Beyond Recommendation Accuracy: A Human-Like Recommender System

Ala'a Nasir Al-slaity

Thesis submitted in partial fulfillment of the requirements for the

Doctorate in Philosophy Degree in Computer Science

Under the auspices of the Ottawa-Carleton Institute for Computer Science

uOttawa

University of Ottawa
Ottawa, Ontario, Canada

February 2021

Abstract

Since the emergence of Recommender Systems (RS), most of the research has focused on improving the capability of a recommender system to predict and provide an accurate recommendation. However, the literature has demonstrated increasing evidence that providing accurate recommendations is not sufficient to increase users' acceptance of the provided recommendations. Hence, it is vital for a recommender system to focus not only on the accuracy of the provided recommendations but also on other factors that influence the acceptance of recommendations and the extent to which these recommendations are convincing or *persuasive*. Consequently, there becomes a need for new research paradigms to help improve the capabilities of recommender systems, which goes beyond the recommendation accuracy. One of the recently emerged research directions that consider this need fosters the idea of adopting human-related theories from the social sciences domain, such as persuasiveness of social communication.

In this context, however, a challenging, non-trivial, and not fully explored issue that arises is: *how to integrate human-related theories into a recommender system to be one of its intrinsic characteristics in order to improve its performance beyond its accuracy?* This thesis aims to address the above issue from two angles: first, it investigates improving recommender systems by increasing users' acceptance of the recommendations. To achieve this, the influence of persuasion principles on users of recommender systems is investigated. Then a reference architecture framework to adapt and integrate persuasion features as a substantial characteristic of recommender systems is proposed. The proposed framework, named **Per**sonalized **Per**suasive RS (*PerPer*), adopts concepts from the social sciences literature, namely personality traits and persuasion principles. In addition, *PerPer* adapts machine learning concepts, in particular, the Learning Automata, to support its learning capabilities.

Second, the thesis discusses evaluating recommender systems beyond their accuracy. Particularly, it proposes two evaluation approaches that aim to evaluate recommender

systems in a comprehensive way that goes beyond evaluating accuracy only. The first evaluation approach is called the **Com**prehensive **Per**formance evaluation (*ComPer*). It adopts concepts from the human learning domain and provides a simple, thorough, and setting-independent evaluation approach for recommenders. The essence of *ComPer* is to consider a recommender system as a human being, and hence the former's outcomes (i.e., recommendations) can be evaluated and validated in a way similar to how humans' learning outcomes are evaluated. The second evaluation approach adopts goal-oriented modeling to provide an evaluation that does not only assess recommenders beyond their accuracy but also considers the multi-stakeholders of RSs. We demonstrate, empirically, and by user studies, the feasibility and usefulness of the proposed approaches.

The **contributions** of the thesis are: (1) A characterization of recommender systems as systems supported with human traits and features, which goes beyond the conventional recommender systems known in the literature. (2) A user study that examines the impact of persuasive principles on users of recommender systems. (3) A ***Per***sonalized ***Per***suasive RS (*PerPer*) reference architecture framework to enrich recommender systems with persuasion capabilities that are personalized and adaptive for different users. (4) A mapping between human's cognitive skills and the recommendation process. (5) The ***Com***prehensive ***Per***formance evaluation (*ComPer*) framework to provide a comprehensive assessment of recommender systems considering multiple evaluation dimensions other than accuracy. And (6) a goal-oriented evaluation approach to assess the impact of multiple alternatives for recommendation approaches on the satisfaction of RSs stakeholders' goals.

Acknowledgment

First and foremost, I thank ALLAH, the almighty, for helping me and giving me the strength and patience to complete this work.

I would first like to express my sincere gratitude, appreciation, and respect to my supervisor, Professor Thomas Tran, for his continuous support of my Ph.D. study and related research. His continuous support, encouragement, inspirational thoughts, valuable guidance and assistance, patience, and careful review helped me in all the time of research and writing of this thesis. He provided me with valuable support and gave his time and experience generously. I could not have imagined having a better advisor and mentor during my journey in the Ph.D.

I am also grateful to the University of Ottawa for providing all the required resources to accomplish this achievement. I also appreciate the generosity of the University of Ottawa for granting me the International Doctoral Scholarship and the Nicole Bégin-Heick scholarship.

Special thanks to my Ph.D. committee members, Prof. Julita Vassileva, Prof. Nancy Samaan, Prof. Tony White, and Prof. Stéphane Somé, for their insightful comments and feedback, which incited me to view my research from various perspectives. Their support and help highly contributed to the success of this work. I am also grateful to Dr. Abdullah Aref, Dr. Sanaa Alwidian, and Haoye Lu for the stimulating discussions and feedback that had a positive impact on my thesis.

My heartfelt thankfulness goes to my lovely and always caring mother, Narjes, and father, Nasir. My mother, the sweetheart, and the peaceful soul in our family. Without her encouragement and prayers, this work would never exist. My father, Nasir, who has always encouraged me towards success, tolerated me during my studies, and always directed me to take the right path. My sincere thanks to my sisters: Areej, Alaa (whose name in English has the same spelling as my name!), Ayat, Asmaa, Shaimaa, and Baraah. You were always with me despite physical distances. Without your understanding and being around me all

the time, I would never complete this hard path. Furthermore, I deeply appreciate the encouragement and support shown by my parents-in-law, friends, cousins, and aunts.

Last but not least (or "*Last a nub least*," as my daughter spells it!), A very big thank you to my small family, Sanaa, Marie, and the little Lina. They were the main motivation for me to continue. My heartfelt thankfulness goes to the life-long companion, my wife, Sanaa. Without her help, support, and encouragement, I would never be here now. She was the wife, the mother, the friend, and the colleague. She played all these rules just perfectly despite her great busyness completing her Ph.D. studies. I will never forget your supportive words, even when you were in very difficult situations. The Tim Horton's coffee that we used to share in our next-doors offices has a special taste that will never be forgotten.

Dedication

Challenging work needs the guidance of elders and the support of those who are very close to our hearts.

For my guides and my supporters:
My parents for whom I did all this, and
My wife who was, is, and will be the main supporter

I dedicate this work

Table of Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Definition
ANOVA	Analysis of Variance
AUC	Area Under Curve
BFI	Big Five Inventory
BRA	Boosted Reinforcement Approach
CALA	Continuous Action-Set Learning Automata
CARS	Context-Aware Recommender Systems
CBRS	Content-based Recommender System
CFRS	Collaborative Filtering Recommender System
ComPer	Comprehensive Performance Evaluation
CRS	Conventional Recommender System
CVM	Comprehensive Veracity Measure
ELM	Elaboration Likelihood Model
ERS	Educational Recommender Systems
ERW	Equal Relative Weights
FALA	Finite Action-Set Learning Automata
FFM	Five Factors Model
FNR	False-Negative Rate
FPR	False-Positive Rate
GRA	General Reinforcement Approach
GUTH	Gamification User Types Hexad
HCI	Human-Computer Interaction
HRS	Hybrid Recommender System
IPIP	International Personality Item Pool
JASP	Jeffrey's Amazing Statistics Program
KBRS	Knowledge-Based Recommender System
LA	Learning Automata
L_{RI}	Linear Reward-Inaction
MAE	Mean Absolute Error
MAP	Mean Average Precision
MRR	Mean Reciprocal Rank
NLP	Natural Language Processing
PA	Persuasion Agent
PE	Persuasion Engine
PerPer	Personalized Persuasive Framework
PerPer–BRA	PerPer Framework with Boosted Reinforcement Approach
PerPer–GRA	PerPer Framework with General Reinforcement Approach
PIS	Product Importance Score
PLSA	Probabilistic Latent Semantic Analysis
PP	Persuasion Pool

PPRA	Persuasive Product Recommendation Agents
PPS	Product Preference Score
PRS	Persuasive Recommender Systems
REB	Research Ethics Board
RMSE	Root Mean Square Error
RS	Recommender System
RTD&C	Round-Table Discussion and Consensus
SDT	Self Determination Theory
TAM	Technology Acceptance Model
UI	User Identity
UM	User Model

Chapter 1. Introduction

This chapter provides a brief overview of recommender systems and their types, discusses the problem specification, the motivation, as well as the objectives and scope of this thesis. Furthermore, the research questions, methodology, and a list of existing contributions and publications that are derived directly from this research are discussed. Finally, the organization of this thesis is outlined.

1.1. Overview

Recommender Systems (RSs) are software systems that assist users in finding information, products, services, and more based on users' personal preferences [1]. An RS aims basically to reduce information overload by estimating relevance between recommended items and users' preferences. Since the mid-1990s, several approaches of recommender systems have been proposed and implemented in a wide range of application domains. Examples of such applications include, but are not limited to, recommending music in *Last.fm*[1], and *Spotify.com*[2], suggesting friends on *Facebook*[3], and *LinkedIn*[4], recommending courses to be taken in *Coursera.com*[5], and *EdX.com*[6]. In these applications and others, RSs have proven their usefulness as supportive tools to recommend highly accepted items despite the enormous amount of information available on the web. Hence, RSs systems have become one of the most useful and critical means for online systems.

According to Jannach et al. [1], RSs can be broadly classified into four main categories based on how recommendations are made. The first category is the *Collaborative Filtering Recommender Systems* (CFRS), which is the most popular approach. The main idea of CFRS is to recommend items that are liked by *other* users with preferences (or

[1] https://www.last.fm/
[2] https://www.spotify.com/ca-en/
[3] https://www.facebook.com/
[4] https://www.linkedin.com/
[5] https://www.coursera.org/
[6] https://www.edx.org/

tastes) similar to the preferences of the currently targeted users. Second, *Content-Based Recommender Systems* (CBRS), in which the system recommends items similar to those that were liked by the *same* user in the past. Third, *Knowledge-Based RS* (KBRS), which makes recommendations based on explicit knowledge about the available items, user preferences, and recommendation criteria. Usually, KBRSs are applied in domains where CFRS and CBRS cannot be applied and where explicit knowledge acquisition is required, such as in cars and home sales websites. Finally, the *Hybrid Recommender System (HRS)* approach, which, as its name indicates, involves any combination of two or more of the above-mentioned recommenders' categories.

In general, an RS may generate non-personalized or personalized recommendations. *Non-personalized* recommendations, usually found in newspapers or magazines, are less popular and much simpler. A typical example of this type of recommendation is to recommend the top-n selection of items (e.g., the most-watched movies) without necessarily targeting a particular user. *Personalized recommendations*, on the other hand, mean that different items are recommended to different users based on their personal preferences.

The majority of RS approaches (whether they are CFRS, CBRS, KBRS, or HRS) make recommendations with the main focus of *accurately* matching the recommended item with users' preferences, without taking into account additional factors such as a user's mood or her personality [2]. According to Ricci et al. [3] however, all components of RS (such as, for example, the recommendation algorithm and the graphical user interface) should be customized to provide useful and personalized suggestions for users. This means that an RS should not only focus on providing accurate recommendations but should also take into consideration other important factors that help users in their decisions.

1.2. Problem Statement

Recommender Systems have gained vital importance in the literature, and they became one of the essential topics discussed in most of the top-tier venues [4]. Most of the RS-related research had focused on one particular aspect, namely on *improving recommender's accuracy*[7]. In the context of this thesis, accuracy means the capability of a recommender to

[7] In the literature, "Accuracy" and "Correctness" are synonyms and used interchangeably to express the same concept.

predict and provide recommendations that are close to users' preferences [5], [6]. Despite its importance, researchers have recognized that providing accurate results is not always sufficient to increase users' acceptance of the provided recommendations and/or to *evaluate* the overall performance of these systems [7]-[10]. For instance, recommending a highly purchased item or an item with the highest ratings is considered an accurate recommendation. Nevertheless, it does not necessarily lead to this recommendation being accepted by a new user. This is because this item might already be seen/purchased by that user, and it is not of her interest anymore, or that the user has already rated the item, and she is no longer interested in buying it again.

In addition, Jannach et al. [11] compared several recommendation algorithms in an offline experiment. Different splitting protocols and metrics were considered. The authors considered two evaluation dimensions, accuracy and coverage. Surprisingly, the results show that some of the commonly used algorithms, which achieve high accuracy, tend to recommend popular items only. These items are probably not very interesting for the users in reality. Therefore, the authors concluded that it is not advisable to compare different algorithms by relying only on accuracy measures. Recommending popular items does not always lead to the desired sales or persuasion effects, although it leads to high accuracy [12]. The recommendation of only popular items can, in addition, lead to limited system-wide diversity and coverage,

Realizing the above-mentioned shortcomings, researchers started to investigate other factors to be considered during the design of recommender systems for the purpose of mitigating the non-sufficiency of accuracy as a sole factor in the designing and the evaluation of RSs. Gkika and Lekakos [13], for instance, indicate that accompanying persuasive explanations with a recommendation has the potential to increase users' acceptance of that recommendation.

To emphasize and summarize the problem: the literature witnesses a consensus that recommendation accuracy is insufficient to enhance users' experience with RSs. Hence, it is vital for RSs to focus not only on the *Accuracy* of the provided recommendation but also on other factors that influence the acceptance of recommendations and the extent to which such recommendations are convincing or persuasive. In addition, considering accuracy as the only dimension for evaluating and comparing the performance of RSs is not

sufficient. Other dimensions should be taken into consideration to evaluate the performance of these systems comprehensively.

1.3. Motivation

This work is motivated by the challenges discussed in Section 1.2. It is also inspired by the recently emerged research direction, which fosters the idea of adopting human-related concepts in the context of RSs. In particular, several researchers suggest that some patterns of human-human interactions can be deployed and utilized for human-computer interactions, even though computers do not have the same communication skills as humans [14]. For instance, Fogg [15] suggested that computer interactions with users become more accepted if the computer (as a source of the interaction) is perceived as credible and likable by the user (as a receiver of the interaction). In addition, users can be more persuaded by systems that possess human-like features, for example, a chatting robot with a human-being name and style of conversation [15].

Wang and Benbasat [16] investigated the social role of RSs and suggested that RSs should be treated as social actors with human characteristics, such as integrity. Zanker et al. [17] argued that interactions with RSs should be viewed not only from a technical perspective but also from social and emotional perspectives. Aksoy et al. [18] found that there is a higher chance that a user accepts recommendations if the RS generates them in a way similar to the user's decision-making process. Thus, they suggest that the similarity rule is also applied when humans interact with RSs. Furthermore, as Bauernfeind [19] stated, building a trust relationship between users and RSs is an important factor to assist users in making a decision, and it increases users' intentions to adopt the RS [16].

Many other studies in the literature also supported the concept of "Recommenders as Social Actors." For instance, Bonhard and Sasse [20] contended that in order to generate more trustworthy, useful, and understandable recommendations, understanding the social embedding of a recommendation is a vital key. Similarly, Al-Natour et al. [21] argued that online shopping assistants could be perceived as social actors, and hence, can be viewed as users with personality attributes and behavioral traits. Finally, According to Yoo et al. [22], applying human-human communication theories to RSs opens up a new avenue for comprehending the role of these systems and their interactions with users.

Despite this increased attention towards the social aspect of RSs, the existing literature still lacks a comprehensive view for understanding RSs as a social actor [22]. Many of the important factors that help developing social RSs have not been examined yet [22]. Motivated by the aforementioned research directions, this thesis investigates some of the unexplored factors of social (or human-like) RS to mitigate the problem mentioned in Section 1.2. In particular, this thesis aims to mitigate the aforementioned issues from two angles, as follows:

Enhancing User's Acceptance

The classical communication theories in general, and the persuasive technology literature in particular, suggest that people would be more convinced or persuaded by a recommendation if the source of communication displays persuasive cues during the communication process [22]. In addition, some studies in the context of human-computer interaction (HCI) suggest that various aspects of human-human communication, when adopted and deployed, would have the potential to improve the quality of interactions between humans and technology [14], [23], [24]. Inspired by these theories and studies, we perceive an RS to be a source of communication whose task is not only to recommend items conventionally but more than that, to persuade users and convince them about why they should accept the recommended item.

We are not the first one to investigate this research direction. Researchers, such as Zanker et al. [17], Nguyen et al. [25], Gretzel and Fesenmaier [26], and Yoo [27] investigated the persuasive role of recommenders and suggested that in order to increase the acceptance of a recommendation, the recommender itself should be persuasive. However, despite the researchers' findings and conclusions, there are still remaining research gaps and challenges that are not fully explored and addressed yet [22]. Among these gaps is the research question: *how to embed and integrate persuasive features into recommender systems' characteristics to increase users' acceptance*?

The first contribution of this thesis is to answer this question by proposing an approach to enrich RSs with persuasive capabilities. The goal of this approach is to increase users' acceptance of recommendations while preserving the system's accuracy. It is worth mentioning that the proposed solution in this direction does not replace the existed

recommendation algorithms. Instead, it supports them by augmenting persuasion features to the recommender. The goal of doing so is, as mentioned above, to increase user's acceptance. Although RSs recommends items based on previous interactions with the user (which normally reflects the user's preferences), the user may not have enough motive to use the recommended items. In such a case, persuasion plays an essential role in influencing the user's decision and, therefore, increasing acceptance. To this end, we can say that the proposed solution can be used with most RSs when the use of persuasion is accepted. This includes a wide range of application domains, such as eCommerce, Movie, Music, Learning, and health.

Evaluating Recommender Systems Comprehensively

As mentioned in Section 1.3, RSs are evaluated and compared based mainly on their accuracy. However, RS's ability to meet its user's requirements (and hence their perceived acceptance of the recommendations) depends on different aspects (other than accuracy) related to the application area or domain, recommender's behavior, and the user's information needs [28]. Thus, there becomes an increasing consensus that a single metric (accuracy in this context) is insufficient to evaluate the actual performance/effectiveness of recommenders [29].

Inspired by the idea that an RS can be seen as a human, we posit that an RS can also be analyzed and evaluated in the same way that human cognitive abilities are evaluated. As will be discussed in Chapter 5, the recommendation process of RSs is a learning process that can be analogized to humans' learning process. The outcomes of this learning process are the recommendations suggested by the system. Accordingly, we imply that an RS learning outcome (i.e., its predictions or recommendations) can be assessed in the same way that human learning outcomes are assessed. Based on that, this thesis proposes a cognitive-based evaluation method that aims to provide a comprehensive evaluation of RSs. To do so, the proposed method considers multiple evaluation dimensions instead of considering recommendation accuracy only. That is, it aggregates different evaluation dimensions into a single, yet comprehensive evaluation criteria.

Besides, an RS is a multi-stakeholders system, where each stakeholder has goals that usually conflict with the goals of other stakeholders. Chapter 7 proposes a goal-

oriented evaluation approach to assess the impact of multiple alternatives of recommendation approaches on the satisfaction of RSs stakeholders' goals. The contribution of this approach is that it does not only consider multiple evaluation dimensions (in addition to accuracy) but also takes into consideration multi-stakeholders goals.

It is worthwhile to mention that this part of the work (i.e., the evaluation of RS presented in Chapter 5 to Chapter 7) proposes two multidimensional evaluation solutions. This part is intended to introduce new research directions for evaluating RS. Therefore, the work is still preliminary, and the proposed approaches still need further analysis and validation to be mature enough. Nonetheless, we hope that this work will be extended further when there are more researchers interested in these new directions.

1.4. Research Questions

This research aims to answer the following research questions:

RQ1: How users' acceptance of recommendations is influenced by different persuasive strategies?

RQ2: How can we deploy human-related theories to enhance users' acceptance of recommendations?

RQ3: How to integrate and embed persuasiveness as intrinsic characteristics of a recommender system to increase users' acceptance of the recommendations?

RQ4: How can we adopt humans-learning evaluation theories to evaluate RSs in a multidimensional manner beyond reliance on the recommendation accuracy alone?

RQ5: How can we evaluate RSs beyond their accuracy while considering all stakeholders' goals?

Chapter 3 discusses a user study that aims to answer RQ1 and RQ2. Chapter 4 build on the results of Chapter 3, and it continues addressing RQ2 in addition to RQ3. Chapter 5 and Chapter 6 consider RQ4, while Chapter 7 addresses RQ5.

1.5. Methodology

This thesis follows the directions of March and Smith [30]. it aims to provide a *construct*, which is a new paradigm of recommender systems that supports more effective,

personalized, persuasive, and highly accepted recommendations. The provided is introduced to address a specific *problem*, which is addressing the lack of persuasive features in RSs, which affects the perceived acceptance of the provided recommendations negatively, and also avoid reliance on accuracy as a single parameter used often in the literature to evaluate the performance of recommender systems, which has been shown to be an insufficient evaluation dimension.

The underlying research methodology for this thesis is inspired by the design science research methodology [31]. This methodology is an approach towards research in information systems. Following this methodology, a literature review in the context of recommender systems and their evaluation is provided first. This review aims to highlight basic concepts, the current state of the art, and gaps in the literature in this context. Then, a conceptual characterization of RSs as systems supported with human traits and features, which goes beyond the conventional RSs known in the literature, is identified (using inspiring analogies and theories from other well-investigated domains).

The problem is then identified and scoped, and further illustrated through representative scenarios and examples. After building a knowledge base for the research, we characterize the requirements of the solution needed to tackle the identified problem. In particular, we propose a new RS paradigm, augmented with persuasive features, to provide more convincing recommendations, which, in turn, increases the user's perceived acceptance of the suggested recommendation. In addition, we propose evaluation approaches that support the comprehensive evaluation of RSs.

We also identified the most crucial features of the proposed solutions that distinguish them from other existing approaches. Finally, the feasibility of the proposed approaches is evaluated empirically and with user studies. The evaluation aims to assess the feasibility of using the proposed solutions to tackle the motivating problems.

The importance of using the design science research methodology lies in the fact that the results of each activity were exploited to improve the next activity and refine its outcomes. For example, in the evaluation step, the obtained results are used as input to the next step, and they are feedback to refine or improve the proposed solutions. For instance, based on the outcomes of the user study presented in Chapter 3, we noticed that following static guidelines is insufficient. Thus, we refined this solution by proposing the adaptive

framework (proposed in Chapter 4). Then, based on the evaluation of the first implementation proposed in Chapter 4, we proposed the second approach that aims to overcome the shortcomings of the first approach. In addition, after proposing the *ComPer* evaluation (Chapter 5) and evaluating it, we noticed that it did not handle the requirements of multiple stakeholders. Thus, we suggested the goal-modelling-based solution (Chapter 6).

1.6. Thesis Contributions

The main goal of this thesis is to investigate the social role of RSs and to show the feasibility of adopting human-related theories to derive the development of recommender systems forward. By addressing the research questions mentioned above, this research provides the following scientific contributions:

- A user study that investigates the relationship between users' characteristics and persuasive strategies used in the context of RS. The study also aims to investigate the effect of different context factors, such as application areas, on this relationship. The results of this user study are used as input for one of the proposed algorithms used to enrich RS with persuasive capabilities. (**RQ1**)
- A reference architecture framework to enrich recommender systems with persuasive features, taking into consideration the importance of making these features personalized. In particular, this research proposes the ***Per***sonalized ***Per***suasive RS (*PerPer*) framework. The *PerPer* framework provides not only persuasive recommendations but also personalized ones tailored and adapted based on each person's personality and preferences. This approach can be used by designers of RSs who aim to add persuasive capabilities to their systems. (**RQ2**, and **RQ3**)
- The ***Com***prehensive ***Perf***ormance evaluation (*ComPer*) framework to provide a comprehensive evaluation of recommender systems considering multiple evaluation dimensions. *ComPer* is a cognitive-based method inspired by the idea that a recommender system can be perceived as a human being with cognitive skills. Hence, the output of the recommendation process can be evaluated the same way as the human learning process is evaluated. (**RQ4**)

- A goal-oriented evaluation approach that assesses the impact of different recommendation algorithms on the satisfaction of RSs stakeholders. The proposed approach not only considers multiple evaluation dimensions (in addition to the accuracy) but also considers multi-stakeholders' goals. These goals are sometimes contradictory or conflicting. (**RQ5**)

1.7. Publications

In relation to my research, this work has led to the following published and submitted articles so far:

Published Papers

- **Alaa Alslaity**, and Thomas Tran. (2020). "*On the Impact of the Application Domain on Users' Susceptibility to the Six Weapons of Influence*." In: Gram-Hansen S., Jonasen T., Midden C. (eds) Persuasive Technology. Designing for Future Change. PERSUASIVE 2020. vol 12064. Springer, Cham. (**Best paper award**). (Discussed in Chapter 3)

  This paper investigates the effect of the application domain on RS users' susceptibility to Cialdini's persuasive principles. It presents the results of a within-subject study that is conducted to draw the relationship between the application domain and the persuasive principles. Two application domains were considered, namely an e-commerce RS and a movie RS.

- **Alaa Alslaity**, and Thomas Tran. (2020). "*The Effect of Personality Traits on Persuading Recommender System Users*." Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, ACM RecSys2020. (Discussed in Chapter 3)

  This paper explores the influence of the Big Five Personalities on recommender system users' susceptibility to six persuasive principles. The study contains two parts; the impact of personality traits as an independent variable and the impact of personality traits in conjunction with the application domain.

- **Alaa Alslaity**, and Thomas Tran. (2019). "*ComPer: A Comprehensive Performance Evaluation Methodology for Recommender Systems*." International Journal of Information Technology and Computer Science (IJITCS). 11(12):1-18. (Discussed in Chapter 5 and Chapter 6)

  This paper introduces and explains the details of the proposed evaluation method, which is called (*ComPer*). It discusses, in detail, all the aspects of the method. Also, it provides the results of our experimental evaluation.

- **Alaa Alslaity**, and Thomas Tran. (2019). "*Towards Persuasive Recommender Systems*." International Conference on Information and Computer Technologies (ICICT, 2019). Hawaii, USA. (Discussed in Chapter 4).

  This paper introduces the framework that we proposed for enriching RSs with personalized persuasive capabilities. It defines the problem, gives the general definitions along with the formal definition, describes the basic ideas, and provides an illustrative example.

- **Alaa Alslaity**, and Thomas Tran. (2018). "*Towards a Comprehensive Evaluation of Recommenders: A Cognition-Based Approach*." Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, Proceedings 31. Springer International Publishing, 2018. (Discussed in Chapter 5)

  This paper presents the foundation of the proposed evaluation method for RSs. It discusses the main idea and describes the inferred mapping between humans and RSs. In particular, this paper shows how the recommendation process can be seen as a learning process, where the outcomes can be evaluated as we evaluate the outcomes of a human learning process.

- **Alaa Alslaity**. (2018). "*A Unified Evaluation Framework for Recommenders*." Graduate Symposium @ 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31. Springer International Publishing, 2018. (Discussed in Chapter 5).

This paper presents my research agenda on proposing a cognitive-based evaluation approach for recommender systems.

Submitted Papers (Under revision)

- **Alaa Alslaity**, and Thomas Tran. (2020) "*The Impact of Persuasive Techniques on Users of Recommender Systems: A User Study*." the Journal of Proceedings of the ACM on Human-Computer Interaction (PACMHCI). (Discussed in Chapter 3)

  This paper diagnoses the extent to which users of different characteristics get influenced by various persuasive principles that a recommender system uses and how the context can affect the influence of these principles. It also provides general guidelines for designing persuasive RSs.

- **Alaa Alslaity**, and Thomas Tran. (2020). "*Personalized Persuasive Recommender System: A Framework and a Machine Learning-based Implementation*." The Journal of Expert Systems with Applications (Discussed in Chapter 4).

  This paper discusses the PerPer framework for augmenting RSs with persuasive features. It also discusses the evaluation results of *PerPer.*

- **Alaa Alslaity**, and Thomas Tran. (2020). "*Goal Modeling-based Evaluation for Recommender Systems*." ACM User Modeling, Adaptation, and Personalization UMAP 2021 Conference. (Discussed in Chapter 7).

  This paper aims to develop the basis of an approach that uses goal modeling to support the selection of a recommendation algorithm. It shows through an illustrative example that it is feasible to model the recommendation alternatives and their contributions to multiple stakeholders' goals so that the selected algorithm is well-aligned with the overall system requirements.

- **Alaa Alslaity**, and Thomas Tran. (2020). "*Evaluating Recommender Systems: A Systemized Quantitative Study*." International Journal of Intelligent Information Technologies (IJIIT).

  The objective of this paper is to assess the uniformity of the evaluation designs of RSs presented in practice and their consistency with the theoretical

sides of recommenders' evaluation. Also, it aims to enclose the boundaries of recommenders' evaluation settings (offline evaluation, in particular). It shows the results of a quantitative survey of articles published in the last decade in the recommendation field. In addition, it introduces the Recommender Evaluation Guidelines (REval) that presents a roadmap for those who are intending to evaluate recommender systems. The work presented in this paper is related to the theme of Chapter 5.

1.8. Thesis Outline

The rest of this thesis is organized as follows:

Chapter 2: presents the theoretical background and definitions of the concepts that are adopted from the social sciences into the RS area and will be used in this work. Also, it provides a literature review about these concepts in the RSs context.

Chapter 3: discusses the user study conducted to examine the impact of Persuasive Recommender Systems (PRS) on different users.

Chapter 4: Builds on the findings of Chapter 3 and discusses the personalized persuasive framework for RSs.

Chapter 5: presents the second main aspect, which is the cognitive-based evaluation method of RSs.

Chapter 6: provides the validation for the cognitive-based evaluation.

Chapter 7: discusses the goal-oriented evaluation approach of RSs.

Chapter 8: concludes the main points of this research, with more elaboration on the thesis' contributions and future work.

Part I

Persuasive Recommender Systems

Chapter 2. Background and Literature Review

This chapter surveys the related work relevant to the main theme of this thesis and its related concepts and situates my work in the literature. Section 2.1 provides a brief introduction to recommender systems. Section 2.2 introduces the review methodology. Section 2.3 discusses background concepts, which include recommender systems-related concepts and other concepts from social sciences and persuasive technology that were initially designed for different domains, but were adopted by, or adapted to, the context of recommender systems. Section 2.4 surveys the state-of-the-art related to persuasive and personality-based recommender systems. Finally, Section 2.5 provides a summary of the chapter.

2.1. Recommender Systems

Recommender systems are software mainly used to assist users in selecting items that are of their interest [1]. System's recommendations are usually referred to using the term "Item"[3]. Normally, each RS focuses on a specific type(s) of items, such as movies or news. RSs have shown great success in a wide range of application domains, which include, but are not limited to, recommending music or movie on streaming websites, suggesting friends on social media, recommending learning materials in learning support systems, and, of course, recommending items in eCommerce systems.

2.1.1 Personalized vs. Non-personalized Recommendations

In a very broad classification, recommender systems generate personalized or non-personalized recommendations. The personalized recommendation is the most common approach in the literature. In such an approach, the system provides suggestions that are tailored to the current user (or group of users). That is, different users receive different items based on their personal preferences Such recommendations are selected based on different factors such as users' similarities or items' properties. Thus, some sort of information about the items and the users is needed to provide personalized recommendations [32].

Non-personalized recommendations, on the other hand, give the same recommendation of items to all the users. Thus, they do not require knowledge about users or items. Instead, these systems provide general recommendations either randomly or based on other factors, such as popularity (i.e., how many times the item was used). Non-personalized recommendations are less popular; they are commonly found in newspapers or magazines. A typical example of this type of recommendation is to recommend the top-n selection of items (e.g., the most-watched movies) without necessarily targeting a particular user. Non-personalized recommendations are not typically addressed in the literature despite their effectiveness in some scenarios.

2.1.2 Categorization of Recommender Systems:

There various types of recommender systems that are different in terms of domain, knowledge used, or, most importantly, the recommendation algorithm used. The literature distinguishes between six different categories of recommendation approaches [3], which are: Content-Based, Collaborative Filtering, Knowledge-based, Demographic-based, Community-based, and Hybrid approaches. According to Jannach et al. [1], RSs can be broadly classified into four main categories only. Therefore, this section focuses on these four categories.

Collaborative Filtering Recommender Systems (CF)

Collaborative Filtering (CF) is the most popular approach [3]. Its main idea is to recommend items that are liked by other users with preferences (or tastes) similar to the preferences of the currently targeted users. The rating history is fed to the algorithm to calculate the similarities between users. One of the common ways to implement CF is the neighbourhood-based method, which focuses on the relationship between items or users. CF is domain-independent. It is typically used when it is difficult to describe the items [33]. CF techniques are further divided into model-based and memory-based [34], [32]. Memory-based, in turn, is divided into user-based and item-based. Examples of the known learning algorithms used in CF are the Association rule [35], Clustering [36], Decision Trees [37], and Artificial Neural Network [38].

The advantage of CF is that it can be used in domains where there no enough information or feature about the items. Also, CF recommenders can recommend items that are relevant to the user even if they are not in the user's profile [39]. However, CF suffers from the Cold-start problem (i.e., new users with very few ratings) [40], and Sparsity of data (i.e., only a limited number of items are rated by users) [34], [41]. Examples of CF applications include GroupLens, Amazon, and Ringo [42].

Content-Based Recommender Systems (CBRS)

Content-based approaches rely on the similarities between items. That is, they recommend items similar to those that were liked by the same user in the past. The similarity between items is mainly calculated based on the attributes or features of the items. For instance, in a movie domain, the movie genre, year, and actors can be used to calculate similarity. Thus, CBRS is domain-dependent [33]. The recommendation process is done based on features extracted from the content of the items the user has evaluated in the past [34], [35]; the system recommends items that are mostly related to the items that are highly accepted by the user.

CBRS uses different types of models to find similarities between documents in order to generate meaningful recommendations. Different techniques are used to extract the similarities between items. The most common techniques are Vector Space Model (e.g., Term Frequency Inverse Document Frequency (TF/IDF) or Probabilistic models such as Naı̈ve Bayes Classifier [43]), Decision Trees [44] or Neural Networks [45]. The advantage of using CBRS is that it does not need knowledge about other users' profiles. However, it needs a deep knowledge of the items. CBRS is widely implemented and studied in the literature (e.g., NewsDude [46], and LIBRA [47])

Knowledge-Based RS (KBRS)

As the name indicates, Knowledge-based approaches rely on knowledge of a specific domain to know how items meet users' requirements. That is, they make recommendations based on explicit knowledge about the available items, user requirements, and recommendation criteria. There are different types of KBRS, such as case-based [48] and Constraint-based [49]. Usually, KBRSs are applied in domains where CFRS and CBRS cannot be

applied and where explicit knowledge acquisition is required, such as in cars and home sales websites.

Hybrid Recommender System

Hybrid recommendation systems, as its name indicates, involve any combination of two or more of the above-mentioned recommenders' categories. The main idea behind using hybrid approaches is that they benefit from the advantages of one approach to overcome the limitation of the other one [3]. For instance, CF approaches suffer from the cold-start problem (recommending items to new users or recommending new items that have not been rated yet). Content-based approaches, on the other side, have the capability to overcome the cold-start problem. Based on that, several hybrid approaches have been proposed. The most common approaches are [33]: 1) Weighted hybridization: combining the results of different algorithms by integrating the scores from each algorithm. 2) Switching hybridization: the algorithm switch between multiple algorithms. And 3) Cascade Hybridization: in this model, the output of each algorithm is input to the next algorithm, which in turn refine it and output the refined results. Examples of Hybrid recommendations include P-tango [50], DailyLearner [51], and EntreeC [34].

2.2. Review Methodology

This thesis investigates the work proposed in the literature that is related to adapting persuasion and personality traits theories from social sciences into the domain of recommender systems. In particular, we articulate the search queries according to the following questions:

- **Question 1**: What work was done that supports the adoption of persuasion theories in the RS domain?
- **Question 2**: What work was done that adopt personality traits theories in the RS domain?

In addition, since the second part of the thesis discusses evaluating RSs, we also investigate the following question:

- **Question 3**: What work was done that supports the evaluation of RSs beyond their accuracy?

This chapter focuses on surveying the literature that addresses the first and second questions. For organization purposes and to enhance comprehension of the subject matter, the third question is discussed in Chapter 5, closer to the related chapters that discuss evaluation approaches of RSs.

The search process was a combination between manual inspection and snowballing strategies. I manually searched the proceeding of the annual ACM RecSys conference, which discusses RSs related issues. Then, I searched the following databases (Scopus, IEEE, ACM digital library, and Google Scholar). For further exploration, I used the snowballing strategy through surfing the referred papers. The search was restricted to peer-reviewed articles published in English-language.

Following are the general search queries for each search question:

- **Question 1:**

  ("Recommend* System?" OR "Recommend* Algorithm" OR "RS?" OR "Collaborative?Filtering") AND
  ("Persuas*" OR "Influence") AND
  ("Strateg*" OR "Principle?" OR "Technique?" OR "Methodolog*")

- **Question 2:**

  ("Recommend* System?" OR "Recommend* Algorithm" OR "RS?" OR "Collaborative?Filtering") AND
  ("Personality?trait?" OR "User*Characteristics" OR "Five?Factor?model" OR "Big?Five?personality")

- **Question 3:**

  ("Evaluat*" OR "Assess" OR "Compar*") AND
  ("Framework" OR "Model" OR "Protocol" OR "Strategy" OR "Method*" OR "Approach") AND
  ("Recommend* System?" OR "Recommend* algorithm" OR "RS?" OR "Collaborative?Filtering")

2.3. Background Concepts

This section introduces the necessary background concepts required to understand this research. In particular, Section 2.3.1 discusses the traditional communication theories and

how they can be connected to the field of recommender systems. Sections 2.3.2 and 2.3.3 provide an overview and theoretical definitions of human-related theories from the social sciences utilized in this thesis, namely *personality types* and *persuasive strategies*.

2.3.1 Recommender Systems and Communication Theories

The literature in communication theories suggests that a persuasive communication process involves the following parts: (1) *source* of a communication, (2) *target*, or destination, of a communication, (3) the *message* to be communicated, (4) the *context*, which affects the whole communication process, and (5) the *effect*, or impact, of the communicated message. These five components, together, as shown in Figure 1, are called the *communication-persuasion paradigm* [52], [53]. The last part (i.e., the effect) represents the outcome of the communication process. The features of all other parts influence this outcome. As it is depicted in Figure 1, each part has multiple factors that influence the extent to which a component is convincing. Consequently, the outcomes of a persuasive communication process are influenced by the combination of multiple factors from all parts.

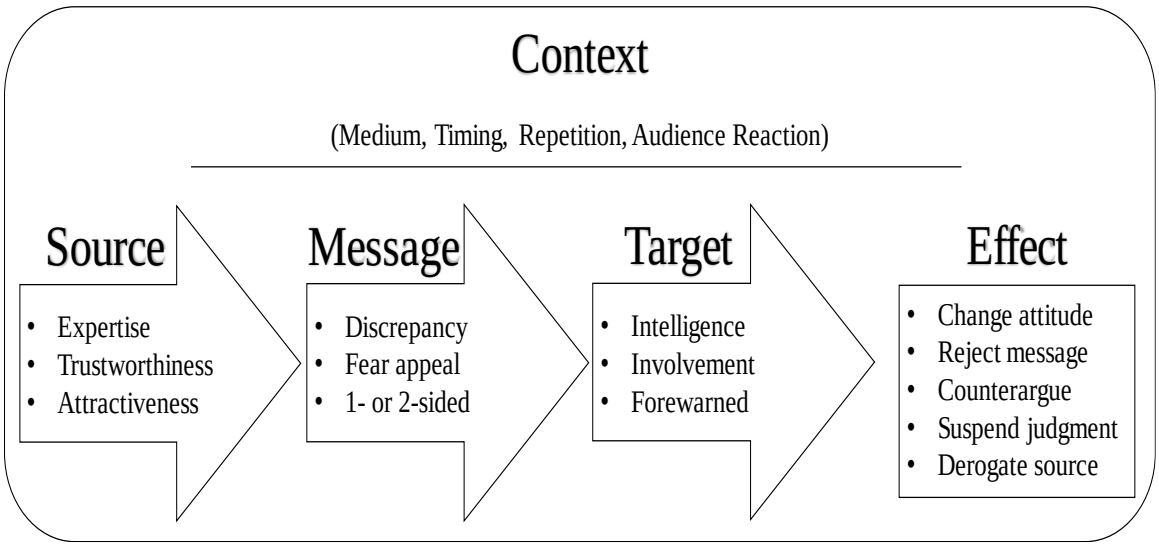


Figure 1 The communication-persuasion paradigm [52], [53]

Applying the communication-persuasion paradigm to the context of recommender systems, a recommendation provided by any recommender system can be seen as persuasive communication. The overall effect of the recommendation process is to persuade or convince the target (i.e., the one for whom a recommendation is created). Hence, and based on the

communication-persuasion paradigm, the ability of a recommendation to convince a receiver depends on four components: the source of the recommendation and the formers' characteristics, the form and the content of the recommendation (i.e., the message), the one who receives the recommendation and his/her characteristics (i.e., the target), and the contextual factors [52]. A mapping (or an analogy) between the communication-persuasion paradigm and persuasive recommender systems is illustrated in Figure 2.

According to Yoo et al. [22], the communication-persuasion factors discussed in the communication literature have recently gained increased attention in a technology-mediated communication context. Researchers have concluded that these factors are also important when people communicate with technology [14]-[17]. Besides, studies indicated that we could make technology-based interactions more persuasive by incorporating the persuasive cues explored in human-human communication [22].

RSs are typical examples of such applications that involve technology-based interactions between users and computerized systems. These systems mainly concern about providing recommendations to their users (i.e., advising users). According to Fogg [14], when computer systems take the role of advising users, it becomes significant to understand the social role of these systems. Since RS is an important example of such systems, traditional persuasive factors could provide a useful framework to study the communication between RS and its users [22].

Yoo et al. [22] adopted the communication-persuasion paradigm to the RS domain to provide a general view of a recommender system as a social actor with persuasion capabilities. The conceptual framework of a Persuasive Recommender System (PRS) is depicted in Figure 2. Analogously to the traditional communication process, this framework suggests that the RS itself is the source of the communication, its recommendation is the message, and the user is the target of that message. All these components reside in a particular context with many properties that can influence the outcome of the communication. The output of this communication process is an effect that encourages or discourages users from accepting the provided recommendation and from deciding to reuse the system again or not.

