L., 2012. A hybrid approach for personalized recommendation of news on the web. Expert Systems with Applications 39 (5), 5806 – 5814.
URL http://www.sciencedirect.com/science/article/pii/S0957417411016332

Wikipedia, 2016. Web widget. Accessed: 2017-06-03.
URL https://en.wikipedia.org/wiki/Web_widget

World Wide Web Consortium, 2000. Platform for privacy preferences (p3p) project. Accessed: 2017-06-05.
URL https://www.w3.org/P3P/

Zawadzinski, M., sept 2016. What is device fingerprinting and how does it work? Accessed: 2017-06-03.
URL http://clearcode.cc/2016/09/device-fingerprinting/

Zotero, oct 2006. Home. Accessed: 2016-12-02.
URL https://www.zotero.org/

Appendix A

Paper I

Itishree Mohallick and Özlem Özgöbek: Exploring Privacy Concerns in News Recommneder Systems, to appear in WI'17: International Conference on Web Intelligence 2017.

ACM ISBN 978-1-4503-4951-2/17/08

http://dx.doi.org/10.1145/3106426.3109435

This page is intentionally left blank.

Exploring Privacy Concerns in News Recommender Systems

Itishree Mohallick
Department of Computer and Information Science, NTNU
Norway
itishrem@stud.ntnu.no

Özlem Özgöbek
Department of Computer and Information Science, NTNU
Norway
ozlem.ozgobek@ntnu.no

ABSTRACT

With the increasing ubiquity of access to online news sources, the news recommender systems are becoming widely popular in recent days. However, providing interesting news for each user is a challenging task in highly-dynamic news domain. Many news aggregator sites such as Google News suggest its users to provide sign in to the system for getting user-specific (relevant) news articles. For more generic news recommendation, the system collects user click history and page access pattern implicitly. Often the users are not sure about the usage of the collected and consolidated data by the recommender systems which they usually trade for receiving the news recommendation. Privacy of user identity, user behavior in terms of page access patterns contributes to the overall privacy risks in the news domain. This review paper discusses the current state-of-the-art of privacy risks and existing privacy preserving approaches in the news domain from user perspective.

CCS CONCEPTS

• **Computer systems organization → Recommender Systems**; *Privacy Concerns and solutions*; → News recommender systems

KEYWORDS

recommender systems, news recommender systems, privacy

1 INTRODUCTION

Recommender systems (RS) have become increasingly popular in the last decades since the internet has emerged as an integral part of the common household. Whether it is a book to buy, a piece of online news to read, or a music to listen on the internet; recommender systems can suggest the relevant items to users depending on their interests and needs. These special applications can filter out and evaluate the overwhelming amount of information available on the web, to predict and recommend the desired ones for its users. Typically, these intelligent and adaptive systems serve as a solution for the information overload problem and extends its' service to provide personalized recommendation in multiple domains. For example, Netflix movie recommender system (Movie domain), Amazon.com (E-commerce), Group Lens Recommender System (UseNet news), Google news personalization system (News domain) are few of the pioneers in their respective domains. These systems traditionally rely on the user-system interaction history to build user profiles and represent relevant suggestion based on the user's interest.

Personalized recommendation has become one of the key features for the online content. This requires the acquisition of user data which can be later used and analyzed by the recommender systems for generating a relevant recommendation. For instance, Amazon.com (e-commerce site) recommends items to its users based on their previous purchases. However, a non-personalized recommendation is much easier to generate and often used in e-magazines and e-newspapers. "Top stories" section in Google News is an example of non-personalized recommendation for its readers. Since most of the recommender systems aim at providing a personal recommendation, more weight is given to the research in this area. Most algorithms underlying recommender systems focus on either collaborative filtering (CF), content-based filtering (CB), or hybrid methods (combination of the prior two methods) for generating recommendations [1].

The recommender systems offering news articles to online newspaper readers, based on their predicted news interest are known as news recommender systems. In order to provide interesting news articles of choice and creating "personal newspaper" for each user, the news recommender system requires accurate user profiles which contains current user interest and detailed user activity. Google News, Yahoo! News, and NewsWeeder are few examples of the most popular news recommender systems of current times. In highly-dynamic news domain, the task of recommending news efficiently from a large corpus of newly published news articles is a challenging task. The increasing volume, unstructured news content, continuous growth rate, and the ubiquity of access, makes news recommendation more difficult compared to the recommender systems in other

WI '17, August 23-26, 2017, Leipzig, Germany
 ACM ISBN 978-1-4503-4951-2/17/08…$15.00
http://dx.doi.org/10.1145/3106426.3109435

domains. The process of aggregating news articles from the abundance amount of available news sources according to user interest is known as news personalization [23].

The scope of such 'personalized' services are not limited to any domain or any specific information content. However, 'personalization' requires more detailed information related to the user attributes and preferences. The accuracy of recommendation depends on the detailed user information and serves as the basis for generating the recommendation. On contrary, the same amount of collected and consolidated user data induces threat to the user privacy in the recommender systems [13; 20; 28].

Users share their preference information with the recommender system to avail tailored services of their taste. Any unwanted exposure of user's personal information by either the recommender system or any third party involved gives rise to privacy concerns for that recommender system. Recommender systems using collaborative filtering techniques mostly rely on user feedback for generating recommendation. While providing feedbacks, users mostly give ratings which can reveal their interest, political view, sexual orientation, and so on. Any potential way of leakage to such information can put the user's privacy at stake [13]. With the benefits of the user-centered recommendation, the collected data also creates the privacy risks for the individuals [28]. In the recommendation context, users are assumed to be honest and curious. However, there are some malicious users (adversaries) who may use the system for influencing the recommendation. These users can deliberately query the database or inject fake information to learn sensitive personal information of users. These malicious users pose privacy threat in the recommender systems by inducing different kind of attack [29].

Recommender Systems can identify a specific user's need and help the user to combat information overload issue. Modern RSs deploy various sophisticated recommendation technologies for generating precise and accurate recommendation but at the same time falling out to provide the required privacy to users. In the past, many researches addressed the privacy breaches with the so called robust recommender systems. So, privacy is an important aspect in the recommender system as personalization is hard to achieve without any loss of privacy. This paper discusses privacy concerns, the various technical challenges, and solutions concerning privacy for the news recommender system from a user perspective. The paper is structured as follows: The domain specific characteristics are included in Section 2. A brief review of the privacy aspects of recommender systems are given followed by the privacy characteristics in news recommender systems in Section 3. The state-of-the-art on the adopted privacy preserving methods in recommender systems are given in Section 4. A brief discussion is included in Section 5 to find out if the privacy preserving techniques stated in the previous section are suitable for the news recommender systems. Section 6 concludes this paper.

2 NEWS RECOMMENDATION

A news recommender system can be defined as a tool for filtering out the incoming news and presenting a ranked list of relevant news articles for an individual user. There are examples of news recommender systems used for both the commercial and research purpose. For instance, Google News, The New York Times, Daily learner, Adresseavisen, and News 360 are the commercial ones whereas NTNU Smart Media [15], and PEN recsys [14] represent the research oriented news recommender systems.

News articles pose unique challenges due to the dynamic nature of the news domain, such as recency and popularity which evolves with time very rapidly. These specific features of the news domain differentiate news recommendation from rest of the application domains (e.g., movie, e-commerce, health and business) [37]. Most of the news readers prefer to stay updated by reading the current news. Such users do not always sign in to a system for accessing the online news articles. In that case, the news service provider can not rely on the detailed (explicit) user profiles for generating personalized news access for its readers. So, user profiles are created based on the implicit feedback containing user behavior, rather than explicit ratings as most of the readers do not provide any feedback or rate the news article they read. These kinds of feedbacks are implicitly observed in user activities and are collected from the logs of users click patterns as discussed in detail in the paper [9; 23].

Characteristics of News Domain

In order to build a successful recommender system, one has to gain insight of the domain. The characteristics of a domain have the potential to affect the availability and utility of different knowledge sources. News domain, being dynamic in nature undergoes constant changes. Hence, it is not possible for the online news readers to rate or experience each of the news sources or articles. Ratings are considered as an important source of knowledge in recommender systems. But in news recommender systems, ratings are not considered as an important characteristic because they are rarely available [9; 17]. This section gives a precise overview of the characteristics of the news domain.

Heterogeneous nature of information sources present in the news domain makes the recommender systems to satisfy different goals while recommending news. The news domain consists of different items (news articles) spaces. The different news stories (from sports, science, health, technology, etc.,) represent unique characteristics and satisfy different user's preferences. For example, Google News classifies the various heterogeneous news articles into different topic categories, such as 'India', 'World', 'Business' and so on [23]. Although, the different news articles come under a common category 'News', the disparate categories like 'Entertainment' and 'Science' which are further sub-classified co-exists under the same item space, i.e. News. In this case, only content knowledge specific to the news articles are not enough as a camera-only site, which recommends only camera [28].

Unstructured format of the news stories makes the recommendation process difficult to analyze and might result in unreliable recommendation. News recommender systems are mostly text centric as the news domain is rich in text and unstructured in nature. A significant amount of item attributes (subjective content or text description) are available in every news article from which the text-attributes are extracted. These text-attributes known as the keywords, are later utilized to identify the specific features of the news articles. This feature extraction is used to provide content-based recommendation in news recommender systems [2].

Large volume is an important property of the domain as multiple news articles overload the web within limited time span. This requires more computation for generating news recommendation.

Greater *item churn* [7; 17] is a key characteristic in the news domain where the items (news stories) enter and leave the system rapidly [29]. For example, Google News has a higher item churn than most of the other recommender systems [7]. The underlying item-set on Google News continually goes through churn i.e. insertion and deletion of news articles in every few minutes. By undergoing churn, Google News keeps track of the most interesting stories which appeared in last couple of hours, in any given period. Although, memory-based methods are adopted to mitigate the item-churn issue, this method fails to handle when the system deals with many users and items. In this scenario, model-based methods are used, but no model older than a few hours can mitigate the item churn as news articles are inserted or deleted at a high frequency in Google News.

Named entities/*entity preference* is important in news domain as most of the news articles describe the occurrence of a specific event. The description of the events includes the time of the event, the place of the event, the entities involved and the information of the event. News readers might have special preferences for news articles with some named entities. So, preferences for certain entities are important while recommending news to individual readers.

Context can be the time of the day or the day of a week when the user is reading the newspaper. More specifically, one typical news reader will access the news headlines in the morning while looking for some entertainment in the evening. This kind of contextual information can be used to alleviate the cold start and data sparsity issues in the existing systems [4; 36]. A detailed study regarding the context aware news recommendation can be found in [19]. Location of the user can be treated as spatial context and can be used to discover the user's changing preferences [35].

Recency is a crucial feature in news domain as the news articles have very short life span [22]. Most of the stories are consumed within few hours of their arrival. Such as a story regarding a rugby match might remain popular for the day of publication of that article. After 2 days, the popularity of the same story changes. As recency of the news articles depend directly on the time, time is considered as an important characteristic of the news domain. Popularity of the news articles and the user preference both change over time. The approach adopted in [34] has considered time factor along with user interests and preference models to recommend a news item.

Filter Bubble [26], term coined by Eli Pariser can be well fitted in the context of news domain. Filter bubble is a special characteristic of the various personalization-based service providers like Google News, Yahoo! News, AOL, Facebook and ABC News. This has changed the way users consume information. Filter bubbles are used in the service providers to present the most pleasant and familiar piece of information to the user community. For instance, users risk to get only news articles that match their previous reading behavior due to the presence of this invisible filters. Therefore, users risk to miss important news stories due to a personalized filter.

Interaction Style in the news recommender systems is unique in the way the news readers access the system. Collecting user data is challenging in personalized news recommendation as the recommender must deal with the unavailability of explicit user ratings [17]. Explicit feedback data (e.g., ratings and votes) are infrequent in the news domain as the user unanimously interacts with the system [9]. So, collecting implicit feedback data is important for creating the user-item matrix for news recommendation. For example, the click events are considered as vote for the news and are used to create the binary matrix for predicting the user's news preferences. Implicit feedback data (e.g., news click history) which is also known as implicit user ratings are collected for the news articles in the form of clicks (Click Through Rates-CTR). The ratings are considered binary in nature: the read articles are considered as the '1' rating and the unread articles are treated as the '0' rating [7]. However, the unread articles do not always express the unlikeliness of the reader for that news article (e.g. in the case where the article is not noticed by the reader). Also, the time spent on the news stories not necessarily express the preferences of the user for that specific article. Examples of news recommender systems with explicit feedback is given in [23]. News Dude, a news recommender system, provides a list of options to the user such as 'interesting', "not interesting" while reading the news stories to collect explicit user opinions about the read article.

Changing user preferences is an important characteristic in the news domain. News domain has a dynamic environment. The latest news articles continuously get arriving while the older news article gets outdated. Users preferences change over time driven by the changing environment of the news cycle. For example, a reader may prefer the news related to politics during the "US Presidential Election" or sports related news during the "FIFA World Cup" which will change once the events are over. This can be referred as short-term interest. Contrary to this, long-term interest reflects a user's actual steady preference for a specific news section [23]. Also, preferences for some type of news will increase or decrease as the user naturally tends towards something specific. Stable preferences are rare in the news domain. Bias can be another factor which may influence the user preferences in the news domain.

Explanation is mostly associated with elevated risk applications where the recommendation is explained with a valid reason. The explanation (*scrutability*) helps the recommendation to get accepted. Although news recommendation is treated as a minimal risk domain as compared to Health or Real-estate, explanation of recommendation can still help in improving the reader's experience [31].

The *privacy risks* in news domain may be seen lower as compared to the risks involved in the health domain (medical diagnosis). But when we consider the diversity of topics mentioned in news articles, learning a user's preferences on news domain can reveal much more sensitive information than expected. The collection of user data, the management of user profiles and the generation of personalized recommendations raise several privacy issues. User's tolerance for false positive recommendation is determined by the risk factor. The tolerance for false positive is going to be high in the low-risk items (news articles) in the highly voluminous and ephemeral news domain [29].

3 PRIVACY IN RECOMMENDER SYSTEMS

The privacy breach of the recommender systems takes place due to the large data collection and inference capacity of the recommender systems. This violation is the result of either the "direct access to existing data" or the "inference of user's preferences data (which is completely new data)". Furthermore, some research has identified the risk of re-identification of individuals' and their attributes from the recommendation outputs by the attacker (intruder). Hence, the recommender systems or any of the user (internal and external entities) can be accounted for the risks involved [13]. Primarily privacy risks constitute the data privacy and the recommender systems privacy.

As the identifiable information is the main source of the potential privacy breach, it is worth describing the several types of information used in the agnostic recommender systems. This diverse information can be verified with their presence and influence in other domain such as news recommender systems in the later part of this section. The user-centered information falls into categories like user attributes (demographic information such as name, age, gender, occupation, and relationship status), user preferences (ratings, tags and comments, or favorite item list), behavioral information (implicit), contextual information (location, time stamp, etc.), information about stereotypical users in a specific domain, item metadata (e.g., genre for movies, artist for music, Top stories for news), purchase history (bought items or used contents), user feedback on recommendation (explicit), recommendations, and social link of the user (e.g., friends on Facebook, specific group membership) [13; 20]. Information based on user preferences and user history is most likely to prone for the breach whereas item metadata and domain knowledge are less susceptible to the privacy threats in the generic recommender systems.

Early in the research [21], privacy risks are identified as the amount of personal information collected by the recommender systems and the exposure risk to this information. Apart from the risk of exposure, the privacy concerns are induced by user bias (a group of human users or software agents) in the form of a "shilling attack". This kind of attack is done by creating a special kind of attack profile within the constraints of the recommender systems. The main potential benefit of such attacks is to manipulate the potential buyer's community. Such attacks are carried out by inducing special opinions (positive or negative) about any products or services with some vested interest in mind to bias the output prediction. One such example attack in the form of human user is discussed in the context of book recommendation in Amazon.com. In this case, the author of a book wrote multiple positive reviews about the book and published them online to increase the prediction output. Lack of individual control over the information due to the accumulated authorization of the service providers leads to similar privacy concerns in the RS. The research work as in [28] shows the privacy breach due to presence of special user who rates products across different types or domains in the system. Presence of such user with eclectic tastes enables the recommender systems to generate serendipitous recommendation. On the other hand, they can be used for revealing the personal information and to identify individuals in the system. These special users known as '*straddlers*' possess the highest risk than the other users. As described in [20], the privacy risks of the recommender systems fall into the given categories irrespective of the domain: i.e. *data collection , data sales, data retention, employee browsing private information, recommendation revealing information, shared devices or services,* and *stranger views private information.*

Privacy Characteristics in News Recommendation

Several studies on the privacy aspects of the recommender systems have been published till date which focus extensively on the privacy concerns. However, research on specific privacy issues while recommending news articles is still younger. Therefore, we are trying to investigate the special characteristics related to privacy concerns in news recommender systems and link them to the existing privacy concerns in the recommender systems.