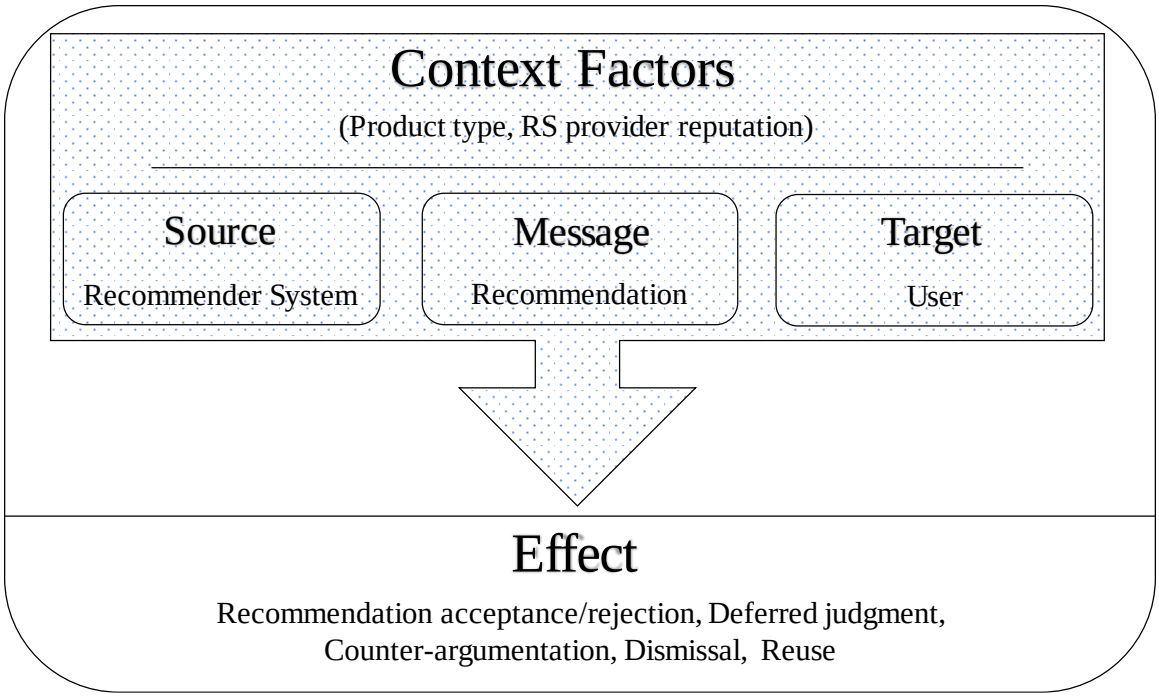


Figure 2 The conceptual framework for persuasive recommender systems (PRS) [22]

This framework is the source of inspiration for the work proposed in Chapter 3 and Chapter 4 of this thesis. According to Yoo et al. [22], this framework outlines the relationships between the key constructs of a PRS. However, this framework is still preliminary and needs more enhancements, as many human-human communication factors are still not applied yet into the RS domain. To this end, this work aims to investigate further factors and to integrate more of the existing, but not applied, human-related theories to the recommendation process to feed more details and advance findings to this framework.

2.3.2 User Personality

User personality can be defined as "*a set of characteristics possessed by a person that influence his or her cognitions, emotions, motivations, and behaviors in various situations*" [52]. Psychologists have extensively studied humans' personalities and their characteristics. As a result, different books have been written [54]-[56], and an abundant number of articles have been published [57], [58]. Also, different theories have been stated, such as the *International Personality Item Pool* (IPIP)[8] [59], the *Self Determination Theory* (SDT)

[8] https://ipip.ori.org/

[60], *Gamification User Types Hexad* [61], and the one of special importance for this work, the *Five-Factor Model (FFM)* [62].

The Five-Factor Model (a.k.a., the Big Five personality traits)[9] was initially introduced in 1961 by Tupes and Christal [63]. FFM is a hierarchical organization of the personality traits of humans. We choose to adopt the FFM model in this thesis due to multiple reasons: firstly, it is an identified and well-established model [64], and it is the most widely accepted model of personality [65]. Secondly, the literature witnesses a consensus that these five factors are able to describe the wide range of personalities [55]. Thirdly, this model categorizes all people into five dimensions only, and its comprehensiveness has been supported theoretically and practically [60]-[62]. Finally, the FFM appears suitable for usage in recommender systems as it can be quantitatively measured.

The FFM contains five dimensions or personality traits, namely *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. These five dimensions represent differences among personalities [66], where each dimension (or trait) represents a relatively stable pattern of frequent interpersonal situations that characterizes human life [56]. According to John and Srivastava [67], each personality trait involves many characteristics, or adjectives, that distinguish it from other traits. The authors stated that: "*the Big Five structure does not imply that personality differences can be reduced to only five traits. Yet, these five dimensions represent personality at the broadest level of abstraction, and each dimension summarizes a large number of distinct, more specific personality characteristics*". Table 1 lists the five personality traits, along with their adjectives [62].

[9] Five Factor Model (FFM), and Big Five personality traits (BF) will be used interchangeably through this thesis.

Table 1 Big Five Personality Traits and their corresponding adjectives

Factor	Adjectives (Facets)
Openness	Artistic, Curious, Imaginative, Insightful, Original, and Wide interest
Conscientiousness	Efficient, Organized, Planful, Reliable, Responsible, and Thorough
Extraversion	Active, Assertive, Energetic, Enthusiastic, Outgoing, and Talkative.
Agreeableness	Appreciative, Forgiving, Generous, Kind, Sympathetic, and Trusting
Neuroticism	Anxious, Self-Pitying, Tense, Touchy, Unstable, and Worrying.

Each of the five personality traits in FFM can be represented as a spectrum, where an individual's personality may fall anywhere on the spectrum, as depicted in Figure 3. The figure shows each trait as a spectrum, where personalities can be either low, high, or anywhere in between that trait spectrum. For instance, a person cannot be said as purely *Extraversion*. Instead, she is placed on the scale between Introverted and Extroverted. A higher place on the scale (e.g., closer to the Extroverted end) means that the Extraversion facets are high. Having said that, an individual may hold properties of multiple traits that collectively form her personality [68].

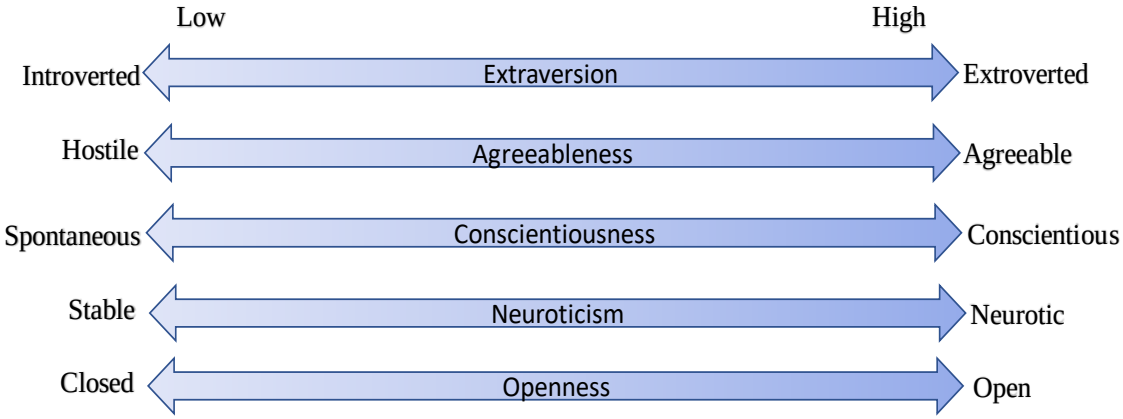


Figure 3 Big Five personality traits as spectrums [69]

The next subsection discusses assessing personality traits and how an individual's personality can be determined according to the FFM.

User Personality Test

In the literature, there are different methods to acquire personality traits. These methods include (1) Story-based method, which involves creating a story from a group of statements that are related to each personality dimension [70], (2) Text-based method, which is conducted by analyzing written texts or spoken languages [57], (3) Keyboard-based method, which involves analyzing people typing features on keyboards or by monitoring how they use the mouse [71], [72], and (4) Questionnaire-based method, performed by asking different questions through which experts can infer a user's personality [73]. The last method (i.e., Questionnaire-based) is the most popular approach to obtain the BF personalities. Hence, we relied on it for this thesis work, as will be discussed in Chapter 3 and Chapter 4.

Several questionnaires have been proposed and deployed to evaluate personality traits. These questionnaires differ in their questions and sizes (i.e., number of questions). Some of these questionnaires focus on measuring the facets of each personality type. Such questionnaires, however, are usually time-consuming due to the number of questions that are required to measure the facets [74]. Thus, we did not consider such tests.

One of the most popular questionnaires is the Big Five Inventory (BFI), which was introduced in the late 1980s [75], [76]. It consists of 44 short-phrase items, the reason why it is known as *BFI-44.* This assessment can be answered in about five minutes. In the beginning, it was debated that such a short questionnaire is sufficient to measure the BF dimensions [75]. This perspective, however, has been changed later, where the BFI-44- questionnaire was no longer considered short; instead, it was viewed to be tediously long. This shift in perspectives is mainly due to researchers are faced with a limited assessment time. Thus, there has been an increasing trend towards using a shorter assessment [75]. As a result, many super short (1 to 10-items instruments) have been introduced in the literature. For instance, The BFI-10 is a shorter version of the BFI-44 assessment that consists of only ten items [75].

Many of the proposed assessments have been discussed, and they have shown promising results suggesting that they can be feasible options for conducting questionnaires. Nonetheless, we decided to use the BFI-44 for several reasons: first, it is a short version of the *Neuroticism-Extraversion-Openness-Personality-Inventory-Revised* (NEO-

PI-R), which is the most accurate and popular inventory in the literature [77]. Second, this questionnaire has been recognized as a well-established measurement of personality traits [78]. Third, the BFI-44 is preferred, especially when the test time is limited. RS is an example of such a case because, usually, users feel bored when asked to do many steps, which in turn consumes their time. It is worth mentioning that BFI-10 is shorter than BFI-44; hence, it takes less time. However, there are tradeoffs between a shorter and more accurate assessment. The more questions asked in the assessment, the more accurate the observation of traits [73]. Given that the standard BFI contains only 44 short phrases, it is not recommended to use the shorter version unless there are exceptional situations [76].

The 44 short-phrase items of the standard BFI are rated on a five-step scale from 1 (Strongly disagree) to 5 (Strongly agree). Then, the individual's rank of each personality trait is calculated using a particular equation. The BFI-44 questionnaire and its calculation instructions are presented in Appendix A.

2.3.3 Persuasive Strategies

In social sciences, persuasion is defined as a style of communication designed to influence the actions and behaviors of individuals [79]. The use of technology to influence the behavior and the attitude of users is called *Persuasive technology* [80]. Behavioral scientists have shed considerable light on how some interactions lead people to cede, comply, or alternate [81]. The research findings imply that there is a wide range of means for persuasion and social influence. Consequently, researchers and practitioners have proposed different taxonomies of persuasion strategies, such as the 40 strategies that have been described by Fogg [14], the six principles of Cialdini [81], and over 100 groups suggested by Kellermann and Tim [82].

This thesis deploys *Cialdini's six persuasion principles* (a.k.a., the *six weapons of influence*[10]) [81]. We considered these principles because they have been widely used in the literature, and they have been verified as global persuasive approaches [68]. In addition, as stated by Gkika et al. [68], these principles provide a solid framework in order to investigate the persuasive power of messages as peripheral cues in recommender systems. The

[10] In this thesis, the terms "six principles of influence", "Cialdini's principles, and "six weapons of influence" are used as synonyms.

six influence strategies, along with their definitions adopted from Cialdini [81], are as follows:

1. *Reciprocity*: "People repay in kind." People have the tendency to return a favor. So, humans are more inclined to follow recommendations made by a person the receiver is in debt to. For instance, if your friend sent you a birthday gift, then you owe that friend a gift to be sent in the future. This principle is used in many computerized systems such that they give new users a gift (such as a voucher or a free service for a limited time)
2. *Scarcity*: "People want more of what they can have less of." People consider whatever is scarce as more valuable. Therefore, they want more of scarce things. Displaying numbers of availability for products is an example of implementing this principle. eCommerce websites (e.g., Amazon) use this principle by providing a limited-time-only promotion, often presented as a countdown timer showing the time remaining before the offer expires.
3. *Authority*: "People defer to experts." People are more inclined to accept recommendations made by a legitimate authority. For instance, it is more likely to give change for a parking meter to a stranger if she wears a uniform rather than casual clothes, and it is more likely to buy a toothpaste if a well-known dentist recommends it.
4. *Social Proof*: "People follow the lead of similar others." Usually, people have a tendency to do what others do. So, people are most likely to accept recommendations if they are aware that others have accepted as well. When people are uncertain, they look to the actions of others to decide. The widespread deployment of this principle is checking the reviews of others. For instance, when people want to book a reservation in a resort, they usually check out the reviews of that resort. This principle is widely implemented in the computerized system, and it is implemented in different ways, such as showing ratings, emphasizing the number of followers or fans, or presenting testimonials.
5. *Liking*: "People like others who like them." People are most likely inclined to accept requests made by somebody (or in a way) they like. For instance, people may prefer one store over the other because they like the employees in that store. In online

communication, the systems also need their customers (or users) to enjoy the service. So, the service or product should be presented attractively.

6. *Consistency* (or *Commitment*): "People align with their clear commitment." People have the tendency to commit to their previous or reported opinion or behavior. The basis of this principle is that if people commit to doing small requests, it will be easier to persuade them to do larger requests. Commitment is implemented online by different means, such as asking the customer to test a new feature in the application for free and to write a review about it. If the customer used the application and wrote a good review of the service, she would be more likely to continue using this service as a kind of commitment.

These principles provide approaches that cause one person to say yes to another one. That is, appropriately implementing these principles can increase the acceptance of your requests or your products. Recently, a seventh principle was added to these six principles [83]. The new principle is called "Unity," and it focuses on leveraging the "We" concept (i.e., acting together and being together) to influence behavior [84]. This thesis focuses on the original six principles.

The essence behind these principles is that different persons can be influenced differently. That is, there is no master persuasive technique, which can influence all people. So, these principles describe different ways of persuasion. Furthermore, each principle can be implemented in many ways. For instance, two different implementations of the *Consistency* principle are either asking customers to start from small things or by rewarding them for investing time and effort in your brand.

2.4. Literature Review

Throughout the review of the recommendation literature, we noticed increasing attention towards the social aspects of RSs. This direction is manifold, such that various theories are used to handle different issues for many recommendation approaches. For instance, user personality has been incorporated into RSs for different goals such as to increase accuracy and diversity, to mitigate the cold-start problem, and to make group recommendations. Moreover, researchers use different aspects (such as personality traits, demographic information, temperament, lifestyle, and emotions) to present different personality styles or

properties [85]. This is just a simple example that demonstrates the high dimensionality of this topic.

To mitigate this issue and narrow down the review, we focused on the most related articles to the thesis topics. Consequently, we divided the literature review into three parts. Each part discusses articles that are related to one of the topics depicted in Figure 4. The rest of the section is organized as follows: Section 2.4.1 discusses personality-based RS, Section 2.4.2 discusses persuasion in the RS domain, and section 2.4.3 discusses works that combine personality and persuasion in the RS domain.

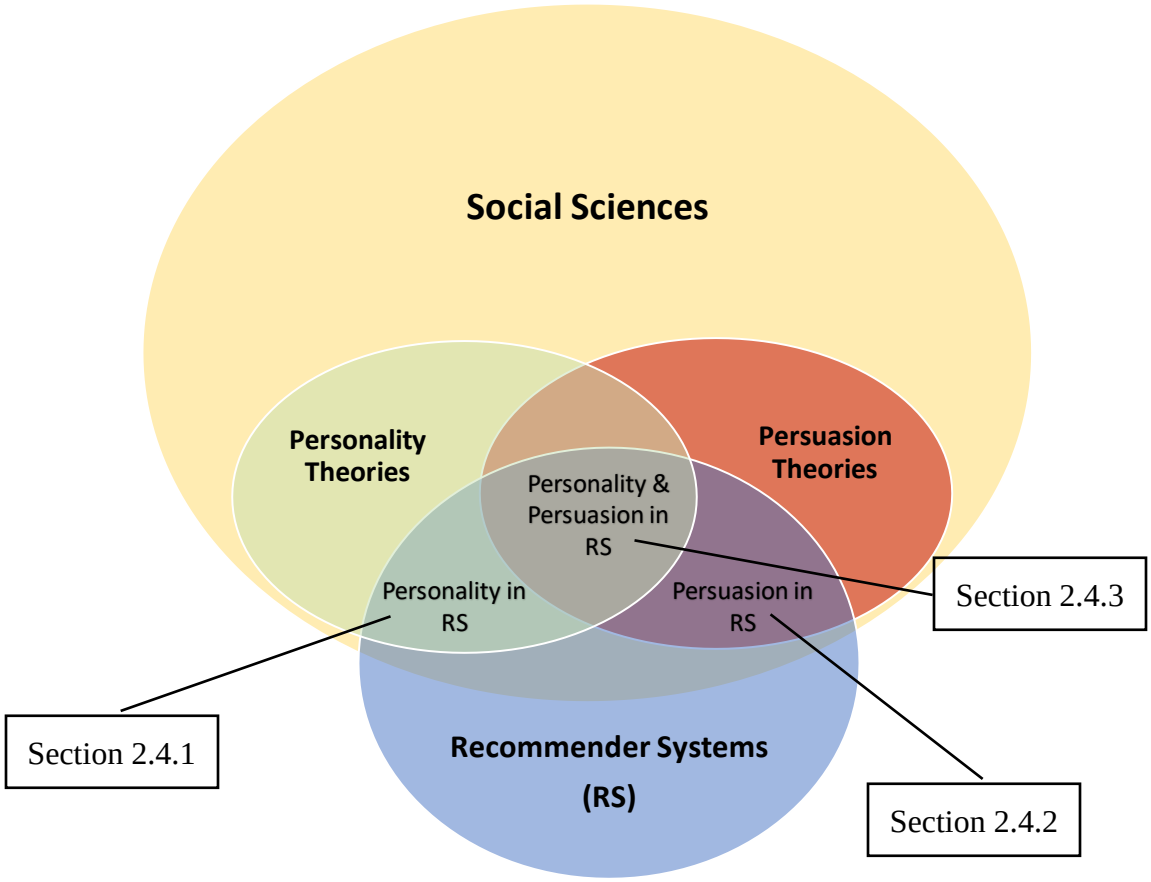


Figure 4 Organization of the literature review

2.4.1 Personality-Based Recommender Systems

Different researchers have studied the importance of personality-based recommendations. These studies consider different aspects as indicators of users' personalities [85]. Personality aspects include demographic information, temperament, lifestyle, emotions, and personality traits. According to Hu & Pu [85], personality traits (which is the aspect of a

special focus of this thesis) are among the most important human psychological aspects used in RSs. Hence, this aspect has received special attention in the recommendation literature.

Burger, J.M [86] defined personality as a "*consistent behavior pattern and intrapersonal processes originating within the individual*." RSs that utilize personality traits to provide personalized recommendations are called *Personality-based Recommenders* (PBR) [87]. These systems have used personality aspects to address different issues, such as the cold-start problem (i.e., recommending items for new users or recommending new items), the diversity of recommendations (i.e., recommending a diverse set of items), and group recommendations (i.e., to provide suggestions for a group of users instead of a single user). This subsection provides an overview of how personality has been utilized in RSs with a focus on those articles that concern the FFM.

To investigate users' acceptance of personality-based RSs, Hu and Pu [87] evaluated and compared a Personality-based recommender with a rating-based one. They relied on the Technology Acceptance Model (TAM), a model that aims to provide an explanation of the determinants of computer technology acceptance. A within-subject user study is conducted to investigate users' acceptance of the systems. The study concluded that, in terms of usability, the personality-based RS is much better than the rating-based one, and it is perceived to be slightly more accurate. Moreover, the results disclose that users prefer the personality-based RS and have a considerably higher intention to reuse such a system.

Rentfrow and Gosling [88] investigated the relationship between music and personalities. They analyzed the music preferences of over 3500 individuals. This study converged to reveal four dimensions of music preferences: *Reflective* & *Complex*, *Intense & Rebellious*, *Upbeat & Conventional*, and *Energetic & Rhythmic*. Then, the authors connect each of these dimensions to users' personalities based on the FFM model. The results show that the music type reveals some of the listeners' personality. For instance, The *Energetic* and *Rhythmic* music dimension was positively related to *Extraversion* and *Agreeableness* personalities. That means individuals who enjoy Energetic and Rhythmic music tend to be talkative, full of energy, and are forgiving.

In another study, Rentfrow et al. [89] extended the work presented in [88] to include general entertainment domains that included books, magazines, films, and TV shows, in addition to music. The contents were categorized into five categories: Communal, Cerebral, Aesthetic, Dark, and Thrilling. Then, as in the previous study, the author mapped these categories to the five personalities presented in the FFM. The results confirmed the previous results such that users' personalities influence their choices. For example, the results show that *Extraversion, Agreeableness,* and *Conscientiousness* are positively related to the Communal category, while *Openness* and *Neuroticism* are negatively related to the communal category. The aesthetic category was positively related to *Agreeableness* and *Extraversion,* and negatively to *Neuroticism*.

Hu and Pu [90] proposed a general algorithm for personality-based music recommender systems. The proposed algorithm relies on matching users' personalities with their musical preferences group to make recommendations. The authors defined these matchings based on the results reported by Rentfrow and Gosling [88]. They also explored whether users' knowledge about the domain affects their appreciation of the personality-based recommender. The results of a user study with 80 participants show that, overall, users enjoyed using personality assessments to get recommendations. Also, the results show that domain knowledge influences users' perception of the system. Novice users (i.e., users who are not familiar with music) generally appreciate personality-based RSs, and they had more positive perceptions than expert users.

Researchers have also found that personality is an important factor in Context-Aware Recommender Systems (CARS). To understand the impact of personality on emotion induction in different social circumstances during the consumption of movies, Odi´c et al. [91] conducted an experiment based on the CoMoDa dataset [92], a contextual movie recommender system dataset. Odi´c et al. aimed to answer the following question: "are different personality profiles influence the emotion induction when users are alone as opposed to when they are with a company while watching movies?". They observed that different social context (i.e., alone vs. with others) leads to different patterns of experienced emotions for *Extraversion, Agreeableness,* and Neuroticism. However, people with *Conscientiousness* and *Openness* personalities did not exhibit different patterns in their induced emotions.

Some researchers rely on personality to enhance recommendations diversity. Chen et al. [93], for example, investigated whether personality may impact the need for diversity. To do so, they conducted a user study with 181 participants. The study obtained users' selections of movies as well as their personality traits. It also considered two levels of diversity: one over each individual attribute of the items (such as the movie's director, country, or actor/actress). The other is over the combination of all attributes (named *overall diversity*). The results show a significant correlation between some personalities and users' diversity preferences. For instance, it shows that users with high *Neuroticism* are more inclined to choose diverse directors, while low *Agreeableness* users prefer diverse movie countries, and high *Conscientiousness* users prefer diverse movie's release time. In terms of overall diversity, less *Conscientious* people prefer a higher level of overall diversity.

Based on the findings of Chen et al.'s survey [93], Wu et al. [94] proposed a personality-based diversity adjusting approach for movie recommendation. They have incorporated personality into a content-based recommender system. The proposed algorithm begins by identifying the user's diversity need based on her personality traits. For example, according to the results of Chen et al. [93], low *Agreeableness* users prefer diverse movie countries. Therefore, if the current user possesses low *Agreeableness* and the "country" is her most important dimension, the recommended movie list will contain movies from different countries (i.e., movies with diverse countries).

Another recommendation issue that is investigated in the personality-based RS is group recommendation. Kompan et al. [95], for instance, used the FFM to model individual users as well as the relationship between group members. They adopted a graph-based approach to model group satisfaction. The authors also introduced the influence graph, which is a graph that represents the influence of users. Vertices represent users, and edges represent user influences (i.e., the influence presence between two users). Edges' weight can be calculated based on relationship, personality, and actual context. A small-scale user study is conducted to evaluate the approach. They concluded that the usage of personality-based group satisfaction recommender generally outperforms the standard group recommender.

These are the most related articles in the context of personality-based RS. All these studies and others demonstrate the benefits of including users' personalities in the recommendation process. However, the main limitation of these studies is that they consider

users' personalities in isolation of other factors, such as the application domain and persuasiveness. The next section presents those approaches that concern with the Persuasion of RSs.

2.4.2 Persuasion in Recommender Systems

In the RS-related literature, most of the studies that discuss persuasion rely on several constructs as measures of persuasion. These constructs include trust toward the RS, the cognitive effort required to get a recommendation, and the presentation of recommendations [96]. Such constructs deal with a recommender as a pure computerized system. In this thesis, however, we look at an RS as a social actor (i.e., a human). Thus, this section sheds light on approaches that adopted persuasive theories, from the social sciences, into the RSs in order to enhance the quality of recommendations and users' acceptance

According to Gkika [96], there is relatively limited research that examines the conditions under which an RS has a persuasive effect. However, the research in this direction is gaining increasing attention. As a result, some approaches have been introduced in various areas of RSs, as follows: Yu et al. [14] introduced the basic principles of Persuasive Product Recommendation Agents (PPRA). The article categorizes persuasive recommenders into three types based on their persuasiveness: *Neutral*, *Deceptive*, and *Persuasive*. Based on this classification, the authors distinguished persuasive recommender from neutral and deceptive ones. Also, the article classified persuasive techniques that can be used in PPRA into three types, based on two factors: First, the components that customers will judge to process persuasion, which includes the source of the message, the content of the message, medium or channel, and the audience. Second, how will persuasion be processed? To answer this question, the authors relied on the Elaboration Likelihood Model (ELM), a general attitude change theory. ELM suggests that there are two ways of persuasion: central (used for individuals who evaluate the content of the message carefully) and peripheral (used for less thoughtful individuals). Based on that, the authors proposed a framework to classify persuasion tactics in PPRA. The framework is presented as a 2*2 matrix that combines persuasion components (Source vs. Message) and persuasion routes (Central vs. Peripheral). For example, the first class of the framework is (Source, Peripheral) focuses on influencing customers' evaluation of the source (the recommender itself) through the

peripheral route. According to the authors, this article is a boot step towards a comprehensive understanding of persuasive recommenders. However, the article has no insights into the applicability of the framework.

One of the good candidates for persuasive recommendations is the route planning RS. These systems aim to persuade users to use recommended routes that are considered better than other routes. For instance, Bothos et al. [97] have implemented a route planning recommender system that aims to nudge the users toward following environmentally friendly routes. The system defines an information message based on a set of seven different persuasive strategies. It considers user-profiles and contextual elements. The user profile combines information about the user's travel history, travel habits, and persuasion attempts logged each time they are presented with a message. The proposed system is tested through Android-based smartphone users in the cities of Vienna and Dublin. The user's feedback demonstrates that the system was found interesting. However, the study could not detect any changes in users' behavior, which could be attributed to the persuasive messages.

Another persuasive route recommendation system is proposed by Sengvong & Bai [98]. The proposed recommender aims to facilitate transport use and support society to mitigate external costs, such as traffic pollution, congestion, and accidents. The system uses rewards as a persuasion technique to influence an individual's behavior to choose the least costly route. To do so, the authors have proposed an algorithm that can be used to balance two conflict parties with utility values and ranks. This algorithm is utilized for balancing the conflict between personal needs and social aims. Experimental results show that the proposed system can promote public-friendly commute options more effectively than the traditional recommendation system. However, the proposed system has ignored important factors, such as demographic data and social relationships among individual users. The study was also limited by a small sample of travel routes collected manually and synthetic user preferences.

Wu et al. [99] proposed the *GreenCommute*, which is also a persuasive recommender for public-friendly commute options. This study aims to handle the limitations of the study of Sengvong & Bai [98], previously mentioned. The proposed system aims at facilitating commuters to use public-friendly travel options to alleviate the external costs

in society. *GreenCommute* relies on a rewarding technique that takes users' influential power into consideration to balance the conflict between personal needs and social aims. *GreenCommute* also considers the impact of social pressure in providing recommendations and in calculating the reward values. Thus, users' influential degrees (i.e., their impacts on other users) in the social network is considered, such that users with higher influential degrees can be 'hired' as influential 'seeds' for promoting public-friendly options. The article concluded that *GreenCommute* can provide more efficient recommendations of public-friendly routes by rewarding individuals and utilizing social influence degrees through experimental results.

In the context of restaurant recommendations, Herse et al. [100] compared the effect of human's persuasion on decision-making compared to a robot and non-social machines (information kiosk). In particular, the authors hypothesized that: "a human is more persuasive than a robot, and a robot is more persuasive than an information kiosk." They conducted an experiment where192 participants were randomly assigned to the three different options (human, a humanoid robot, and information kiosk). A restaurant RS is implemented with similar interfaces for both the robot and the kiosk. Their preliminary results imply that embodiment type significantly affects the persuasiveness of RSs. This result suggests that human-like features of agents may support boosting RSs persuasiveness. The results were consistent with the first part of the hypothesis (i.e., a human was more persuasive than a robot), but, surprisingly, the robot was no more persuasive than the kiosk.

Persuasion has also been discussed in the domain of Health RS, where persuasion is used to change or quit non-healthy behavior or to admit or engage in a healthy routine. As an example, Pilloni et al. [101] discussed the persuasiveness of a health recommender system. They proposed an approach to monitor behavioral changes to predict athletes' loss of motivation that leads to stopping training. Upon the detection of such loss, the system refers the athlete to her coach, along with a concise explanation of the detected behavioral changes. Also, the system helps coaches by recommending persuasive techniques to persuade an athlete to continue training.

Another domain that is related to the health domain is online food shopping. For instance, Adaji et al. [102] studied the impact of Cialdini's principles on eCommerce shoppers based on their online shopping motivation. The article relied on a game-based

approach to identify users' susceptibility to the persuasive principles. In particular, a shopping game named *ShopRight* was developed, which simulates a retail store. It recommends the healthiest products to the players. It also combines the suggested product with a persuasive message. Players were asked to use the application for at least three rounds, where their responses to the persuasive messages are recorded. The authors classified the players based on their shopping motivation into four categories: *convenience shoppers*, *variety seekers*, *balanced buyers*, and *store-oriented* shoppers. Then, the change in participants' attitude, intention, self-efficacy, and perceived price of products are identified using pre and post-game surveys. The results show that persuasive principles influence shoppers differently based on their online shopping motivation.

An exploratory study on the outcomes of influence strategies in mobile application RS is conducted by Unal et al. [103]. They analyzed and compared the effect of visible and semi-visible persuasive messages. According to the authors, a semi-visible modality is used to avoid resistance, which is likely to happen when the persuasive message is presented obviously and visibly. The article also discussed users' compliance with these messages and how their compliance is affected by their persuadability[11]. A two-phase study has been conducted; first, a questionnaire-based study about users' behavior in a mobile environment and their susceptibility to persuasive strategies (the six influence strategies of Cialdini). Second, an experimental survey regarding visible and semi-visible persuasive messages. The study results provided evidence that the perception of semi-visible persuasive messages is considerably higher than visible messages. An important conclusion of this study is that the use of persuasive messages should be tackled cautiously. That is, the persuasion messages should be used in the correct context because using them incorrectly may negatively impact their benefits. Also, when using persuasion, we must be cautious and aware of the trust and privacy-related issues.

Several researchers have investigated the feasibility of using explanations as an implementation design of persuasive strategies in RSs. Gkika & Lekakos [13], [96] investigated the application of implementing Cialdini's influence principle on users' intention to

[11] According to the American Heritage Dictionary, "Persuadability" means "To cause (someone) to accept a point of view or to undertake a course of action by means of argument, reasoning, or entreaty." Retrieved December 14 2020 from https://www.thefreedictionary.com/persuadability

use a recommendation. The work was conducted in the context of a movie recommender. In the first part of the study, the authors suggested 30 different explanations, five for each influence strategy. Then, based on the opinions of 17 experts, they selected the best matching explanation for each strategy. In the second part of the study, they conducted a user study using a collaborative filtering movie RS. A total of 184 subjects participated in the study. The results demonstrate that, in general, incorporating a persuasive explanation to recommendations that are close to users' preferences will affect their behavior in terms of accepting/rejecting recommended items.

Another study conducted by Heras et al. [104] explored the use of argumentation (or explanation) as a persuasive technique in the context of Educational Recommender Systems (ERS). The proposed system recommends learning objects for students, taking into account students' profiles, preferences, and learning needs. Along with the recommendations, the system produces several types of explanations. The selection of these explanations is based on three factors, as follows: (1) students' learning profile (personal info, language, topic preferences, educational level, and learning style), (2) learning history (i.e., learning objectives that they already used), and (3) students' similarity to each other's. The results of a user study show that explanations can be used to improve recommenders' persuasion power. Particularly, 72% of students' ratings for the recommended learning objects are improved with the explanations.

Zanker and Schoberegger [105] also explored the persuasion potential of explanations. They conducted a supervised online user study that compares the persuasiveness of three different styles of explanations. Particularly, this study compares fact-based explanations (i.e., explanations of the form of separated facts "A, B, …") with argument style (i.e., explanations of the form "A and B, therefore C"), as well as with sentence-based explanations that need more cognitive effort to be understood. The study has been conducted in the scope of personalized route planning for hikers in the Alpine regions. One hundred thirty-six participants have completed the user study. The results indicated that an argumentative explanation style would influence users more than the two other styles. And Fact-based explanation has stronger persuasion power than sentence-style.

Hyan and Gretzel [106] reviewed the existing literature on the influence of source characteristics in the context of human-human, human-computer, and human-

recommender system interactions. The purpose of their study is to identify potential influence factors to create more credible and persuasive recommender systems. The authors concluded that many social cues that have been identified as influential in human-human and human-computer contexts have yet to be implemented and tested in recommender systems context to make the latter more persuasive. In particular, the authors suggested viewing recommender systems as *social actors* with socializing and persuasion capabilities.

Furthermore, Benbasat and Wang [16] concluded that users not only perceive recommender systems as social actors who possess human characteristics but also that these social perceptions have the potential to influence system evaluations. In particular, the experiment conducted by the authors showed that users perceived human characteristics (such as benevolence and integrity) when they interact with online recommendation systems. As a consequence, perceiving integrity while interacting with recommender systems is important to establish trust between users and RSs, which increase the persuasiveness of RSs, and in turn, increases the recommendation acceptance.

Other studies supported the need for a social focus in recommender system research and investigated the influences of system characteristics when users evaluate systems as well as recommendations [25], [107]-[109]. In particular, these studies emphasized on understanding recommender systems as communication sources to which theories developed for human-human communication apply. These approaches researched the impact of *source characteristics* on persuasion likelihood and outcomes, with a core assumption that credible sources are more effective persuaders.

As mentioned above, trust is considered a facet of the persuasive process that is considered out of the scope of this work. However, given the extensive trust literature associated with RSs, and because of its relation to persuasion, we will shed some light on it. Trust and persuasion are tightly related. According to Petty and Cacioppo [110], trust is inversely related to persuasion *intention*. That is, a source whose intent is to persuade is perceived as less trustworthy than one without persuasive intent. This relationship is confirmed in the context of RS by Unal et al. [103], who found that the perception of semi-visible persuasive messages is considerably higher than visible messages. The authors suggest that a semi-visible modality (which hides the intention to persuade) is used to avoid resistance, which is likely to happen when the persuasive message is presented obviously

and visibly. On the other side, a more trustworthy source is more persuasive. Bauernfeind [19] stated that building a trust relationship between users and RSs is essential to assist users in making a decision, and it increases users' intentions to adopt the RS [16]. To summarize, a trustworthy relationship increases the recommender's persuadability and, therefore, enhances users' acceptance [16]. However, persuasion should be used cautiously so as not to negatively affect trust.

2.4.3 The Combination of Personalization and Persuasion

The previous two sections explored the work that shows the applicability of personality-based RSs (section 2.4.1) and persuasion in RSs (section 2.4.2). The articles discussed in the previous two sections examine personalization and persuasion in isolation. However, Berkovsky et al. [111] suggested that both aspects (i.e., personalized and persuasive) technologies share the same goal, which is influencing users or affecting their interactions. This section investigates the articles that discuss the relationship between users' personality characteristics and Cialdini's principles of persuasion to personalize persuasive principles.

In this research direction, researchers consider various personality determinants. Oyibo et al. [112], for instance, investigated whether the personality of users who came from different cultural backgrounds (i.e., origins) would affect the extent to which they got persuaded by Cialdini's persuasive principles. The authors used "originality" as a personality determinant. Particularly, the study provided a comparative analysis of the Nigerians' susceptibility to the persuasive strategies compared to Canadians. A study of 248 total responses (88 Nigerians and 196 Canadians) was conducted. The results revealed that Nigerians' susceptibility is different from the Canadians for all strategies except for the *Commitment* strategy. In particular, *Authority* and *Scarcity* were found to be the most effective on Nigerians. On the other hand, *Reciprocity* and *Liking* were found the most effective on Canadians.

In another study, Oyibo et al. [113] conducted a study that aims to determine Africans' persuasion profiles. Nigeria is considered as a case study, and Cialdini's principles were considered for the persuasion profiles. The article also investigated if gender has any impact on the responses. The results of 88 responses show that Nigerians are susceptible to all principles but at different levels. Specifically, the level of susceptibility to the six

principles was ordered, from the highest to the lowest, as follows: *Commitment, Reciprocity, Authority, Liking, Consensus*, and *Scarcity*. The study also concluded that gender plays a role in Nigerians' responses, with males found to be more susceptible to *Commitment* and *Authority* than females. The article, however, involved a small sample, and the participants are all Nigerians.

Regarding the relationship between personality traits (namely, the FFM) and Cialdini's principles of persuasion, the research in this direction is relatively limited [114], especially in the area of RSs. To the best of our knowledge, only three prior studies have discussed this correlation. Gkika et al. [68] extended the study of Gkika & Lekakos [96] as follows: in addition to investigating the persuadability of the six influence strategies, the study discussed the effect of personality in the acceptance of recommendations. The BFI-44 has been used to infer users' preferences. The persuasive strategies have been incorporated into the systems as explanations (text messages) with the recommended items. The authors conducted a within-subject experiment, such that participants were exposed to six different explanations (one for each persuasive principle). After that, participants were asked to rate recommendations in order to evaluate which strategy, if any, influenced the user's intention to watch a movie. The results revealed that all of the six influence principles positively change users' acceptance of the recommendations. Particularly, *Authority* and *Social Proof* were found to be the most effective persuasive principle. In addition, the results indicated that users' personality impacts the effectiveness of persuasion approach.

Oyibo et al. [114] studied the relationship between the Big-Five personality traits and Cialdini's persuasive strategies. The authors conducted their study in Canada, with participants of Canadian origins. The study used a 32-item scale to measure how susceptible people are to Cialdini's principles. A total of 216 responses were considered in the study. The results revealed that *Conscientiousness*, *Agreeableness*, and *Openness* are the most consistent predictors of Cialdini's persuasive strategies. *Neuroticism* turns out to be the least predictor of the persuasive principles, as it only predicts *Social Proof*. Also, they noticed that none of the personality traits predicts *Scarcity* among Canadians.

Alkış and Temizel [115] also conducted a study similar to Oyibo et al. [114]. However, this study considered a different sample. That is, the study is conducted with 381 participants; all of them are Turkish undergraduate students. To collect data, the authors

conducted a structured questionnaire that comprised the Big Five Inventory (BFI) personality traits scale and the Susceptibility to Persuasion Strategies scale. The study found that personality traits are important in the selection of influence strategies. *Agreeable* people were found the most susceptible to Cialdini's principles compared to the other traits.

Kaptein [116]-[118] studied the impact that social influence strategies (such as Cialdini's principles) have on users of Ambient Intelligence (AmI) systems. The author found that there are substantial differences in individual susceptibility to different influence strategies. Therefore, the author posited that persuasive technologies should be adaptive to these differences. In particular, the author suggested that designers of persuasive technologies should use *persuasion profiles* to adapt to individual differences in users' responses to influence strategies. According to the author, a persuasion profile is defined as: "*a collection of expected effects of different influence strategies for a specific individual*"[119]. In addition, Kaptein [116] proposed a persuasive system that adapts its methods for behavioral change to the responses of its individual users. To achieve this, the author proposed to use persuasion profiles in an online commerce setting and evaluate its performance using Bayesian learning. The proposed approach is deemed feasible in terms of increasing revenues for the system that utilizes persuasion profiles.

2.4.4 Limitations of Closely-Related Work

The literature review revealed increasing attention towards the personality-based and persuasive RSs. The work in this area provided promising findings regarding the feasibility of adopting human-human theories to human-recommender communication. Nonetheless, personalization still has a huge untapped potential to maximize the impact of persuasive applications [111]. This section discusses the major gaps and limitations observed through the literature review, and we try to fill through this work, which can be summarized as follows:

- Despite the existing work in both personality-based and persuasive recommendations, most of the studies (as mentioned in sections 2.4.1 and 2.4.2) discussed these two concepts separately. These two concepts, however, share the same goal, which is influencing users [111]. Also, the study of Gkika et al. [68] has shown that users' personality affects the effectiveness of persuasive strategies. Thus, it is important

to consider both concepts when we design a persuasive RS. Chapter 3 investigates this issue.