News consumption pattern among the readers has evolved a lot in the last decades. With the emergence of several news aggregator sites, consolidated news is presented to the readers. Often the readers are either suggested or insisted upon to have a login by the system to avail relevant recommendations. But, mostly the readers are not obliged to such requests to sign in and hence the recommender system cannot have a persistent reader profile or any identifiers. In such scenario, the user log data is the only mean of generating the recommendation. Some recommender systems track the browsing pattern of readers by setting cookies on their devices. In this process, the reader's reading history or the list of the visited websites are easily stored by the system for further use. In most of the cases, the users are least aware of the accessed information and the future usage of the data. In a news recommendation scenario, the user fears for the privacy of his/her identity and do not want to disclose his/her page access pattern among others. This "personally identifiable information" and other related information which can be linked

together to identify the readers (in a later time) are coined as the primary privacy threats in the news recommender systems.

News context may play a vital role in revealing privacy of users. Users often provide their location details while using the online news if they are interested in local news. It is easier for the service providers to collect this contextual detail through the user's mobile devices (through GPS or Wi-Fi) to provide location specific news. In this way, the current location or the neighborhood of the users can be revealed if these user clicks are disclosed.

As no explicit feedback is expected from the readers, the only relevance data available is the clickstream data. The readers can access the news sites from different devices or multiple users can use the same shared device. Although this complicates the ability of recommender systems to track the user's browsing patterns, these shared devices can further lead to privacy breach for the readers. For example, in case of a mobile news recommender, one member's browsing pattern or recommended news can cause a privacy breach if accessed by another member of the family (in the same shared device: i.e. mobile). This can be a risk when the members of the family have a different political inclination or news preference (news related to crime, sexual orientation, and religious opinion). The same holds true when the news is accessed from different news sources from different devices and exposed to strangers. The news recommender systems in the context of social networking can lead to such privacy risks.

The news recommender systems active in the social networks such as Facebook, Twitter, Myspace, Google+ and LinkedIn also possess privacy risks for its users. News recommendation on a user's profile page or news tagging can reveal the user's preference which may add to the violation of user interest. So, the user is more prone for privacy as the user's personal history, friend list, interest, etc., is readily available to the recommender systems. Although this type of privacy concern raised due to the social networking sites is not within the scope of this paper, several privacy aspects lies in the social news domain.

Another kind of privacy risk to the user lies within the service provider itself. Most of the news recommender systems clearly state their privacy policies regarding the usage of personal information. But in case the system expands or shuts down due some unforeseen circumstances, the future of the personal information gathered by the system remains in doubt. The data can either be sold or used for other purpose without any knowledge of the user. Even though the system claims to sale the anonymized data, but the re-identification risk cannot be overlooked. In this context of service providers, their lies another concern from the authorized employees of the system. As they have the access to the personal information, employees with malicious intention can take advantage of this situation. However, this is against the work ethics and the privacy policy provided by the system. The other kind of privacy risk lies due to the possibility of online data retrieval. Erasing the old data without affecting the recommendation quality can be a solution to decrease the risk of privacy. But intentionally or unintentionally, it is not always guaranteed that the data is completely deleted [20]. Although the system claims to erase the data once the user is no longer registered within the system or due to "forget me" right of the user, the data can still be available from somewhere (e.g., backup) in the system. This kind of user data works as a potential threat to user privacy if available to any malicious user.

4 PRIVACY PROTECTION TECHNIQUES IN RECOMMENDER SYSTEMS

Recommender systems possess a unique tradeoff between the utility and privacy. Most of the recommendation perform well without considering privacy into account. The utility can be measured in terms of efficiency and accuracy while recommending the services. However, an ideal recommender system can hold on to the utility factor while taking care of data privacy as well. Preserving privacy means to prevent information disclosure caused by legitimate access to the data in the context of recommender systems. Several prior research has been done for preserving user privacy in recommender algorithms [3; 25]. As discussed in [25], k-anonymity model has been developed for protecting privacy in a dataset. Various kinds of privacy preserving algorithm such as perturbation, decision trees, clustering, cryptography based techniques are discussed in data mining [32]. A graph-theoretic model has been discussed in [28] where privacy concerns arise due to the presence of the *straddlers* in the recommender systems. Apart from the technological solutions, data protection laws and guidelines are used to protect the data privacy for users. This section will provide a brief description regarding the state-of-the-art of the privacy preserving techniques in the recommender systems from both the technical and nontechnical perspective. Later a closer look for potential application of these techniques in the news recommender systems is discussed.

Anonymization approach in the dataset helps to replace or remove any identifiable information from the data, while the other structure of the data remains intact. For example, the anonymized Netflix Prize dataset are published and allowed for re-identification where the identities of the users were replaced with random numbers. Two major challenges in this kind of techniques are high sparsity and a large volume of the data. The sparsity of the data can later lead for re-identification of the records in the anonymized dataset [25].

Agent based approach works in the same way as anonymization except that the users must rely and trust the agent rather than the recommender systems. The users can remain anonymous as the agent (either hardware or software) acts as an intermediary between the users and the recommender systems. So, the user can easily hide their personal details and the rating from the recommender systems [5].

Perturbation (obfuscation) approach adds a certain amount of noise to the actual data. In this method, the original ratings get replaced by different values before the ratings are submitted to a central server in collaborative filtering approach [13; 20]. In case the disguised user profiles become accessible to any of the untrusted third parties still, the real data can remain safe from misuse or manipulation. This altered data can offer "plausible

deniability" to users where they can deny the accuracy of the data if they suspect that the data has been compromised [33]. The privacy of the user is enhanced but users have to rely on the centralized, domain specific server for receiving the recommendation [27].

Aggregation approach takes place when the different user information is aggregated without any direct interaction of the recommender systems. A degree of uncertainty is added to user's actual information so that it becomes difficult for the recommender system to identify and link the aggregated and actual user data [30].

Differential privacy is a reliable trend for preserving privacy in the recommender systems. In this process, the users in the dataset are kept computationally indistinguishable from the users in the already released dataset. This is achieved by adding adequate amount of noise to input or output of the recommender systems. The amount of noise determines the level of accuracy of output recommendation and privacy of the input user information. Differential privacy framework was first applied with collaborative filtering techniques in the recommender systems [24]. Here noise is added to the input ratings and then a differentially private item covariant matrix is computed. The major drawback of this method is adding right level of noise, as too much of noise can have an adverse effect on the output recommendation and less noise fails to hide the contribution of the user [13].

Cryptographic procedures are helpful in addressing the privacy risks when the data is exposed or shared by third parties either by purpose or by force. Secure multiparty computation works well for offline recommendation. Whereas in homomorphic encryption (multiplicative or additive), one operation is allowed on the encrypted value followed by another on the ciphertexts. A basic function on the encrypted value is calculated without the prior knowledge of the actual data. The result of the function is then obtained by decryption [13; 20]. This technique can work with or without a centralized server. However, the later structure (decentralized) is less preferred by the recommender systems as it does not strengthen the business model. Also, the need of more user involvement for generating recommendation can lessen the accuracy of output in the recommender systems.