- Section 2.4.3 explored the research that examines the correlation between Cialdini's principles and users' characteristics, with a focus on personality traits. We noticed that the work in this direction is still limited, especially in the recommendation domain. Up to our knowledge, three studies (Gkika et al. [68], Oyibo et al. [114], and Alkış, N., & Temizel [115]) are the only studies that have recently studied the relationship between Cialdini's principles and the FFM. Only one of these studies (Gkika et al. [68]) has addressed this topic in the context of RSs. Furthermore, these studies discussed the impact of users' personalities in isolation of other factors. Chapter 3 also explores this issue by examining the correlation between Cialdini's principles and both user characteristics and the context of the RS domain.
- Another issue related to section 2.4.3 is that most of the literature conducted generic studies (i.e., studies that are not designed to a particular application domain or context). These studies used general evaluation items to assess the susceptibility of users to persuasion principles. Oyibo et al. [113] suggested that the actual users' susceptibility to the persuasive principles may differ when implemented and evaluated for real persuasive application. This issue is taken into consideration in Chapter 3 by tailoring the persuasion questionnaire to the RS domain. This issue is also considered in Chapter 4. Specifically, we evaluate the effect of Cialdini's principles in a prototype for a real domain (persuasive RS).
- The literature raised increasing calls to the importance and the benefits of integrating persuasive features to recommendation systems. As it is discussed in section 2.4.2, most of the articles (such as [96], [100], and [103]) discussed the feasibility of persuasive recommenders. That is, they discuss the applicability of incorporating persuasive features to RSs. However, there is a lack of approaches that provide solutions to translate and integrate persuasiveness into RSs in a personalized and dynamic manner. Chapter 4 proposes a solution to this issue.

Table 2 summarizes the main differences between this thesis and the most related work. It is divided into three sections according to the categorization of the literature review.

Table 2 Major differences between previous work and this thesis

Article	Main Limitations and gaps
Personality-related (section 2.4.1)	• Does not consider persuasion (i.e., they discuss personality in isolation of persuasion). • Domain-specific solutions (e.g., [87], [88], and [94])
Persuasion-related (section 2.4.2)	• Does not consider personality traits (i.e., they discuss persuasion in isolation of personality effect). • Domain-specific solution (e.g., [97] - [101]) • Discuss the feasibility of persuasion, but lack the solutions to integrate persuasion (e.g., as [96], [100], and[103])
Personality-persuasion- related (section 2.4.3)	• Does not consider the FFM. (e.g., [112], and [113]) • Ignores some other factors, such as gender, age, culture, or context). (e.g., [112], [68]) • Is not designed for RS. (e.g., [112] - [115]) • Does not provide adaptive solutions.

2.5. Chapter Summary

This chapter discussed the concepts that are not familiar in the RS area, namely, the personality traits (Section 2.3.2) and persuasive strategies (Section 2.3.3). It also surveyed the research related to the proposed work. Particularly, Sections 2.4.1 discussed personality in RSs, and Section 2.4.2 discussed the persuasiveness in RS. Section 2.4.3 discusses the most related articles that study the relationship between personality and Cialdini's six principles.

The next chapter presents a user study about the susceptibility of RS users to Cialdini's principles and how their persuadability is affected by various factors. Based on the findings of this study, Chapter 4 proposes a personalized persuasive recommendation framework, which is an adaptive framework that aims to increase users' acceptance of recommendations by providing persuasive capabilities to the recommender.

Chapter 3. Impact of Persuasive Recommender Systems on Users

This chapter discusses the user study that I have conducted to examine the impact of Persuasive Recommender Systems (PRS) on different users. In particular, the purpose of the study is to diagnose the extent to which users of different characteristics get influenced by the different persuasive principles that a recommender system uses during the process of recommendation. The rest of this chapter is organized as follows: Section 3.1 discusses the motivation behind this study. Section 3.2 describes the study methodology. The data collection and analysis processes are presented in Section 3.3. Design guidelines and discussion are provided in Sections 3.4. Finally, Section 3.5 summarizes the chapter. Most of the content of this chapter were published in [120] and [121].

3.1. User Study Motivation

There is a wide range of persuasive principles introduced by the psychologists; among them are the six weapons of influence proposed by Cialdini [81] (as discussed in Section 2.3), which are among the most commonly used principles in the persuasive technology area. Researchers have recently started investigating how users respond to or interact with these six principles. However, despite the increasing research interest, there are still several gaps when it comes to recommender systems and persuasion (as mentioned in section2.3.3). These gaps can be highlighted as follows: (1) A limited number of researches studied the influence of persuasive principles on users while taking the users' characteristics into consideration. (2) The few studies that discussed the impact of users' characteristics neglected other important factors, such as the application domain (or the context). (3) Most of the studies (such as the ones conducted by Oyibo et al. [114] and Alkış et al. [115]) were generic and were not designed in particular for the RS area. And (4), the considered study samples were limited, in the sense that they either considered single culture

(participants from a single origin such as [115] and [114]), or considered a particular community (such as university students [68] and [115]), or a particular age range [115].

The user study conducted in this thesis aims to fill these gaps by investigating the susceptibility of RSs users towards the six weapons of influence and the extent to which users' *attributes* and the application *domain* affect this susceptibility. To achieve this, I conducted an online questionnaire that consists of two main parts: the *personality test* and the *persuasion test*. The persuasion test, in turn, is divided into three subparts: the eCommerce domain, the movie domain, and the generic (no domain) parts. In each subpart, participants were asked to rate six sentences that represent Cialdini's persuasive principles. More details about the design of this user study are discussed in Section 3.2

Our study is different from the previously mentioned studies in the following aspects: (1) It considers both users' characteristics in conjunction with context characteristics (i.e., application domain). (2) It is designed and deployed for a particular domain application, namely recommender systems. And (3) the study sample is heterogeneous in the sense that it involved participants from different continents, ages, and gender. Besides, this study provides numerical mappings between personality traits and Cialdini's persuasive principles, with the aim to provide a potentially better perception of the relationships between the two factors (Personality and Persuasive principles). This numerical mapping is not provided by any of the already existing studies. More details about this mapping are discussed in Chapter 4.

3.2. User Study Methodology

The design and evaluation of this user study were guided by the Goal, Question, Metric (GQM) framework of Basili [122]. Following this framework, I first identified the goals of the study. Then for each identified goal, I articulated a set of questions reflecting the goals and a set of measures to assess the achievement of the goals (more details are provided in Section 3.3). The main goals of this user study are to:

- **G1:** Evaluate the impact of different factors, namely: age, gender, culture, personality traits, and application domain, on RS users' susceptibility to persuasion principles.
- **G2:** Provide guidelines for designing persuasive recommender systems (PRS).

- **G3:** Evaluate the relationship or the correlation between users' personalities and the different persuasion principles. The purpose of this evaluation is to guide the development of the right PRS in real life to suit individual personalities in order to increase users' acceptance of recommendations. More details about achieving this goal are discussed in Chapter 4.

The following subsections explain the study design and other components, such as the ethics approval, participants, data collection and analysis, and the results.

3.2.1 Study Design

This study follows a within-subject study design; wherein, the same person tests all the conditions (i.e., all the sections of the questionnaire). The study starts by asking questions about the demographic background. Then, participants are required to answer two questionnaires. The first one (depicted in Appendix A) is about testing the personality traits according to the FFM, in which we deployed the BFI-44 items (discussed in section 2.3). The second questionnaire is the persuasion questionnaire (presented in Appendix B), which aims to test users' susceptibility to the six persuasive principles. This questionnaire is divided into three parts based on the application domain. The first part is the generic part, which does not reflect any particular domain. The second part is designed to represent the eCommerce RS domain, while the third part represents the movie domain. We used six different cues (sentences) in each part, where each cue represents one of Cialdini's persuasion principles. Participants were asked to rate these sentences based on a seven-step Likert scale from -1 to 5, with -1 being the least level of susceptibility (or negative impact) and 5 being the maximum level.

It is worthwhile to mention that the persuasive sentences that we used in the eCommerce and the movie parts of the persuadability questionnaire were inspired by, and designed based on, the study of Gkika et al.[68]. We adopted the same sentences for the movie RS, as suggested by the authors. Then, we followed the same convention to formulate persuasive statements for the eCommerce RS. Table 3 depicts the persuasive sentences that we used for both domains. The rationale behind adopting Gkika et al.'s methodology is that the persuasive statements used in their study are well-designed and thoroughly tested

by seventeen experts to choose the best matching sentence for each persuasion principle. In addition, Gkika et al.'s study is one of the few studies conducted for the RS domain, which is the same as the domain considered in this thesis.

Table 3 Persuasive sentences presented in the questionnaires.

Principle	**Persuasive Cue**
Authority	*eCommerce RS*: The recommended item won 3 prizes as the best-manufactured product! *Movie RS*: The recommended movie won 3 Oscars!
Commitment	*eCommerce RS*: This item belongs to the kind of items you usually buy. *Movie RS*: This movie belongs to the kind of movies you enjoy watching.
Social Proof	*eCommerce RS*: 87% of users rated the recommended item with 4 or 5 stars! *Movie RS*: 87% of users rated the recommended movie with 4 or 5 stars!
Liking	*eCommerce RS*: Your Facebook friends bought this item! *Movie RS*: Your Facebook friends like this movie!
Reciprocity	*eCommerce RS*: A friend of you who bought the item you suggested to him/her in the past recommends you this item! *Movie RS*: A Facebook friend, who saw the movie you suggested to him/her in the past, recommends this movie!
Scarcity	*eCommerce RS*: The recommended item will be available for two months only! *Movie RS*: The recommended movie will be available for two months only!

The questionnaires have a Likert scale, multiple choices, and open-ended questions. The Likert questions have a seven-step scale, ranging from -1 to 5. Each degree in the scale represents the extent to which a statement influences users' decisions. For instance, whether to purchase an item or to watch a movie. The indications of the seven scale ratings are as follows: 1 to 5 ratings are scaled from very low to very high positive effect, zero (0) is the

neutral value, which indicates that there is no impact on the decision, and (-1) indicates a negative impact. The (-1) rating is included in the questionnaire because some of the statements may cause resistance towards the recommendation. For instance, Kaptein and Eckles [123] found that persuasive strategies may not achieve their goals to influence users. The authors stated that a negative response might, indeed, be generated if the user receives an inappropriate message. For example, a *Scarcity* strategy may have a negative impact on a user who does not like to decide under stress.

The questionnaire was available online using *SurveyMonkey*[12], a well-known platform that provides online services to create and share surveys, collect responses, and analyze the results. *SurveyMonkey* is a secure, powerful, and intuitive website. Also, it offers a wide range of functionalities and services. Besides, the University of Ottawa has a university-wide license for the enterprise version of the *SurveyMonkey*. By this license, most of the services can be accessed and used by the University of Ottawa staff for free.

Participants were asked to complete the questionnaire only once (i.e., the questionnaire can be completed in one session). The session takes approximately 15 to 20 minutes. The time factor was, sometimes, an obstacle towards the completion of the survey. To mitigate this issue, the questionnaire was made available with no restrictions on the time of accessibility, and the participants were given a choice to complete the questionnaire at their own convenience.

3.2.2 Ethics Approval

For this study to be compliant with the Canadian ethical standards, an application was submitted to the Research Ethics Board (REB) at the University of Ottawa. The REB helps ensure that this research meets the highest ethical standards and that the greatest protection is provided to participants who serve as research subjects. This research was approved by the REB. The ethics approval certificate is included in Appendix C.

[12] https://www.surveymonkey.com/

3.2.3 Participants

Participants were recruited or invited to participate by different means of communication, including email-based invitation letters and paper-based posters. Our study respects participants' privacy, such that none of their information (e.g., names and email addresses) are included or posted with the study. A total of 329 participants responded to the study. After filtering the responses (based on the exclusion criteria described next), we retained a total of 279 responses. Table 4 summarizes the demographic information of the participants.

Table 4 Participants' demographic information (N=279)

Subject	**Category**	**Count**	**[Percentage]**
Age	16-25	53	[19%]
	26-35	129	[46%]
	36-45	59	[21%]
	46+	38	[14%]
Gender	Male	177	[63%]
	Female	98	[35%]
	Undefined	4	[2%]
Continent	Asia	75	[27%]
	Europe	13	[5%]
	North America	186	[67%]
	South America	2	[1%]
	Australia (Oceania)	3	[1%]

3.3. Data Collection, Analysis, and Results

According to the Goal, Question, Metric (GQM) framework, which guided our evaluation, we conducted the analysis with several questions in mind to achieve the goals stated in Section 3.2. These questions are as follows:

- ***Q1***: *How do RS users' characteristics (such as age, gender, and culture) affect their responses to Cialdini's persuasion principles?*

- ***Q2:*** *How do RS users' personality traits affect their responses to Cialdini's principles of persuasion?*
- ***Q3:*** *Is RS users' susceptibility to Cialdini's principles affected by the context in which the RS is deployed?*
- ***Q4***: *To what extent is a user's susceptibility to different persuasion principles affected by the combination of the user's characteristics and the recommender's application domain?*

The mean ratings collected from the study are used as metrics to analyze the above questions. The answers to the study questions are analyzed as follows:

- Demographic-related information is used to infer the impact of the user characteristics (i.e., age, gender, and culture) on the extent to which users got influenced by a particular persuasion strategy (to answer Q1 and Q4).
- Answers to the Likert scale questions associated with the personality test are used to infer relationships between users' personality types and persuasive principles (to answer questions Q2 and Q4).
- Dividing the persuasion questionnaire into two parts (representing eCommerce and Movie domains) is considered to answer Q3.
- Open-ended questions, where participants can write additional comments, are used to investigate unexplored insights on how users perceive different persuasive strategies.

To conduct statistical analysis of data, we used an open-source tool called *JASP* [124]. JASP stands for **J**effrey's **A**mazing **S**tatistics **P**rogram in recognition of the pioneer of Bayesian inference Sir Harold Jeffreys. It is a cross-platform and free statistics package. JASP helps in conducting statistical analyses through its easy-to-use GUI.

To determine whether the differences between groups are statistically significant, we performed the ***An****alysis* ***of*** ***Va****riance (ANOVA).* ANOVA is used to analyze the differences between the means of multiple groups in a sample. In this study, groups are defined based on multiple factors, such as gender, age, and culture. According to Gelman [125], ANOVA is an important exploratory and confirmatory data analysis method. For ANOVA analysis, the persuasive principles were included as the dependent variable, while other

factors (such as gender and age) were included as independent (or between-subject) variables. The significance level (α) is set to be 0.05 for all ANOVA analyses, such that a *P-value* value less than (0.05) is considered significant.

In addition to the ANOVA analysis, we also calculated the Effect Size [126], which shows whether an intervention or experimental manipulation has an effect greater than zero or how big the effect is. Reporting the effect size is useful for various reasons [126]. First, it presents the magnitude of the reported effects in a standardized metric. Such standardized effect sizes help to communicate the practical significance of the results. Second, effect size compares standardized effect sizes across studies, which helps meta-analytic conclusions. Third, previously reported effect sizes (i.e., from previous studies) could feed the design of a new study. For instance, previous analysis can provide an indication of the average sample size needed for a new study. We deployed the *Cohen's d*, one of the most widely used effect sizes [127].

It is worth mentioning here that the focus of these analyses presented in this chapter is to solely report on our findings regarding the considered factors, more than to provide explanations or justifications of each factor's impact. Based on our findings, we aim to develop general guidelines that could help tailor the development of RSs, with the user-related factors taken into considerations.

3.3.1 Data Filtering

To ensure the reliability of the results, the responses are filtered based on different measures. First, I removed incomplete responses (where some questions are left unanswered). Second, I considered the questionnaire's completion time, which gives a true indication about whether participants have indeed read/answer the entire question or not. Based on a pilot study, I noticed that participants, on average, need about 20 minutes to complete the questionnaire fully. Hence, I discarded responses that took less than 10 minutes. Third, some participants simply do not put in the effort required to respond accurately or thoughtfully to all questions asked. The data collected from such responses are known as careless or inattentive data [128]. To detect and remove these responses, I used the *attention questions technique*. Attention questions (or items) are defined as "Items placed in scale with explicit correct response" [128]. Particularly, I injected some irrelevant

questions (i.e., questions that are not related to the context of the questionnaire) and clearly indicated what the participant should provide as an answer to this question. For instance, I associated the following text with one of the attention questions: "*You must give this question the rate 3*). Responses from participants who got the attention questions incorrect were also discarded.

3.3.2 Descriptive Analysis

To evaluate RS users' susceptibility to the six persuasive principles, I computed the overall mean rating (μ) for each principle. The overall mean rating is the average of all ratings received from the whole sample. The higher the mean rating, the more influence the principle. In general, participants perceived all principles as persuasive. This can be inferred from Figure 5, where all mean values are greater than the neutral value (i.e., Zero), which indicates that all strategies are influential. It can also be implied from the figure that participants perceived *Social Proof* as the most influential strategy (with μ= 3.03), followed by *Commitment* (μ = 3.0) and *Reciprocity* (μ = 2.96). On the other hand, *Scarcity* is perceived as the least influential strategy (μ = 2.12). These observations indicate that the majority of participants believe that any of the six persuasive principles can influence their decisions to some extent. This is in line with the conclusion of Gkika and Lekakos [13], who stated that accompanying persuasive explanations with a recommendation can increase users' acceptance of that recommendation.

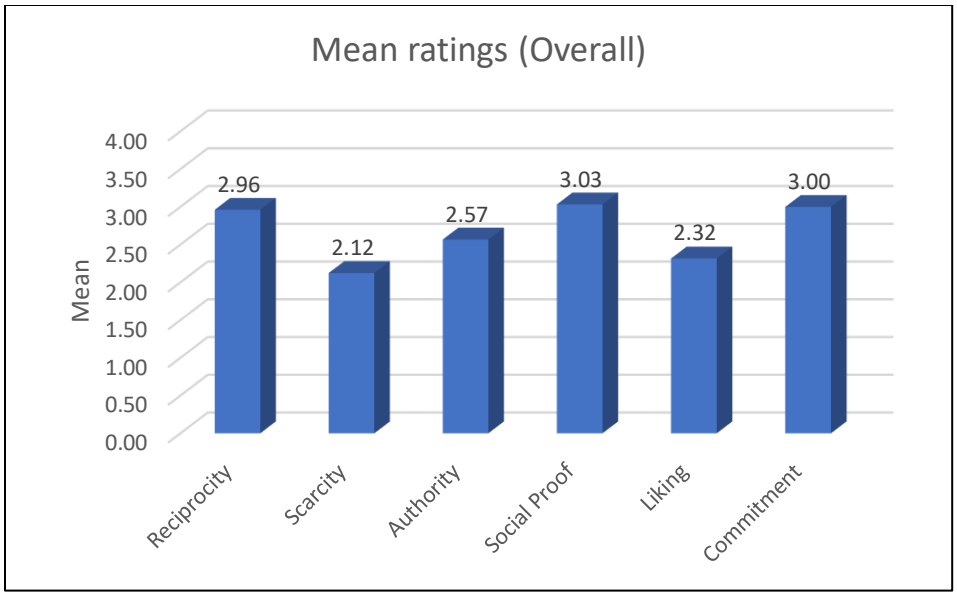


Figure 5 Mean rating values for the whole sample

Despite the observations mentioned above, which are generic by nature, the particular question arises here: *how the responses of RS users differ based on various factors, such as user-related factors and the context*? The two subsequent sections aim to answer this question.

3.3.3 User-related, Context-independent Analyses

This section focuses on the impact of user-related aspects/characteristics and how they affect users' susceptibility to the six weapons of influence. It discusses these factors in isolation of other aspects. Particularly, this section investigates how RS users differ in their reactions to the six persuasion principles based on four aspects: gender, age, culture, and personality traits, *without* considering the context at which users receive recommendations.

The Impact of Gender

The first aspect examined in this analysis is the users' gender. Figure 6 compares the mean ratings of females compared to males. The figure shows that *Reciprocity* is the most effective principle for both genders. On the other hand, *Scarcity* is the least effective for males, and *Authority* is the least effective for females. The figure also shows that males and

females responded differently to all principles. However, the difference is marginal in regard to *Reciprocity* and *Scarcity*.

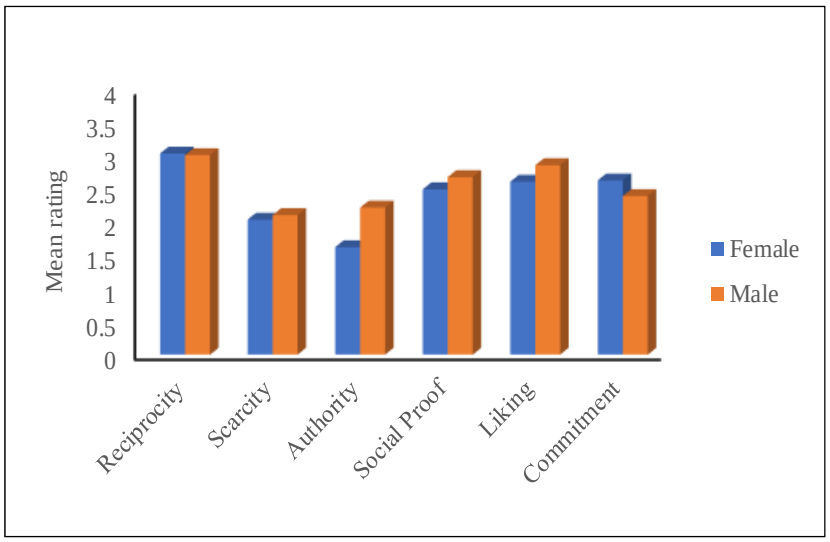


Figure 6 Mean rating based on gender (Female:35%, Male: 65%)

The significance analyses (ANOVA) reveal that the difference is significant with respect to the *Authority* principle, with males being more persuadable than females. *Cohen's P* shows that the effect size in regard to the Authority is close to moderate (*Cohen's P* = 0.4). These results are, to a high extent, in line with the findings of Oyibo et al. [113], who did not find gender differences with respect to *Reciprocity*, *Liking*, *Scarcity*, and *Social Proof*. In addition, the authors found a significant difference between males and females in terms of *Authority*, where males found to be more persuadable. Vargheese et al. [83] also found that the effectiveness of the persuasive principles does not differ based on the gender of the participants. The details of the ANOVA results and the effect size for this section and all subsequent sections are presented in Appendix D.

The Impact of Age

This section investigates the effect of users' age. Following the methodology of Abdullahi et al. [129], I defined four age groups as follows: 16-25, 26-35, 36-45, and older than 45 years. Figure 7 shows the average ratings for the six persuasion principles for each age group. The figure shows that participants of all ages agree that *Reciprocity* is the most influential principle. It also shows some differences between the age groups. *Authority* is found to be the least influential for all groups, except the third group (ages 36-45), where

Scarcity is found to be the least influential. Overall, younger users (ages 16-35) were found to be more susceptible to all principles except *Commitment*.

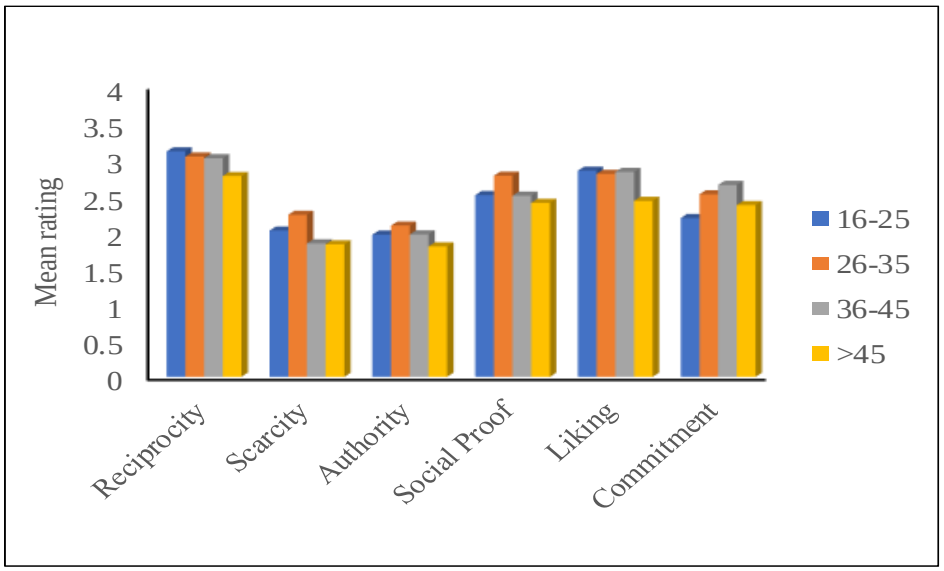


Figure 7 Mean rating based on age (16-25: 20%, 25-36: 40%, 36-45: 22%, >45: 18%)

The significance analysis shows that there is no significant difference between users of different ages and how they got affected by all persuasion principles. This finding is in line with a recent study conducted by Oyibo et al. [130], who found that the differences between younger and older people are more significant in collectivist[13] cultures than individualist[14] cultures. Oyibo et al.'s findings support the results illustrated in Figure 7 since the majority of participants in this study are from individualist continents (as shown in Table 4). Also, a recent study by Vargheese et al. [83] confirms these findings. Vargheese et al. stated that there is no actual relationship between the effectiveness of persuasive strategies and the ages of participants. In addition, studies in social sciences found that people's personalities are known as relatively stable. That is, they are stable throughout individuals' lives, with some slight exceptions in the facets rather than the essential traits [131]. Hence, people's decisions do not extensively change from one age to another.

[13] A *Collectivist culture* (e.g., China, and Japan) is one that emphasizes the needs of a group or a community over the individual.
[14] An *Individualist cultures* (e.g., United States and Western Europe) emphasizes personal achievements regardless of the expense of group goals.

The Impact of Culture

To study the impact of different cultures on users' susceptibility to persuasion principles, responses are categorized based on the continents. As shown in Table 4, where I received responses from five continents. However, since the number of participants from each of South America and Australia is low (two and three participants, respectively), I omitted these responses from all culture-based analysis.

Figure 8 depicts the mean ratings of the three continents for the six principles. It shows that participants from all continents are differently susceptible to all principles, where the differences between the mean ratings are noticeable. According to the figure, Asians are the most susceptible to all principles except the *Liking* principle. On the other hand, Europeans are the least susceptible to *Reciprocity*, *Scarcity*, *Authority*, and *Social Proof*, while North Americans are the least susceptible to *Liking* and *Commitment*. Hence, it could be implied that Asian participants were the easiest to be persuaded, followed by North Americans, and Europeans.

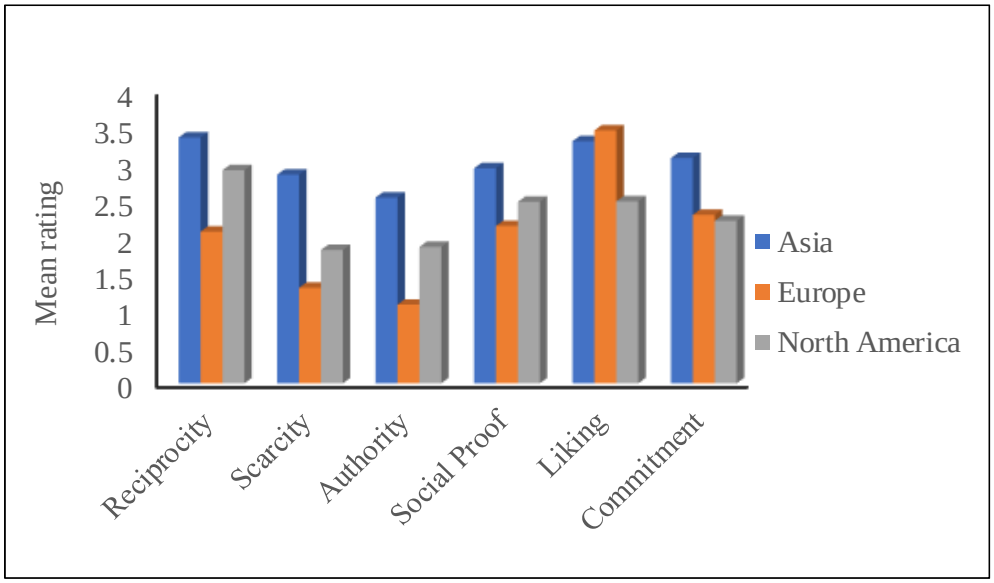


Figure 8 Mean rating based on the culture (Asia: 27%, Europe: 6%, N. America: 67%)

According to the ANOVA significance test, there is a significant difference between the three groups in terms of all principles. All details about ANOVA results and the effect size are presented in Appendix D (Table 18, and Table 19, respectively). According to Oyibo et al. [130], culture has an impact on the difference between persons in terms of their

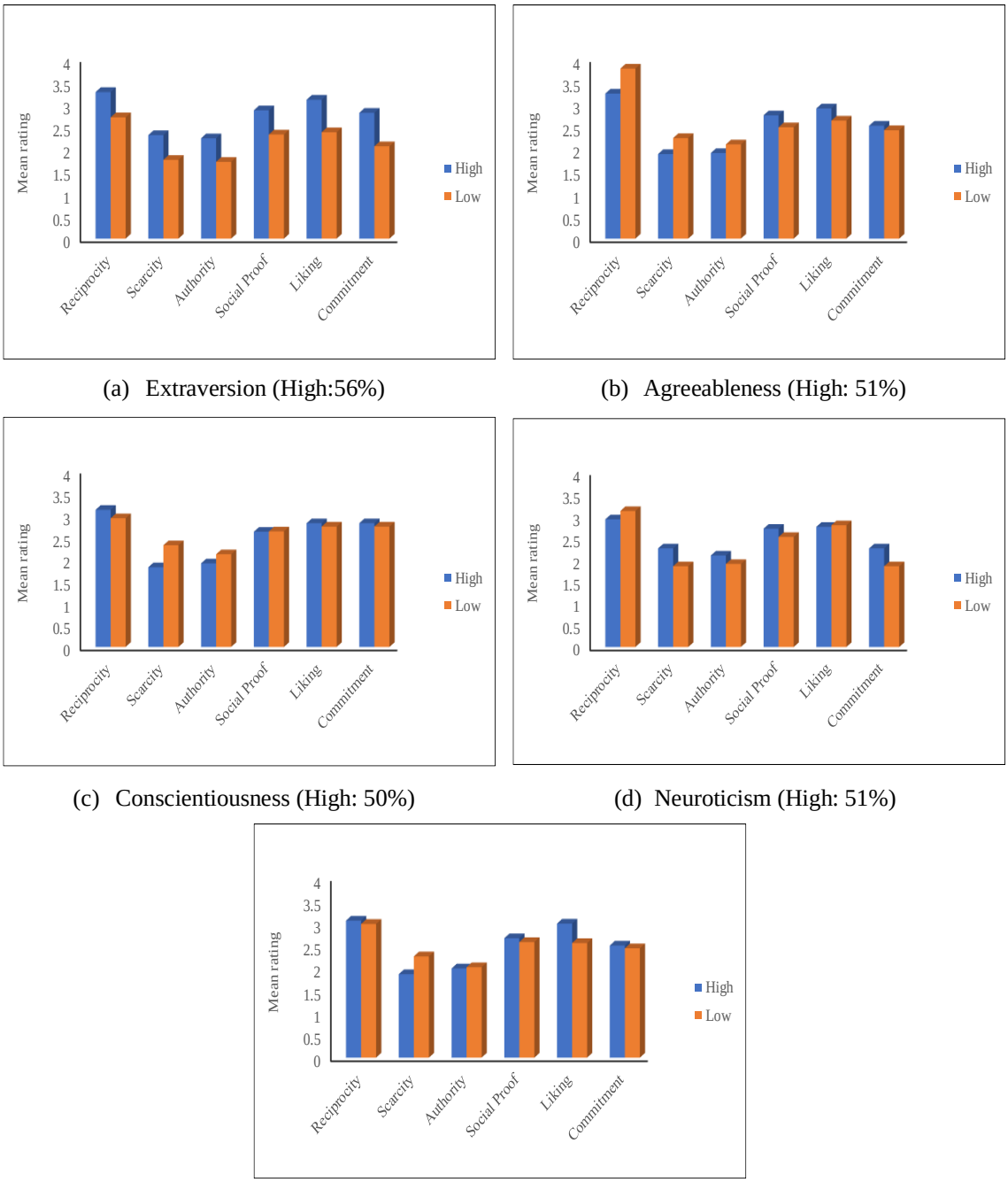
responses to persuasive strategies. This finding is consistent with the results presented in this section.

The Impact of Personality Traits

This section discusses the relationship between the personality traits of RS users and the six persuasive principles. Figure 9 depicts the mean ratings of the persuasive principles grouped based on each of the five personality traits (as discussed in Section 2.3). In this analysis, participants were categorized based on the strength of their corresponding personality traits into two groups: *High* and *Low*. High (respectively Low) means that the participant's personality falls in the high (respectively low) end of the personality traits spectrum (where each personality trait can be considered as a spectrum graduating from a high degree to a low degree of that trait). In this study, a person *P* is considered high in a personality trait *PT* if *P*'s score for trait *PT* is greater than the average score of all participants for that trait. Otherwise, *P* is considered to be low in that *PT* trait.

Figure 9 (a-e) show that the mean ratings for each of the five personality traits are larger than the neutral value (Zero), which means that all personalities are susceptible to the six persuasion principles. The degree of susceptibility, however, varies from one personality trait to another and also varies within the same personality trait, depending on the level of that trait (i.e., high or low). The figure also shows that *Reciprocity* is the most influential principle on all personality types.

Figure 9-(a) compares the participants' degree of persuasion based on their level of *Extraversion* trait. Participants who are highly extraversion are called *extroverts*, while low-level extroversion participants are called *introverts* [132]. According to the figure, *extroverted* people are more susceptible to all persuasive principles than *introverted* people, with *Reciprocity* being the most influential principle and *Authority* being the least influential one. The significance test indicates that both extraversion groups (i.e., High and Low) are significantly different (with $P < 0.05$) in their responses to all of the six persuasion principles. Also, the effect size is moderate for all principles (as presented in Table 21).



(a) Extraversion (High:56%)

(b) Agreeableness (High: 51%)

(c) Conscientiousness (High: 50%)

(d) Neuroticism (High: 51%)

(e) Openness (High: 50%)

Figure 9 Mean rating based on the personality traits

Figure 9-(b) reports on the results related to the *Agreeableness* trait. The figure shows that *Scarcity* is the least influential principle for participants who are high in *Agreeableness* (known as *Agreeable* persons), while *Authority* is the least influential principle for those who are low in *Agreeableness* (known as *Hostile* persons). Moreover, *Reciprocity*,

Scarcity, and *Authority* principles affect *Hostile* people more than *Agreeable* people. The significance test shows that both *Agreeableness* groups are significantly different in their responses to *Reciprocity* and *Scarcity*, where the effect size is found to be higher in regard to reciprocity (0.33) comparing to Scarcity (0.22)

Regarding *Conscientiousness* trait, Figure 9-(c) shows that *Reciprocity* is the most effective principle for both groups, while *Scarcity* is the least influential principle for people who are high in *Conscientiousness* and *Authority* is the least influential principle for people who are low in *Conscientiousness*. It can also be noticed from the figure that the differences are marginal, especially in the term of *Social Proof, Liking*, and *Commitment* principles. However, the difference is significant in terms of *Scarcity* with low effect size (0.3).

The Neuroticism personality trait analyses are illustrated in Figure 9-(d), which indicates that *Neurotic* persons (i.e., high in *Neuroticism*) are influenced the least by the *Authority* principle. On the other hand, *Stable* persons (i.e., low in *Neuroticism*) are influenced the least by *Scarcity* and *Commitment* principles. The figure also shows that Neurotic participants are more vulnerable to *Scarcity*, *Authority*, *Social Proof*, and *Commitment* than Stable participants. According to the ANOVA significance test, the responses of the two groups are significantly different in regard to Scarcity and Commitment ($P = 0.04$) with a small effect size for all principles.

Finally, Figure 9-(e) depicts the mean rating regarding the *Openness* trait. It can be noted from the figure that *Scarcity* and *Authority* are the least persuasive principles for *Open* persons (i.e., high in *Openness*) and *Closed* persons (i.e., low in *Openness*), respectively. Moreover, *Open* people are more susceptible to *Reciprocity*, *Social Proof*, *Liking*, and *Commitment* principles. The difference between both groups is minor in regards to *Reciprocity*, *Authority*, *Social Proof*, and *Commitment*, but they are significant in terms of *Scarcity* (with $P = 0.48$, *Cohen's* $P = 0.24$) and *Liking* ($P = 0.012$, *Cohen's* P=0.31).

To recap, the results provided in this section indicate that the impact of Cialdini's persuasion principles on RS users varies from one person to another based on users' personal characteristics and traits. The next section investigates another factor, namely the *context* (or the application domain). It discusses the impact of the context as a sole factor

and the role it plays on the persuasion process when it is combined with the aforementioned user-related factors.

3.3.4 Context-dependent Analysis

As mentioned in Section 2.3, the Communication-Persuasion Paradigm states that user-related aspects are not the only factors that affect the influence of a message. Other factors, such as the context, are also important. This section investigates the effect of the context on RS users' acceptance of Cialdini's six weapons of influence. The context is represented here by the *application domain*. Two recommendation domains were considered in this study, namely a movie RS and an eCommerce RS. The section begins by discussing the effect of the application domain in isolation of other factors. Then, it investigates the impact of users' characteristics (discussed in Section 3.3.3) when the application domain is taken into account.

The Impact of Application Domain as a Sole Factor

This section reports on how participants rated each persuasion principle in the eCommerce RS domain and the movie RS domain. Figure 10 depicts the mean ratings for each principle in both domains. It shows that all principles have an influence on users in both domains, as indicated by mean ratings being greater than zero. The figure also shows that *Social Proof* and *Reciprocity* are the most and the second most influential principles in the e-commerce domain, while *liking* is the least influential principle. On the other hand, the same users perceived *Commitment* and *Social Proof* as the most and the second most influential principles in the movie RS, and *Scarcity* as the least influential. From another perspective, the figure indicates that movie RS users are more vulnerable to persuasive principles than eCommerce RS users (with four out of six principles being more influential).

The significance analysis and the effect sizes are presented in Table 22 (Appendix D). The table shows that the responses were significantly different (with $p<0.001$) in terms of three principles, namely *Reciprocity*, *Liking*, and *Commitment*. This means that the application domain has a significant impact on users' susceptibility to these persuasive

principles. Overall, the data shows that the application domain plays a role in affecting users' vulnerability to the six principles.

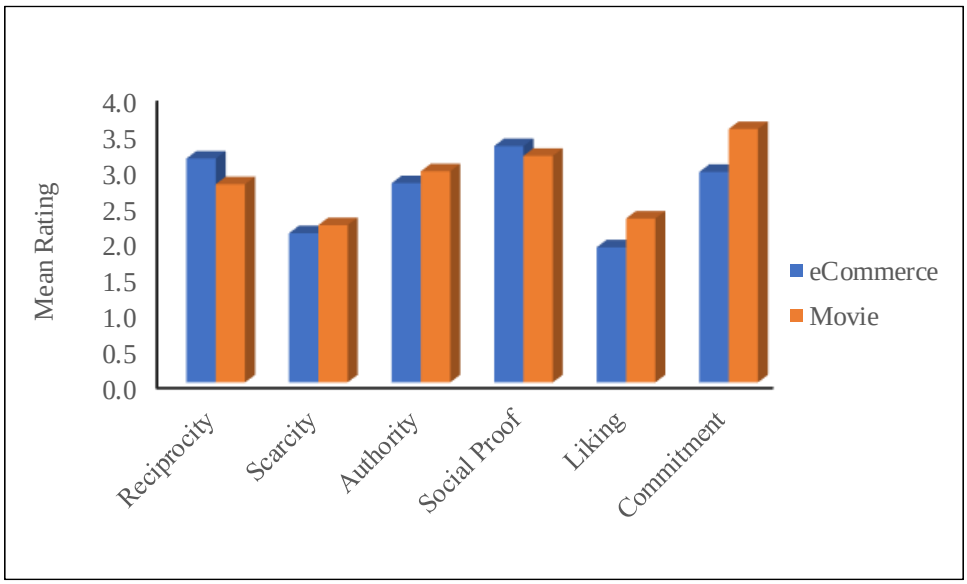


Figure 10 Mean rating based on the application domain

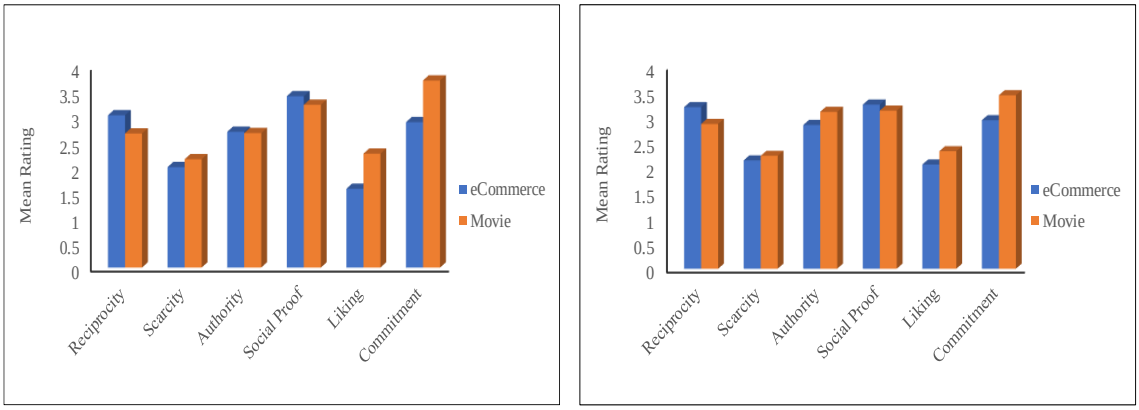
The Impact of User-related Aspects in Combination with Application Domain

This section reports on a study similar to the one discussed in Section 3.3.3, but this time with taking the application domain (or context) into consideration. The purpose of these analyses is to examine how users' susceptibility to persuasion principles is affected by user-related aspects (i.e., gender, age, culture, and personality traits) while taking into consideration the impact of the context.

The Impact of Gender and Context

This section presents the results about how the responses of males and females vary from one domain to the other. Figure 11 depicts the mean ratings for the six principles for both gender groups. Figure 11-(a) indicates that females perceived *Reciprocity* and *Social Proof* as more persuasive in the eCommerce domain, while they perceived *Scarcity*, *Liking*, and *Commitments* as more persuasive in the movie domain. The figure also shows that, in the eCommerce domain, *Social Proof* is the most influential principle for females, while *Liking* is the least influential principle. In the movie domain, however, the most influential principle is *Commitment*, whereas *Scarcity* is the least influential for females.

Regarding male participants, Figure 11-(b) shows that *Reciprocity* and *Social Proof* are more influential in eCommerce than in the movie domain, while the other four principles are more influential in the movie domain. In the eCommerce domain, males perceived *Social Proof* as the most influential principle and *Liking* as the least influential. However, males perceived *Commitment* and *Scarcity* as the most and the least influential principles in the movie domain, respectively.



(a) Female (b) Male

Figure 11 Mean rating based on gender for both application domain

An important observation from Figure 11 is that the order of the influence of persuasion principles (from the most influential to the least influential), which is also known as the *persuasion profile* [112], are different across domains for the same gender group. Based on all of the above results, it can be deduced that the responses of females and males are affected by the application domain.

The significance analysis shows that the difference between users' responses in different domains is statistically significant with respect to the *Reciprocity*, *Liking*, and *Commitment*, where *Reciprocity* is more influential in the eCommerce domain, while *Liking* and *Commitment* are more influential in the Movie domain. This observation is true for both males and females. More detailed results are depicted in Appendix D (Table 23 and Table 24). The *Cohen's P* values (Table 25) show moderate effect sizes in terms of *Liking* and *Commitment*.

The Impact of Age and Context

This section discusses how RS domains affect users' responses from different age groups to the six weapons of influence. Figure 12 depicts the overall mean ratings of each principle for each age group based on the application domains. The figure contains four charts; each one summarizes the results regarding one age group. As the figure shows, the six persuasive principles are all influential in both domains (with mean values > 0) for all ages. Also, *Reciprocity* is perceived by all ages as more influential in the eCommerce domain. While *Liking* and *Commitment* are perceived as more influential in the movie domain for all ages.

The figure also indicates that participants from all age groups perceived *Social Proof* as the most influential principle and *Liking* as the least influential in the eCommerce domain. On the other side, in the movie domain, *Commitment* is perceived as the most influential principle for all ages, while *Liking* is the least influential for groups (16-25, and 36-45), and *Scarcity* is the least influential for the others.

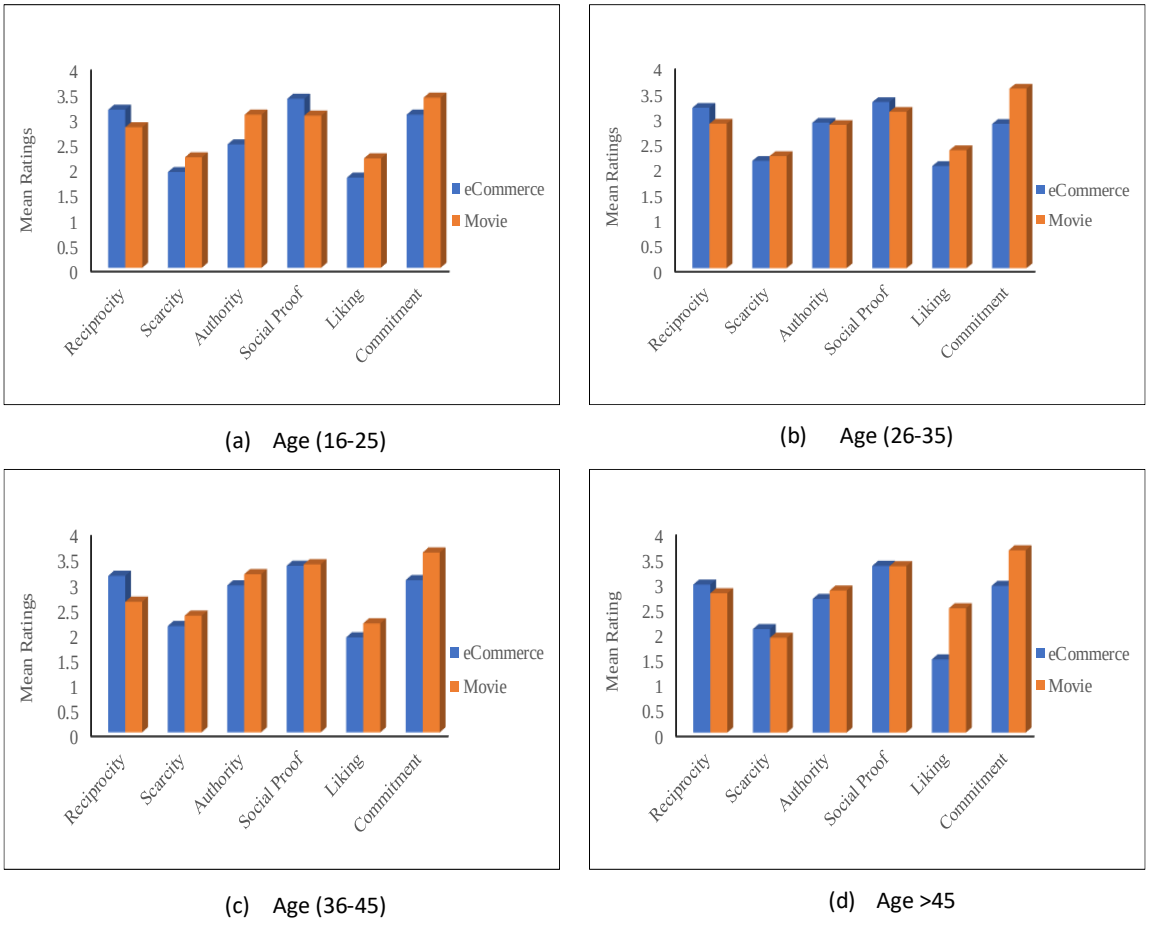


Figure 12 Mean rating based on the age and the application domain

From Figure 12, we can also infer that the persuasion profiles for each age group are different from one domain to the other. For instance, Figure 12-(b) shows that the persuasion profile of participants from the second group (i.e., age 26 to 35) in the eCommerce domain is (*Social Proof, Reciprocity, Authority, Commitment, Scarcity,* and *Liking*). However, the persuasion profile for the same group in the movie domain is (*Commitment, Social Proof, Reciprocity, Authority, Liking, Scarcity*), which quite different from the eCommerce persuasion profile. The above inference is true for all cases, where the persuasion profiles are different to a high extent from one domain to the other.

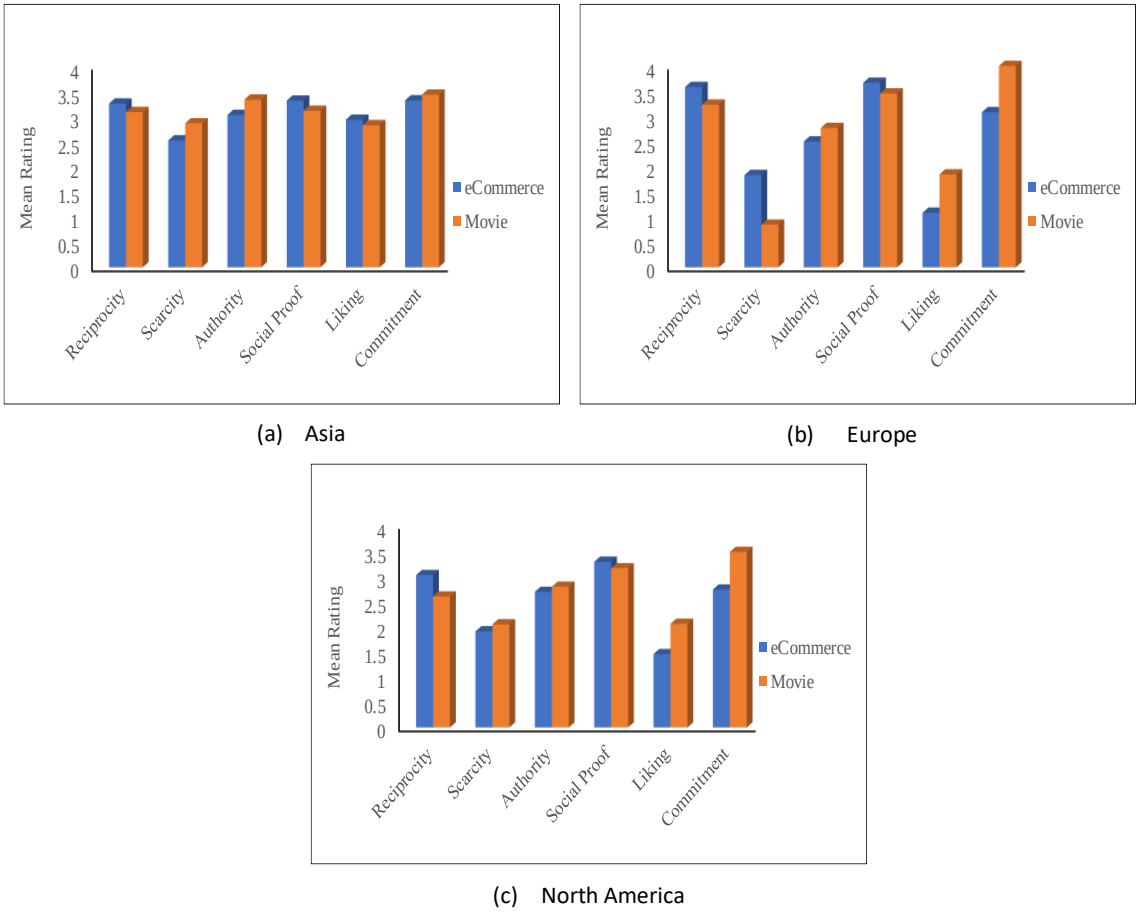
The ANOVA results (presented in Appendix D) demonstrate that some of these differences are significant. First, the responses of users from the first age group (16-25) were significantly different in terms of *Reciprocity, Authority,* and *Commitment* with small effect size. *Reciprocity* is found more influential in the eCommerce domain, while *Authority* and *Commitment* are more influential in the Movie domain. Second, users of ages (26-35) responded totally differently to *Reciprocity, Liking,* and *Commitment*, with small effect size in terms of *Reciprocity* and *Liking* and medium effect size in terms of *Commitment*. Third, the responses of the third age group (36-45) were significantly different with regards to *Reciprocity* and *Commitment*., with medium effect size for both Finally, older users (45 years old and more) responded significantly differently to the *Liking* and *Commitment* principles, medium effect size for *Commitment* and high effect size for *Liking.* Complete ANOVA results and Effect sizes are presented in Table 26 - Table 30).

Overall, Figure 12 shows that the susceptibility of users from the same age group to persuasion principles varies from one domain to the other. The most important observation in this context is that users of the same age respond differently to the persuasive principles if they interact with them in different contexts.

The Impact of Culture and Context

The mean ratings of the six principles based on the culture and context are depicted in Figure 13. The first observation from this figure is that people from the three continents are differently influenced by the six principles of persuasion. *Commitment* is the most influential for Asians in both domains, while *Scarcity* and *Liking* are the least influential in the eCommerce and the Movie domains, respectively. For European and Northern

Americans, *Social proof* and *Commitment* are the most influential in eCommerce and Movie domains, respectively, while *Liking* and *Scarcity* are the least influential in the eCommerce and the Movie domains, respectively.



(a) Asia (b) Europe

(c) North America

Figure 13 Mean rating based on the culture and the application domain

Figure 13 also indicates that the persuasion profiles of participants from one culture in the eCommerce domain are different, to a very high extent, from their persuasion profile in the movie domain. Particularly, persuasion profiles for North Americans are different, where none of the principles occupies the same order in both profiles. The profiles of the Europeans are different except that *Authority* is the fourth most influential in both profiles. Asians also have different profiles except for *Social Proof*, which is the third most influential in both domains.

The ANOVA analyses and the effect sizes (Appendix D, Table 31 - Table 34) show that Asians are the least affected by the context. The differences in their responses to the persuasive principles were significant for the *Authority* principle only, while the effect sizes

were small for all principles. The Europeans responses were significantly different for both *Scarcity* and *Commitment* principles, with medium and large effect size, respectively. Finally, the responses of the Northern Americans were significantly different in terms of three principles, which are *Reciprocity, Liking,* and *Commitment*, with medium effect sizes for Liking and Commitment and small effect sizes for the others.

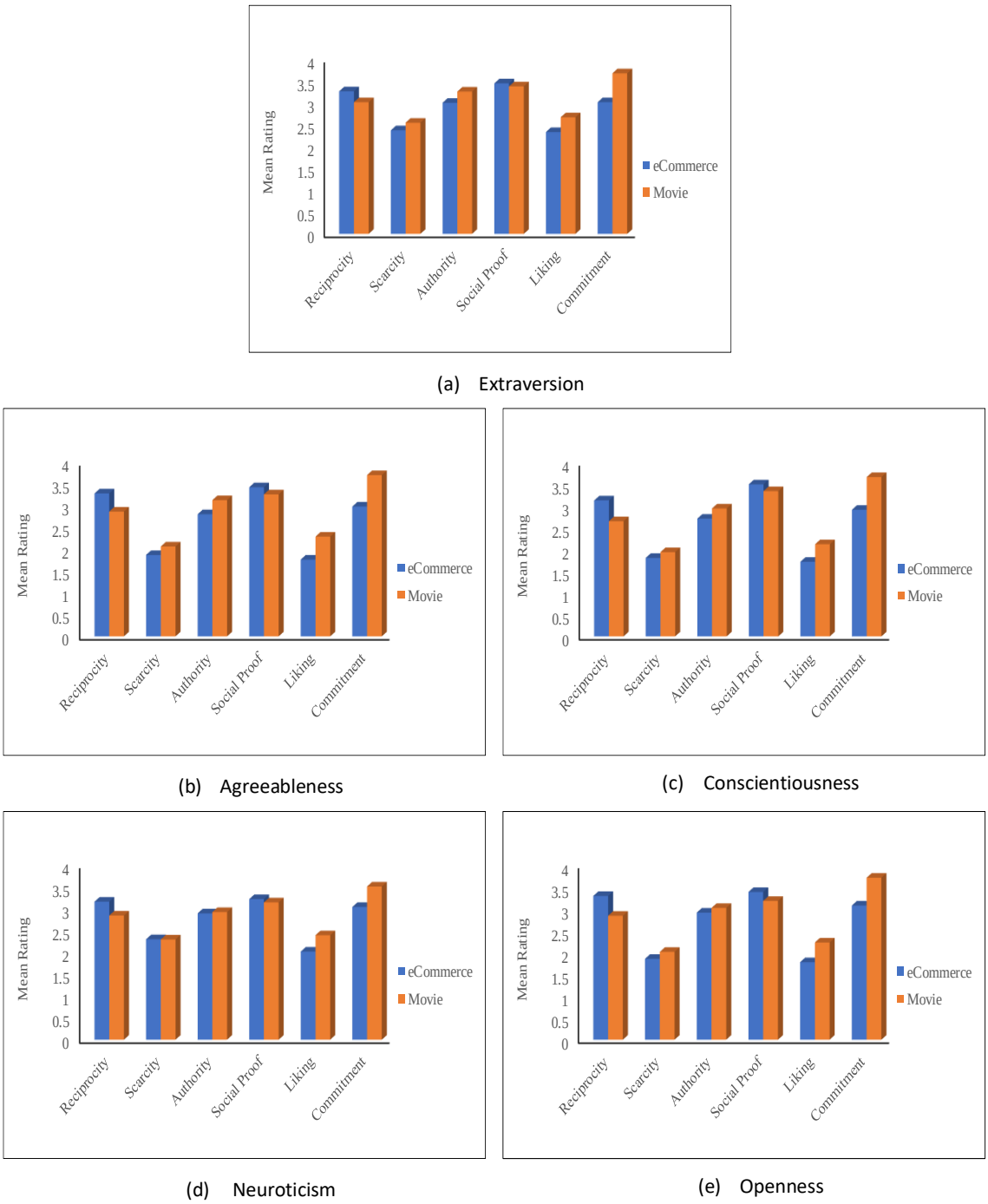
The Impact of Personality Traits and Context

This section discusses the effect of the context on users' persuadability based on their personality traits. Specifically, it investigates how users with a particular personality trait perceived the persuasive principles in the eCommerce domain comparing to the Movie domain.

Figure 14 depicts the average ratings for the persuasive principles based on the personality traits and the context. The figure contains five charts; each chart represents the results of one personality trait. For instance, Figure 14-(a) shows the mean ratings given by participants who are high in Extraversion traits. As a general observation, the figure shows that the means for all principles vary from one domain to another, which holds true for the five personalities. It can also be inferred from the figure that in the eCommerce domain, *Social Proof* and *Liking* are the most and the least influential principles for all personalities, respectively. In the movie domain, however, *Commitment* and *Scarcity* are the most and the least influential principles, respectively.

The ANOVA analyses demonstrated significant differences in the following cases: 1) Extroversion, Conscientiousness, and Neuroticism personalities perceived *Reciprocity*, *Liking*, and *Commitment* significantly different in the Movie domain compared to the eCommerce domain. These personality types perceived *Reciprocity* as more influential in the eCommerce domain, while *Liking* and *Commitment* are more influential in the Movie domain, 2) all principles, except *Social Proof*, were perceived significantly differently by Agreeable participants. *Reciprocity* is found more influential in the eCommerce domain, while the other four principles are found more influential in the Movie domain, 3) participants who have Openness personality perceived four principles significantly differently in the Movie domain compared to the eCommerce domain. These principles are *Reciprocity, Social Proof, Liking,* and *Commitment*. Medium effect sizes were found in terms of

Commitment for all personalities, while a large effect size is found in terms of Liking for *Agreeable* people.



(a) Extraversion

(b) Agreeableness

(c) Conscientiousness

(d) Neuroticism

(e) Openness

Figure 14 Mean rating based on the personality traits and the application domain

The above observations demonstrate an actual effect of the context on all personality types. Besides, this effect is significant in many cases. Specifically, at least three out of the six principles were perceived significantly different by each personality trait.

3.4. Discussion and Design Guidelines

As mentioned earlier in the chapter, the major purpose of the user study is to investigate the impact of persuasive principles, when associated with recommender systems, on users of different personalities, ages, genders, cultures, and different application domains. While conducting the study, our focus was, on one side, to report on the answers and ratings provided by the participants regarding their experience with persuasive recommender systems (without providing any justification about the nature of the results). On the other side, based on the reported results, we aim to come up with empirical findings or insights, including several guidelines that could help in tailoring the design of persuasive RSs to users with different traits, ages, genders, and cultures and also with taking different application domains into considerations.

Despite the conservative nature of the obtained results, which might threaten their generalizability to actual persuasive recommender systems, we can still provide the following recommendations/guidelines for designers of such systems.

- Associating persuasive strategies (Cialdini's six persuasive principles, as an example in this thesis) with recommender systems shows efficacy in impacting users' decisions to a high extent. This is demonstrated by the results presented in Figure 5 and Figure 10.
- Section 3.3.3 shows that the persuadability levels of users vary depending on their characteristics or personality features. Hence, we believe that tailoring the design of RS to individuals of different traits is a promising design direction. In this thesis, the big five personality traits are taken as an example to distinguish the different personalities of users. The following insights are observed based on the results presented in Figure 9:
    - *Reciprocity* is the most influential principle for all personality types. So, it could be used as a generic persuasion strategy for all types of users. This is

particularly useful in contexts where users' personalities cannot be obtained.

- On the other side, *Authority* and *Scarcity* are the least or the second least influential principles for all personality types. Thus, they should be given the lowest priority in static (non-personalized) contexts.
- Despite being applied in some applications, *Scarcity* is found to be the least influential strategy for users who are *Agreeable*, *Conscientious*, and *Open* people (as depicted in Figure 9). Having said that, designers are recommended to adopt a new style for attracting users' attention, other than the threat of item's scarcity.
- All persuasive principles are more influential for *Extrovert* users comparing to *Introvert* users.

- Gender, in general, is not a crucial factor to be considered alone when examining the impact of persuasive strategies on users (as the results presented in Figure 6 showed). However, our study revealed that females and males respond significantly differently to the *Authority* principle, with males being more persuadable than females. Therefore, if the *Authority* principle is to be used, we recommend using it with males more than females.
- Age does not have a significant effect on the persuadability process if it is considered in isolation of other factors.
- On the other hand, the results depicted in Figure 8 shows that culture is a very important factor. RS users from different continents were significantly different in their responses to the persuasive principles, with Europeans being the most susceptible to the *Liking* principle and Asian are the most susceptible to all other principles.
- Finally, the application domain is an important factor to be considered. So, our general recommendation is that if the RS seeking more personalized persuasive capabilities, the application domain should be considered. Following are some other designing tips related to the application domain and based on the domains under study, as discussed in section 3.3.4:

- The responses to the *Reciprocity*, *Commitment*, and *Liking* were significantly different. *Liking* and *commitment* are more effective in the Movie domain, while *Reciprocity* is more efficient in the eCommerce domain.
- *Social Proof* can be used as a non-personalized strategy in the eCommerce domain, as it is found to be the most influential principle (as depicted in Figure 10). On the other side, *Commitment* is the most influential principle in the Movie domain.

It is worthwhile to mention that other studies such as [68], [115], and [112] also provided some insights based on their studies on the relationship between user characteristics and persuasive principles. Nonetheless, as mentioned in section 3.1, our study is different from other related studies in the sense that it considers user-related aspects in combination with context-related aspects, it involves a heterogeneous sample, and it is tailored to the recommender systems area.

The study of Gkika et al. [68], which is the most related study to our user study, suggested some solutions (which the authors called "*paths*") that lead to an acceptance of the persuasion principles. The paths are represented as a combination between the presence and the absence of some personality traits. This work, however, is still limited, such that it does not consider all the possible combinations of personality traits, and it did not consider other factors. These limitations, in turn, affect the generalizability of the study. As mentioned before, we are planning to follow the methodology of Gkika et al.'s study to provide more comprehensive paths related to all the aspects considered in our study.

It is important to highlight that the generalizability of our findings to a real RS might be threatened due to some limitations of our study. The main limitation stems from the fact that the study is based on a self-report questionnaire and not a real recommender system. A second limitation is that the diversity of the cultures is limited to three continents, with only 13 participants from Europe. This, in turn, may threaten the generalizability of our findings to other populations. Accordingly, an important insight to be mentioned here is that the guidelines provided by our study and by the aforementioned related studies are general tips. Also, the guidelines did not comprehend all possible design cases (i.e., the combinations of every factor with all other factors) and could not provide absolute or decisive rules for deploying persuasive strategies in RSs. Nevertheless, the provided

guidelines are generic enough to be used with several persuasive RSs that are not completely personalized. Besides, our near future work will involve mitigating the generalizability issue by expanding the study to a larger number of participants and considering more RS's areas.

Having said that, a one-size-fits-all design is insufficient for deploying persuasive systems in a personalized manner. Also, following static guidelines is not the most efficient approach, although it helps in enhancing the persuasion capabilities of RSs. Therefore, our last insight is that an adaptive personalized approach is required to provide the best persuasive capabilities for RSs. Chapter 4 introduces a potential solution to mitigate this issue. It discusses an adaptive framework to deploy the persuasion principle in RSs.

3.5. Chapter Summary

This chapter elaborates on the user study conducted to examine the impact of Persuasive Recommender Systems (PRS) on different users. The chapter reported on the results of how users of different characteristics (age, gender, culture, and personality traits) get influenced by the different persuasive principles that a recommender system deploys and how different contexts may affect users' persuadability. The chapter also suggested a set of design insights and guidelines as a takeaway for designers who wish to empower recommender systems with persuasive features. The next chapter builds on the observations obtained from this chapter and investigates further the design and development of persuasive, personalized recommender systems.

Chapter 4. Personalized Persuasive Framework for Recommender Systems

One of the most important takeaways from the user study that we conducted in Chapter 3 is that there is a need for a recommender system that provides recommendations, which are not only persuasive but also tailored towards, and customized for, the needs of the person who uses the recommender (i.e., personalized). This chapter introduces the ***Per****sonalized* ***Per****suasive* Recommender System (***PerPer***), an adaptive framework to enrich RSs with persuasion capabilities that are personalized for different users.

PerPer is designed with two main goals in mind. The first goal is *personalization*. As demonstrated in Chapter 3, it is insufficient to deploy persuasive principles in a one-size-fits-all manner. Instead, persuasive approaches should be *personalized* in order to increase their efficacy. *PerPer* is personalized in the sense that for each user (or user group), a different persuasive strategy (or set of strategies) is used. The second goal is *adaptivity*. The user study conducted in Chapter 3 is useful in revealing that users perceived the persuasive principles differently. However, the study did not draw clear borderlines between these differences. To explain, the results of the study did not decisively indicate that a particular user X has been *only and absolutely* persuaded by a persuasion principle Y. Instead, the results provide generic guidelines about which persuasion principles could fit the most for a person X (and everybody with similarities to X) without taking into consideration the particularities of each person.

To this end, we propose *PerPer* as a reference architecture framework that establishes the basis for an adaptive personalized persuasive recommender system. *PerPer* is adaptive in the sense that it adjusts its selections of the most influential persuasion principles for each user. *PerPer* conforms to the requirements of adaptive persuasive systems stated by Kaptein [117], which can be debriefed as (1) *Identification* (i.e., the system shall be able to identify its users), (2) *Representation* (the system should be able to represent different persuasive strategies), and (3) *Measurement of success* (i.e., the system should be able to measure the effectiveness of the persuasive principles).

To the best of our knowledge, among the few approaches that deploy persuasion capabilities into RSs (such as route planning RSs [97]-[99], restaurant RSs [100], and Educational RSs [104]); none of them provide a reference architecture framework for adaptive, personalized and persuasive RS. *PerPer* aims to fill this gap by standing to be a generic framework that can be augmented with any recommender system and empower it with personalized persuasion capabilities in an adaptive manner.

The rest of this chapter discusses the *PerPer* framework, its main components, operations, formalization, implementation, and empirical validation. The theoretical foundations of this chapter have been published in [133].

4.1. *PerPer* Main Components

This subsection explains the main components and features of the proposed framework. The *PerPer* framework is designed with a goal in mind of deploying it as a generic framework that can be attached to, and work with, any kind of RS that uses any recommendation method, such as for example, collaborative filtering and content-based methods.

Figure 14 provides a conceptualization of the *PerPer* main components, namely the *Recommender System* part and the *Persuasion Engine* part. The Recommender System part represents any kind of RS that incorporates persuasive capabilities into its tasks to increase users' acceptance of recommendations. To achieve this goal, the Recommender System interacts with the Persuasion Engine part by sending feedback from users and receiving persuasive cues from the persuasion engine. It is worth mentioning that *PerPer* can be plugged into any RS regardless of the used recommendation method (e.g., Collaborative Filtering, Content-based method, etc.).

The Persuasion Engine (PE) part, on the left side of the figure, is the most crucial component of the *PerPer* framework. Its main goal is to provide the appropriate persuasion cues to the Recommender System. The Persuasion Engine consists of three components: The *Persuasion Pool (PP)*, the *User Model (UM)*, and the *Persuasion Agent (PA)*. The subsequent sections provide details about these three components.

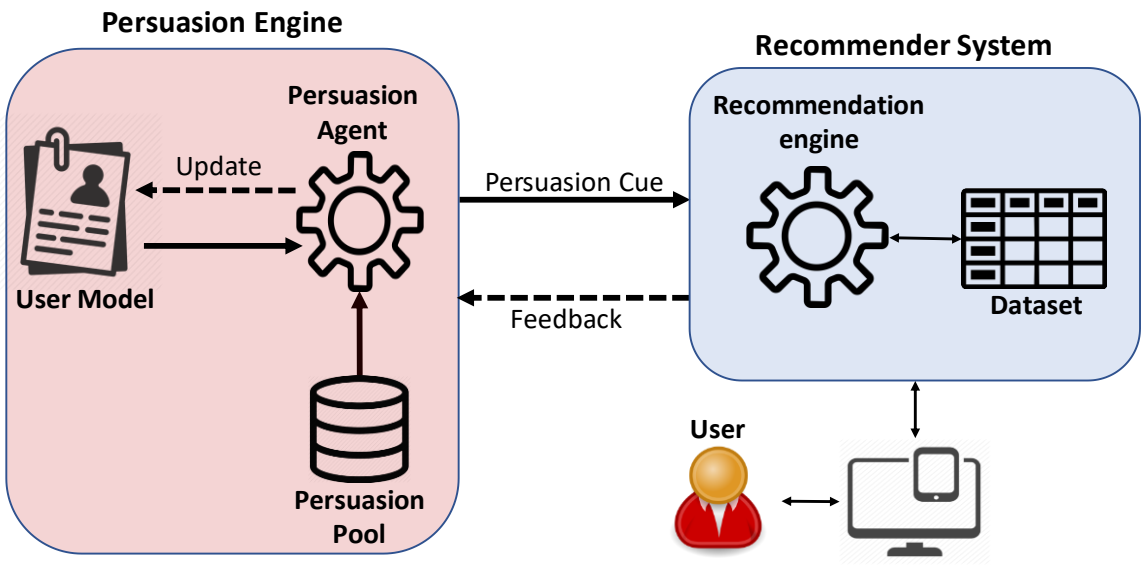


Figure 15 Conceptualization of the proposed PerPer framework

Persuasion Pool (PP)

The *Persuasion Pool* contains a set of persuasion approaches or directions, which we refer to as "*persuasion paths."* In this thesis, we define a *persuasion path* (P_{path}) as *a tactic that leads to persuading users or changing their behaviors*. For instance, Cialdini's six principles discussed in Chapter 3 can be considered as persuasion paths. The pool contains several persuasion paths, where each path can be achieved by one or more *Persuasion Cues*. A *Persuasion Cue* (P_{cue}) is defined in this thesis as *the actual mean of performing or deploying a persuasion path*. Examples of persuasion cues are the sentences introduced in Section 3.2.1 Table 3).

It is important to pinpoint that the proposed *PerPer* framework and its components are generic and are not tailored for any Persuasion Path or Persuasion Cue. Hence, any persuasion strategy can be considered by this framework. The implementation of *PerPer* in this thesis, however, considers Cialdini's persuasion principles as Persuasion Paths. The Persuasion Cues of each path are deployed as persuasion statements (presented in Table 3).

User Model (UM)

The *User Model* is the representation of users inside the system. It contains a collection of data that identify users. A *UM* consists of two parts: *User Identity* and *Persuasion Profile*.

The *User Identity* (*UI*) part contains data about users' characteristics. This data may involve a wide range of information, depending on the requirements of the application. It can be as simple as a unique identifier or composite information about the user's personality traits, demographic information, etc. *User Identity* is the component that *PerPer* relies on to personalize its selection of the provided persuasive cues.

The second part of the UM, the *Persuasion Profile* ($P_{profile}$), is the user's record of preferences regarding Persuasion Paths. It shows how users perceived the different *Persuasion Paths* throughout their interactions with the system. The *Persuasion Profile* of any user can be initialized in different ways, such as randomly, equally, or based on experimental studies (more details in Section 4.4). Normally, the *User Identity* can be used in the construction of the *Persuasion Profile*. It is worth noting that system designers may consider different persuasion profiles for each user, or they may consider similar persuasion profiles within a group of users who are recognized as similar users. All these details can be determined based on the system requirements.

It is worthwhile mentioning that we are not the first one who proposes the use of persuasion profiles. Kaptein [118] was one of the pioneers who reviewed the findings of social science on the social influence strategies and explained how these findings lead to the idea of persuasion profiling. In addition, the author was the first one to implement a system that uses persuasion profiles, which were described in detail by Kaptein & Eckles [119].

Persuasion Agent (PA)

At the heart of the *PerPer* framework is the *Persuasion Agent (PA)*. The Persuasion Agent performs an iterative process of updating the Persuasion Engine components and feeding the Recommender System part. In particular, the PA relies on *UMs* (with both *User Identities* and *Persuasion Profiles*) to decide on the P_{path} that should be used. Based on this decision, it selects a suitable P_{cue} from the pool and feeds it to the Recommender Systems in order to be used with the provided recommendations. In addition, the *PA* updates *UMs* based on the responses received from the Recommender System. More details about this process are discussed in the next section.

Table 5 summarizes these components and provides a brief definition for each of them.

Table 5 *PerPer* components

Component	Description
Persuasion Engine (*PE*)	The container of all other components. It is a concept more than an actual component.
Persuasion Path (P_{path})	Any approach (or method) that leads to persuading users or changing their behavior
Persuasion Cue (P_{cue})	The actual mean of deploying a persuasion path.
Persuasion Pool (*PP*)	The component that contains a set of persuasion cues related to the "persuasion paths."
User Model (*UM*)	The representation of users inside the system. It is divided into two parts: User Identity, and Persuasion Profile
User Identity (*UI*)	The part of UM that contains data by which the system identifies the users.
Persuasion Profile ($P_{profile}$)	The part of UM that contains user's preferences regarding the P-paths
Persuasion Agent (*PA*)	The heart of the *PerPer* framework, which is responsible for performing the iterative process of *PerPer*.

4.2. *PerPer* Operations

This subsection discusses how the components of the *PerPer* framework interact with each other to achieve their tasks. Specifically, it shows how *PerPer* selects a particular persuasion strategy and provides it to a specific user and how it proceeds throughout the system's lifecycle.

The working of the *Persuasion Engine* (*PE*) can be viewed as a four-step process that consists of *Initialization*, *Selection*, *Injection*, and *Updating* (as depicted in Figure 16). The *Initialization* step is the first step in the process and is done only once, while the other steps are recurring. During the *Initialization* step, the engine initializes all components that need to be initialized. Particularly, the components of the *UM* are initialized at this stage. This process is normally completed at the beginning, and it can be done in different ways. For instance, the user may be asked to fill out particular information when she registers. This information represents the seeds for the *User Identity* part of the *UM*. In addition, another important part to be initialized at this step is the $P_{profile}$. Initializing the persuasion

profile can be done following simple or advanced approaches. As an example of a simple *Initialization*, a system designer may give an equal level of precedence or preferences to all P_{paths}, or she may give them random precedences. Such simple approaches of *Initialization* are useful if the *UI* is not constructed yet. On the other hand, more advanced approaches based on experimental or user study results can be used to initialize the persuasion profile, where users' involvement is required in these approaches. In section 4.4, we present two approaches; one uses a simple initialization method, while the other uses an advanced and personalized initialization method.

The *Selection* step involves choosing a P_{path} from the persuasion pool. This operation is triggered by the Recommender System, and it happens when a recommendation is about to be presented to the user. Once the *UM* is constructed, the *PA* can select persuasion paths based on the corresponding $P_{profile}$. For each persuasion path, the pool may contain one or more cues. Similar to the *Initialization*, the *Selection* step is performed using a simple algorithm (e.g., random selection) or advanced algorithms (e.g., machine learning algorithms).

After selecting the persuasion paths and their associated cues, the P_{cue} is *injected* into the Recommender System, where the *PA* sends the corresponding cue so as to be integrated with the recommendations. It is worth mentioning that P_{paths}, expressed by means of P_{cues}, can be injected in different ways. For instance, if a cars company wants to recommend a particular car product to its customers, then it may deploy the *Authority* persuasion principle and augment the recommendation with persuasion cues in different ways, as follows: 1) presenting a picture of a famous Formula One driver while driving his car (which is manufactured by that company). Or 2) write a statement indicating that their cars are the first choice of a Formula One driver. In this thesis, we deploy P_{paths} as explanations (or statements) alongside the recommended items.

Finally, the *Updating* step focuses on updating the $P_{profile}$ of the *UMs*. After selecting a P_{path} and injecting its P_{cue} to the Recommender System, the *PA* waits for a response back from the Recommender System. This response represents users' feedback (i.e., acceptance or rejection) about the persuasion cue they received. Upon the receipt of the feedback, the *PA* updates the $P_{profile}$ of the corresponding user. Similar to the *Initialization* and *Selection* steps, the update algorithm can be simple or advanced, depending on the

designer's decision. Section 4.4 provides more details about the use of a machine learning-based algorithm to implement the updating operation.

It is worthwhile to mention that the feedback received from the recommender could be implicit or explicit. In the implicit feedback, users' preferences are captured without the direct intervention of the users [134]. For example, they can be captured by monitoring users' mouse movements or the time spent on a web page. Explicit feedback, on the other side, requires more effort from the user; it can provide more reliable data, though. This reliability leads to increased users' acceptance and more confidence in the recommendations [135]. Thus, explicit feedback is more common in the recommendation area than implicit feedback; the most common approach for explicit feedback is giving a rating on a scale. In the implementation proposed in section 4.4, we deployed the *PerPer* framework using explicit feedback.

A visualization of the *PerPer* iterative operations is depicted in Figure 16. The cycle of selection, injection, and updating are iteratively repeated until one of the P_{paths} dominate.

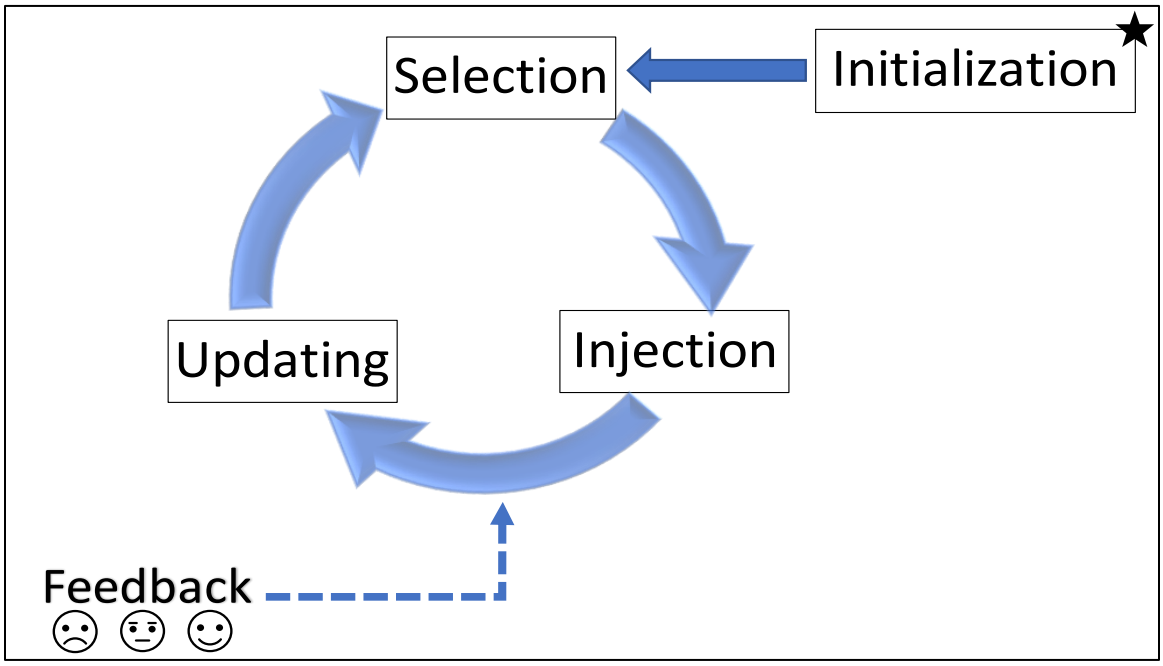


Figure 16 Main Operations of PerPer

4.3. *PerPer* Formalization

This section provides a formalization of the fundamental components of the proposed framework. Formalization reduces the ambiguity of abstract concepts, which leads to a better understanding of the problem. This, in turn, helps researchers to illustrate and analyze problems and solutions efficiently. Besides, formalization enables the development of robust algorithms and tools through solving problems that are exact by nature and that have provable properties.

I formalize *PerPer* components and their operations/interactions as a quintuple (P, R, W, U, S), where,

- P: is the set of possible persuasion paths that existed in the persuasion pool. That is $P = \{p_1, p_2, \ldots, p_r\}$, where p_i denotes the persuasion path (i), and r denotes the number of persuasion paths.
- R: the set of possible responses or feedback that can be sent back from the user of the RS to the persuasion engine. The set R depends on the application domain and the design requirements of the system. For instance, a system designer may ask the user to give feedback on a five-step Likert scale. In such a case, we say that $R = \{1,2,3,4,5\}$. Normally, the highest value of R indicates that the selected persuasion path has influenced the user (i.e., the user accepted it).
- W: the weight vector of the persuasion paths (i.e., it represents the Persuasion Profile, which is part of the User Model component described in section 4.1). This vector determines the importance (or the precedence) of each persuasion path to the corresponding user. It is presented as follows: $W(k) = [w_1(k), w_2(k) \ldots w_r(k)]$, where $w_i(k)$ denotes the weight of persuasion path (p_i) at a time instant k. This weight may be presented in various shapes. For instance, it can be an actual weight that represents the significance of each path, or it can be a probability by which each path can be selected.
- U: the updating scheme by which the persuasion agent updates W. This scheme is used during the *Updating* step of *PerPer*. It updates the weight vector (W) based on the response received from the recommender.

- S: the Selection Scheme by which the persuasion agent selects a path. This scheme concerns selecting a persuasion path ($p_i \in P$) based on the weight vector (W). More details about how we implemented S, and U are presented in section 4.4.3

4.4. Machine Learning-based Implementation of *PerPer*

This section discusses the implementation details of *PerPer* in practice. As previously mentioned, *PerPer* is a general framework that can be implemented in different ways and with any RS. Hence, the implementation presented in this chapter is just one suggested implementation approach to deploy *PerPer*. As a general view, Figure 17 depicts the main steps to deploy *PerPer*, which can be summarized as follows:

1. Select the list of persuasion strategies to be used as persuasion paths (P) and define the P_{cues} that will represent each path.
2. Define the set of possible responses (R) that can be received from the users as feedback to the injected persuasion cue. Also, define how the feedback will be elicited.
3. Determine how the *User Model* will be initialized. In particular, this step answers three questions: 1) what is the information required to build the user model? 2) how will this information be gathered? And 3) how will the persuasion profiles be initialized?
4. Define the Selection scheme (S) and the Update scheme (U).

The rest of this section discusses details of the implementation of *PerPer* to provide personalized and adaptive P_{cues} for each user based on her responses. The proposed approach is inspired by a well-known machine learning-based concept, namely the *Learning Automata* (*LA*), which is discussed in the next subsection.