Laws and regulations are adopted by many countries to regulate the user privacy and the industries' function. For instance, the revision of EU data protection rules "Regulation (EU) 2016/679" and "Directive (EU) 2016/680" ensures a more stricter privacy guideline for the European consumers across Europe and outside as well [12]. The Regulation and Directive are adopted by the European Commission and European Parliament in April, 2016 and going to be effective from May, 2018. These rules will help the users to gain more control over their personal data. These regulations will address the consumer's privacy concerns through a set of new guidelines such as "right to be forgotten", "easier access to one's data", "the right to know when user's data is hacked" and so on. This reform of EU data protection rules will ensure the safety of the personal data inside EU and wherever the EU user's data will be accessed or processed. This will have a greater impact on the companies like Google and Facebook who are processing EU data outside of EU. Another example of a privacy law is EU-U.S. Privacy Shield which is adopted on July, 2016 for safeguarding EU data from being transferred to U.S. [11]. This regulation aims at bringing more clarity on transborder data flows by implying "strong data protection obligations on companies receiving personal data from EU". The legal approach is helpful for prevention of any problem raised after the violation of personal information takes place whereas the technical solutions prevent the violation itself.

Awareness and user control can aid and enhance the user's knowledge regarding their privacy. Users are given tools for managing their privacy, enabling them to easily realize the conditions and policies of their information usage. For instance, the W3C Platform for Privacy Preferences [6] recommendation allows recommender systems to inform their users about the privacy policies implemented on their data use. These practices allow users to define their privacy preferences, enabling them to restrict the use of their information and hide or obfuscate the information registered about them.

5 DISCUSSION

News recommendation is different from rest of the recommender systems due to its unique characteristics which are discussed in Section 2. There are many challenges associated with news domain such as cold start, data sparsity, recency, scalability, serendipity and unstructured content. Privacy is still considered as a major concern in the news recommendation context. The various privacy preserving techniques addressed in the previous section is relevant for recommender systems in general. This section presents a brief discussion concerning the above-given privacy solutions and their possible application on news recommendation.

In order to overcome the limitations of one privacy preserving technique, multiple approaches can be combined while generating recommendation for a personalized system. In Google news, the recorded click histories are kept secure by using anonymization techniques [23]. As discussed in [8], news recommendation has been generated for the users without revealing their identities to the recommender systems through diversification. In this process, the user must select his preferred publisher and no other user history is considered while building profile.

Data perturbation techniques can help the users to secure their privacy with the received news recommendation. A similar scenario has been proposed in [20] where random perturbation is combined with a peer-to-peer structure. In this dynamic random perturbation scenario, the user can control the data for each request.

Cryptographic procedures are not very suitable for news recommender systems as the computational difficulty can create delay in generating recommendation and the cost to maintain the framework can be high. Contrary to this concept, the research work done in [10] has proposed a cryptographic protocol to generate recommendations without revealing any sensitive user data (preference or ratings) in a highly dynamic environment where the number of user keeps on changing. To overcome the

computational difficulties, a two-server model has been proposed where one server acts as Service Provider and the other server acts as the Privacy Service Provider (PSP). The feasibility of this system needs to be tested in the news domain to realize a complete news recommender system.

User control can act as a useful tool for dealing with privacy in the news domain. Transparency tools and user control can yield a more satisfied user who can control their individual privacy [16]. A recent work [18] has considered these two concepts while engineering the mobile news recommender systems where the users can control their news stream recommendation via a user interface. Hence, retaining their own privacy while receiving the news service.

As a rule of thumb, awareness of the issue and more clarity (understanding) of how the news recommender systems deal with reader's personal data are ideal to deal with the privacy concern. Primarily, individual news aggregator sites, e.g. Google News, Adresseavisen and Schibsted, should clearly state about the policies and methodologies they apply with the recommender system instead of providing some vague description. The way personal data and the output (recommendation) is handled within or outside (trusted or not-trusted third party) the framework should be clearly stated by the news web sites.

6 CONCLUSIONS

Privacy in recommender systems holds a prominent place for the successful evaluation of such intelligent and adaptive systems. In this paper, we have presented the state-of-the-art of privacy concerns and available solutions in news recommender systems while discussing the special characteristics and privacy features of the news domain. The scope of this paper remains limited as very few literatures address the various privacy concerns related to the news recommender systems directly. The privacy protection techniques can be combined to protect privacy and at the same time to maintain the level of accuracy and efficiency in news recommender systems. A more detailed research can help to build a robust news recommender system which complies with policy, user aspect, and technical perspective while considering privacy.

ACKNOWLEDGMENTS

This work is a part of the master thesis which is supported by the NTNU SmartMedia program on news recommendation.

REFERENCES

[1] ADOMAVICIUS, G. and TUZHILIN, A., 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng. 17*, 6, 734-749. DOI= http://dx.doi.org/10.1109/tkde.2005.99.

[2] AGGARWAL, C.C., 2016. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated.

[3] BERKOVSKY, S., EYTANI, Y., KUFLIK, T., and RICCI, F., 2005. Privacy-enhanced collaborative filtering. In *Proceedings of User Modeling Workshop on Privacy-Enhanced Personalization*, 75-83.

[4] BRAUNHOFER, M., CODINA, V., and RICCI, F., 2014. Switching hybrid for cold-starting context-aware recommender systems. In *Proceedings of the Proceedings of the 8th ACM Conference on Recommender systems* (Foster City, Silicon Valley, California, USA2014),ACM,2645757,349-352. DOI= http://dx.doi.org/10.1145/2645710.2645757.

[5] CISS, R. and ALBAYRAK, S., 2007. An agent-based approach for privacy-preserving recommender systems. In *Proceedings of the Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems* (Honolulu, Hawaii2007), ACM, 1329345, 1-8. DOI= http://dx.doi.org/10.1145/1329125.1329345.

[6] CRANOR, L., LANGHEINRICH, M., MARCHIORI, M., MARTIN PRESLER-MARSHALL, and REAGLE, J., 2002. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification *2017*. https://www.w3.org/TR/P3P/.

[7] DAS, A.S., DATAR, M., GARG, A., and RAJARAM, S., 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the Proceedings of the 16th international conference on World Wide Web* (Banff, Alberta, Canada2007), ACM, 1242610, 271-280. DOI= http://dx.doi.org/10.1145/1242572.1242610.

[8] DESARKAR, M.S. and SHINDE, N., 2014. Diversification in news recommendation for privacy concerned users. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, 135-141. DOI= http://dx.doi.org/10.1109/DSAA.2014.7058064.

[9] DOYCHEV, D., LAWLOR, A., and RAFTER, R., 2014. *An Analysis of Recommender Algorithms for Online News*. CLEF.

[10] ERKIN, Z., VEUGEN, T., and LAGENDIJK, R.L., 2013. Privacy-preserving recommender systems in dynamic environments. In *2013 IEEE International Workshop on Information Forensics and Security (WIFS)*, 61-66. DOI= http://dx.doi.org/10.1109/WIFS.2013.6707795.

[11] EUROPEAN COMMISSION, 2016. The EU-U.S. Privacy Shield *2017*,, June 23, . http://ec.europa.eu/justice/data-protection/international-transfers/eu-us-privacy-shield/index_en.htm.

[12] EUROPEAN COMMISSION, 2016. Reform of EU data protection rules *2017*, June 23. http://ec.europa.eu/justice/data-protection/reform/index_en.htm.

[13] FRIEDMAN, A., KNIJNENBURG, B.P., VANHECKE, K., MARTENS, L., and BERKOVSKY, S., 2015. Privacy Aspects of Recommender Systems. In *Recommender Systems Handbook*, F. RICCI, L. ROKACH and B. SHAPIRA Eds. Springer US, Boston, MA, 649-688. DOI= http://dx.doi.org/10.1007/978-1-4899-7637-6_19.

[14] GARCIN, F. and FALTINGS, B., 2013. PEN recsys: a personalized news recommender systems framework. In *Proceedings of the Proceedings of the 2013 International News Recommender Systems Workshop and Challenge* (Kowloon, Hong Kong2013), ACM, 2516642, 3-9. DOI= http://dx.doi.org/10.1145/2516641.2516642.

[15] GULLA, J.A., FIDJESTØL, A.D., SU, X., and MARTÍNEZ, H.N.C., 2014. Implicit User Profiling in News Recommender Systems. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies* DOI= http://dx.doi.org/10.5220/0004860801850192.

[16] HANSEN, M., 2008. Marrying Transparency Tools with User-Controlled Identity Management. In *The Future of Identity in the Information Society: Proceedings of the Third IFIP WG 9.2, 9.6/11.6, 11.7/FIDIS International Summer School on The Future of Identity in the Information Society, Karlstad University, Sweden, August 4–10, 2007*, S. FISCHER-HÜBNER, P. DUQUENOY, A. ZUCCATO and L. MARTUCCI Eds. Springer US, Boston, MA, 199-220. DOI= http://dx.doi.org/10.1007/978-0-387-79026-8_14.