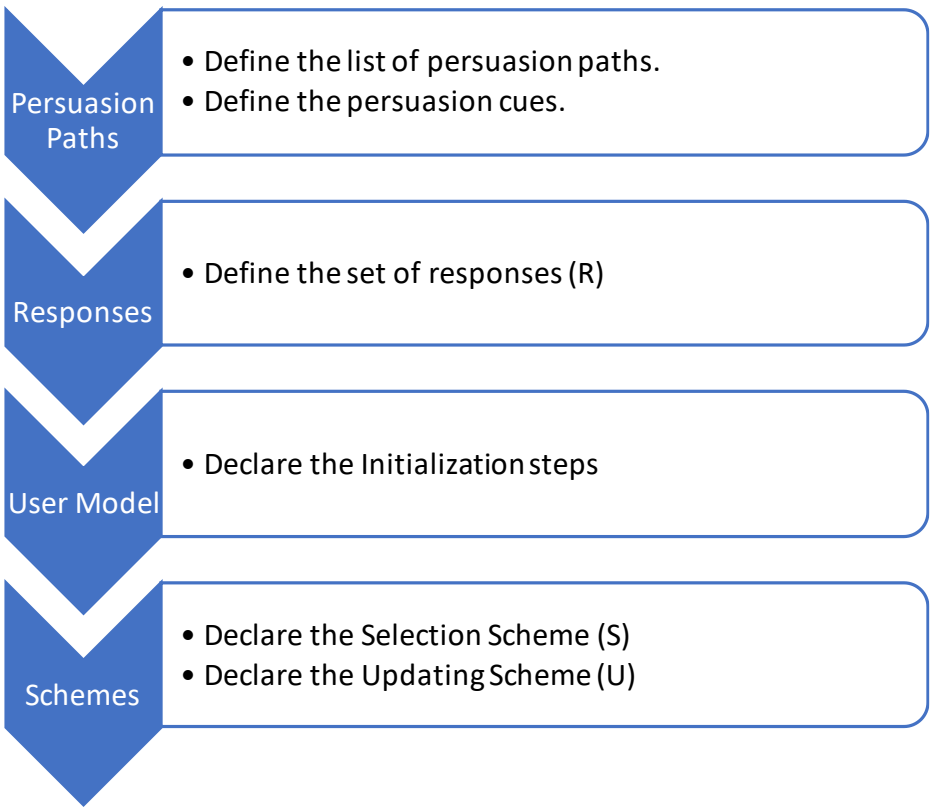


Figure 17 Main steps of deploying PerPer

4.4.1 Learning Automata

Learning Automata (LA) is an adaptive decision-making approach that tries to learn the best action from a set of alternatives in a random environment [136]. The main idea of LA is that the current action is selected based on former experiences from the environment. Basically, the learning process starts by giving *equal probabilities* to all possible actions. In each step, one action is selected from the action set, where the selection of actions occurs randomly according to their probabilities. The selected action is deployed in the environment. The environment evaluates the selected action and produces a reinforcement signal (or a response). Based on this response, the automata update the actions' probabilities according to a learning algorithm. This process is continued until one of the actions dominates [137].

A main strength point of LA is the generality of the action set concept [136]. That is, LA can be applied to any action set. Examples of action sets include queues in a queuing network, alternative routes in a communication network, or parameter values in a system.

In addition, the action set can be finite or continuous, which makes the LA approach universe such that it applies the same mathematical framework to a wide variety of learning problems. Based on the action set, LAs are divided into two categories: *Finite Action-Set Learning Automata* (FALA) and *Continuous Action-Set Learning Automata* (CALA) [138]. As the names indicate, the action set in FALA is finite, while it is continuous in CALA. My work focuses on the FALA because the set of actions is finite, as discussed in Section 4.4.3.

An abstraction of a learning automata's operations is depicted in Figure 18. A *learning automaton* is the decision-making unit in LA. The main operation of an automaton is to update the probabilities of actions according to the response of the environment. By *environment*, we mean the medium where the automaton run. Interactions with the environment that leads to updating action probabilities proceed in discrete time steps (or stages). In each step k, the automaton picks an *action* at random (according to the current action's probability distribution). The selected action represents the input to the environment, which in turn responds with a reaction (or *reinforcement*). The response of the environment is the input to the automaton that is used to update action probability distribution. The cycle of selecting an action, eliciting reinforcement, and updating action probabilities, is repeated until one of the actions dominates.

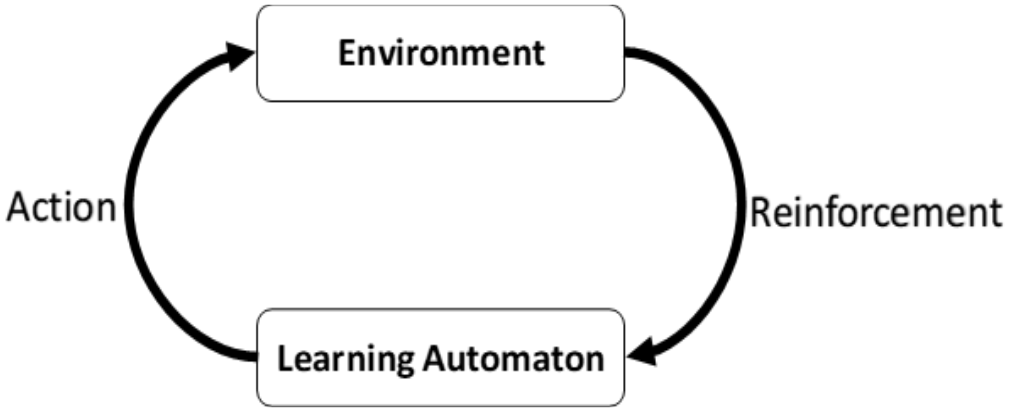


Figure 18 An abstraction of learning automata ([137])

4.4.2 Adapting Learning Automata in *PerPer* Framework

According to Figure 18, the learning automata consists of automaton, environment, actions, and reinforcements. Analogously, Figure 19 depicts the mapping of these components into *PerPer*. The *Persuasion Agent* in *PerPer* corresponds to the automaton in the learning automata. Following the same analogy, *Persuasion Cues* corresponds to the set of actions. A recommender system with its users represents the main components of the environment. Finally, users' feedback is the reinforcement signal that is received from the environment. The learning process in the context of *PerPer* produces identification of the best persuasion strategy that influences a particular user.

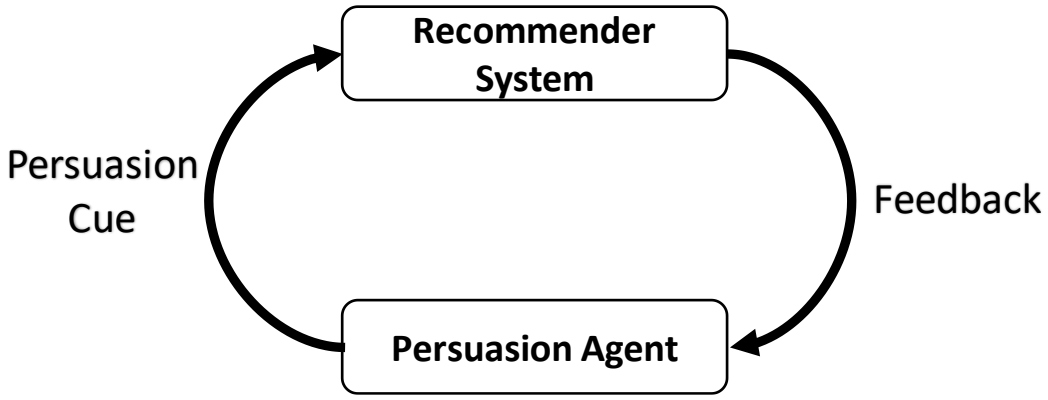


Figure 19 Proposed Automata Model for *PerPer*

Thathachar and Sastry [136] define an automaton as a quadruple (A, B, F, P) where A is the set of possible actions (or outputs of the automaton), where each action of an automaton at instant k is denoted by $\alpha(k)$. The set of all possible inputs (also called reinforcements) from the environment to the automaton is denoted as B. An element of the reinforcement set, denoted by $B(k)$, can be either finite set or an infinite set. F, is the learning algorithm by which the automaton update action probabilities. Finally, P, is the action probability vector that is used by the automaton to select an action at random.

Based on the above definitions and the formalization of *PerPer* (discussed in section 4.3), we map the learning automata into *PerPer* as follows: the actions set are the persuasion paths. That is, an element $p_i \in P$ is an action in the action set. The set of possible reinforcement signals is the set of responses (R) that can be received from the environment. The action probability vector represents the weight vector (W). As mentioned before,

the weight may be presented in various shapes. In the LA context, the weight is presented as a probability to choose an action. Finally, the learning algorithm can be mapped to the updating scheme (U) by which the weight vector is updated.

4.4.3 The Deployment

This subsection discusses the implementation of *PerPer* as a LA. Specifically, it shows what the components (*P*, *R*, and *W*) are and how the schemes (*U* and *S*) are implemented.

I considered Cialdini's six principles as the set of persuasion paths (P). That is, each principle represents a path (p_i). Accordingly, the set of persuasion paths, $P = \{Reciprocity, Scarcity, Authority, SocialProof, Liking, Commitment\}$, where $p_1 = Reciprocity$, $p_2 = Scarcity$, and so on. The number of actions ($r = 6$). These paths are expressed using persuasion cues as statements that are augmented with the recommendations.

As mentioned before, explicit feedback is more reliable than implicit feedback. Thus, this implementation elicits users' feedback explicitly. The users are supposed to give a rating on a five-step Likert scale. Therefore, the responses set (R) consists of five elements such that $R = \{1,2,3,4,5\}$. The weight vector (W) consists of six entries, where each entry is expressed as percentages that represent a probability. That is, the entry w_i indicates the probability to choose p_i as the next P_{path}.

Based on the main idea of *LA*, I consider a random selection scheme (S). At time instance (k), a P_{path} is selected randomly based on the probabilities in $W(k)$. This means that a path (p_i) will be selected randomly in the probability (w_i). It is worthwhile to mention that for each p_i, the pool may contain one or more cues. In such a case, the selection scheme should also discuss how to choose from this set of cues. Nonetheless, this work focuses on the case of having one cue for each persuasion path.

The Updating Scheme (U)

The reinforcement scheme is the most crucial factor of any learning automaton [136]. In general, the idea of the reinforcement scheme is as follows: suppose that p_i is the selected action at time instant k. If a good input (known also as a non-penalty response) occurs, the probability $w_i(k)$ of p_i is increased (i.e., a reward is given), and all other probabilities are

decreased. On the other side, if a penalty input is received, $w_i(k)$ is decreased (i.e., a penalty is given), and all other probabilities are increased accordingly.

A wide array of reinforcement schemes has been proposed in the literature, such as for example, the *Linear Reward Penalty* (L_{RP}) and the *Linear Reward Inaction* (L_{RI}). The one of interest in this thesis is the *Linear Reward-Inaction* (L_{RI}) algorithm. It is one of the earliest FALA algorithms. According to Narendra et al. [137], it is well known in the literature that L_{RI} is effective in many domains. Besides, it is a simple algorithm, and it is absolutely expedient [136]. Hence, it has been extensively applied in the literature.

The general idea of (L_{RI}) algorithm is that if the action p_i succeeds (i.e., selected), it is rewarded by increasing its probability $w_i(k)$ and all other probabilities are decreased accordingly. On the other side, if a penalty input (or a negative response) is received, the probability vector is not updated. Based on these definitions, the updating scheme (U) includes two procedures, namely *rewarding* p_i (calculated by equation 1) and *penalizing* all other actions (calculated according to equation 2).

$$w_i(k+1) = w_i(k) + \sum_{j=1, j\neq i}^{r} g\left(w_j(k)\right) \quad (1)$$

$$g\left(w_j(k)\right) = \delta w_j \quad (2)$$

$$w_j(k+1) = w_j(k) - g\left(w_j(k)\right) \quad (3)$$

In equation 1, $w_i(k+1)$ is the probability of the rewarded action (p_i) at time instant ($k+1$), $w_i(k)$ is the current probability of action (p_i). The upper bound of the summation (r) is equal to the number of actions and $g\left(w_j(k)\right)$ is the reward function calculated in equation 2, where δ is a factor used to calculate the rewarding value in the reward function. The penalty values of the rest of the actions are calculated according to equation 3. The penalty values may vary from one action to the other. According to equation 1, the reward value of p_i is the summation of the penalties of all other actions. This equation preserves the property that the increment (or decrement) of $W(k)$ should be consistent with the following requirement [136]: "$W(k+1)$ *should be probability vector whenever* $W(k)$ *is. Note that for* $W(k)$ *to be a probability vector, we should have* $0 \leq w_i(k) \leq 1, \sum_{i=1}^{r} w_i(k) = 1$". [136].

This is the idea we used as a base to deploy the updating scheme in the implementation of *PerPer*. For the initialization step (i.e., to initialize the weight vector), $W(0)$, we

propose two implementations, which we refer to as the *General Reinforcement Approach (GRA)* and the *Boosted Reinforcement Approach (BRA)*, as follows:

General Reinforcement Approach (GRA)

The General Reinforcement initialization approach follows the straightforward implementation of the L_{RI} reinforcement scheme. The main idea is that at the beginning, all actions are chosen with equal probabilities. We express this as:

$$w_i(0) = \frac{1}{r} \quad \forall\, i \leq r \tag{4}$$

Where r is the number of actions. Since we have six persuasion paths ($r = 6$), then $w_i(0) = 1/6$ for all paths. That is, the initial state of any *Persuasion Profile* is that all persuasion paths are evenly distributed at random.

Starting from an even distribution of the probabilities of the action set is a popular learning algorithm that has been applied widely in the machine learning domain. In the context of this thesis, however, an expected issue is that users may not be motivated to continue giving the feedback required for the persuasion technique. After a couple of iterations, this process may cause boredom for users, which, in turn, may lead users to resist reusing the system. This issue is actually common when we rely on explicit feedback. As a potential solution, we suggest the *Boosted Reinforcement Approach*, as explained in the next section.

Boosted Reinforcement Approach (BRA)

In the *Boosted Reinforcement Approach* (BRA), we suggest that a faster convergence can be reached if a more efficient algorithm is followed. The rationale behind this hypothesis is that, in general, the initial probability of an action decreases as the number of actions increases. This is because the GRA initializes all actions with similar probabilities, which equals ($\frac{1}{r}$). The larger the (r) value, the smaller the probability. Consequently, more iterations are required for the best action probability to converge (i.e., to increase form $\frac{1}{r}$ to unity) [136]. Conversely, if the initial state of a particular action increases, the number of iterations required for that action to dominate decreases. For this, we suggest the BRA, which is inspired by hypothesis 1.

Hypothesis 1*: The best persuasion path can converge faster if it is given a higher weight in the initial state.*

To explain, suppose that a persuasion path p_i is the best path among all others. If $w_i(0)$ is initialized to be larger than $w_j(0)\ \forall\ j \neq i$, then p_i will have a higher chance of being selected. Therefore, if n and m are the times required for p_i to converge using the GRA and the BRA, respectively, then $w_i(\mathrm{m})$ will converge faster than $w_i(\mathrm{n})$, and we say that $\{w_i(n) = 1 | \ n < m\}$.

In this context, however, the question that arises is, *how can we expect the potential best persuasion path for a particular user?* One answer to this question is by relying on some user characteristics to build a sort of user model. The goal of this modeling is to have an initial persuasion profile that is expected to be close to the actual profile. To build these initial profiles, I considered users' personality traits. Particularly, I conducted a study to investigate the relationship between the big five personalities and the six weapons of influence. The goal of this study is to provide numerical results that show the probabilities by which a persuasion principle can be the most suitable for each user's personality. The result of this study is a matrix that shows the efficiency of each persuasive principle on each personality trait. More precisely, the matrix shows the percentage of users, among the entire study sample, who are highly influenced by each persuasive principle. This matrix represents the third goal (G3) that is mentioned in Section 3.3.

To achieve the aforementioned goal, I relied on the data collected from the user study (discussed in Chapter 3). Based on the collected data, I calculated the ratios of persons who are highly vulnerable to each persuasion principle. Table 6 depicts the resulted matrix, where rows represent the big five personality types, while columns represent the six persuasion principles. Each record (i.e., a full row) is the persuasion profile for the corresponding personality trait. Numbers in the cells are the percentages of the participants who are highly influenced by the corresponding principle. For example, the first row shows that out of all participants who are high in *Extraversion* traits, 39% exhibit strong influence by the *Reciprocity* principle, 29.9% are highly influenced by the *Consistency* principle, 49.4% are highly influenced by the *Social Proof* principle, and so on.

Table 6 Mapping the big five personalities and the six weapons of influence

	Reciprocity	**Consistency**	**Social Proof**	**Liking**	**Authority**	**Scarcity**
Extraversion	0.390	0.299	0.494	0.487	0.338	0.636
Agreeableness	0.362	0.227	0.447	0.496	0.262	0.652
Conscientious-ness	0.317	0.194	0.388	0.511	0.230	0.683
Neuroticism	0.387	0.232	0.394	0.437	0.261	0.613
Openness	0.358	0.212	0.460	0.453	0.234	0.672

It can be noticed in Table 6 that the numbers in each row (i.e., persuasion profile) do not sum up to one (1). This is actually non-surprising because the same person can be highly influenced by multiple strategies at the same time. For example, Gkika et al. [68] found that an *Open* and not *Agreeable* person is highly influenced by *Scarcity* and *Consistency*. Nonetheless, as it has been mentioned previously, the action probability vector should sum up to (1) at any time instance (i.e., $0 \leq p_i(k) \leq 1 \; and \; \sum_{i=1}^{r} p_i(k) = 1$). For this, I normalized these numbers by dividing each value over the summation of all values in the corresponding vector. Table 7 depicts the normalized values of the matrix.

Table 7 Mapping the big five personalities and the six weapons of influence (Normalized values)

	Reciprocity	**Consistency**	**Social Proof**	**Liking**	**Authority**	**Scarcity**
Extraversion	0.147	0.113	0.187	0.184	0.128	0.241
Agreeableness	0.148	0.093	0.183	0.203	0.107	0.267
Conscientious-ness	0.136	0.084	0.167	0.220	0.099	0.294
Neuroticism	0.167	0.100	0.170	0.188	0.112	0.264
Openness	0.150	0.089	0.193	0.190	0.098	0.281

These mapping values are used to initialize the persuasion profile (or the weight vector) in the BRA approach. That is, depending on the *strongest* personality trait of the current user, the algorithm initializes her profile according to Table 7. For instance, if the personality

test shows that the current user's personality is high in *Extraversion*, *Agreeableness*, and *Openness*, but *Extraversion* is the highest (i.e., the strongest trait), then the *Extraversion* trait is considered. Accordingly, the initial weight vector, $W(0)$, will be $[0.147; 0.113; 0.187; 0.184; 0.128; 0.241]$, which is the first row in Table 6. That is, if the current user is categorized as an *Extrovert* person, then the *Reciprocity* path will be selected with a probability of (14.7%), while the *Consistency* will be used with a probability of (11.3%), and so on. The same convention is followed for all other personalities.

It is worth mentioning that for each user, there is a different weight vector (i.e., one vector for each user). These vectors are stored, along with users' other information, to construct the user model (UM), and they are updated according to the responses received from that particular user. This implies that *PerPer* is a fully personalized and adaptive approach.

4.5. Validation

This section discusses the validation of the *PerPer* framework to assess its feasibility in increasing users' acceptance of recommendations when augmenting persuasive capabilities to recommender systems. We compared RS that deployed PerPer to a non-persuasive counterpart RS. In particular, the evaluation aims to answer the following two questions:

1. Does deploying *PerPer* increase users' acceptance of recommendations in comparison to the acceptance of non-persuasive recommendations?
2. Is the boosted reinforcement approach (BRA) better than the general reinforcement approach (GRA) in terms of achieving a faster convergence of the best strategy and increasing users' acceptance?

The following sections provide more details about the experimental design, the user study that we conducted to answer the above questions, and the evaluation results.

4.5.1 Experimental Design

To validate the proposed approach, I implemented a prototype movie recommender system using a random recommendation-based algorithm. This kind of recommendation algorithms suggests movies randomly based on their production year and genre. Specifically, the movies were categorized, based on the year, into three groups: Old,

Moderate, and New. The algorithm selects one of these categories with equal probability. Then, it chooses a movie (with production year within the selected category) based on its genre. The genres are also selected equally at random.

Despite the abundant number of other alternatives of recommendation algorithms, we choose a random recommendation-based algorithm because of the following reasons; First, the goal of our evaluation is to be generic as much as possible by assessing the impact of the persuasion cues in isolation of other factors. Other algorithms, however, consider other factors (such as users' preferences or movie similarities) while building their recommendation list. This, in turn, could affect users' responses and, therefore, we will be less certain about the results. Second, since our user study is relatively long, we need to reduce the study time as much as we can so as to get more accurate responses. Providing random recommendations does not require building user models in order to begin recommending movies, which, in turn, reduces the study time. Third, random recommendation-based algorithms are easy to implement and do not require additional information or resources.

To evaluate the notion of persuasion and personality awareness in a recommender system setting, I implemented three variations of the movie recommender system (with two persuasive RSs, and one non-persuasive RS. Then I tested the persuasive systems against the non-persuasive counterpart. The implemented recommender systems are as follow:

- A movie recommender system without persuasion capabilities, which I refer to as "*Conventional Recommender System*" (CRS).
- A movie recommender system deploys the *PerPer* framework, with the User Model being initialized following the *General Reinforcement Approach*. We refer to this recommender as the *PerPer–GRA*.
- A movie recommender system deploys the *PerPer* framework, with the User Model being initialized following the *Boosted Reinforcement Approach*. We refer to this recommender as the *PerPer–BRA*.

The implemented recommender systems are Java web applications. The Eclipse IDE (version 2019-12 4.14) and Java SDK 13.0.1 are used as the development environment. The developed recommenders are released as an online application available on the cloud

using *Microsoft Azure*[15], or simply *Azure* as it is commonly known. *Azure* is a complete cloud computing platform that hosts the existed applications and streamlines new application development. It provides a wide range of software, platform, and infrastructure cloud services. To get our application available online, we used *Azure "App Service*" to host the application. For the database, we connect the application to the *phpMyAdmin* database through *Azure*'s "*MySQL In App*" service.

The implemented recommender systems are evaluated through a user study. Participants of the user study were asked to interact with the online application through one session. A participation session lasts for 20 to 30 minutes divided into three consecutive phases, where each phase represents one of the aforementioned implementation variations. In each phase, participants were asked to rate thirty-five movies (one movie at a time). The ratings represent the user's degree of intention to watch the recommended movie. The main difference between the first phase (which evaluates CRS) and the other two phases (which evaluate the *PerPer–GRA* and the *PerPer–BRA*) is that the former does not include persuasion cues alongside the recommendations as depicted in Figure 20.

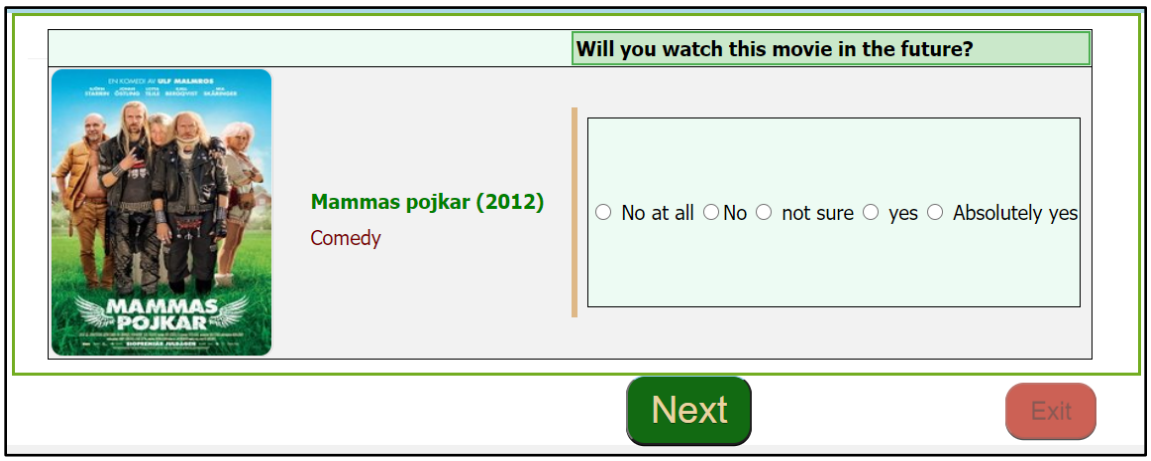


Figure 20 User interfaces of the CRS

In the second and the third phases, the system provides persuasion cues together with each recommendation, as depicted in Figure 21. The interface in these two phases looks similar so that the user obtains the same look-and-feel experience. The main difference between the second and third phases is that in addition to the rating step, the PerPer–BRA phase has

[15] https://azure.microsoft.com/en-ca/

one extra step: the personality test. For this test, we deployed the BFI-44 (discussed in section 2.3.2). It is worth mentioning here that the persuasion cues used in this implementation are similar to the cues presented in Table 3, Chapter 3.

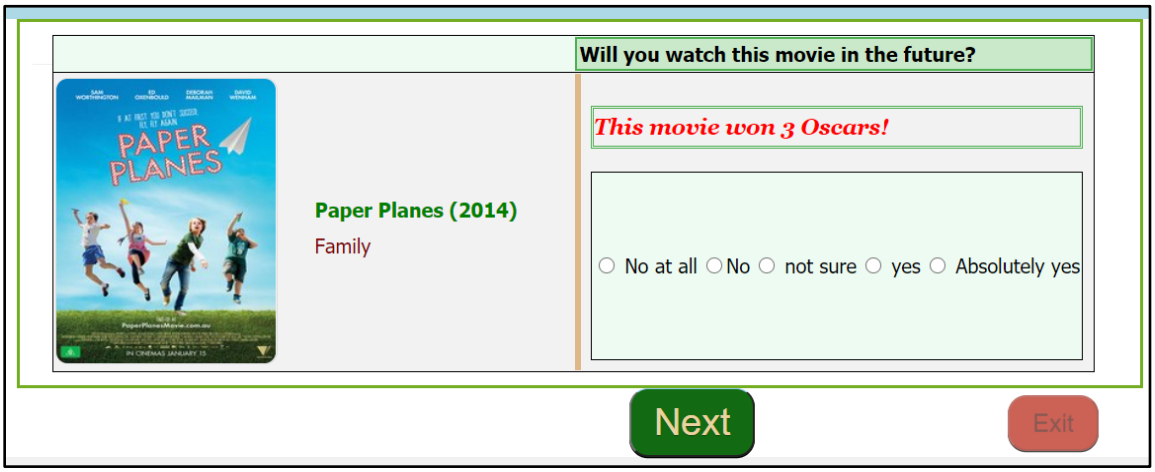


Figure 21 The user interface of the *PerPer–GRA* and the *PerPer–BRA*

4.5.2 Participants and Data Collection

Participants were recruited through different means, including emails, in-class announcements, and paper-based posters. A link to the application was given to participants who showed interest in participating. To preserve users' privacy, no personal information was collected from participants, nor their information were shared with any other website. In total, 58 persons participated in the study. We filtered the data such that we removed incomplete records. For example, we did not consider the answers of participants who completed one phase only or provided an incomplete response. After filtering the data, we retained 44 responses. Table 8 depicts the participants' demographic information.

It is worth noting that we followed a within-subject design in this study, where participants were asked to complete all phases. Due to the nature of this study, 44 responses are considered sufficient as they would be equivalent to 88 participants if the study was of a between-subject type. In addition, a within-subject study is able to control user-related factors (such as the participant's mood or stress levels), which in turn reduces noise in the data and makes it easier to detect a relationship between the independent and dependent variables [139].

Table 8 Participant's demographic information (N=44)

Subject	Category	Count [Percentage]	
Age	16-25	17	[39%]
	26-35	18	[41%]
	36+	9	[20%]
Gender	Male	24	[55%]
	Female	19	[43]
	Undefined	1	[2%]

4.5.3 Data Analysis and Results

The data collected from this study is analyzed as follows: The BFI-44 test is used to identify users' personality traits, which is then used to initialize the persuasion profiles of users. The ratings for the recommended movies are essential information for the evaluation. These ratings express the intention of users to watch the recommended movie. That is, they represent users' acceptance of the recommendation. It is worth mentioning that participants could have made their ratings because of the movie itself or because of the persuasive explanations. This is a challenge for our study. To mitigate this challenge, we relied on a neutral recommendation algorithm (namely, random recommendation). The essence behind this selection is, as mentioned before, to reduce the effect of other factors that may impact users' ratings. Also, we used the same random recommendation algorithms for the three versions of the implemented recommender in order to make the results comparable.

Based on the evaluation questions mentioned in Section 4.5, the analysis of results is divided into two sections: the first one answers the first evaluation question by comparing the conventional recommender CRS to the *PerPer–GRA* version. The second section compares *PerPer–GRA* with *PerPer–BRA* to answer the second question.

Users' acceptance of the recommendations is used as the essential mean of comparison. Inspired by the study of Gkika et al. [68], we considered the ratings as an indication of persuasion. That is, if the user gives a high rating to the recommended movie, it is considered as an indication of acceptance, and hence, it indicates that the user is persuaded to accept the recommendation. The following two subsections discuss the results.

Comparing *PerPer* to the Conventional Recommendation

This section evaluates whether deploying *PerPer* enhances users' acceptance of the recommendations or not. To achieve this, it compares the responses obtained from the *CRS* part with the responses of the *PerPer–GRA* part.

Figure 22 shows the mean ratings of the *PerPer–GRA* and the *CRS*. The x-axis shows the number of interactions between the system and the user. The number of interactions represents the number of recommendations and the corresponding ratings to such recommendations. For instance, 15 interactions mean that the system recommended 15 movies, and the user responded by rating each one of these movies. The y-axis of the figure represents the mean ratings of all users for the corresponding number of interactions.

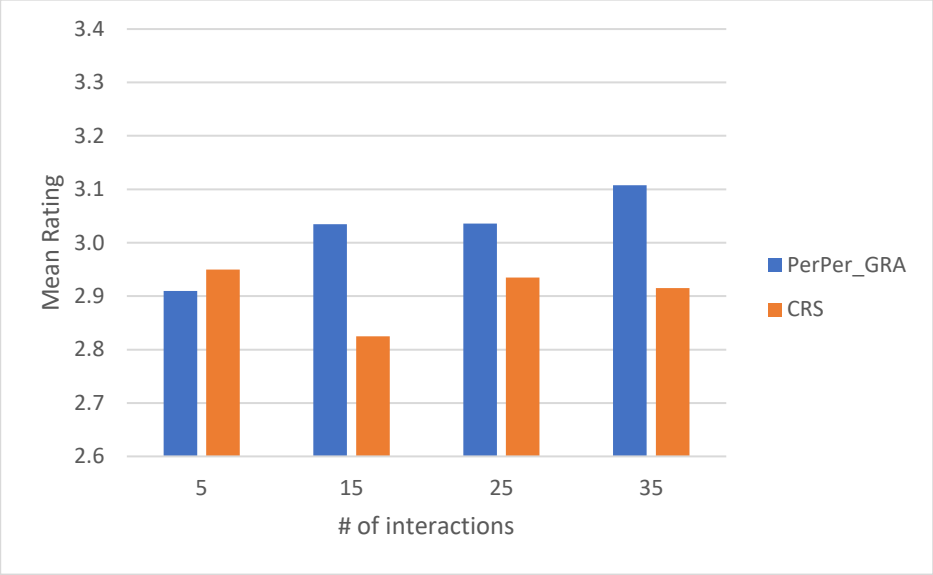


Figure 22 Mean ratings for CRS comparing to *PerPer–GRA*

The figure shows that the mean rating for the *PerPer–GRA* is better than that for the *CRS*. for all cases except the first case when the number of interactions = 5. This is partly due to the fact that the *PerPer-GRA* assigns equal probabilities to all persuasion paths. This means that any persuasion cue could be provided to the user along with the recommendation, without any kind of personalization. As a result, there becomes a strong chance that the provided recommendation might be rejected or low-rated by the user (as a result of resistance due to using an inappropriate persuasion path). However, as the number of interactions increases (i.e., >5), the *PerPer–GRA* builds a better knowledge base about users

and their personal preferences. Consequently, the chance to select more appropriate persuasion path increases, and hence users accept recommendations and assign them higher ratings than the CRS. Another observation that can be drawn from the figure is that user ratings for the *PerPer-GRA* increase as the number of interactions increases.

It can be noticed from the figure that the differences between means are around (0.2 – 0.4). Such a difference, although it seems to be small, is good enough because of various reasons. First, the values presented in this figure are averaged over 44 participants and 35 interactions for each participant. Second, taking into consideration the limited scale of the rating values (which is 1 to 5), such seem-small numbers are, in fact, relatively good. For instance, the (0.3) difference in the mean rating is equivalent to (6%) enhancement in a five-scale value.

The enhancement of the average rating means that the users gave a higher rating to the recommended items, which, in turn, means that the user accepted the recommendation. Having said that, we can say that an enhancement in the average rating implicitly reflects an enhancement in users' acceptance. Therefore, and based on the above observations, we conclude that *PerPer* enhances user's acceptance of the recommendations.

Comparing PerPer–GRA to PerPer–BRA

The first part of the analysis compared the conventional recommendation approach CRS with *PerPer–GRA*. This section tries to answer the second evaluation question by comparing the two versions of *PerPer*, namely the *PerPer–GRA* and the *PerPer–BRA*. It considers the results of phases two and three of our user study.

Figure 23 depicts the mean ratings for both approaches. The figure shows that the users' mean rating is better in the case of *PerPer–BRA* compared to the *PerPer–GRA* for all numbers of interactions, even when the number of interactions is low. This observation is unsurprising because, in the *PerPer–BRA*, the persuasion profiles are initialized from the beginning to be personalized. Hence, users accept the provided recommendations, which is reflected by their high ratings. Another important observation from Figure 23 is that the minimum rating in the case of *PerPer–BRA* is 3.2. This is more than the maximum rating achieved in the case of *PerPer–GRA*. This indeed reflects the superiority of the boosted reinforcement approach over the general reinforcement one.

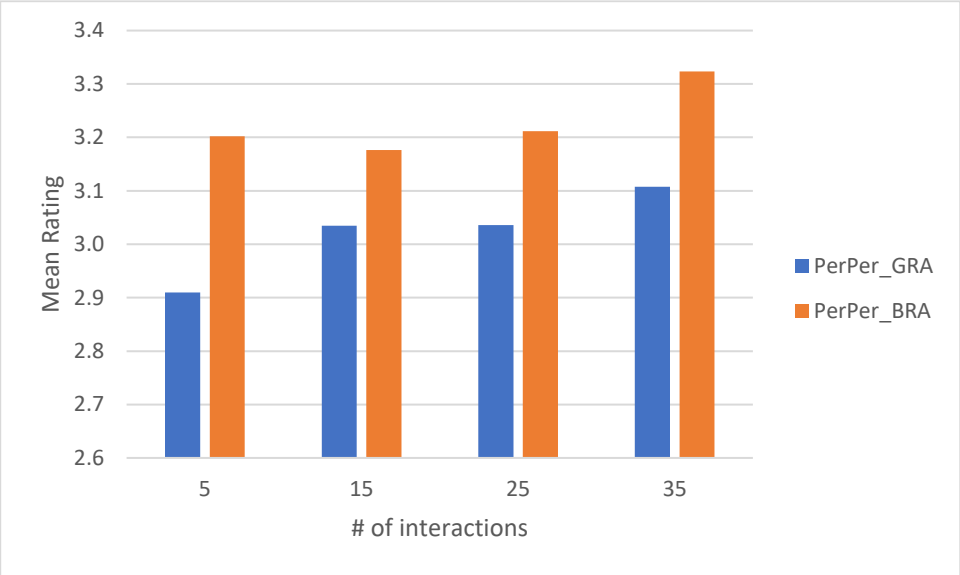


Figure 23 Mean ratings for PerPer–GRA comparing to PerPer–BRA

To assess how *PerPer–BRA* contributed to accelerating the convergence of the best strategy, we relied on the status of the persuasion profiles at the end of the interactions for each participant. Particularly, for each participant, we compared his/her persuasion profiles after completing the *PerPer–GRA* part and the *PerPer–BRA* part. The highest probability in each profile is considered as an indication of how close to the convergence the profile is. The results revealed that, out of the 44 participants, only eight profiles were closer to convergence in the *PerPer–GRA* profile, while the remaining were closer in the case of *PerPer–BRA*.

To summarize, the observations demonstrate the superiority of *PerPer–BRA* such that it was able to increase user's acceptance even in the initial interactions. Also, the enhancement is getting better as the number of interactions is increased. In addition, PerPer–BRA was able to accelerate the convergence of the best persuasion path while maintaining improvement in users' acceptance.

4.6. Chapter Summary

This chapter presented the ***Per****sonalized* ***Per****suasive* RS (*PerPer*), a framework for augmenting RS with persuasive capabilities. The chapter discussed *PerPer's* main components, operations, and interactions between components. Furthermore, the chapter

discussed the mechanism of dynamically selecting a particular persuasion strategy that is tailored or personalized for a particular user. To implement and validate *PerPer*, we employed a machine learning-based approach by adopting the Learning Automata concept. Based on this approach, *PerPer* is demonstrated to be feasible in terms of increasing users' acceptance of recommendations.

PerPer is evaluated using a user study where we implemented a prototype of a movie RS. The user study involved three parts, namely, the Conventional Recommender System (*CRS*), the General Reinforcement Approach (*PerPer-GRA*), and the Boosted Reinforcement Approach (*PerPer-BRA*). The analysis of the results obtained from 44 participants shows that *PerPer-GRA* was able to enhance users' acceptance of the recommendations in comparison to CRS. The results also show that the *PerPer-BRA* outperforms the *PerPer-GRA* in terms of accelerating the convergence of the best persuasion path while maintaining improvement in users' acceptance.

Up to this point, it is worth mentioning that the state-of-the-art approaches in the literature that focus on RSs with persuasion features are either domain-specific (such as, for instance, the route planning RS [69] and educational RS [75]) or do not adapt their solutions by considering personality-related aspects while providing persuasive recommendations (such as [97] and [98]). The deviation between these approaches and our proposed approach (which is by nature, domain-independent, adaptive, and personalized) makes it infeasible to conduct a fair and consistent comparative analysis. However, in the near future, we are planning to deploy the *PerPer* framework into a particular context, for example, the educational domain, and then compare its performance against the performance of their counterpart non-personality-aware persuasive RS that is applied already in the educational domain. By this, we would have a better characterization of the circumstances under which *PerPer* achieves better performance gains. Furthermore, we envision implementing *PerPer* using different approaches, for example, using Bayesian learning, as proposed by Kaptein [116], and compare these approaches (e.g., Bayesian learning-based implementation) against the Learning Automata-based implementation.

The next two chapters discuss the second part of this work, which is evaluating RSs. It presents the cognitive-based approach to evaluate RSs beyond their accuracy.

Part II

Evaluating Recommender Systems

Chapter 5. Cognitive-Based Evaluation for Recommender Systems

This chapter introduces the second angle of the research considered in this thesis, namely *evaluating RSs beyond their accuracy*. In particular, the chapter proposes an evaluation approach called ***Comp****rehensive* ***Per****formance* evaluation (*ComPer*). This approach is inspired by the main theme of the thesis, where we envision to view RSs as systems that are not only capable of providing recommendations but also supported with persuasion and cognitive features, which makes them, in principle, close to human beings. *ComPer* evaluation approach considers an RS as a human being with cognitive skills, and hence, the recommendation process is analogized to a human's learning process. Following this principle, the outcomes of the recommendation process can be evaluated in the same manner as we evaluate the outcomes of the human learning process.

The contents of this chapter have been introduced in two papers [140], [141] and expanded in a third one [142].

5.1. Context

Evaluating Recommender Systems is a nontrivial task due to many reasons [9], [144]. First, RSs are applied in a wide range of applications, including e-commerce, healthcare, education, etc. Second, different types of RSs use different algorithms, and therefore, have different inputs and outputs. Third, recommendation results are of different types, such as rating prediction and a recommendation list. Based on the nature of the results, the same metric could be evaluated differently. Finally, due to the multifaceted nature of recommender systems, there is a variety of evaluation metrics (or dimensions) introduced to the field.

Earlier in the literature, it was common to evaluate RSs using only one evaluation dimension, with a particular focus on using *accuracy* (or correctness)[16] as the main

[16] The terms "Accuracy" and "Correctness" are used interchangeably in the literature.

evaluation metric. Later on, there became an increasing consensus that the accuracy of recommendations is insufficient to evaluate the actual performance/effectiveness of recommenders [29], [144]. For instance, recommending an item with the highest ratings is not necessarily a useful recommendation because the user might have already seen this item and that he/she is no longer interested in it. This issue triggered several calls in the literature to start considering more evaluation dimensions other than accuracy. To respond to this requirement, researchers proposed a plethora of evaluation dimensions, which results in a vast variety of dimensions being introduced in the literature.

Having multiple evaluation dimensions is a double-edged sword. On the one hand, it enriches the evaluation process of RSs; but on the other hand, the wide variety of dimensions makes it challenging to select the most appropriate dimension. In addition, the existence of multiple evaluation options requires significant time and effort to design a proper experiment [6]. As a matter of fact, the abundance of evaluation options makes it easy to find an evaluation design that suits a particular algorithm but ineligible for others, which represents an increasing impediment to fairly compare different studies [172]. For example, good *scalability* means that if the system scales up to the point where there exists a large number of items to be recommended, it can make recommendations within a reasonable time. However, to increase the system's scalability, other dimensions might be affected negatively. For instance, the algorithm may recommend items based on only a part of the item set instead of considering the whole dataset. As a result, the processing time (i.e., which represents *scalability* here) will be increased, but the *coverage* and *correctness* will be decreased[6]. Said et al. [4] argued that, due to the existence of multiple evaluation methods, it becomes necessary for these methods to be "*critically analyzed to ensure improvement in the field*" [4]. According to Avaspour et al. [6], a better framework or standard for understanding the relationship between dimensions is required.