[17] ILIEVSKI, I. and ROY, S., 2013. Personalized news recommendation based on implicit feedback. In *Proceedings of the Proceedings of the 2013 International News Recommender Systems Workshop and Challenge* (Kowloon, Hong Kong2013), ACM, 2516644, 10-15. DOI= http://dx.doi.org/10.1145/2516641.2516644.

[18] INGVALDSEN, J.E., GULLA, J.A., and ÖZGÖBEK, Ö., 2015. User Controlled News Recommendations. In *IntRS@RecSys*, 45-48.

[19] INGVALDSEN, J.E., ÖZGÖBEK, Ö., and GULLA, J.A., 2015. Context-Aware User-Driven News Recommendation. In *INRA@RecSys*.

[20] JECKMANS, A.J.P., BEYE, M., ERKIN, Z., HARTEL, P., LAGENDIJK, R.L., and TANG, Q., 2013. Privacy in Recommender Systems. In *Social Media Retrieval*, N. RAMZAN, R. VAN ZWOL, J.-S. LEE, K. CLÜVER and X.-S. HUA Eds. Springer London, London, 263-281. DOI= http://dx.doi.org/10.1007/978-1-4471-4555-4_12.

[21] LAM, S.K.T., FRANKOWSKI, D., and RIEDL, J., 2006. Do You Trust Your Recommendations? An Exploration of Security and Privacy Issues in Recommender Systems. In *Emerging Trends in Information and Communication Security: International Conference, ETRICS 2006, Freiburg, Germany, June 6-9, 2006. Proceedings*, G. MÜLLER Ed. Springer Berlin Heidelberg, Berlin, Heidelberg, 14-29. DOI= http://dx.doi.org/10.1007/11766155_2.

[22] LI, L., WANG, D.-D., ZHU, S.-Z., and LI, T., 2011. Personalized News Recommendation: A Review and an Experimental Investigation. *Journal of Computer Science and Technology 26*, 5, 754-766.

DOI= http://dx.doi.org/10.1007/s11390-011-0175-2.

[23] LIU, J., DOLAN, P., and PEDERSEN, E.R., 2010. Personalized news recommendation based on click behavior, 31. DOI= http://dx.doi.org/10.1145/1719970.1719976.

[24] MCSHERRY, F. and MIRONOV, I., 2009. Differentially private recommender systems: building privacy into the net. In *Proceedings of the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (Paris, France2009), ACM, 1557090, 627-636. DOI= http://dx.doi.org/10.1145/1557019.1557090.

[25] NARAYANAN, A. and SHMATIKOV, V., 2008. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the Proceedings of the 2008 IEEE Symposium on Security and Privacy* (2008), IEEE Computer Society, 1398064, 111-125. DOI= http://dx.doi.org/10.1109/sp.2008.33.

[26] PARISER, E., 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.

[27] POLAT, H. and WENLIANG, D., 2003. Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Third IEEE International Conference on Data Mining*, 625-628. DOI= http://dx.doi.org/10.1109/ICDM.2003.1250993.

[28] RAMAKRISHNAN, N., KELLER, B.J., MIRZA, B.J., GRAMA, A.Y., and KARYPIS, G., 2001. Privacy Risks in Recommender Systems. *IEEE Internet Computing 5*, 6, 54-62. DOI= http://dx.doi.org/10.1109/4236.968832.

[29] RICCI, F., ROKACH, L., SHAPIRA, B., and KANTOR, P.B., 2010. *Recommender Systems Handbook*. Springer-Verlag New York, Inc.

[30] SHOKRI, R., PEDARSANI, P., THEODORAKOPOULOS, G., and HUBAUX, J.-P., 2009. Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. In *Proceedings of the Proceedings of the third ACM conference on Recommender systems* (New York, New York, USA2009), ACM, 1639741, 157-164. DOI= http://dx.doi.org/10.1145/1639714.1639741.

[31] TINTAREV, N., 2007. Explanations of recommendations. In *Proceedings of the Proceedings of the 2007 ACM conference on Recommender systems* (Minneapolis, MN, USA2007), ACM, 1297275, 203-206. DOI= http://dx.doi.org/10.1145/1297231.1297275.

[32] VERYKIOS, V.S., BERTINO, E., FOVIN, I.N., PROVENZA, L.P., SAYGIN, Y., and THEODORIDIS, Y., 2004. State-of-the-art in privacy preserving data mining. *Sigmod Record 33*, 1 (Mar), 50-57. DOI= http://dx.doi.org/10.1145/974121.974131.

[33] WALTON, D., 1996. Plausible deniability and evasion burden of proof. Argumentation. *Argumentation 10*, 10--47.

[34] WEN, H., FANG, L., and GUAN, L., 2012. A hybrid approach for personalized recommendation of news on the Web. *Expert Systems with Applications 39*, 5 (4//), 5806-5814. DOI= http://dx.doi.org/https://doi.org/10.1016/j.eswa.2011.11.087.

[35] YUNSEOK, N., YONG-HWAN, O., and SEONG-BAE, P., 2014. A location-based personalized news recommendation. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, 99-104. DOI= http://dx.doi.org/10.1109/BIGCOMP.2014.6741416.

[36] ZHU, X. and HAO, R., 2016. Context-aware location recommendations with tensor factorization. In *IEEE/CIC International Conference on Communications in China (ICCC)* IEEE, Chengdu, China 1-6. DOI= http://dx.doi.org/10.1109/ICCChina.2016.7636832.

[37] ÖZGÖBEK, Ö., GULLA, J.A., and ERDUR, R.C., 2014. A Survey on Challenges and Methods in News Recommendation. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies :WEBIST*, 278-285. DOI= http://dx.doi.org/10.5220/0004844202780285.

Appendix B

Survey Questionnaire

This page is intentionally left blank.

Privacy Concerns in Recommender Systems

Recommender systems aim to recommend the most suitable items to the users, based on their personal preferences. When we think about the huge number of items available online, recommender systems are one of the best ways to find suitable/interesting items for us. Recommender systems take place in most of the online services. During an online shopping, movie watching, music listening or news reading experience, we come across to recommended items. Recommender systems work based on the user information. When the user purchase, browse or read an item, the recommender system learns from these actions and starts to build a user profile in order to generate better personalized recommendations. Recommender systems are a way to provide personalized services.

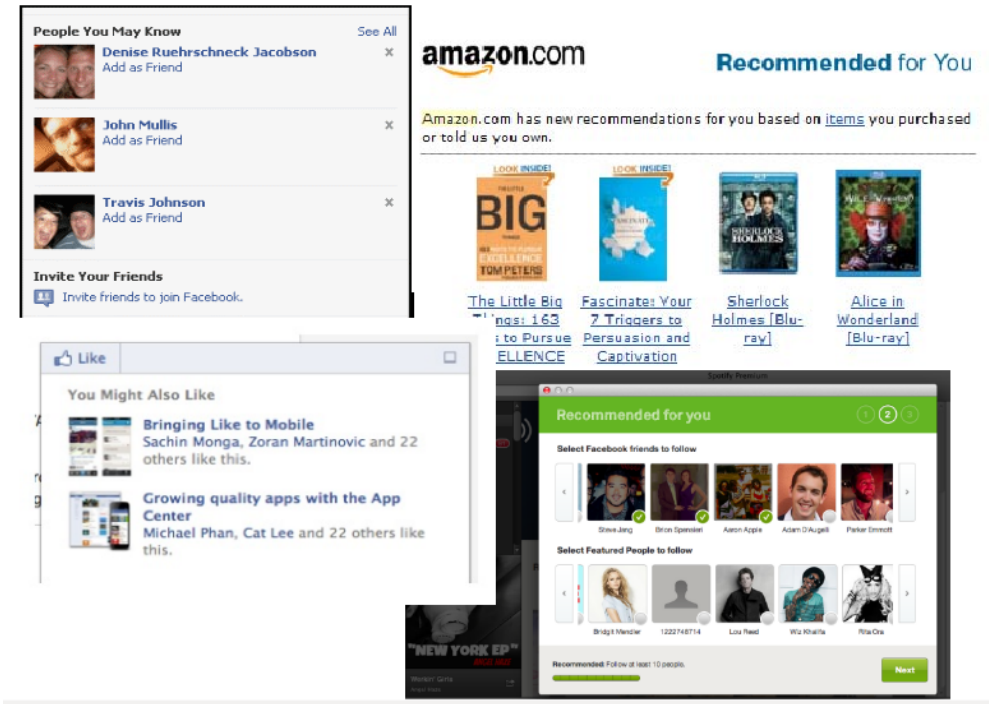
In this questionnaire we aim to find out more about people's opinions about privacy issues in recommender systems. All the information we collect in this questionnaire will ONLY be used for research purposes and will NOT be shared with third parties.