For the reasons given above, and the gaps identified in the literature, we cannot deny that there is a lack of uniformity in the current evaluation methods. Hence, a unified evaluation methodology becomes desired to compare recommenders [145]. As a potential solution to this issue, this chapter proposes an evaluation methodology called ***Com****prehensive* ***Per****formance* evaluation (*ComPer*). *ComPer* is a cognitive-based method inspired by the idea that a recommender system can be viewed as a human being with cognitive skills,

and so, the recommendation process can be analogized to a human learning process. Consequently, the outcomes of a recommendation process can be evaluated in the same manner as we evaluate the outcomes of the humans learning process. For this, *ComPer* innovates a mapping between Bloom's taxonomy (discussed in section 5.2.1) and the main phases of RSs (i.e., information collection, learning, and recommending). Based on this mapping, a correlation between the most common evaluation dimensions and Bloom's cognitive dimension is inferred. The name *ComPer* stems from the main goal of this approach, which is to provide a **com**prehensive-enough evaluation of the **per**formance of RSs.

ComPer aims to achieve three goals: *thoroughness*, *simplicity*, and *independency*. To achieve the first goal, thoroughness, *ComPer* considers *multiple* evaluation dimensions instead of considering only the accuracy, and in fact, this is the essence of *ComPer*. The second goal, simplicity, is satisfied by representing the final evaluation result using one *single value* that embeds all the considered dimensions. Finally, the *ComPer* approach conducts evaluation and produces results that are independent of data and evaluation settings. That is how the third goal (i.e., *independency*) is seemed to be achieved by *ComPer*. Independency leads to a fair[17] comparison between different recommenders. In Section 6.3, we show through experimental evaluation how *ComPer* fulfills these three goals.

5.2. *ComPer* Preliminaries

As previously mentioned, the proposed evaluation approach is built based on a mapping between humans' learning process and RSs' recommendation process. This mapping is extrapolated based on a study of both parties (i.e., human's learning and RS process) that led to a numeric correlation matrix. The essence of the proposed approach is to obtain a correlation between RSs (presented by their main phases and common properties) on one side and humans' cognitive skills (presented by Bloom's taxonomy and its cognitive dimension) on the other side.

This section discusses the components that *ComPer* relies on to infer the mapping, namely Bloom's taxonomy, phases of RSs, and the evaluation dimensions of RS.

[17] In the context of this chapter and Chapter 6, "fairness" means evaluating and comparing recommenders without being biased to a particular experimental settings (i.e., different setting provides comparable results)

Declaration: In the rest of this section, the six cognitive dimensions of Bloom's are written using *Calibri italic font*. The RS phases are written using a ***Times New Roman Italic Bolded*** font, and the evaluation dimension of RS are written using *Times New Roman Italic font*

5.2.1 Bloom's Taxonomy

Bloom's Taxonomy is a model that was introduced to the education field in 1956 by Benjamin Bloom, an educational psychologist at the University of Chicago [146]. It classifies learning behaviors into three domains: cognitive domain, affective domain, and sensory domain. These domains can be thought of as the goals of the learning process. That is, for each learning session, learners should have acquired one or more of three goals, which are: mental skills or knowledge (Cognitive domain), growth in feelings or emotional area (Affective domain), and physical skills (Sensory domain). The taxonomy also defines assessment dimensions for each domain to evaluate different educational outcomes.

The *cognitive domain*, which is the domain of special interest of this work, involves knowledge and the development of intellectual skills [146]. This domain underlines six intellectual skills (or outcomes). Following are the six dimensions of Bloom's cognitive domain, ordered from the simplest to the most complex one:

- *Remember*: the ability to retrieve knowledge from memory. It concerns recall definitions, facts, etc., or to recite previously learned information.
- *Understand*: The ability to determine meanings from all kinds of messages, such as interpreting, classifying, or comparing.
- *Apply*: The ability to use learned material by implementing a procedure in a given or new situation.
- *Analyze*: The ability to break material into constituent parts and detect how the parts are related to each other, as well as being able to distinguish between the components or parts.
- *Evaluate*: the ability to make judgments and comparisons based on criteria or standards.
- *Create*: The ability to put elements together and recognize them into a new pattern.

Bloom's six dimensions of cognition are considered as evaluation dimensions to assess the learning outcomes of human beings in terms of acquired cognitive skills. These dimensions are used by educators to design and evaluate their courses. In this thesis, these dimensions are adapted to explain the mechanism of our proposed evaluation method, *ComPer*.

5.2.2 Phases of Recommender Systems

According to Isinkaye et al. [135], a recommendation process is divided into three main phases: Information collection, learning, and prediction (or recommendation). The first phase, ***information collection,*** is all about collecting users' information to build their profiles that will be used later by the other two phases. In general, the performance of any RS is strongly associated with the amount and type of information collected. This information can be collected explicitly (as inputs from users) or implicitly (by inferring users' preferences indirectly, for example, through user's behaviors). In the second phase, ***learning***, the recommender system focuses on applying learning algorithms to extract or infer users' features from the information gathered in the information collection phase. The last phase is the ***prediction/recommendation***. This phase focuses on predicting which items the user may (or may not) prefer. The result of this phase is affected by the previous two phases. That is, the more the information collected, and the better the learning algorithm used, the more useful the recommendation. Figure 24 depicts these three phases.

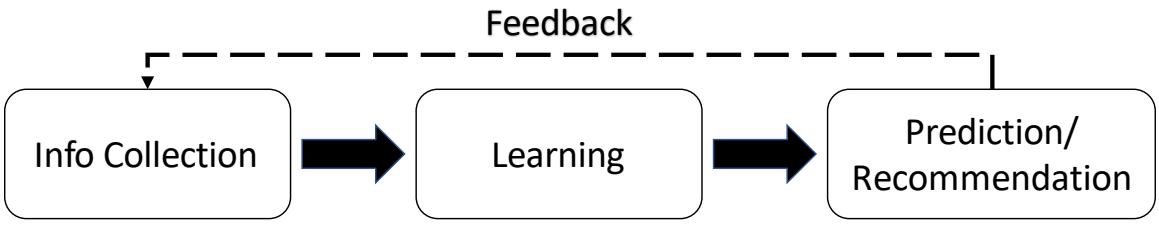


Figure 24 Main phases of recommender systems

5.2.3 Evaluation Dimensions of Recommender Systems

In the literature, there are more than sixteen evaluation dimensions or metrics[18] that have been introduced to the RS field. Avazpour et al. [6] classified these evaluation dimensions

[18] In the literature, the terms "properties", "dimensions", and "metrics" are used interchangeably to represent the RSs' dimensions that we want to assess.

into four categories: *recommendation-centric*, *system-centric*, *user-centric*, and *delivery-centric*. A summary of dimensions, along with their definitions, is listed in Table 9. For a more detailed discussion about evaluation dimensions and how they can be assessed, the reader is advised to refer to [6] and [7].

As previously stated, *ComPer* emphasizes the cognitive side of RSs, represented by the skills required to learn users' preferences and provide recommendations as learning outcomes. That is, *ComPer* focuses on the recommendation process itself, represented by its main phases (Section 5.2.2), rather than focusing on other aspects of recommendation, such as usability or risk. Having said that, *ComPer* puts emphasis on *recommendation-centric* and *system-centric* evaluation dimensions. In particular, *ComPer* considers nine evaluation dimensions: *Correctness*, *Coverage*, *Diversity*, *Confidence*, *Robustness*, *Learning Rate*, *Scalability*, *Stability*, and *Privacy*. The purpose of these dimensions is to assess a recommender and its recommendation process in a manner analogous to assessing humans' learning and their learning outcomes.

It is worth noting that evaluation dimensions in Table 9 are not all of the same popularity or importance. For example, in code suggestion domains, the accuracy of the recommended code options is more important than diversity. Nevertheless, there is no consensus in the literature about the most important evaluation dimensions that can ultimately be used to assess a particular domain. This issue motivated us to investigate the literature to build an understanding of the most commonly used evaluation dimensions among the nine dimensions mentioned above.

Our study surveyed 135 articles published in the last decade in the context of evaluating RS. Our survey revealed that *Accuracy* is the most considered dimension. This popularity is mainly due to the fact that accuracy is the first dimension introduced to the RS field, and it is easy to be measured [147]. *Coverage* and *Diversity* are the second most popular dimensions, followed by *Robustness* and *Scalability*. The survey also indicated that *Confidence*, *Adaptability*, and *Stability* had not been evaluated by any of the articles under study (i.e., they are not commonly used in the literature). So, we decide not to consider them further in our proposed evaluation framework. In addition, we excluded the *Privacy* evaluation dimension since measuring privacy is a non-trivial task that is still not fully explored [6]. Accordingly, after excluding the non-common dimensions, we ended

up focusing on the five most popular evaluation dimensions[19], namely: *Correctness*, *Coverage*, *Diversity*, *Robustness*, and *Scalability*.

5.3. Related Work

The literature has recently witnessed increasing attention to the evaluation methodologies of recommender systems. Particularly, there is almost agreement on the necessity of a common methodology to evaluate the multifaceted nature of RSs. Hence, several researchers highlighted the need for such an evaluation methodology. This section presents the evaluation approaches that have been proposed recently, along with a discussion about their limitations.

A methodological description framework and a formalization of the offline evaluation procedure of Time-Aware RSs (TARS) are provided in [148]. Through their survey and analysis of the TARS literature, the authors found that there are clear divergences in the evaluation protocols and methodologies. Consequently, they identified a set of key conditions, and based on these conditions, they introduced a categorization of evaluation protocols for TARS. Finally, the authors suggested methodological guidelines that may help researchers select the proper combination of evaluation conditions. This study focuses on one type of recommender systems, namely the Time-Aware RS. Also, it does not provide a comprehensive enough evaluation framework; instead, it provides some general evaluation guidelines that can be followed to evaluate a single domain only, which is TARS. This issue is shared with other proposed frameworks, such as for example, the work of Trattner et al. [151], which studied the evaluation of group RSs. They show how to use the evaluation methods deployed for single-user RS in the context of group-based recommenders. Other examples of such domain-specific approaches include the work of Monti et al. [152] on evaluating *Sequence-based Recommender System* and Ludewig et al. [153] on evaluating *session-based recommendation algorithms*. All these studies and solutions do not provide generic approaches.

[19] For simplicity, the term "evaluation dimensions" is used throughout this chapter to indicate these five dimensions unless mentioned otherwise.

Table 9 Recommender System evaluation dimensions and their categories [6]

Category	Dimensions	Definition
Recommendation-centric	Correctness (Accuracy)	The closeness of the recommendations or predictions to the actual user preferences
	Coverage	The proportion of items (or users) that the system can recommend (or the system can generate recommendations for)
	Diversity	The dissimilarity between recommended items
	Confidence	How confident is the recommender in its results (recommendations or predictions)
System-centric	Robustness	How stable is the RS in the presence of fake (or false) information
	Learning Rate (Adaptivity)	How adaptable (fast) is the system to new information
	Scalability	How scalable is the system under extreme conditions, such as a huge dataset?
	Stability	The consistency degree of the recommendations over time
	Privacy	Is there any potential risk to users' privacy
User-centric	Trust	To what extent do users trust the system's recommendations
	Novelty	Recommending items that are new to the user (she did not know about them)
	Serendipity	How surprising yet successful are the recommendations?
	Utility	The value gained from the system for different actors, such as users and system owners
	Risk	How risky is the acceptance of the recommendation for the user?
Delivery-centric	Usability	How usable is the recommender (i.e., how easy is it for users to adopt it)?
	User preferences	How do users perceive the recommendation system?

Meyer et al. [29] provided a methodology to analyze the efficiency of RSs in an industrial context. They categorized RSs into four structuring functions: help to decide, help to compare, help to discover, and help to explore. Based on these functions, the authors

proposed an offline evaluation protocol; the main idea of this protocol is to map each one of these four functions to the appropriate evaluation measures. Although the proposed protocol aims to consider different recommendation functions for RSs in general, it fails to consider different evaluation dimensions. That is, the protocol focuses mainly on evaluating the recommendations' accuracy using different measures based on the aforementioned functions.

Some researchers focus on introducing protocols that serve as supportive solutions to enhance traditional evaluation settings rather than introducing an evaluation methodology. Najjar and Wilson [150], for instance, developed a group testing framework that aims to provide means to enhance evaluating group recommendation. The main goal of the proposed framework is to generate synthetic groups that are modeling actual group preferences automatically. The generated groups are parameterized to test different group contexts. The proposed framework contains two main components: First, group modeling, which defines specific group characteristics that will be used to form the synthesized groups. Second, group formation, which is a mechanism that identifies compatible groups from a dataset of single users by applying the defined model (i.e., the group model that is identified in the first component).

Another supportive solution has been proposed by Said et al. [154]. The goal of this model is to propose a method for fair data splitting that can give more accurate results for recommenders' correctness. In particular, the model was designed specifically for the recommendation task (i.e., the task of recommending an item or a set of items). The protocol has been developed with a "find good item" scenario in mind. The authors suggested criteria to choose candidate items (or users) for the training and testing sets. This splitting method aims to mitigate the issue that the accuracy values are biased to the dense users' records (i.e., users with many ratings in the dataset). Also, it aims to make sure that all available data for each user is used for training the algorithm. This protocol is limited in terms of dimensionality such that it considers the accuracy dimension only.

A dataset-focused framework is introduced by Dooms et al. [155]. They provided a framework for introducing and benchmarking new datasets to RS. The proposed framework focuses on the traditional Movie rating datasets (such as MovieLens), which are static datasets, and they become old. The article aims to utilize social media and smartphones to

create more valuable datasets by combining new data sources. To do so, the authors proposed a five-step framework to introduce and benchmark new datasets. The applicability of the framework is demonstrated on a new movie rating dataset called MovieTweetings, which analyzed different aspects following the proposed framework. The authors concluded that MovieTweetings was interesting for user-centric and simulation-focused experiments and could be used as a complement to the MovieLens dataset for offline comparative experiments.

Another evaluation approach is proposed by Sohail et al. [143]. The goal of this approach is to avoid biased and fake ratings in the dataset. The paper introduced a user's sincerity measure that focuses on eliminating feedback of insincere users. This measure is calculated based on different factors, including user visits to the review page and the time spent on that page. Based on this information, the authors defined two scores: the *Product Importance Score* (PIS) and the *Product Preference Score* (PPS). Then, they obtained the *Comprehensive Veracity Measure* (CVM), which is defined as the average of different evaluation measures. Although this evaluation measure considers multiple measures, it is considered limited because it only considers accuracy-related measures (such as Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), false-positive rate (FPR), and false-negative rate, FNR); As mentioned above, Accuracy does not reflect the actual performance of RSs.

Other researchers focus on evaluating the usability aspects of RSs; Pu et al. [156], for instance, proposed a user-centric model called *ResQue*. The model was built based on usability-oriented research in the RS field, as well as principles from popular usability evaluation models. *ResQue* is a user study-based model that consists of thirteen constructs and a total of sixty questions. It categorizes the questions into four higher-level constructs, which are: Perceived system qualities, users' beliefs, users' subjective attitudes, and their behavioral intentions. The model aims to assess the perceived qualities of recommenders, such as usability, usefulness, and interface quality. This framework presents a unified approach to user-driven evaluation. However, it is time-intensive for participants and may be overly costly for focused hypothesis testing. Also, it focuses only on the usability aspects of RS.

AppFunnel [157] is another example of a usage-centric evaluation protocol; it adopts the concept of conversion funnels to the mobile app RSs. The main goal of *AppFunnel* is to extend the evaluation of mobile app RSs from only considering the instant use of the app to considering the long-term use of it. To do so, it allows the usage-centric evaluation to consider application engagement stages along the applications' lifecycle beyond installation. *AppFunnel* suggests that users go through four stages in order to reach the final conversion. The final conversion represents the conversion of the application from a recommended app to one that is used in the long-term. The four stages are View, Installation, Direct Usage, and Long-term usage. Like many other models, *AppFunnel* is not applicable to all kinds of RSs; it can be applied for recommenders that recommend items that can be used for a long time (applications recommendations, in particular). Moreover, it considers only one dimension, which is the user's engagement, as a measure of RS quality.

Erdt et al. [149] aim at providing a hybrid evaluation approach that aims to alleviate the gap between the accuracy of online evaluation and the simplicity of offline experiments. Hence, Erdt et al. investigated the use of crowdsourcing[20] (such as microWorkers and CrowdFlower websites) as a supporting source of willing users to evaluate Technology Enhanced Learning (TEL) RSs. The authors aimed to evaluate the novelty, relevance, and diversity of a proposed TEL recommender. They used crowdsourcing as a source to recruit participants. Their results show that using workers from crowdsourcing can be a potential alternative for evaluating TEL. This work is tailored for a single domain of RSs, and it is a questionnaire-based approach.

Olmo and Gaudioso [145] propose an objective-based framework for the standardization of recommendation system evaluations. They view RSs as applications that consist of two main subsystems: interactive (called the Guide) and non-interactive (called the filters). The guide concerns of when and how the recommended items should be presented to users. The filter focuses on what to recommend (i.e., which item should be shown to the user). Accordingly, they categorize RSs based on these two functions. Consequently, a performance metric (P) has been introduced as the quantification of the final performance

[20] Crowdsourcing: "an open call to online users from a very large community to contribute to solve a problem or to perform a human intelligent task in exchange for payments, social recognition or entertainment"[149]

of a recommendation system over a set of sessions. (P) is defined as the number of selected relevant recommendations that have been followed by the user over a recommendation session. This work is comprehensive enough, such that it covers a wide range of RSs. However, it concerns only one evaluation dimension. Also, an issue related to the introduced measure (P) is that it does not treat the usual problems that traditional metrics do. For instance, if, in one session, the user followed ten items (i.e., ten items were indeed useful) out of a million recommendations, this is considered as good as if the system displayed ten recommendations, all of which the user followed (i.e., all of which were useful).

Schröder et al. [158] emphasizes the idea of defining evaluation goals before starting the evaluation process. Particularly, the paper suggests that evaluators should define RS's objectives, tasks (e.g., prediction, ranking, and classification), and user's preferences data to be collected. Based on this information, an evaluation could be set. The authors also analyzed the relationship between objectives for applying recommender systems and metrics used for their evaluation. In particular, they discuss three main classes of accuracy-related evaluation metrics, namely predictive accuracy metrics, classification accuracy metrics, and rank accuracy metrics. Based on that, they provide some guidelines to choose the best metric. Unfortunately, this article only discussed accuracy-related measures while it neglects others.

These are the evaluation methods (or models) that are most relevant to our work. We observed that the literature suffers from three main limitations comparing to our proposed approach. These limitations can be summarized as follows: Firstly, *Limited domain*, such that the proposed model is application-based (i.e., it is proposed to evaluate RSs of a particular application area), such as [148] and [157]. Secondly, *Limited dimensionality*, where the proposed method considers only one or two evaluation dimensions, this issue exists in most of the proposed protocols (e.g., [29] and [154]). Finally, *highly costly*; some protocols are focusing on properties that cannot be evaluated offline. These protocols rely on the user study evaluation model (e.g., [149], [156], and [159]). The problem with user-centric evaluations is that they are considered costly regarding time, effort, and money.

5.4. *ComPer* Main Idea

This section discusses the proposed *ComPer* evaluation approach. Specifically, it provides a detailed description of our innovative analogy between human's cognitive skills and RSs (section 5.4.1). Then, it presents the correlation between the cognitive dimension of Bloom's Taxonomy and both the main phases and evaluation dimensions of RSs as discussed in Section 5.4.2 and Section 5.4.3, respectively.

5.4.1 Recommender Systems and Humans: An Analogy

This section illustrates the core of *ComPer*, which is the mapping that we propose between the cognitive skills of human beings and the main phases of RSs. To exemplify the mapping, this section provides a simple analogy between a salesperson, *Alice*, and an arbitrary recommender system *RecSys*, as follows:

To give recommendations, *RecSys* collects information about users to build their profiles and then exploits users' characteristics and information to predict their preferences. Analogously, a salesperson, *Alice*, tries to recognize her customers by collecting information about them in order to become aware of their preferences and to suggest items that may be of their interest. Up to this point, it can be noted that both *RecSys* and *Alice* tend to ***collect*** information about their customers/users to build enough knowledge base and to ***learn*** about their preferences, then they use this knowledge to suggest or ***recommend*** items that match customers' preferences. We notice that both *Alice* and *RecSys* can perform the same tasks while going through the same phases: collecting information, learning, and recommending or predicting. As a human being, Alice deploys her cognitive skills to perform these tasks as follows:

> **Alice as a RecSys**: during her work, Alice, the salesperson, tries to *remember* her customers and *understand* their needs in order to ***collect information*** about their purchasing trends. After that, she *analyzes* the customer's attitudes based on the already collected information. Such an analysis leads Alice to ***learn*** her customers' preferences. By the end of this learning process, Alice reaches a level where she becomes able to *link* pieces of information together and *creates* patterns for each customer. Therefore, she is able to ***predict*** what may/may not attract a

customer, and more than that, to ***recommend*** the best product that suits each customer.

The above analogy suggests that Alice, as a human being, performed the same tasks as an RS. From the other side of the analogy, RSs could be viewed as a human being with some levels of comprehension skills that humans possess. Consequently, we can posit that an RS can be analyzed and assessed in the same way human cognitive abilities are assessed. In particular, since the recommendation process (Figure 24) has similarities to the humans learning process (as discussed in the above analogy), we imply that RS learning outcomes (represented by its recommendations) can be assessed in the same way that human learning outcomes are assessed. To elaborate further with our mapping, we suggest using Bloom's cognitive dimension and its six learning objectives as a basis to evaluate the recognition abilities of an RS.

5.4.2 Mapping Between Bloom's Cognitive Domain and RS's Phases

According to the aforementioned analogy, and based on the definitions of Bloom's learning objectives (Section 5.2.1) and RSs phases(section 5.2.2), we inferred the following mappings: to ***collect information***, an RS should be able to retrieve information and construct meanings from different types of messages, which means that an RS should be able to *remember* and *understand*. To ***learn***, an RS should be able to *apply* prior knowledge in different situations and *analyze* the interrelationships between different components of a system. Finally, an RS's ***recommendation/prediction*** depends on its ability to make a judgment (i.e., *evaluate*) and recognize elements into new structures (i.e., *create*).

Figure 25 visualizes the aforementioned mapping. At the top of the figure, we see the fundamental goal of the recommender system, which is to provide a high-quality recommendation. The middle level of the figure shows RS's main phases, while the bottom level shows human's learning objectives. As the figure indicates, the high quality of a recommendation is achieved by the recommendation process presented by its three main phases. In addition, the Recommend phase is affected by both the amount of information collected and the efficiency of the learning process (as mentioned in section 5.2). The lower layer of the figure shows the association between the learning objectives and the RS phases, with a causal association relationship. This means that the quality of each phase of RS is

affected by the quality of the learning objectives. For example, the recommender's ***Learning*** process gets better as the recommender's ability to *Analyze* or *Apply* is enhanced.

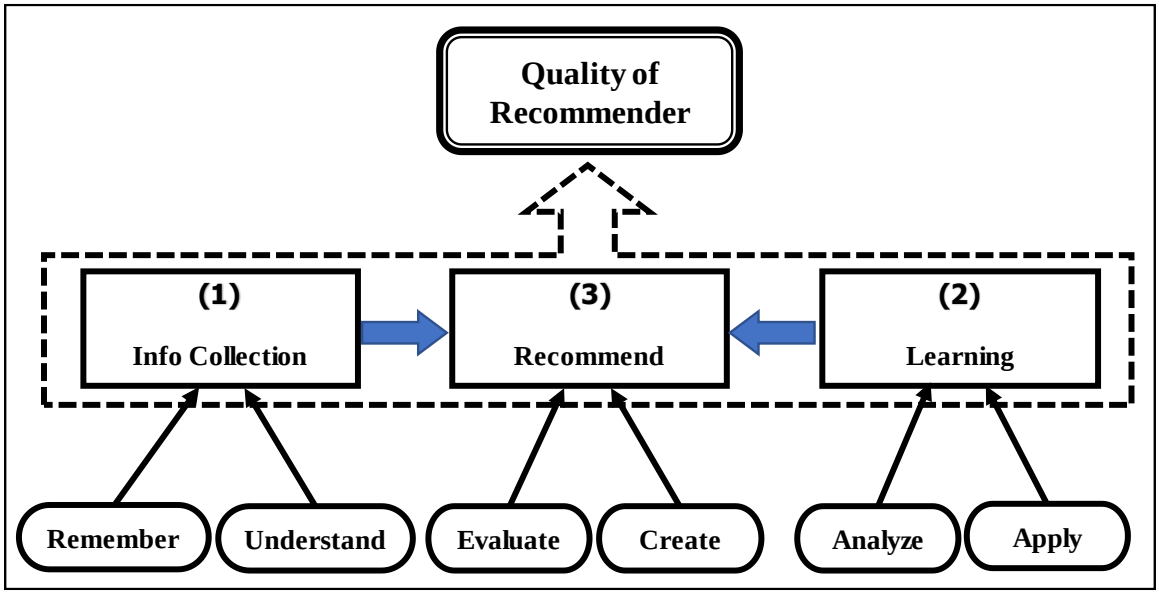


Figure 25 The mapping between recommenders' main phases and Bloom's learning objectives

To conclude, Figure 25 shows that the overall quality of an RS is proportional to its recommendation process, which in turn, is proportional to the recommender's cognitive skills, presented by Bloom's learning objectives. That is, the quality of the recommender increases as its recognition skills increase. From this perspective, we suggest adopting the evaluation dimensions of humans' cognitive skills to evaluate the cognitive skills of RSs. These dimensions, however, are unfamiliar in the recommendation context. Thus, the next section establishes correlations between Bloom's learning objectives and the five most popular RS evaluation dimensions, such that each learning objective is captured and assessed by these evaluation dimensions. The subsequent section describes the process by which we inferred these correlations.

5.4.3 Correlating Bloom's Learning Objectives and RS Evaluation Dimensions

The correlation between learning objectives and evaluation dimensions is built based on the definitions of both sides. This mapping is performed through a multi-step process,

which involves domain experts' inferences, Natural Language Processing (NLP), and the matching of the results obtained from the first two steps. The correlation aims to infer numerical mappings between each one of the evaluation dimensions and the corresponding learning objective.

The correlation is inferred based on the definitions and descriptions of the learning objectives and the evaluation dimensions discussed in the literature. For the results to be more accurate, especially in the NLP step (which strongly depends on these definitions), we tried to find as many related descriptions as possible. We elicit definitions of the learning objectives and the evaluation dimensions from different resources, including research papers and web resources. The text used to infer the mapping contains a total of 7255 words. After excluding repeated and stop words, the document retained 1947 words. As an example, Appendix E shows the text related to the *Scalability* dimension along with its inferred vector. The following two subsections describe the process of inferring the final correlation matrix.

Natural Language Processing-based Correlation

The purpose of this step is to provide numerical means that demonstrate how close (or how important) an evaluation dimension is to a learning objective. These numerical values help establish the weights of each evaluation dimension in the final value of *ComPer*, as described in section 5.5. To infer these values, we adopted concepts from the Natural Language Processing (NLP) domain following the approach presented by Duwairi [160]. In particular, we presented the problem as an NLP *classification* task. Each learning objective is considered as a document category, and the evaluation dimensions are considered as documents that need to be classified under these categories.

The learning objectives and the evaluation dimensions are presented as feature vectors using the definitions that we have collected. From these definitions, we created a bag-of-words for each of the objectives and the dimensions. These words were pre-processed by removing stop words and repeated words, and a word-stemming is done for the remaining words. Then, the *Similarity Dice* measure is used to find the similarity between every two vectors [160], as defined by equation 5.

$$Sim_{Dice} = \frac{2 \times |F_x \cap F_y|}{|F_x| + |F_y|} \quad (5)$$

Where F_x is the feature vector of a learning objective x and F_y is the feature vector of an evaluation dimension y. The similarity Dice (Sim_{Dice}) value represents the strength of correlation between an evaluation dimension and a learning objective, as illustrated in Table 10. For instance, *Correctness* is more related (or important) to *Remember* (0.075) than *Coverage* (0.047).

Table 10 Correlating Evaluation dimensions & Learning objectives (NLP results)

	Remember	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
Correctness	0.075	0.094	0.079	0.079	0.096	0.038
Coverage	0.047	0.040	0.078	0.093	0.103	0.091
Diversity	0.067	0.090	0.102	0.135	0.107	0.111
Robustness	0.052	0.077	0.045	0.113	0.094	0.086
Scalability	0.060	0.081	0.159	0.136	0.138	0.101

Human-based Correlation

In this step, three experts were given the definitions of the learning objectives and the evaluation dimensions. The experts were asked to infer the correlations between both concepts. Specifically, they were asked to give one of the labels (*N*, *P*, or *S*) to describe the correlation. These labels represent the strength of the correlation between the corresponding learning objective and evaluation dimension, and they are interpreted as follows, (*N*) means "No-Correlation," (*P*) means "Partial correlation," and (*S*) means "Strong correlation." The experts were not familiar with the thesis content, and they were asked to infer the correlation without explaining why we need to infer this correlation.

To explain how the experts inferred the correlations, let's take, for example, the learning objective *Remember* and the evaluation dimension *Coverage*. As mentioned previously in section 5.2, *Remember* represents the ability to retrieve and recall relevant knowledge from long-term memory. *Coverage* represents the proportion of available information for which recommendations can be made (i.e., to what extent the RS covers items or users' space). Based on these definitions, the experts had the following two observations: 1) the more the information available, the better the remembering skill will be; and 2) the more the item/user space is covered (or retrieved), the higher the coverage value will

be. According to these observations, the experts indicated that *Coverage* is strongly correlated (*S*) to *Remember*. Following the same rationale, experts have inferred the rest of the correlations.

Each expert was asked to infer the correlation in isolation of the others. That is, the experts did not communicate with each other. Every expert provided his/her own correlation table. Then we aggregated the received correlation tables from all the experts as follows: if at least two experts gave the same label, that label is considered as the final correlation label. If each expert gave a different label (i.e., if the labels of the three experts were N, P, and S), then the P label (as a median value) is considered. Table 11 shows the correlation table resulted after aggregating the obtained evaluation from the experts.

Table 11 Correlating Evaluation dimensions & Learning objectives (experts results)

	Remember	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
Correctness	P	P	P	S	S	S
Coverage	S	P	S	P	S	S
Diversity	S	S	P	S	P	S
Robustness	P	P	S	S	S	P
Scalability	S	N	S	P	P	N

Matching the NLP-based and the Human-based Correlations

To assess the consistency between experts' results and the NLP results, we created a mapping between *Dice* similarity values and the correlation labels (*N*, *P*, or *S*) obtained from experts. To ensure that this mapping is representative, we calculated the difference (d) between the maximum and the minimum Sim_{Dice} values (presented in Table 10). Then we divided the range of (d) into three intervals that correspond to *N*, *P*, and *S*, as depicted in Figure 26 and equation 6.

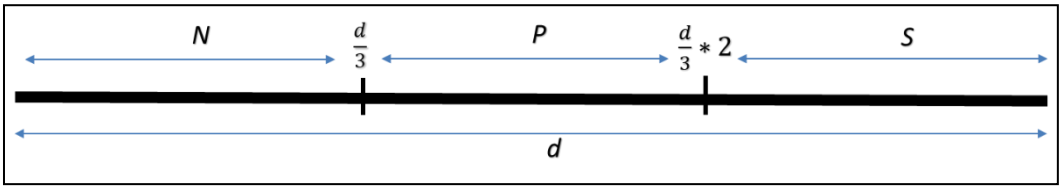


Figure 26 Mapping Dice similarity values to experts' correlation labels

$$SimTOCorr = \begin{cases} N, & Sim_{Dice} \leq d/3 \\ P, & \frac{d}{3} < Sim_{Dice} < \frac{d}{3} \times 2 \\ S, & Sim_{Dice} \geq \frac{d}{3} \times 2 \end{cases} \quad (6)$$

According to Table 10, the maximum Sim_{Dice} value is (0.159), and the minimum value is (0.038). The difference between these two values is (d = 0.121). Accordingly, and based on equation 6, any value that is <= 0.121/3 is considered to be *N*, a value greater than 0.121/3 and less than 0.121/3*2 is considered as *P*, and finally, a value that is >= 0.121/3*2 is considered as *S*. Table 12 shows the results obtained by equation 6; It shows the (*N*, *P*, and *S*) labels that correspond to each Sim_{Dice} value presented in Table 10.

Table 12 Mapping $\boldsymbol{Sim_{Dice}}$ values to the correlation labels (*N*, *P*, or *S*)

	Remember	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
Correctness	P	S	P	P	S	N
Coverage	P	N	P	S	S	S
Diversity	P	S	S	S	S	S
Robustness	P	P	P	S	S	S
Scalability	P	S	S	S	S	S

By comparing the results of the NLP step as presented in Table 12 with the results obtained from the experts (depicted in Table 11), we found that the experts completely confirmed thirteen cases (44%) of the NLP correlations (i.e., the experts gave the same correlation label). Fourteen cases (46%) of the cases were partially confirmed (i.e., the two results were close to each other). An example of partially confirmed cases is the correlation between *Coverage* and *Remember*, where the experts indicate that they are strongly correlated, while the NLP result indicates that they are partially correlated. The remaining three cases were not confirmed (i.e., the experts and the NLP gave completely different correlations). Specifically, these are the cases where the NLP step gives the label (*S*), while the experts give the label (*N*), or vice versa. With the help of a fourth expert, we re-evaluated the correlation cases that are not completely confirmed by the experts. After this re-evaluation, we confirmed the correlation provided by the NLP step of nine cases.

Table 13 depicts the consistency table that is obtained after the correlation steps. The cells with the letter "O" represent the consistent values (i.e., where the experts' confirmed the

NLP results), while the cells with "X" represent inconsistent results. For instance, the first cell in Table 5 (i.e., the correlation between *Remember* and *Correctness*) has an (O) because the three experts' and the NLP step evaluated this correlation similarly. Also, the next cell (the correlation between *Remember* and *Coverage*) has an (O) because the fourth expert and the NLP evaluated it as a partial correlation. For readability purposes, cells that show consistency are grey-shaded.

Table 13 Consistency table between the experts and the NLP correlations

	Remember	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
Correctness	O	O	O	O	O	O
Coverage	O	O	O	X	O	O
Diversity	O	O	X	O	X	O
Robustness	O	O	O	O	O	X
Scalability	O	X	O	X	X	X

As Table 13 shows, (22) correlations out of (30) in total have been confirmed. To resolve the remaining (8) cases of conflicts, we compared their values that have been obtained as results of the three steps (the first group of experts, the NLP, and the last expert). The mean of these three values is considered as the final correlation. For example, since the three steps evaluated the correlation between *Understand* and *Scalability* as *N*, *S*, and *N* labels, respectively, the value *N* is considered as a final result of this correlation.

The Final Correlation Matrix

As mentioned before, the Sim_{Dice} values are considered as a basis for establishing the weights of the evaluation dimensions in the final value of *ComPer*. These values are considered as follows: for correlations that show consistency with the NLP results (i.e., consistent cells in Table 13), the corresponding Sim_{Dice} values are used. For the inconsistent cells (i.e., the eight "X" cells in Table 13), we used the average of all Sim_{Dice} results that belong to the corresponding correlation label. For example, for "*N*" labels, we took the average of all Sim_{Dice} values (resulted from the NLP step, and presented in Table 10) that were categorized (according to equation 6) under the "*N*" label. According to equation 6 and Table 10, two Sim_{Dice} values are mapped to the (*N*) label, these values are (0.038) and (0.04). The average of these two values is (0.039). Hence, since the correlation label

between *Scalability* and *Understand* is "*N*" (as it is explained in the previous paragraph), it is given the value (0.039).

The final correlation results obtained after this process is presented in Table 14. The table shows the correlation strengths between Bloom's learning objectives and each of the evaluation dimensions of RSs. Rows represent RS dimensions, while columns represent Bloom's learning objectives. The values in each cell indicate the correlation strength, where the higher the value, the stronger the correlation. For instance, the correlation between *Remember* and *Correctness* is stronger than its correlation with *Coverage*. Therefore, *Correctness* has a higher weight than *Coverage* in terms of evaluating the *Remember* objective.

Table 14 The Correlation Matrix (Learning objectives & Evaluation dimensions)

	Remember	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
Correctness	0.075	0.094	0.079	0.079	0.096	0.038
Coverage	0.047	0.040	0.078	0.066	0.103	0.091
Diversity	0.067	0.090	0.066	0.135	0.066	0.111
Robustness	0.052	0.077	0.045	0.113	0.094	0.066
Scalability	0.060	0.039	0.159	0.066	0.066	0.066

These values are used as inputs to the equation used to calculate the final value of *ComPer*. As we will discuss in section 5.5, Each value in the table represents the weight factor of the corresponding evaluation dimension. To clarify, take as an example the first column, which represents the weight factors for the evaluation dimensions in correspond to *Remember*. For instance, the weight factor for *Correctness* is 0.075, and for *Coverage* is 0.047. As described in section 5.5, each factor is multiplied by the value of the corresponding dimension to calculate the final value obtained by *ComPer*.

5.4.4 Visualization of the Correlation Between *ComPer* Components

To give a general overview of the relationship between *ComPer*'s components, Figure 27 depicts a full visualization of these components. At the center of the figure, we can see the three main phases of a recommendation process, along with the learning objectives related to each phase (as described in section 5.4.2). The core of the model is divided into three parts based on RS's main phases; each part contains two corresponding learning objectives.

The evaluation dimensions are sorted based on their correlation strength with the corresponding objective; the closer the dimension to the center, the stronger the correlation with the objective. For instance, *Correctness* has the strongest correlation with *Remember*, and *Coverage* has the weakest correlation.

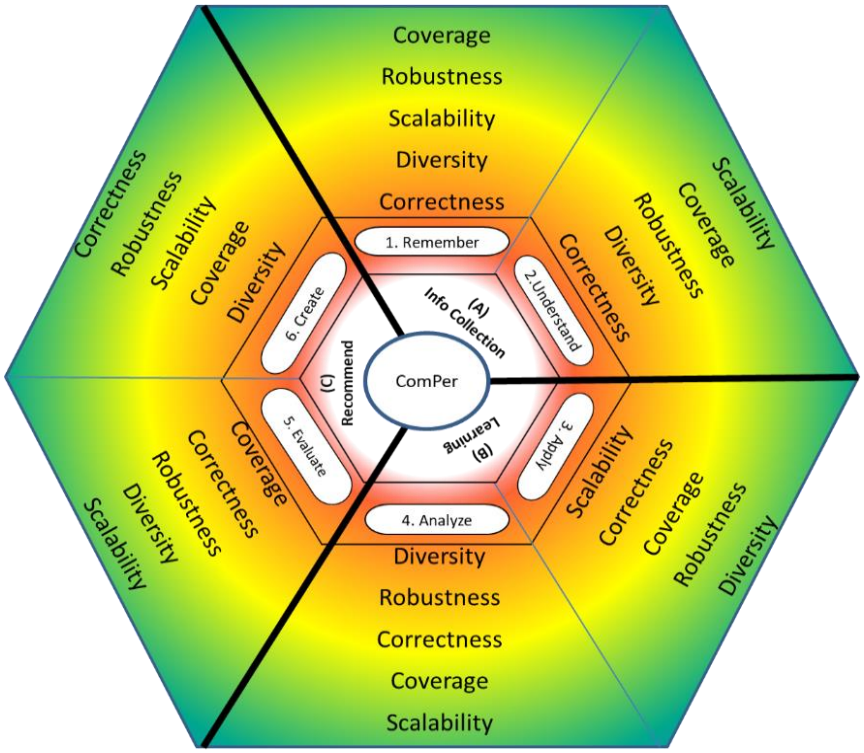


Figure 27 *ComPer's* main components

5.5. Formal Definitions

This section formally defines the components of *ComPer* as follows: the set of all learning objectives is defined as $O = \{Remember, Understand, Apply, Analyze, Evaluate, Create\}$. The set of the evaluation dimensions of RSs is defined as $D = \{V_{Correctness}, V_{Coverage}, V_{Diversity}, V_{Robustness}, V_{Scalability}\}$, where V indicates the values of the corresponding dimension. For instance, $V_{Correctness}$ is the value obtained when we evaluate correctness, $V_{Coverage}$ is the coverage value, and so on. We denote the final evaluation value obtained from *ComPer* as $ComPer_{value}$.

$ComPer_{value}$ embeds the summation of multiple values, where each value represents a learning objective. In particular, to calculate the $ComPer_{value}$, we adopt the Weighted Sum Model (WSM) [161], as follows:

$$ComPer_{value} = \sum_{i=1}^{|O|} A_i^{wsm} \text{ , where} \quad (7)$$

$$A_i^{wsm} = \sum_{j=1}^{|D|} Corr_{ij} \; X \; d_j \quad (8)$$

A_i^{wsm} is the weighted some score of objective ($o_i \in O$), $Corr_{ij}$ is the weight of dimension ($d_j \in D$) with respects to objective ($o_i \in O$), as presented in Table 14, d_j is the value of the j[th] dimension. Equation 7 represents the final evaluation value obtained from *ComPer*, which gives the overall performance of a recommender. The higher the $ComPer_{value}$, the better the recommender.

It is noteworthy to mention that *ComPer* depends mainly on dimensions that already exist in the literature to evaluate the considered dimensions (d_j). The rationale behind this is that these dimensions have already been validated, and they have been used in the literature. The aggregation of these dimensions, however, introduced two issues; firstly, not all dimensions have the same scale of results (i.e., not all of them get a value between 0 and 1, for instance). Secondly, the highest values do not necessarily mean better results. For example, when evaluating *Scalability* in terms of time complexity, the higher the time, the worse the recommender. Another example is regarding *Robustness*, which shows how tolerant to biased or fake information the recommender is. So, when calculating *Robustness* as the difference in accuracy before and after injecting false information, the highest values indicate the lowest *Robustness*. We refer to such dimensions as the "*negative dimensions*." Out of the five dimensions that *ComPer* considers, two of them are negative dimensions, which are *Scalability* and *Robustness*.