* Required

1. *Mark only one oval.*

   ◯ Option 1

Examples of recommender systems



2. **Gender** *

   Mark only one oval.

   ◯ Female

   ◯ Male

3. **What is your age?** *
Mark only one oval.

- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55-64
- ◯ 65+

4. **What is your nationality?** *

_____

5. **How often do you think you are using recommender systems/personalized services (e.g. news/movie/music/book recommendation, targeted ads, etc.)?** *
Mark only one oval.

- ◯ Several times a day
- ◯ Every day
- ◯ Every week
- ◯ Every month
- ◯ Less frequent

6. **Have you ever requested to see your user profile or any other information the provider has about you?** *
Mark only one oval.

- ◯ Yes
- ◯ No

7. **Do you think recommender systems you have used respect laws and regulation on privacy and security?** *
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Don't know / Not sure

8. **If not, how do you think they may violate user privacy?**
Check all that apply.

- ☐ They collect more data than what has been approved?
- ☐ They use the data for other purposes than what has been approved?
- ☐ They share data with third parties
- ☐ They combine several data sources to extract information that has not been approved
- ☐ Other: _____

9. **To what extent do you think the existing recommender systems violate user privacy?** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
None	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	A lot

How important is it for you to get recommendations in the following domains?

10. **News** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

11. **Music** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

12. **Movies** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

13. **Books** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

14. **Other products (Shopping)** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

15. **Tourism (travels, hotels, excursions, etc.)** *
Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

User Profile

User profiles for recommender systems are build based on the user's interaction with the system. Within time these profiles can get quite detailed.

16. **Would you be willing to let your user profile be shared across applications?** *
Mark only one oval.

- ◯ Not at all
- ◯ Yes, but only within the same domain (e.g. between newspapers)
- ◯ Yes, also across domains (e.g. news user profile used for movie recommendations)
- ◯ Yes, also across domains (e.g. news user profile used for movie recommendations) but I would like to choose the applications

17. **Would you be willing to let your user profile be shared across applications if it is a service provider that you trust?** *
Service provider is the party that generates recommendations (e.g. Amazon, Google, local newspaper)
Mark only one oval.

- ◯ Not at all
- ◯ Yes, but only within the same domain (e.g. between newspapers)
- ◯ Yes, also across domains (e.g. news user profile used for movie recommendations)
- ◯ Yes, also across domains (e.g. news user profile used for movie recommendations) but I would like to choose the applications

18. **What makes a service provider trusted for you?** *
Check all that apply.

- ☐ Option to modify and delete my user profile
- ☐ Show respect to privacy regulations
- ☐ Ask permission from me when they want to use or share my data
- ☐ If it is a well known or popular brand (like Google or Amazon)
- ☐ If they don't share my data with third parties
- ☐ If they are transparent to me about their usage of my data
- ☐ The public opinion about the service provider's reliability
- ☐ Clear, short and understandable description of their privacy policies
- ☐ Other: _____

19. **If you recently bought (or browsed) a book about skiing, how much would you like to get news articles related with skiing?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Very much

20. **If you recently clicked on an advertisement about skiing equipment, how much would you like to get news articles related with skiing?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Very much

21. **If you recently read a news article about a book review, how much would you like to get product offers on that book (from finn.no, ebay.com etc.)?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Very much

22. **If you recently browsed (or bought) a new video game on ebay.com or finn.no, how much would you like to get ads (from different sellers) on similar games?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Very much

23. **If you recently clicked on an advertisement about a smart phone, how much would you like to get product offers on that phone (from finn.no, ebay.com etc.)?** *
Mark only one oval.

	1	2	3	4	5	
Not at all	◯	◯	◯	◯	◯	Very much

24. **If you were able to inspect, modify and delete your user profile yourself, would this…** *
Check all that apply.

- [ ] … make you less worried about privacy risks in recommender systems
- [ ] … make it more acceptable to share profiles across applications
- [ ] ... increase the trust to the service provider
- [ ] Makes no difference
- [ ] Other: _____

25. **What does it mean for you to own your online data?** *

Your online data includes your user profile and your actions (browsing, clicking etc.) while using a service.
Check all that apply.

- [ ] I can modify and delete my data
- [ ] I store my data in my device
- [ ] I decide how my data is shared

26. **How important for you to own your data in recommender system domain?** *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Not important	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very important