To overcome the first issue (i.e., dimensions have different scales), *ComPer* relies on the ad hoc normalization technique using Equation 9:

$$d' = 1 \; - \frac{1}{d+1} \quad (9)$$

Where d' is the normalized value of d. In this way, we normalize all the results to be in the range [0, 1]. We used this ad hoc normalization rather than other normalization methods

(such as the min-max normalization) because it is a general method, and it is applicable for any single value even if the range of values is unknown.

To overcome the second issue (i.e., the negative dimensions), We find (d''), as follows ($d'' = 1 - d'$). Algorithm 1 shows the complete steps of the proposed evaluation approach.

ALGORITHM 1: *ComPer* calculation process

Input: *O, D, correlation matrix (Corr)*
Output: $ComPer_{value}$

Begin:

$ComPer_{value} \leftarrow 0$

1) **for each** d *in* D **do**

$d' \leftarrow 1 - \frac{1}{d+1}$

$d \leftarrow d'$

If $d \in NegativeDimensions$

$d'' \leftarrow 1 - d'$

$d \leftarrow d''$

2) **for each** o *in* O **do**

$o \leftarrow \sum_{d \in D}(Corr_{od} * d)$

3) $ComPer_{value} \leftarrow \sum_{o \in O} o$

End

5.6. Chapter Summary

This chapter proposed the *Comprehensive Performance* (*ComPer)* evaluation approach, which provides a comprehensive assessment of recommender systems. *ComPer* aims to be *thorough* (by combining multiple dimensions), *simple* (by presenting the final result as a single value), and *independent* (by providing comparable results for different settings). The chapter proposed an analogy between RS and humans with cognitive skills. It also discussed how we inferred the correlation matrix between RS dimensions and Bloom's taxonomy. The next chapter shows an empirical evaluation of *ComPer* that demonstrates its feasibility in terms of its ability to provide results that are comparable over different evaluation settings.

Chapter 6. Validation of the Cognitive-Based Evaluation

This chapter discusses the experimental validation conducted to validate the cognitive-based evaluation approach discussed in Chapter 5. It aims to demonstrate the usability of *ComPer* and its ability to provide results that are independent of evaluation settings. In addition, it empirically illustrates the feasibility of using *ComPer* in terms of conducting comparisons between recommendation algorithms, as opposed to the conventional evaluation of RSs. The content of this chapter is presented in [142].

6.1. Methodology and Configurations

All the experiments were executed on a laptop with 6 GB RAM, 500 GB hard disk, Intel Core i5-4200M 2.5 GHz quad-core CPU, running Windows 10-x64. All algorithms, evaluation measures, and other settings are implemented using *Librec* [162], an open-source Java library for recommender systems. *Librec* is one of the modern Java-based recommendation libraries that provides a large number of recommendation algorithms. Moreover, just like *MyMediaLite*, *Liberc* implements state-of-the-art algorithms in addition to the classical ones [163]. However, *Librec* runs faster than *MyMediaLite*, according to a benchmark evaluation provided by Guo et al. [162]. Many researchers found *Librec* a reliable library to implement and evaluate their algorithms [164]-[166].

Since the experiment's goal is to show the applicability of *ComPer* rather than to demonstrate the superiority of a particular recommendation algorithm over the others, it does not matter which algorithms are evaluated. The experiments compare two collaborative filtering algorithms, namely the *AspectModel* [167] and Probabilistic Latent Semantic Analysis (*PLSA*) algorithm [168]. Both algorithms are ranking algorithms (i.e., they provide a ranked list of top-k items). The performance of these algorithms has been assessed

using three datasets, namely *FilmTrust, Movielens 100-k* (or ML-100k, for simplicity), and *CiaoDVD*. We downloaded all the datasets from the *Librec* official website[21].

FilmTrust is the smallest dataset compared to the other two datasets. It was crawled by the *Librec* team from the movie website, *FilmTrust*. It contains 35497 data records, each of which has three attributes: user Id, movie Id, and a rating. The ML-100k is a well-known dataset that is provided by the *GroupLens* research lab for public use. It contains 100,000 ratings provided by 1000 users on about 1700 movies. Like *FilmTrust*, *ML-100k* has three attributes (i.e., user Id, item Id, and rating). The last and the largest dataset that we used is the *CiaoDVD* dataset. The data were crawled by the *Librec* team in December 2013. The dataset contains users' ratings of DVDs on a scale from 1 to 5, timestamps for ratings, and trust information. Table 15 summarizes the properties of these three datasets, where Sparsity indicates the portion of items that the user interacted with (or rated) [169].

Table 15 Datasets properties

	Users	**Items**	**Ratings**	**Sparsity**
Filmtrust	1508	2071	35497	1.14%
ML-100k	943	1682	100000	6.3%
CiaoDVD	7375	99746	278483	0.038%

Each of the evaluated algorithms has been assessed on the three datasets mentioned above. The data is split into a training set and a testing set randomly based on ratings. A ratio-based data split has been followed, with a training set ratio varied from 0.2 up to 0.8 (with steps of 0.2). That is, we have four different data splits (0.2, 0.4, 0.6, 0.8) for each dataset. The number of generated items for the *Top-N* recommendation has been fixed to be *N*=10.

6.2. Evaluation Dimensions

As it is mentioned in section 5.2.3, *ComPer* considers five evaluation dimensions. It is worthwhile to mention that some of the evaluation dimensions can be evaluated using different measures or evaluation strategies. For this, it is necessary to clarify the measures used to assess each dimension, especially because *ComPer* depends heavily on the results

[21] https://www.librec.net/

of these measures. The following are brief explanations of the measures used in the experiments for each of the evaluation dimensions.

- *Accuracy (Correctness)*: correctness is the most commonly used dimension in the literature. Different measures have been introduced to evaluate it, such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Area Under Curve (AUC), etc. Since our experiments consider ranking algorithms, we used the AUC to assess accuracy. The AUC measure is provided as a class (named "AUCEvaluator") within the *Librec* library.
- *Coverage*: according to Avazpour et al. [6], coverage refers to either the item-space (catalog coverage) or user-space (prediction coverage). In this thesis, we consider the catalog coverage, which is implemented as the percentage of recommended items to all users within an experiment [170].
- *Diversity*: different measures have been introduced to evaluate diversity. For our experiments, we used the "diversityEvaluator" class that is already implemented in *Librec*. It calculates diversity as the average dissimilarity of all pairs of items in the recommended list at a specific cut-off position [171].
- *Robustness*: this dimension has been calculated as the Average Hit Ratio (*Average-HR*) shift after attacking the dataset [6]. To do so, we implemented a nuke attack (i.e., to inject fake profiles that try to nuke or destroy some of the items). Then we calculated the Average-HR for both the original dataset (dataset without false information) and the attacked dataset (i.e., after injecting false information). The shift of the Average-HR is the difference between the two hit ratios. For instance, suppose that the Average-HR before the attack was 0.23 and after the attack becomes 0.15, then the *Robustness* value is the difference between these two values, which equals 0.08.
- *Scalability*: according to Avazpour et al. [6], recommendation time is an important indication of the system's Scalability. Thus, we relied on the recommendation time (i.e., the time spent by a recommender to learn and recommend) as a measure of recommenders' *Scalability*.

6.3. Results

This section presents and discusses the evaluation results obtained from the experimental evaluation. It is divided into two sections as follows: Section 6.3.1 discusses the results over a single dataset. Section 6.3.2 shows the effect of changing the experiment setting (i.e., using multiple datasets) on *ComPer*'s results compared to the conventional evaluation method.

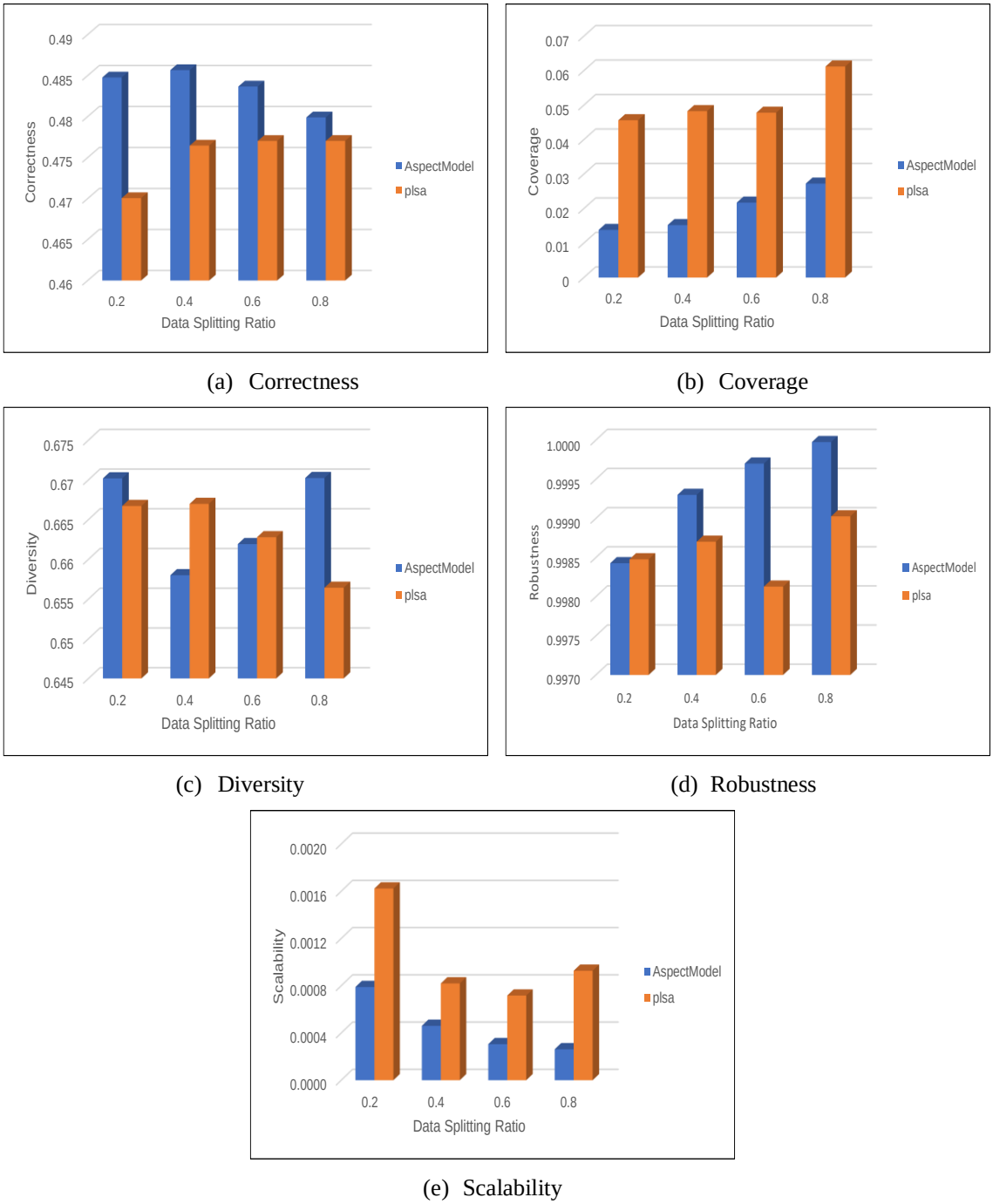
6.3.1 Results for a Single Dataset

This subsection shows the difference between the traditional evaluation of RS, which involves presenting the results of each evaluation dimension, *one at a time*, compared to evaluating RSs using *ComPer*, which involves presenting the results of evaluation dimensions, *all at once*. It discusses the results obtained from the experiments on the *FilmTrust* dataset. The section sheds light on the issues that may face an evaluator when she compares two recommendation algorithms over multiple dimensions. Figure 28 shows the traditional evaluation results, where each dimension is evaluated separately and presented in a figure on its own. Each chart shows the results of one dimension using the *FilmTrust* dataset, using four different splitting ratios (0.2, 0.4, 0.6, and 0.8).

The charts show noticeable fluctuations in the superiority of one recommendation algorithm over the other. *Correctness* and *Robustness* charts show the superiority of the AspectModel algorithm, while *Coverage* and *Scalability* charts show the superiority of the PLSA algorithm. Moreover, fluctuations could appear over the same dimension, as shown in the *Diversity* chart. These fluctuations show that the plentiful evaluation options make it easy to find an evaluation design that suits one algorithm but not the others, which in turn leads to an unfair comparison between different algorithms [172].

$ComPer_{value}$, on the other side, comprehends all these results in a single yet thorough value. Figure 29 depicts ComPer's results, which we got by executing Algorithm 1. The comparison between this figure and Figure 28 demonstrates the simplicity of using *ComPer* results. Also, it exhibits how *ComPer* mitigates the fluctuation issues that were apparent in the conventional evaluation (Figure 28). Solving the fluctuation issue shows that *ComPer* results are independent of evaluation settings (i.e., they are not affected by the settings change). As mentioned previously, the abundance of evaluation settings makes

it easy to find evaluation settings that suit one algorithm but not the others. Following such purposely-aligned evaluation settings negatively impact the fairness of the comparison between alternatives. To this end, we say that, by mitigating this issue, *ComPer* has the potential to support fairness (in terms of different evaluation settings) of the evaluations..



(a) Correctness (b) Coverage

(c) Diversity (d) Robustness

(e) Scalability

Figure 28 Conventional evaluation results using the *FilmTrust* dataset

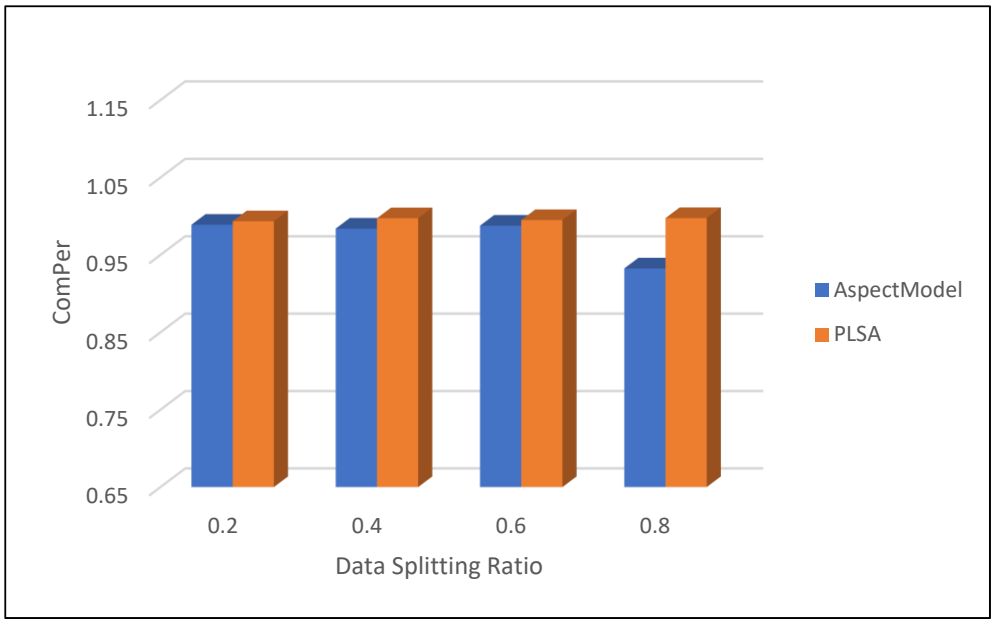


Figure 29 *ComPer* results over the FilmTrust dataset

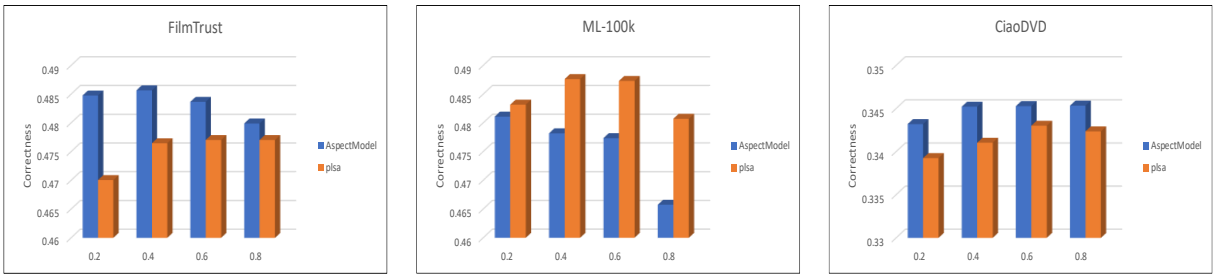
To this end, it is noteworthy to mention that the $ComPer_{value}$ measure provides an overall evaluation of the recommendation method. *ComPer* aims to simplify decision-making by managing the conflict between multiple criteria. $ComPer_{value}$ reflects the quality of the recommender in terms of the five considered evaluation dimensions. Such a comprehensive measure is very useful, especially that the literature witnesses a consensus that evaluating recommenders based on single dimensions is insufficient to comprehend the recommender's overall quality. A single dimension evaluation could be sufficient for small applications when the recommender is designed for a special purpose and tailored for a special group of users. In such a case, the system designers and decision-makers could have previous knowledge about the expected users and other parties of the system. Therefore, they can implicitly weigh the multiple criteria and be comfortable with the consequences. However, for large-scale applications that involve multiple parties and a large number of users with various interests, the decision becomes more complicated, and the consequences of such a decision are magnified comparing to simple applications. Thus, it becomes essential to rely on a non-conventional evaluation method to select the best recommendation approach.

6.3.2 The Effect of Multiple Datasets

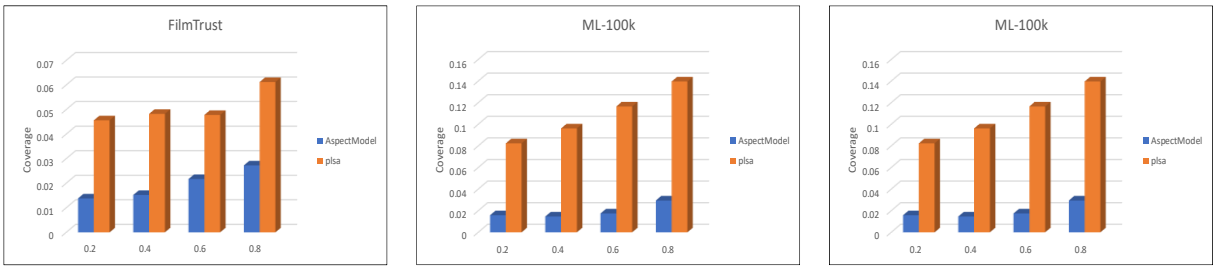
This subsection investigates the effect of using different datasets on the *replicability* of the results. That is, whether the results over one dataset are comparable to the results over another dataset. Figure 30 shows the results obtained from comparing the two recommendation algorithms, AspectModel and PLSA, over the three datasets (*FilmTrust*, *ML-100k*, and *CiaoDVD*). The figure is organized according to the five evaluation dimensions (*Correctness*, *Coverage*, *Diversity*, *Robustness*, and *Scalability*).

As can be inferred from the figure, the issues we noticed over a single dataset also exist over multiple datasets. Figure 30 shows fluctuations of the results over a single dataset. For instance, over the *CiaoDVD* dataset, the AspectModel algorithm overcomes the PLSA algorithm in terms of *Correctness* and *Robustness*, whereas PLSA dominates in terms of *Coverage* and *Scalability*. In addition, the fluctuations are visible for the same dimension over different datasets. In terms of *Correctness*, for example, the AspectModel algorithm surpasses the PLSA over two datasets (*FilmTrust*, and *CiaoDVD*), but not over the third dataset (*ML-100k*).

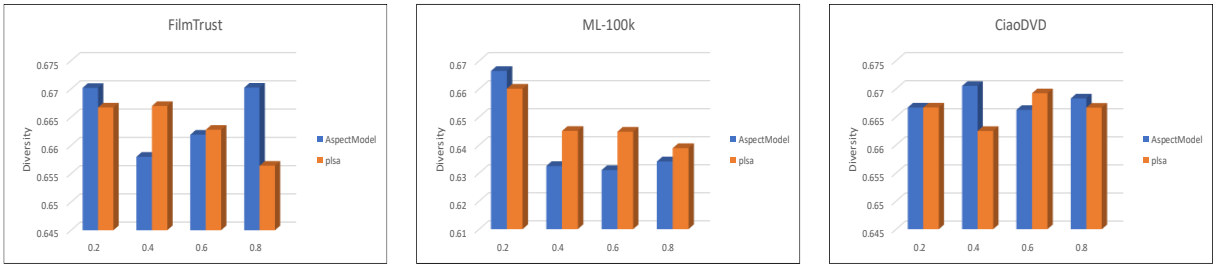
An important observation in this context is that analyzing and reasoning about the five evaluation dimensions over the three datasets requires fifteen (15) charts, which makes it tedious for RS designers to decide which algorithm is better than the other. In contrast, Figure 31 demonstrates the clearness of the results that are obtained by *ComPer*. Instead of using (15) charts, *ComPer* regroups the evaluation results using only three charts. Another issue is that conventional evaluation is highly affected by evaluation settings, such as using different datasets or different data splits. For instance, the results of the same dimension could be changed when the evaluation uses a different dataset. According to Bellogin et al. [172], this issue makes it easy to find an evaluation design that suits an algorithm but ineligible for others, representing an increasing barrier to compare different studies fairly. It is sophisticated to infer a conclusion from Figure 30. It does not show an absolute superiority of one algorithm over the other for all datasets, neither it indicates the superiority of an algorithm over a single dataset. On the other side, we can easily draw conclusions from Figure 31, using *ComPer* results, which show that PLSA is always better than the AspectModel algorithm in the Filmtrust, Ml-100k, and CiaoDVD dataset.



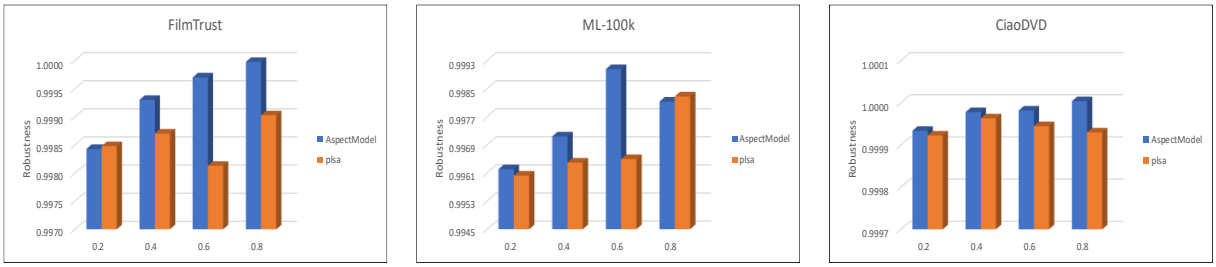
(a) Correctness



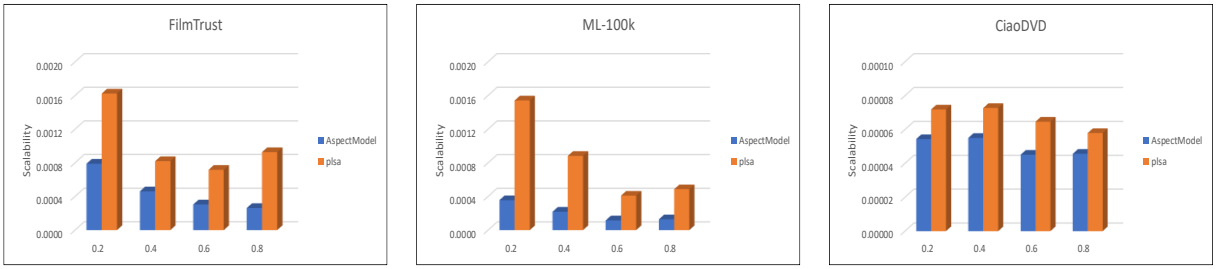
(b) Coverage



(c) Diversity

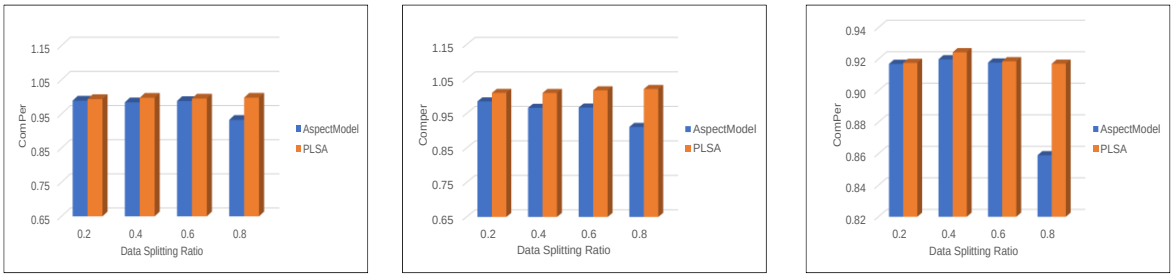


(d) Robustness



(e) Scalability

Figure 30 Conventional evaluation results over the three datasets.



(a) FilmTrust (b) ML-100k (c) CiaoDVD

Figure 31 ComPer results over three datasets

This section conforms with the conclusions of the previous section, such that *ComPer* mitigates the dependency issue (i.e., the effect of changing settings on the obtained results). It also shows how using *ComPer* results simplifies the comparison between different algorithms, and it mitigates the fluctuations that appeared with the conventional results. Besides, it warned us that the inconsistency issue increases as the options for evaluation settings increase. Nevertheless, it shows the effectiveness of *ComPer* to overcome these issues.

It is noteworthy to mention that aggregating multiple evaluation dimensions into a single dimension may cause information loss. That is, it may downgrade the importance of individual dimensions. Nonetheless, relying on a single evaluation dimension is insufficient, as mentioned before. Also, considering multiple single dimensions is a very complicated task, especially when many alternatives are available. In such a situation, aggregation becomes a useful solution to simplify the comparison. Using such aggregated value is helpful and can be used to reduce the number of alternatives (i.e., to select a *set of the best alternatives*). Then a detailed evaluation setting can be used to put more focus on the single dimensions. To this end, we proposed the goal-modelling-based evaluation, which is discussed in Chapter 7.

6.4. Chapter Summary

This chapter evaluated the applicability of *ComPer* (the evaluation method discussed in Chapter 5) empirically using three different datasets. Our results are promising in the sense that *ComPer* is independent of evaluation settings. That is, it is able to provide consistent

results despite changing experimental factors, such as dataset and splitting methods. The results also demonstrated that drawing conclusions based on *ComPer* results is more convenient and easier than the conventional multidimensional evaluation.

By comparing the final results of *ComPer* with the results obtained when evaluating each dimension separately, we notice that *ComPer* achieved its three goals: thoroughness, simplicity, and independency. Thoroughness is presented by the *ComPer* evaluation methodology, in which it combines the results of the dimensions while preserves comparable results. Simplicity is clear if we compared the final results of *ComPer* with the results of the five dimensions in a conventional way. Moreover, these results show that *ComPer* is unbiased to a particular evaluation setting; That is, it allows for a fair evaluation for different recommenders.

Chapter 7. Goal Modeling-based Evaluation for Recommender Systems

In Chapter 5, we proposed *ComPer* as a potential solution to evaluate recommender systems based on the idea that a recommender system can be viewed as a human being with cognitive skills, and so, the recommendation process can be analogized to a human learning process. Inspired by this idea, we proposed mapping between the learning objective of human beings and the evaluation dimensions of RSs. Then, we inferred correlation values between these concepts. These values are then used to calculate the final results of the proposed framework, which we called $ComPer_{value}$. This chapter suggests another evaluation approach based on viewing a recommender system as *a multi-stakeholder system*, with each stakeholder having a set of goals that are different from, and often contradictory to, the goals of the other stakeholders.

It is worth mentioning that the goal modeling-based approach presented in this chapter is not mature and not validated enough. This chapter, however, contributes to the literature in various directions; First, it raises awareness towards the adaptation of this approach in the context of RS. To the best of our knowledge, this is the first work that investigates the use of a goal modelling approach to evaluate RSs. Second, using such an approach has the potential to help system designers to select the most appropriate recommendation approach at the early stages of system design. Third, goal modelling-based evaluation allows designers to evaluate different aspects of the system, such as functional and non-functional requirements as well as multistakeholders' goals. Finally, establishing a well-designed goal model for RS has the potential to improve the recommendation algorithm itself, not only evaluate the system.

7.1. Introduction

This chapter provides a Goal-Oriented evaluation approach that evaluates RSs beyond their correctness. The proposed approach does not only consider the evaluation dimensions or metrics of RS (as in the case of *ComPer* that we discussed in Chapter 5) but also takes into

consideration the different parties or stakeholders (such as end-users, providers, and the recommender system itself) and their goals/objectives. The purpose of suggesting this framework is to take stakeholders' goals as a basis to evaluate recommendation algorithms and select, among the set of alternative algorithms, the best algorithm that satisfies stakeholders and achieve their goals.

Despite the existence of various evaluation methods in the literature (as discussed in section 5.3), most of these methods and evaluation dimensions assess RSs from the users' perspective only [173]. In general, focusing on the user perspective is important as the user is an essential party of any system. Although such a concentration on the users is accepted, it ignores other parties (or stakeholders) of RSs. For instance, in eCommerce scenarios, the recommendation algorithm may follow a procedure that selects items based on the best price for the customer. Such an algorithm ignores other parties' requirements, such as increasing the business's profit, for example. In such a situation, considering users' requirements only is insufficient. Other parties should be considered, as well.

Dealing with a recommender system as a multi-stakeholder system is a relatively new research direction [174]. After a careful review of the state-of-the-art related to recommender systems, and to the best of our knowledge, we found a lack of approaches that consider the requirements of RS's multiple stakeholders while selecting the best recommendation algorithm among a set of alternatives. This observation is supported by the work proposed by Burke et al., [175] that confirms the lack of comprehensive inclusion of the multistakeholder concept into the recommendation settings. Also, Sohrabi et al. [176] raised awareness towards the issue of lacking a systematic methodology that is able to evaluate and compare a set of alternative RSs based on several criteria. Hence, an evaluation approach that can handle the multiple requirements of all parties in the recommender system context is desired [176], [177].

Very limited studies have tailored their evaluation methodology to consider multiples stakeholders. One of the earliest research in this direction is proposed by Said et al. [177]. The authors proposed a prototype of an evaluation model called A 3D Recommender System Benchmarking Model. The main goal of this approach is to take into consideration the non-functional requirements that come from business goals and technical requirements. The model captures three main evaluation aspects, which are user requirements, business

requirements, and technological constraints. Each aspect is represented by a set of qualities. A more recent approach is the utility evaluation models for RSs, which is proposed by Burke et al. [175]. The authors proposed a utility function that takes into consideration the multistakeholder of RSs. The overall system utility is proposed as a tuple of utilities for all users, all owners, and the system. Although this approach considers three stakeholders, it considers only one evaluation aspect (namely, the utility).

To the best of our knowledge, the use of goal-oriented modeling approaches has never been used to assess recommender systems with multi-stakeholders in order to evaluate their goals, which are often conflicting, and to evaluate the best means or alternatives to achieve these goals.

This chapter aims to develop the basis for an evaluation approach that uses goal modeling to support the selection of a recommendation algorithm. Through an illustrative example, we show that it is feasible to model the recommendation algorithm alternatives and their contributions to the satisfaction of stakeholders' goals so that the selected algorithm is well-aligned with the overall system requirements.

7.2. Goal Modeling for Recommender Systems

Goal modeling approaches typically define the goals of systems and stakeholders, along with their relationships such as contributions, dependencies, and decompositions. Goal models are mainly used to analyze systems to determine satisfactions of goals, to discuss potential trade-offs, to share the understanding of who wants what, and to identify and answer what-if questions early before the system being actually deployed [178].

In the context of recommender systems, stakeholders include but are not limited to the end-users who use the system, the system itself, and the service provider. The early identification of stakeholders' goals has the potential to help the recommender system's designers to better understand and evaluate the role of a particular recommendation algorithm in improving specific qualities, such as increasing recommendation accuracy. The overall purpose behind identifying stakeholder's goals early in the design phase is to try to satisfy, as much as possible, the needs of multiple stakeholders, who are different by nature, and often contradictory.

In addition, goal modeling has the potential to help decision-makers, such as systems providers, in ensuring that the provided recommendation algorithms or technologies are aligned with the goals of recommender systems and their main users. Furthermore, goal modeling approaches provide the ability to represent and model the different perspectives and goals of stakeholders in an abstract manner, independently from the actual system being assessed.

The requirements engineering community has proposed many goal-oriented modeling languages over the past years, such as i* [179], NFR framework [180], TRPOS [181], KAOS [182], and GRL [183]. In this thesis, we consider the use of the Goal-oriented Requirement Language (GRL) because it is the only one of these modeling languages that has been standardized [178]; it is part of ITU-T's User Requirements Notation standard [183]. The following subsections introduce the main concepts associated with the GRL goal modeling language.

7.2.1 Stakeholders of Recommender Systems

The term *"stakeholder*" is a well-known term in business. It is defined as any individual or group that can affect, or be affected by, the actions and the objectives of the organization [184]. Abdollahpouri et al. [173] adopted this definition for the recommendation area as follows: "*recommendation stakeholder is any group or individual that can affect, or is affected by, the delivery of recommendations to users.*"

Researchers have defined three primary stakeholders of RS, namely Consumers, Providers, and Systems [173]. Other researchers, such as Said et al. [177], called them Consumer, Vendor, and Service Provider. This thesis uses the definition proposed by Abdollahpouri1 et al., [173], as follows:

- *Consumers* are the end-users of a recommender system. That is, the persons who receive and consume recommendations.
- *Providers* are the suppliers of the recommended items (any kind of items), such as the merchandisers who sell items on Amazon or the owners who list houses for rent on a renting platform.
- *System* (or *Vendor*) represents the entity that owns (or created) the recommendation platform.

We believe that considering the "multi-stakeholder" facet of the RS research area is of vital importance that could pave the way for a new research paradigm. This paradigm may affect the focus, objectives, and priorities of most aspects of RS research, which includes designing, optimizing, implementing, and evaluating RSs. For instance, as a multistakeholder RS incorporates the intentions of all stakeholders, then in the evaluation setting, the best RS is the one that maximizes the benefits of these stakeholders.

7.2.2 Goals of Recommender System

Identifying the goals of recommender systems is a non-trivial task. This is because RSs are implemented in a wide array of domains, each of which has different stakeholders with various requirements or goals.

For instance, Figure 32 depicts an example of a local hosting recommender system (Airbnb as an example), with its stakeholders and their goals. It is worth mentioning that the example illustrated in the figure is a virtual, proof-of-concept example that does not necessarily reflect any real settings.

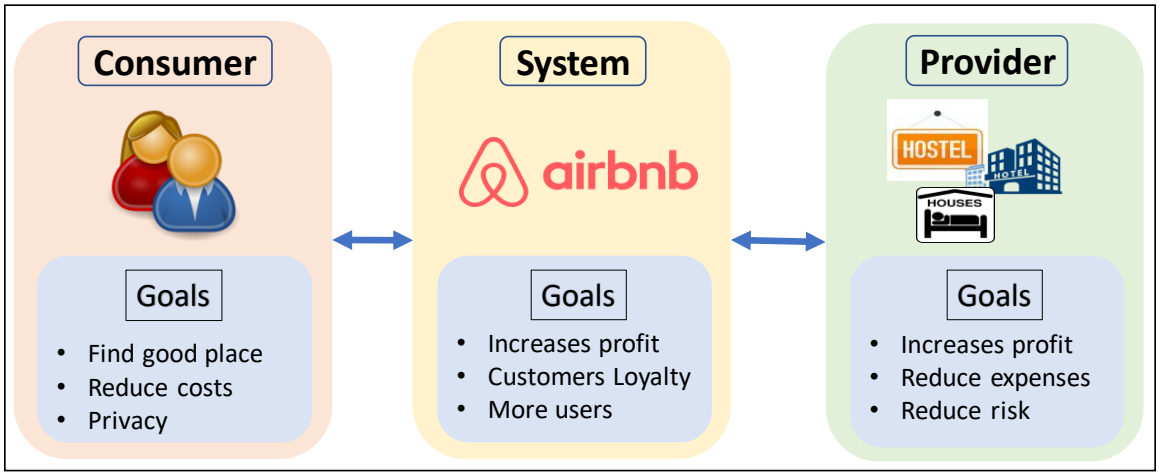


Figure 32 An example of goals for a local hosting recommendation system

From Figure 32, it can be noted that stakeholders could have similar goals or conflicting goals. For instance, one of the *consumers'* goals is to "*find a good place"* while *"reducing the cost."* This goal, to some extent, conflicts with the *System*'s and the *Provider*'s shared goal, namely the "*Increase profit"* goal. In this situation, a good RS is not necessarily the one that provides the highest accuracy; rather, it is the one that can achieve a balance

between stakeholders' requirements and tries to satisfy as many goals as possible. The subsequent section discusses our proposed approach to consider such a situation. It describes our goal-oriented evaluation approach that considers the multistakeholders' goals while evaluating recommender systems.

7.3. Generic Goal Model Structure for Recommender System Evaluation

The overall structure of a goal model for RS evaluation using GRL [183] is illustrated in Figure 33. An evaluation model should include all relevant stakeholders (represented in GRL as actors) and their goals regarding the system under study. Goals in GRL are either hard goals ( ) that can be fully satisfied and measured in a quantifiable way (usually related to functional requirements), or soft goals ( ), which are related to non-functional (or qualitative goals) that has no clear objective measure of satisfaction, such as security or privacy. In addition to goals, an evaluation model should also show the general services, activities, or solutions (presented in GRL as tasks) that the system can/should provide to meet hard goals and soft goals, and the resources ( ) that are needed in order to achieve soft goals, hard goals, and tasks. All these components (i.e., actors, goals, soft goals, tasks, and resources) can be interrelated or structured through different types of links. GRL provides the following links: (1) Decomposition links ( ) are used to connect elements and sub-elements using AND, OR, and XOR decomposition relationships, (2) Contribution links ( ) indicate the positive or negative impact of the satisfaction of one element on the satisfaction of another element, and (3) Dependency links ( ) model dependencies between elements, often across two different actors, to show that an actor depends on another actor to meet a goal, satisfy a soft goal, perform a task, provide a resource, etc. In the case of recommender evaluation, Dependency links connect between tasks/goals and resources required for a particular recommendation algorithm to be selected.

To evaluate a specific recommendation algorithm, a GRL *strategy* is used to describe a particular configuration of alternatives in the GRL model by assigning an initial qualitative or quantitative evaluation value (an integer value between -100 and 100) to some of the elements in the model. A GRL evaluation mechanism then propagates evaluation values to other high-level elements through links and compute their satisfaction

values. Strategies can be compared with each other to help decide on the most appropriate trade-offs among conflicting goals of stakeholders and to choose the best alternative that satisfies these goals. In addition, GRL supports an *Importance* attribute for elements inside actors (also quantitative or qualitative). This attribute is considered when evaluating strategies for the goal model, resulting in satisfaction levels measured at the actor level.

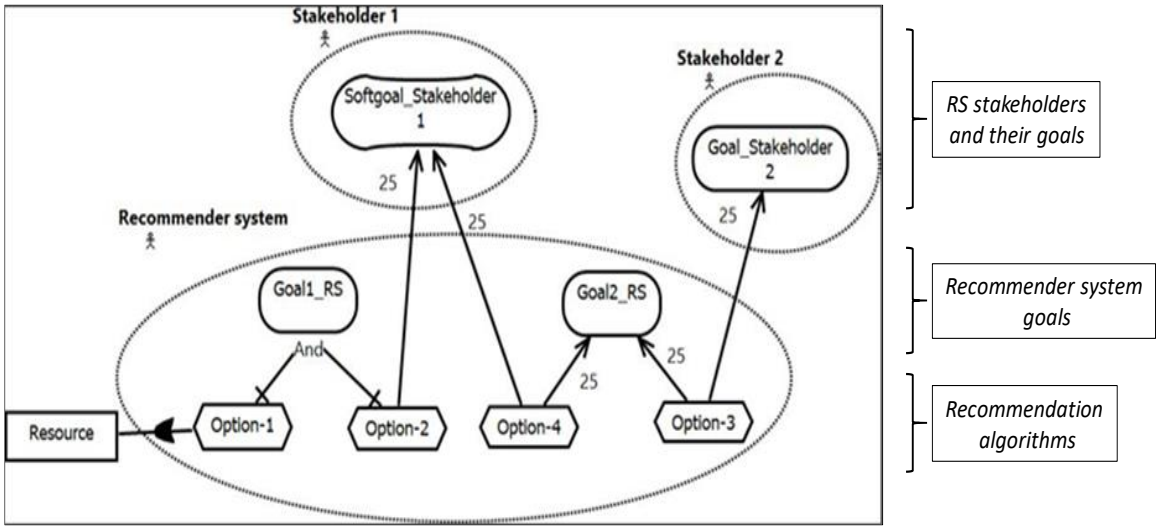


Figure 33 General goal model for recommender systems evaluation

An important benefit about using goal models is that a modeler can adapt the model to the current recommender's context (e.g., a specific recommender system, such as e-commerce or movie recommender) by changing the values of (1) the weights of the contribution from indicators to an alternative algorithm, and (2) the importance values of the goals to their containing actor. Despite these benefits, using such a goal model has some challenges; one main challenge is to determine appropriate contribution weights. Fortunately, many approaches now exist to mitigate this challenge [185]. Among these approaches is the group-based decision approaches such as the Round-Table Discussion and Consensus (RTD&C) or the Equal Relative Weights (ERW) [185]. More details about these approaches are discussed in section 7.5.

7.4. Implementation

This section provides a synthetic goal model for comparing different alternatives of a generic RS and evaluating the effectiveness of each alternative. It provides an example that is built based on the generic structure illustrated in Figure 33. This synthetic implementation aims to explain how to use the proposed approach. The resulting model was suggested based on our interpretations of the literature about recommender systems and its evaluation dimensions surveys, such as [6] and [7].