27. **Do you have other comments about recommender systems and privacy?**

_____

_____

_____

_____

_____

Appendix C

Survey Responses

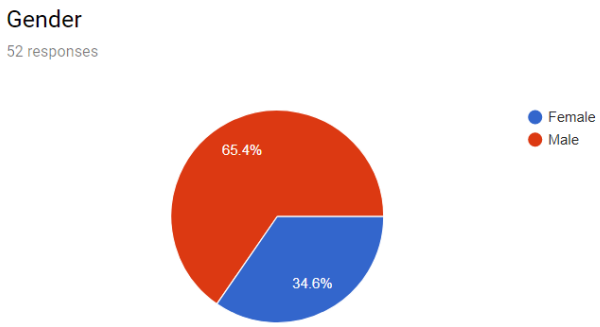


Figure C.1: Survey Response 1

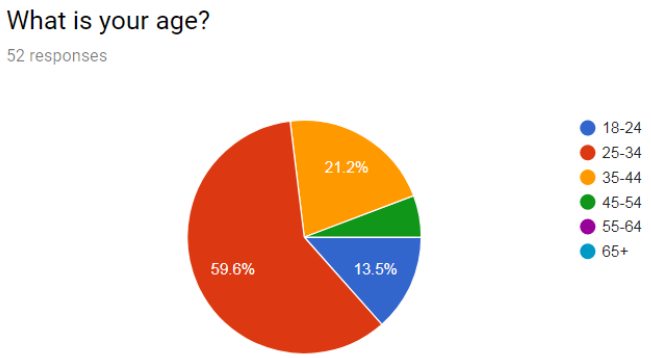


Figure C.2: Survey Response 2

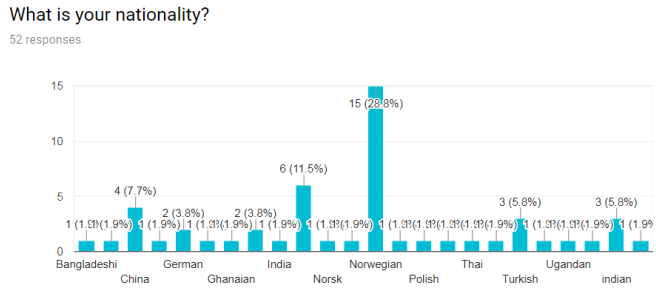


Figure C.3: Survey Response 3

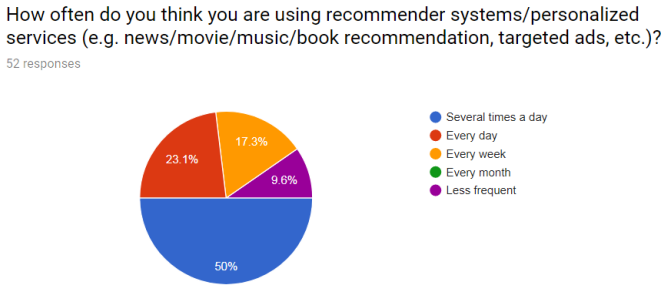


Figure C.4: Survey Response 4

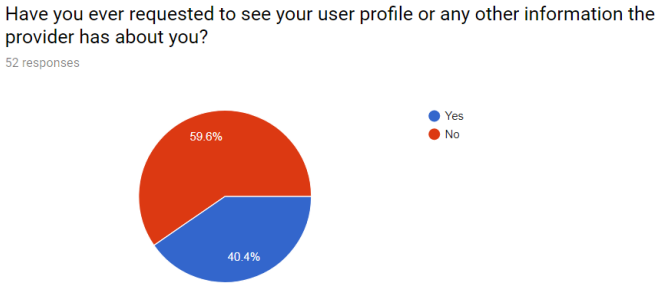


Figure C.5: Survey Response 5

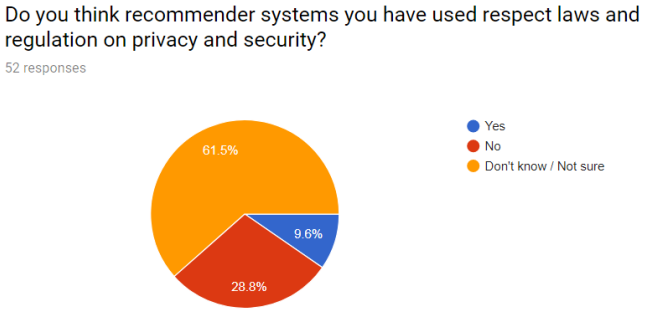


Figure C.6: Survey Response 6

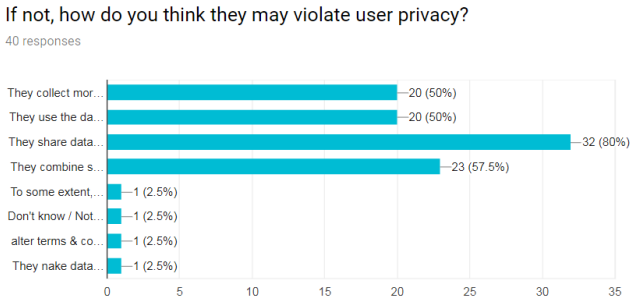


Figure C.7: Survey Response 7

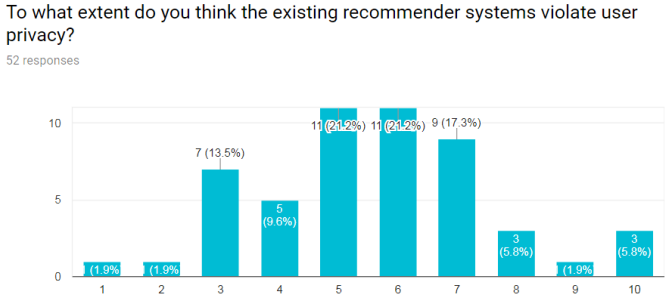


Figure C.8: Survey Response 8

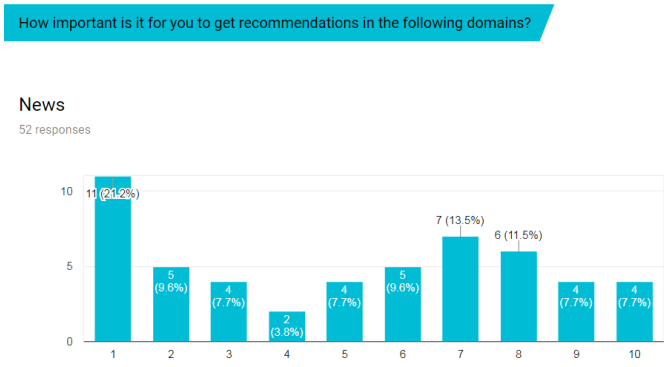


Figure C.9: Survey Response 9

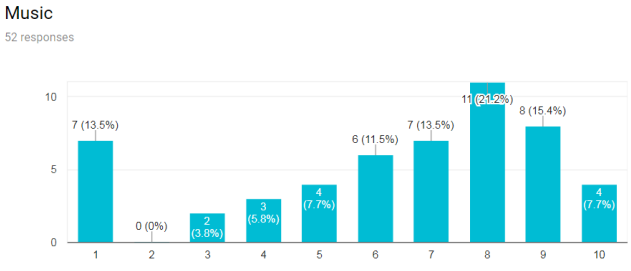


Figure C.10: Survey Response 10

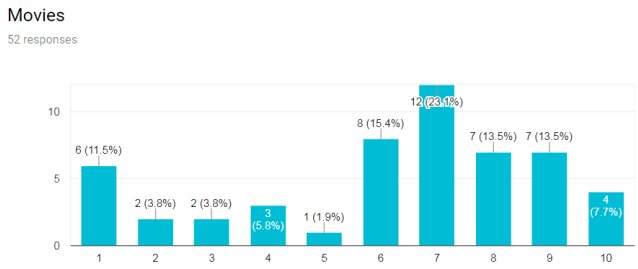


Figure C.11: Survey Response 11

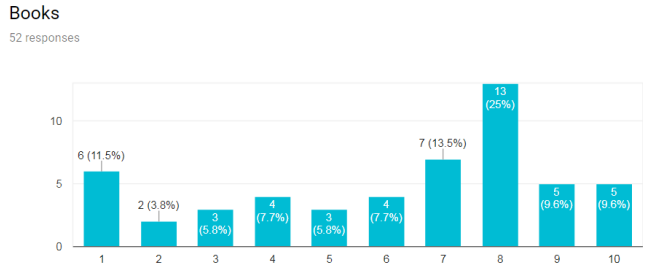


Figure C.12: Survey Response 12

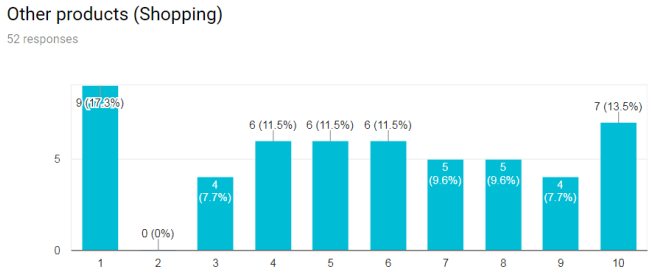


Figure C.13: Survey Response 13

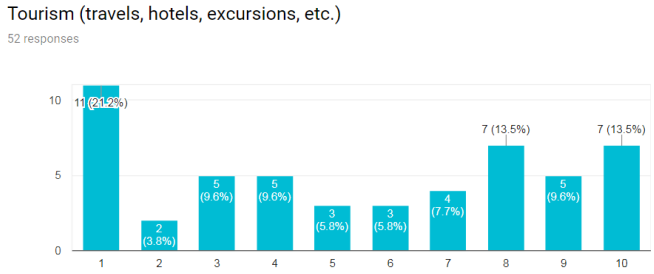


Figure C.14: Survey Response 14

User Profile

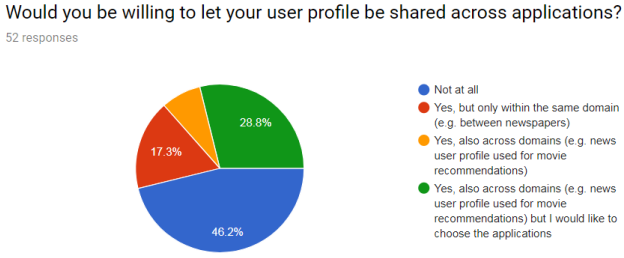


Figure C.15: Survey Response 15

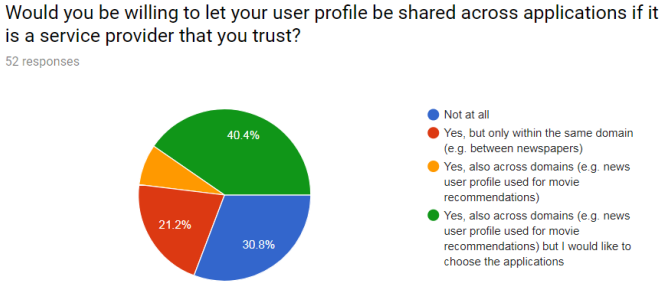


Figure C.16: Survey Response 16

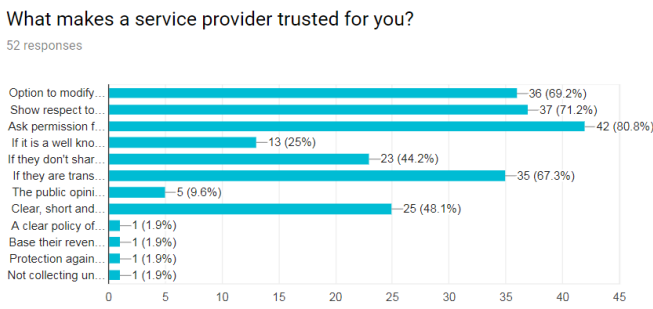


Figure C.17: Survey Response 17

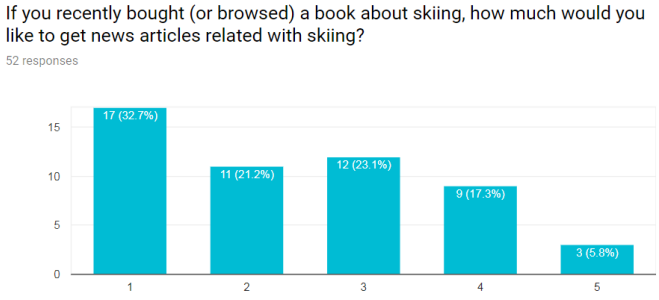


Figure C.18: Survey Response 18

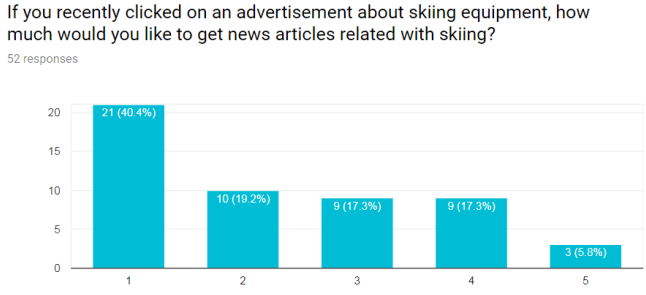


Figure C.19: Survey Response 19

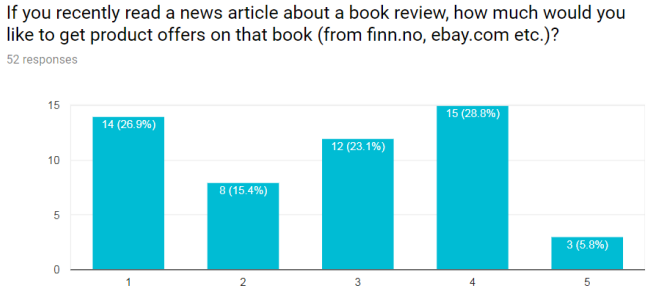


Figure C.20: Survey Response 20

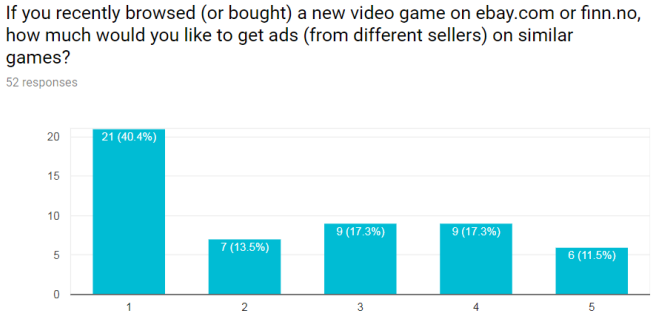


Figure C.21: Survey Response 21

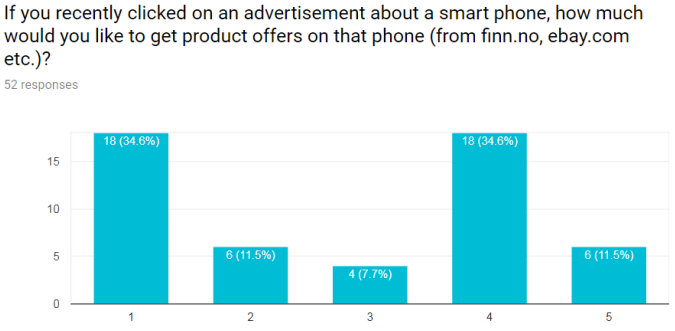


Figure C.22: Survey Response 22

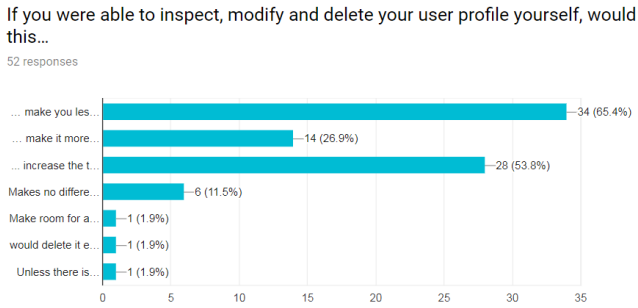


Figure C.23: Survey Response 23

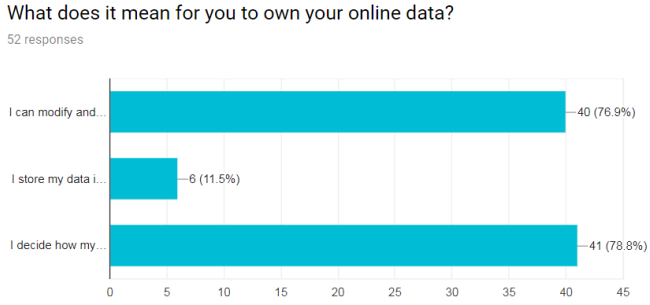


Figure C.24: Survey Response 24

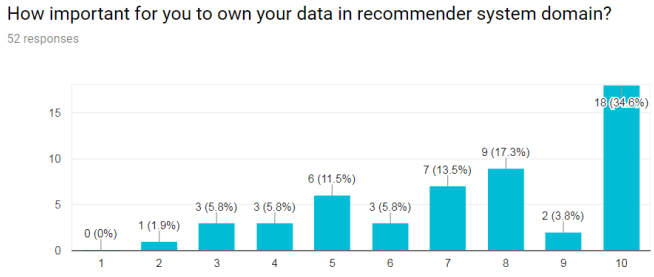


Figure C.25: Survey Response 25

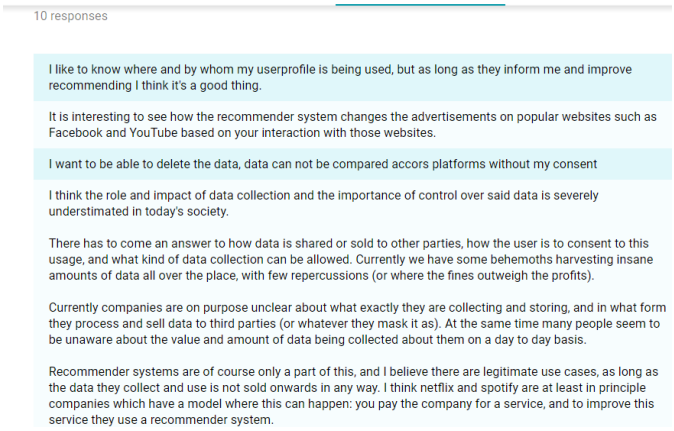


Figure C.26: Survey Response 26(a)

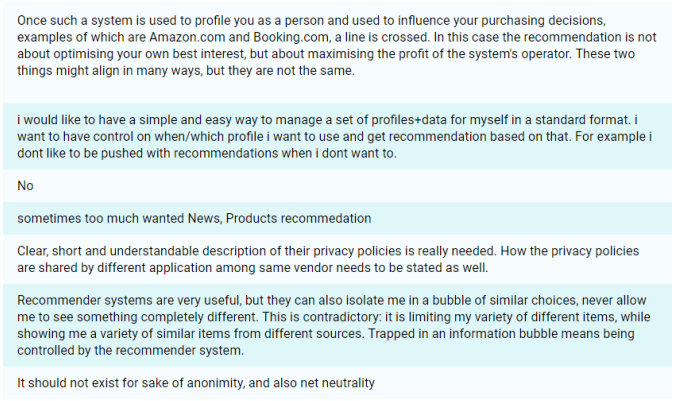


Figure C.27: Survey Response 26(b)