Figure 34 depicts the suggested example and shows the major stakeholders of RSs (i.e., Consumer, Provider, and the System). In this example, the Recommender System itself is represented as a stakeholder. It also shows goals that we identified for each stakeholder and the relationship between these goals. It is important to mention that the RSs alternatives illustrated in this example can be replaced by any recommendation algorithm. Also, there can be more than two alternatives.

To illustrate how we synthesized the model, take an example contribution link between the goals "reduce risk" and "Increase diversity." The *Risk* evaluation dimension is defined as the risk associated with the user's acceptance of the recommendation. To reduce risk, a recommendation algorithm could focus on providing *Accurate* (i.e., very close to user preferences) recommendations. Providing accurate results, in turn, may negatively impact *Diversity*. Thus, the link between the goals "reduce risk" and "Increase diversity" is labeled with contribution weight equal to (-30) to indicate that the source goal negatively impacts, or dissatisfies, the destination goal. Another example is the contribution link between the first alternative (the KBRS algorithm) and the goal "increase privacy." It is known that KBRS does not require many details and information about the users. That means KBRS preserves users' privacy to some extent. Thus, we indicated that the use of the KBRS algorithm affects the satisfaction of the goal "increase privacy," with a contribution weight=60, which means that using KBRS positively impacts the goal "increase privacy."

To adapt the goal modeling approach into the context of evaluating recommender systems, we used the well-known evaluation dimensions (i.e., coverage, correctness, diversity, etc.) to represent the goals of the "Recommender System" stakeholder. For instance, we used the goals "increase coverage" and "reduce risk" to indicate the

coverage and the risk evaluation dimensions, respectively. In order to make the example readable, we used only ten of the RS evaluation dimensions. These dimensions are *Privacy*, *Correctness*, *Coverage*, *Usability*, *Diversity*, *Serendipity*, *Trust*, *Risk*, *Novelty*, and *Fairness*. Each dimension is presented as goals or soft goals. For instance, the *Coverage* dimension is presented by the hard goal "Increase Coverage."

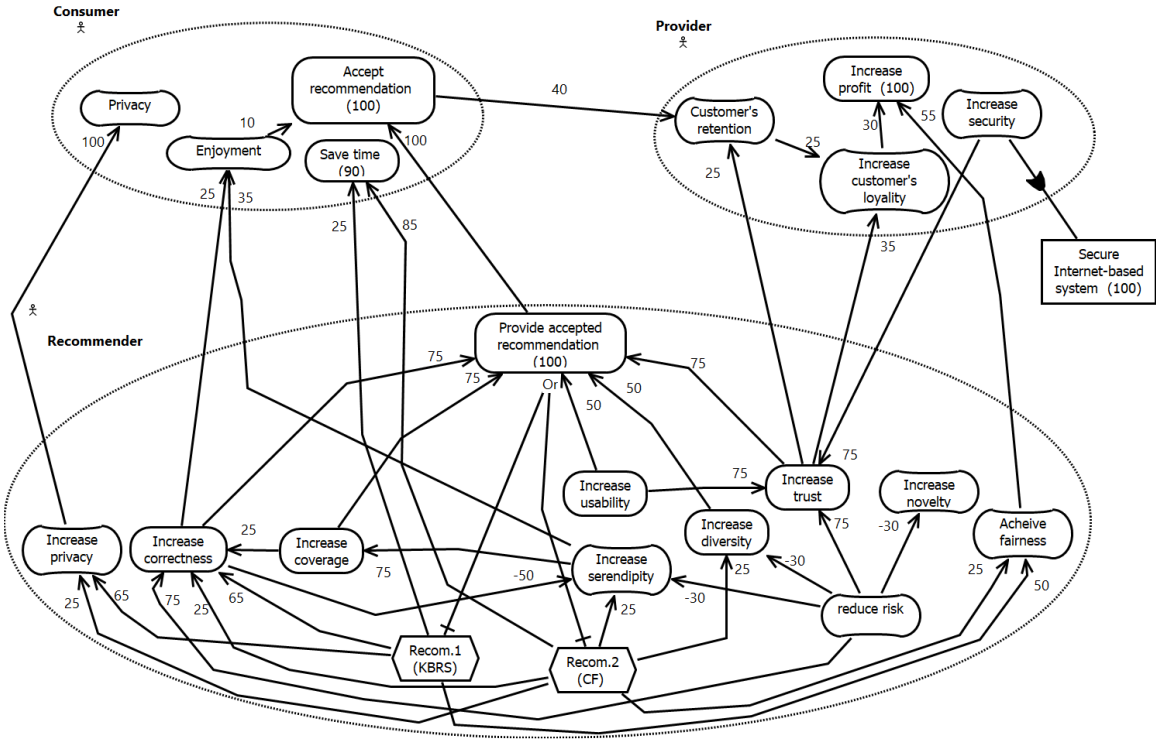


Figure 34 A goal model for recommender systems

From Figure 34, it can be noticed that "Accept recommendations," "Increase profit," and "Provide accepted recommendation" are very important goals for the Consumer, Provider, and Recommender systems stakeholders, respectively. This can be inferred from the quantitative importance value 100 (written between parentheses) associated with each one of these goals. In addition, the goals "Increase trust," "Increase coverage," and "Increase correctness," for example, have a very positive impact on the satisfaction of the "Provide accepted recommendation" goal, which is the main goal of the Recommender System stakeholder. This can be indicated through the high weight of the contribution links (i.e., 75) that connect between these goals. On the other hand, a low value of a contribution weight between any two goals indicates that the source goal has a

low impact on the satisfaction of the destination goal. For example, the goal "Increase coverage" contributes partially to the satisfaction of goal "Increase correctness," with a contribution weight equal to 25. Following the same rationale, a negative contribution weight (as in the link between "reduce risk" and "Increase diversity," with contribution weight equal to (-30) indicates that the source goal negatively impacts, or dissatisfies, the destination goal.

The Recommender System stakeholder has the "provide an accepted recommendation" goal as one of the very important goals, with importance equals to 100. This goal can be realized by "increasing correctness" of the recommendation, "increase coverage," "increase usability," "increase trust," "achieve fairness," "increase diversity," and "increase serendipity." These goals, in turn, can be achieved by using several alternatives of recommendation algorithms proposed in the literature. In this example, and to make it clear, we considered two general options for RS, namely: The *Knowledge-Based RSs* (KBRS) and the Collaborative Filtering RS (CFRS). The two options are presented as two recommendation algorithms, namely: "Recom1. KBRS" which is a KBRS algorithm, and "Recom2. CF" which is a CF algorithm. We studied the impact of using these two types of RS on the achievement of the above-mentioned goals. In particular, Figure 34 illustrates the impact of RS alternatives on some goals, where the impact is expressed in terms of weighted positive or negative contribution links in the goal model. For instance, the use of the KBRS algorithm affects the satisfaction of the goal "increase privacy," with a contribution weight=60. On the other hand, using the same algorithm (i.e., Recom.1 KBRS) has the potential to moderately increase the correctness of the recommendation, as indicated by a contribution link of weight 65 from "Recom.1 KBRS" task to the "Increase correctness" goal.

The use of different recommendation alternatives not only has different impacts on the satisfaction of goals within the same stakeholder (e.g., Recommender System) but also across different stakeholders. This can be noticed clearly from the multiple contribution links (either positive or negative) that are directed from the two RS options towards the goals of the Consumer and Provider stakeholders. For example, the use of "Recom.2 CF" has a strong impact (85) on the satisfaction of the goal "save time" owned by the Consumer stakeholder. On the other hand, the first option (i.e., Recom.1 KBRS) has

only a slight impact (25) on the same goal. The reason behind this is due to the fact that using a collaborative filtering RS does not require the user to provide information about items to be recommended, and hence, it saves users' time efficiently. On the other hand, the knowledge-based RS requires users' intervention to provide information about their preferences or about the specifications of the items they want to be recommended. Having this requirement, the use of "Recom.2 CF" saves the user time slightly.

7.4.1 Evaluating Recommender Systems Alternatives

The goal model illustrated in Figure 34 is useful for identifying stakeholders and for gaining a common understanding of their concerns/goals. In addition, this model assists in documenting decision rationales and enables trade-off analysis based on GRL what-if strategies and tool-supported satisfaction propagation algorithms. For example, the model can help analyzing and understanding the trade-offs between using "Recom.1 KBRS" or "Recom.2 CF" as two alternatives, and the impact of either alternative on achieving the "Provide accepted recommendation" goal. The model analyst does this by defining different *GRL strategies* and evaluating their impact on the satisfaction of higher-level goals and stakeholders. To evaluate the *strategies*, a model analyst can use the jUCMNav (Juicy-em-nav) tool [186]. jUCMNav is an open-source application available under the Eclipse Public License (EPL), and it is developed at the University of Ottawa.

In this thesis, we use the usual GRL quantitative algorithm to assess the impact of using recommenders alternatives on the satisfaction level of the goals connected to them. For this purpose, the goal model of the recommender system (presented in Figure 34) is re-illustrated in Figure 35 and Figure 36, but this time, with GRL evaluation strategies, where the figures represent the evaluation of using recommendation alternatives, one at a time.

To evaluate the impact of choosing the first alternative, namely the "Recom.1 KBRS", we set its satisfaction value to 100 (to indicate that it is of a major focus), and then a GRL evaluation strategy that corresponds to the observations of "Recom.1 KBRS" is performed, as illustrated in Figure 35. Goal satisfaction levels are obtained by propagating the value of this alternative to the rest of the model. Then the satisfaction level of stakeholders can be obtained from the level of goal satisfaction and the importance of these goals to their stakeholders. Figure 35 illustrates the result of a strategy where the

satisfaction value of "Recom.1 KBRS" is initialized to 100. In addition, the GRL strategy used for the evaluation here focuses on the satisfaction of the goals "Increase correctness," "Increase coverage," "Increase trust," "Increase security," and "Accept recommendation." Hence, the satisfaction value of these tasks is initialized to 75. In the Eclipse-based modeling tool we use, jUCMNav, initially selected elements in the evaluation strategy are displayed with dashed contours. jUCMNav also uses a color-coding scheme to highlight satisfaction levels —red for denied, yellow for neutral, and green for satisfied (with various shades for values in between).

As can be seen from Figure 35, running this strategy eventually leads to a dissatisfied (-37) "Increase serendipity" *soft goal*, a partially satisfied (65) "Increase privacy" hard goal, and a completely satisfied (100) "Provide accepted recommendation" hard goal in the recommender system stakeholder.

Figure 36 assesses the impact of using the second recommendation alternative, namely the CF recommender. The figure illustrates the result of a strategy where the satisfaction value of "Recom.2 CF" is initialized to 100 and also, as in the case of the "Recom.1 KBRS", the goals "Increase correctness," "Increase coverage," "Increase trust," "Increase security," and "Accept recommendation" are assumed to be important, and hence, their satisfaction values are initialized to 75. Form Figure 36, it can be observed that running this strategy leads a partial satisfaction (25) of both the "Increase diversity" and "Increase privacy" goals. It also causes the "Provide accepted recommendation" goal to be fully satisfied (100).

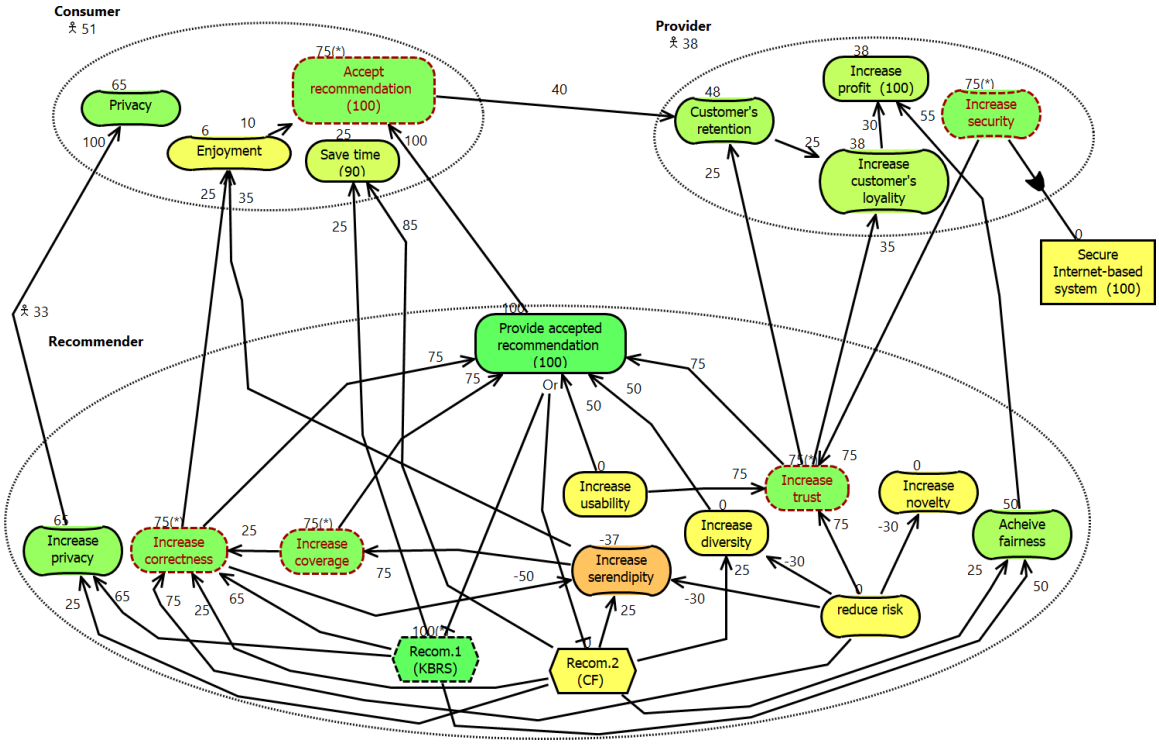


Figure 35 Assessing the Impact of Selecting KBRS alternative

7.4.2 Observations

According to Figure 35 and Figure 36, the following observations can also be made about the impact of using KBRS and CF, respectively, on other stakeholders and their goals:

1. Consumers find that using the CF algorithm is more efficient than using the KBRS in terms of saving their time (with a level of satisfaction = 85). This is because collaborative filtering approaches do not require users' intervention during the recommendation process. Hence, using CF saves users' time compared to the knowledge-based recommender systems, which ask the user to provide information necessary to build evidence to be used in the recommendations.

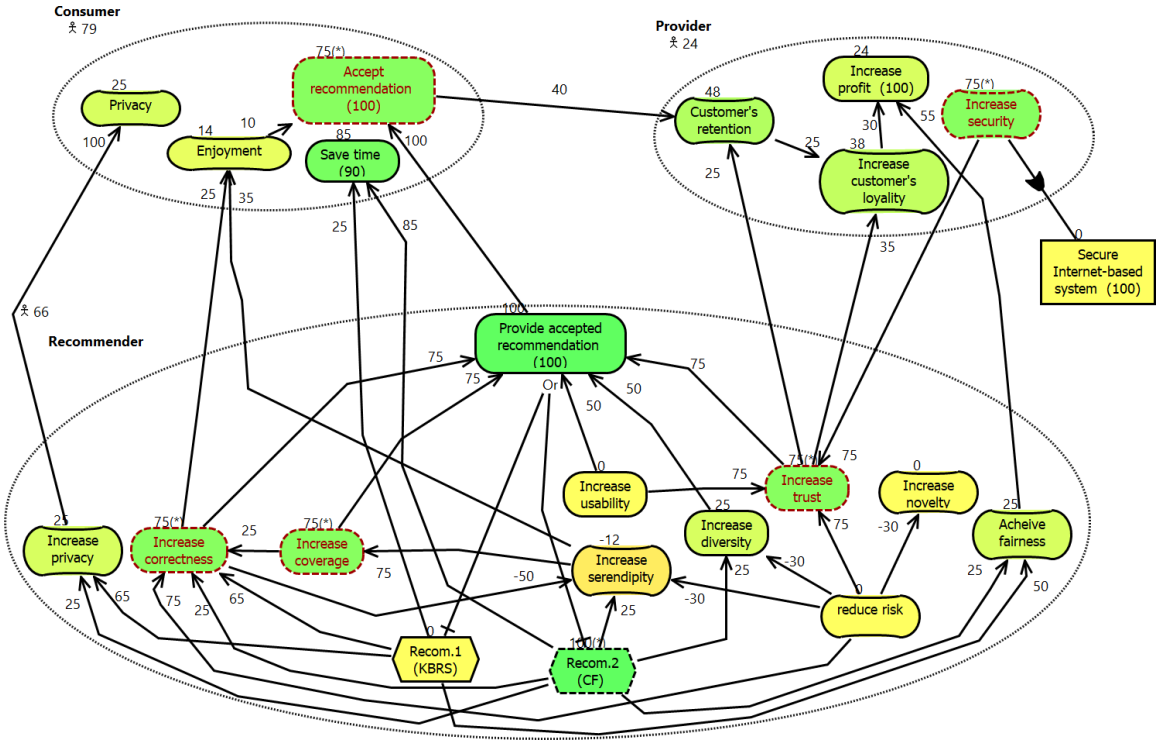


Figure 36 Assessing the Impact of Selecting CF alternative

2. Saving consumer's time using the CF recommenders comes with a price. In particular, the use of CF satisfies the "Increase privacy" goal only partially (25) compared to the KBRS, which satisfies the same goal with a satisfaction level=65. This is not surprising as the user intervention required by collaborative filtering approaches compromises user's privacy to some extent.
3. Providers are weakly satisfied with using both algorithms (with a satisfaction less than 25% when using CF algorithm), while their goal "Customer's retention" is %48 satisfied.
4. Consumer and Recommender stakeholders are satisfied with using CF (with satisfaction levels of 79 and 66, respectively) more than with using the KBRS, which achieves satisfaction levels of 51 for Consumer and 33 for Recommender.
5. Taking into consideration that the CF algorithm does not take into consideration the fair distribution of recommendations among several providers; this, in turn, will have an impact on the satisfaction of the goal "Achieve fairness," where this goal is marginally satisfied (25). On the other hand, the same goal is moderately satisfied

(50) when KBRS is used, due to the fact that the latter supports a fair consideration of multiple providers.

6. Based on the above observation, it can be justified why the provider stakeholder gets satisfied when KBRS is used (38) more than when the CF is used (24).

It is worth mentioning that other evaluation strategies can be defined to evaluate different sets of options and find the most suitable trade-off. The resulting satisfaction values of stakeholders and their goals should be used to compare different strategies and find suitable trade-offs between any recommendation algorithm used by recommender systems, rather than to be interpreted as some sort of satisfaction percentage. Moreover, the model illustrated in Figure 35 and Figure 36 and the values of goal importance, contribution weights, and relationships between goals represent one context with particular settings. The modeler can eventually adapt the model to the other context by changing the importance values of the goals to their containing actor, the weights of contribution links, and so on.

7.5. Discussion

This section discusses how the proposed goal-oriented approach represents a step forward in the evaluation of recommender systems, as well as the main limitations of the approach.

The main contribution of this chapter is introducing the goal-modeling-based approaches to evaluate multiple recommendation alternatives while considering the multi-stakeholder goals. Up to our knowledge, this is the first work that introduces the goal-oriented approaches using the GRL language to the evaluation of RSs. The use of goal models has the potential advantages for documenting stakeholders, goals, tasks, and relationships between all components of the model. Moreover, these models can be used to assess the impact of, and identify trade-offs for, specific alternative recommendation algorithm.

By exploiting the early analysis and evaluation capabilities of goal modeling with GRL, we describe and evaluate the impact of multiple recommendation platforms. This is done through modeling alternatives of recommendation algorithms and their contributions to the goals of different stakeholders so as to perform trade-off analysis between the different recommendation options and to assure that the selected option is well-aligned with

the needs of such systems. In addition, the proposed goal-oriented modeling approach contributes to recommender systems designers to better reason about RS selection and to document the rationale of their decisions. There are, however, many remaining challenges in the generalization of the approach to other cases and in its validation in practice.

The main limitation of our proposed approach is that the goal models illustrated in this chapter are identified and constructed manually based on our own understanding and interpretation of the recommender system domain and its literature. Also, the relationships (i.e., contribution links and decomposition) defined between the various components of goal models are set arbitrarily. Since contribution values could have a significant impact on linked intentional elements (positive, negative, neutral, etc.) and their evaluation level may affect the entire goal model, we acknowledge that our work needs further validation in this regard.

To enhance the accuracy of the proposed model, we will consult experts from the recommender system domain to elicit more real-time perspectives about the domain to be modeled. Also, to obtain a more appropriate justification of contribution levels that a particular recommender alternative has on the satisfaction of stakeholders' goals, we envision the use of other group-based decision approaches such as the Round-Table Discussion and Consensus (RTD&C) or the Equal Relative Weights (ERW) [185]. In the RTD&C approach, a group of experts meets together, and their related contributions/opinions are put up on a screen where experts are asked to discuss and assign relative weights to each choice in each group. ERW approach, on the other hand, contributions that target the same intentional element are considered to have equal weight, despite the fact that some contributors might be more important than others.

It is worth noting that the goal model presented in Section 7.4.1 illustrates a very general context of recommender systems, with general recommendation approaches (i.e., collaborative filtering and knowledge-based recommender systems in this example). This model is broad enough, and it is used as a proof of concept example. That means this goal model can be used as a basis for more specific models that are tailored for specific contexts. As mentioned in Chapter 8, we are planning to conduct a more comprehensive study about the specificities of recommender systems in order to create more representative models.

7.6. Chapter Summary

This chapter proposes a new research direction and raises awareness about the research related to the evaluation of the recommendation techniques while considering RS's multi-stakeholders' goals. The chapter also pinpoints the potential opportunities for contributions associated with the use of goal-oriented requirements modeling, which goes beyond the conventional techniques of evaluating recommender systems. The proposed approach aims to evaluate RS not only beyond their accuracy but also by considering the goals of all parties, which are sometimes contradictions. The chapter showed, through an illustrative example, the applicability of the proposed goal-oriented evaluation approach. The example demonstrated that the use of goal modeling with GRL could be adopted to help RS's designers and decision-makers to choose between various recommendation algorithms based on the requirements of their recommender.

Part III

Conclusions & Future Work

Chapter 8. Conclusions and Future Work

This chapter recalls the potential contributions of the thesis and connects them to the research questions. It also discusses threats to the validity and limitations of the presented work. In addition, it discusses ongoing work and future research opportunities.

8.1. Summary and Contributions

Since the emergence of Recommender Systems (RSs), most of the related research body has focused on improving recommender's *predictability* and *accuracy*. However, the literature has recently witnessed increasing evidence that providing accurate recommendations only is not enough to increase users' acceptance of the system [16]. Hence, it is vital for RSs to focus not only on the accuracy of the used matching algorithm but also on other factors that influence the acceptance of recommendations and the extent to which such recommendations are influential and persuasive.

Having said that, there became a desire, and even a demand, for new research paradigms to help improve the interactions that a recommender system facilitates. One of the recently emerged research directions that consider this need fosters the idea of adopting human-related concepts. This research direction suggests that human-human communication principles can be utilized in human-computer interactions [14].

This thesis attempts to address this demand by adopting theories from the social sciences literature into the recommendation systems domain. The **contributions** of this thesis are:

- A user study of the persuasiveness of recommender systems presented in Chapter 3. The study discussed the effect of both user-related and context-related aspects on users' susceptibility to persuasive principles (**RQ1** and **RQ2**).
- An adaptive Personalized Persuasive Recommender Systems framework, *PerPer*, is presented in Chapter 4. *PerPer* aims to increase the acceptance of recommendations by enriching RSs with persuasive features. The framework is also deployed based on learning automata concepts (**RQ2** and **RQ3**).

- An innovative analogy between the RS's recommendation process and human beings' learning process (Chapter 5). The analogy suggests that a recommendation process is like a learning process, and so, the recommendations (as outputs) can be evaluated in a way similar to the evaluation of learning outcomes (**RQ4**).
- A cognitive-based Comprehensive Performance evaluation approach, *ComPer*, (discussed in Chapter 5 and Chapter 6), which is inspired by the aforementioned analogy. The basis of *ComPer* is a correlation matrix between the Learning dimensions of Bloom's taxonomy and evaluation dimensions of RS. (**RQ4**).
- A goal-oriented evaluation approach that considered multiple stakeholders and their various goals. (**RQ5**)

8.2. Threats to Validity

In our work, there are several threats to validity that must be assessed to better characterize the limitations of the approaches and results presented in this thesis. According to Perry et al. [187], the following are the most relevant categories of threats to validity for our work.

8.2.1 Construct Validity

Construct validity aims to assess to which extent the experiments, tests, and/or case studies actually answer our research hypothesis. An important issue here is the difficulty of testing the proposed approaches on real systems.

A major source of construct threats stems from the potential issues associated with human experts providing correlations, where the deviation between experts' opinions can be large; hence, in some cases, the correlation might not be representative enough. In addition, the reliance on a fourth expert to make a final decision about the conflicts between the NLP correlations and the other three experts is limited and insufficient.

8.2.2 Internal Validity

Internal validity examines any bias and other confounding factors and assesses the degree to which conclusions about causal relationships can be made based on the test settings and

measures obtained. One obvious threat in this thesis is that bias might be introduced by having the thesis author perform the implementation, tests, and analyze the results.

A major source of internal threats stems from the potential issues related to the arbitrary setting of the goal models introduced in Chapter 7, such as the values of goals importance, the weights of contribution links, and the types of decomposition links between the various components of goal models. These values were mainly inferred based on the author's sole understanding and interpretation of the literature.

8.2.3 External Validity

External validity verifies the extent to which the evaluation results can be generalized to other cases or contexts. One important external threat encountered here is that the study samples (Chapter 3 and Chapter 4) may be limited, especially that we are dealing with human-related aspects. Also, for the validation of *ComPer*, other evaluation settings and analysis techniques may not result in efficiency improvements as positive as the ones observed here.

One important external validity that we encountered in this thesis is related to the generalizability of the results to a real RS (as discussed in Section 3.4). In particular, the generalizability of our findings might be threatened by the following limitations: 1) the study is based on a self-report questionnaire and not a real recommender system, 2) the diversity of the cultures is limited to three continents, with only 13 participants from Europe. This, in turn, may threaten the generalizability of our findings to other populations, and 3) the guidelines provided by our study and by the other related studies are general tips and do not comprehend all possible design cases (i.e., the combinations of every factor with all other factors) and could not provide absolute or decisive rules for deploying persuasive strategies in RSs.

8.3. Future Research Directions

There are many opportunities to follow up on the thesis work, including the following directions:

8.3.1 Future Work for Persuasive Recommender Systems

Chapter 3 investigated how RS users' persuadability is affected by their characteristics and context. Following are possible enhancements for this research direction:

- ***Persuasive Principles:*** the study considered Cialdini's six principles of persuasion. Many other principles have also been proposed in the social science literature. So, we suggest studying the effect of other persuasion principles in the context of RS. Also, the literature lacks a study that investigates the efficiency of these multiple principles in the recommendation area (i.e., which principles are the most effective among others).
- ***Persuasive Principles and Recommendation Domain:*** the results of the user study revealed that the application domain might affect RS users' vulnerability to different persuasive principles. Accordingly, a thorough study is required to explore the relationship (if any) between the persuasion principles and different domains of RSs. The goal of this study will be to provide the community with clear guidelines that help them deploy the right persuasion principles in each particular domain. Also, this study will help us mitigate the generalizability issue of our study (mentioned in section 3.4).
- ***Investigation of Other Recommender System-related Aspects:*** the communication-persuasion paradigm (discussed in Section 2.3) suggested that the persuasiveness of a communicated message is affected by four main factors (the source, the destination, the message, and the context). In our study, we considered two aspects, namely user characteristics (the destination) and the recommendation domain (the context). Further studies are still required to explore the effect of other factors, as well as the effect of the combination between these factors.
- ***Justification of the results.*** The focus of the data analyses provided in Chapter 3 was to solely report on our findings regarding the considered factors, more than to provide explanations or justifications of each factor's impact. We believe that the justification of the results is a purely social science. Therefore, we aim to collaborate with social scientists (personality and behavioural change scientists, in particular) to provide deeper insights into the acquired results. The justification of

the results will help us discover unexpected issues that could be associated with the results. This, in turn, could support the generalizability of the study.

- ***Investigating the relation between persuasion and user's preferences***. The user study (discussed in Chapter 3) investigated how users perceived persuasion based on different factors. But the study did not consider other important factors, such as how close the recommendation to the user's preferences is. Users' responses to persuasion may vary from one recommendation to another based on the closeness of the recommended item to the users' preferences. For instance, persuasion may not be very helpful for an item that is 80% matching users' preferences, compared to an item that is only 60% match. A user is more likely to be persuaded to accept the 80%-matched item than the 60% matched one. On the other side, persuasion for an item that is only 20% matching users' preferences may not be useful at all because the item is significantly out of users' interest. To further investigate this effect, an in-depth study is required to investigate the aforementioned issue.

There are many future directions related to the *PerPer* framework. Particularly, the components of *PerPer* should be further investigated, as follows:

- ***The number of iterations required to achieve convergence***: learning automata, like other reinforcement algorithms, depends on inputs received from the environment. This input is presented in our framework as Feedback from users. It is mentioned in section 4.4 that our deployment of *PerPer* acquires feedback explicitly because it is more reliable than implicit feedback. Nonetheless, explicit feedback requires effort from users, which may cause boredom for users, especially when the number of iterations is high. As a result, users may lose the motive to provide feedback at some point. To mitigate this issue, the value of δ in equation 3 could be increased so as to accelerate the convergence. That is, the larger the δ value is, the faster the convergence is [136]. However, as the value of δ being larger, the chance of getting incorrect outcome is increased. So, this value needs to be studied carefully because it is an essential factor that controls the learning process.
- ***Accelerating the Convergence to Reduce Users' Interactions:*** we used Learning Automata as an adaptive approach to deploy *PerPer*. Our experimental results

demonstrated the feasibility of this approach. Nonetheless, this approach requires many interactions between the user and the system for the best action to converge. This is fine for RSs areas where users interact with the system frequently, such as movie and music recommendations. In other domains (such as KBRS), however, the interactions may be less frequent. In such cases, other approaches that can accelerate the convergence of the best persuasion technique should be applied. Thus, as future work, we need to discover other approaches for deploying *PerPer*. One potential suggestion to overcome the above issue is to use *crowdsourcing*. For instance, we can rely on the user's social network, such that the system maintains one persuasion profile for multiple users. These users should have common characteristics so that their preferences are close to each other. In such a way, the interactions with all users in a single group are considered to update the persuasion profile.

- ***Presentation of Persuasive Paths***: in the proposed implementation of *PerPer*, we deployed persuasive strategies as explanations attached to the recommendations. As it has been mentioned, however, persuasive paths can be presented (or implemented) using various cues. Thus, we need to discover other formats of the persuasion paths, study the efficiency of these formats, and to compare them with each other.
- ***Deploy a Multi Automata System***: another direction for future work is related to the persuasion pool of *PerPer*. In the proposed deployment, we supposed that each persuasive strategy is presented by a single cue (or explanation). According to Unal et al. [103], users may resist persuasive messages if the system keeps presenting them visibly. Accordingly, we suggest that considering multiple presentations for each strategy has the potential to avoid resistance. As a potential solution, we suggest adopting multi automata concepts [136]. That is, instead of using single automata, Multi-level learning automata can be implemented as follows: the main automaton is responsible for selecting the next persuasive path to be used, while other automatons are assigned to select appropriate cues of the selected path. That is, the action set of the main automaton is the set of persuasive paths, while the action set of other automatons is the set of cues for the corresponding path.

- ***Investigate Other Deployments of PerPer***: Furthermore, we envision to implement *PerPer* using different approaches. For example, we will adopt the approach proposed by Kaptein [116] that uses Bayesian learning. Also, we will compare our proposed Learning-Automata based implementation with other implementations (such as the Bayesian learning-based implementation).
- ***Implementing PerPer in different application areas:*** in the near future, we are planning to deploy the proposed PerPer framework into a different context, for example, the educational domain, and then compare its performance against the performance of the non-personalized (or static) persuasive RS. By this, we would have a better characterization of the circumstances under which PerPer achieves performance gains.

8.3.2 Future Work for Evaluating Recommender Systems

Chapter 5 proposed the evaluation approach, which we call *ComPer*. The chapter presented the way by which we implemented this approach. Also, Chapter 6 provided an experimental evaluation of *ComPer* by which it demonstrated the feasibility of using such an approach. Despite the work that we have already done in this direction, further efforts are still required to improve the *ComPer* framework, which can be summarized as follows:

- ***Inferring the Correlation Matrix***: A more robust way should be considered regarding the methodology of inferring the correlation between Bloom's dimensions and RS evaluation dimensions (section 5.4.3). One potential suggestion for this enhancement is the number of experts. The more the number of experts involved in the evaluation, the more accurate the results will be. So, we will consider recruiting more experts to infer the correlation, and we will rely on statistical approaches to validate the degree of agreement.
- ***Natural Language Processing Stage***: Another related improvement is about the NLP stage (i.e., step II of the correlation). As it is described in section 5.4.3, this step adopted an approach that depends on the concept of *bag-of-words* similarity. Although this approach is familiar in the NLP literature, its results are sensitive to the text from which we create the bag-of-words. Also, this approach is unable to capture the actual semantic similarity between such definitions. To mitigate this

issue, a more accurate representation would be the word embedding, which provides a vector representation of each word. The advantage of using word embedding is that it can capture the context of words. So, we suggest that using word embedding has the potential to provide more accurate correlations.

- ***The Relation between Evaluation Dimensions and Application Area***: As mentioned in section 5.2, the considered evaluation dimensions have been selected based on their popularity in the literature (i.e., how many times each dimension has been evaluated). Although this factor is a reasonable indication of the importance of different dimensions, it is insufficient because popularity does not necessarily mean importance. Also, the same dimension has different significance for different application areas. For instance, while *Robustness* is so important in applications where there is a high chance of false information (e.g., e-commerce websites), it is less important for other areas (e.g., code recommendations). Hence, an investigation of the relationship between each dimension and RS application areas is required. By doing so, we add another dimension to the proposed approach.
- ***Experimental Evaluation***: since users' acceptance of the recommendations is a crucial main goal of RSs, we will study the level of consistency of *ComPer* results with users' perceived acceptance. To do so, a user study will be conducted. The result of this user study will be compared with *ComPer* results. Related to this point, we also invite researchers and businesses who run actual recommenders to validate *ComPer* results on real systems in order to generalize our findings.
- ***ComPer$_{value}$:*** As mentioned before, *ComPer* combines multiple evaluation dimensions in a single value. Aggregating multiple values into a single one may lead to information loss. That is, it may downgrade the importance of some metrics over others. Thus, a deeper analysis is still required to investigate the possible negative impact of aggregating multiple dimensions into the *ComPer* single value (i.e., *ComPer$_{value}$*). Besides, *ComPer$_{value}$* relied on an ad hoc normalization to mitigate the issue of different scales between the considered dimensions. Although normalization mitigates the issue, there is still a possibility that the final *ComPer* measure is skewed more toward some of the dimensions. As future work, we are planning to investigate the applicability of using ranking instead of normalization.

- ***Enhancing the Accuracy of the Goal-Oriented evaluation approach***: to enhance the accuracy of the proposed model, we need to consult experts from the recommender system domain to elicit more real-time perspectives about recommender systems goals. Also, domain experts will be consulted to obtain a more appropriate justification of contribution levels that a particular recommender alternative has on the satisfaction of stakeholders' goals. We envision the use of other group-based decision approaches such as the Round-Table Discussion and Consensus (RTD&C) or the Equal Relative Weights (ERW) [185].
- ***Domain-Specific Goal-Model evaluation:*** the goal model presented in Section 7.4.1 illustrated a general context of recommender systems, with general recommendation approaches (i.e., collaborative filtering and knowledge-based recommender systems in this example). This model is broad enough, and it is used as a proof-of-concept example. Therefore, more domain-specific models are required to leverage the benefit of the proposed approach. Having such domain-specific goal models will help us find real use cases and implement the model to further validate it.

8.3.3 Other Research Directions

In addition to the above future work, the following are other research directions that are related to persuasion and evaluation of RSs:

- ***Increasing users' ratings:*** Sparsity is a common issue in the evaluation of RSs. The issue is that offline evaluation depends, to a high extent, on the historical data of the users that are stored as datasets. These datasets are usually sparse such that, in many cases, the users do not provide ratings for the recommended items. So, to reduce the sparsity (i.e., to make the dataset denser), users should provide more ratings. To do so, we need to motivate the users to rate more recommendations. One solution to this issue is by using persuasive techniques. For instance, we can use social persuasive strategies such as *Competitive* and *Cooperative* Social Persuasion to influence users to change their attitude by providing more ratings. Also, Gamification can be adopted to mitigate this issue by motivating users to rate the recommended items.

- ***Personality-based Recommendations***: Chapter 3 and Chapter 4 demonstrated how users' personality traits affect the persuadability of users. We believe that personality traits have the potential to enhance user-RS interactions. Therefore, we will investigate the role of personality traits in tailoring the top-N recommendation lists. That is, we will study the relationship between users' personalities and RS's properties such as Diversity, Accuracy, Coverage, etc. Based on the inferred relationships, we can adapt the recommended list such that it involves items that satisfy the most important factor for the user.

References

[1] Jannach D, Zanker M, Felfernig A, Friedrich G. Recommender systems: an introduction. Cambridge University Press; 2010 Sep 30.

[2] Adomavicius G, Jannach D. Preface to the special issue on context-aware recommender systems. User Modeling and User-Adapted Interaction. 2014 Feb 1;24(1-2):1-5.

[3] Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. In Recommender systems handbook 2011 (pp. 1-35). Springer, Boston, MA.

[4] Said A, Bellogín A. Comparative recommender system evaluation: benchmarking recommendation frameworks. In Proceedings of the 8th ACM Conference on Recommender systems 2014 Oct 6 (pp. 129-136).

[5] Mahmood T, Ricci F, Venturini A, Höpken W. Adaptive recommender systems for travel planning. In ENTER 2008 Jan 24 (Vol. 8, pp. 1-11).

[6] Avazpour I, Pitakrat T, Grunske L, Grundy J. Dimensions and metrics for evaluating recommendation systems. In Recommendation systems in software engineering 2014 (pp. 245-273). Springer, Berlin, Heidelberg.

[7] Shani G, Gunawardana A. Evaluating recommendation systems. In Recommender systems handbook 2011 (pp. 257-297). Springer, Boston, MA.

[8] Champiri ZD, Asemi A, Binti SS. Meta-analysis of evaluation methods and metrics used in context-aware scholarly recommender systems. Knowledge and Information Systems. 2019 Nov 1;61(2):1147-78.

[9] Ge M, Delgado C, Jannach D. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In Proceedings of the fourth ACM conference on Recommender systems 2010 Sep 26 (pp. 257-260).

[10] McNee S, Riedl J, Konstan J. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In CHI'06 extended abstracts on Human factors in computing systems 2006 Apr 21 (pp. 1097-1101).

[11] Jannach D, Lerche L, Kamehkhosh I, Jugovac M. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. User Model. User Modeling and User-Adapted Interaction. 2015 Dec;25(5):427-91.

[12] Jannach D, Hegelich K. A case study on the effectiveness of recommendations in the mobile internet. In Proceedings of the third ACM conference on Recommender systems 2009 Oct 23 (pp. 205-208).

[13] Gkika S, Lekakos G. The Persuasive Role of Explanations in Recommender Systems. In BCSS@ PERSUASIVE 2014 May 22 (pp. 59-68).

[14] Yu T, Benbasat I, Cenfetelli R. Toward deep understanding of persuasive product recommendation agents. ICIS 2011 Proceedings. 8. (2011).

[15] Fogg BJ. Persuasive technology: using computers to change what we think and do. Ubiquity. 2002 Dec 1; 2002(December):2.

[16] Wang W, Benbasat I. Trust in and adoption of online recommendation agents. Journal of the association for information systems. 2005;6(3):4.

[17] Zanker M, Bricman M, Gordea S, Jannach D, Jessenitschnig M. Persuasive online-selling in quality and taste domains. In International Conference on Electronic Commerce and Web Technologies 2006 Sep 5 (pp. 51-60). Springer, Berlin, Heidelberg.

[18] Aksoy L, Bloom PN, Lurie NH, Cooil B. Should recommendation agents think like people? Journal of Service Research. 2006 May;8(4):297-315.

[19] Bauernfeind U, Zins AH. The perception of exploratory browsing and trust with recommender websites. Information Technology & Tourism. 2005 Jun 1;8(2):121-36.

[20] Bonhard P, Sasse MA. "I thought it was terrible and everyone else loved it"—A New Perspective for Effective Recommender System Design. In People and Computers XIX—The Bigger Picture 2006 (pp. 251-265). Springer, London

[21] Al-Natour S, Benbasat I, Cenfetelli RT. The role of design characteristics in shaping perceptions of similarity: The case of online shopping assistants. Journal of the Association for Information Systems. 2006;7(12):34.

[22] Yoo KH, Gretzel U, Zanker M. Persuasive recommender systems: conceptual background and implications. Springer Science & Business Media; 2012 Aug 17.

[23] Fogg BJ, Lee E, Marshall J. Interactive technology and persuasion. Persuasion handbook: Developments in theory and practice. 2002:765-97. London: Sage

[24] Reeves B, Nass CI. The media equation: How people treat computers, television, and new media like real people and places. Cambridge university press; 1996.

[25] Nguyen H, Masthoff J, Edwards P. Persuasive effects of embodied conversational agent teams. In International Conference on Human-Computer Interaction 2007 Jul 22 (pp. 176-185). Springer, Berlin, Heidelberg.

[26] Gretzel U, Fesenmaier DR. Persuasion in recommender systems. International Journal of Electronic Commerce. 2006 Dec 1;11(2):81-100.

[27] Yoo KH. Creating more credible and likable recommender systems (Doctoral dissertation). Texas A&M University, College Station, USA. (2010).

[28] Kille B, Albayrak S. Modeling Difficulty in Recommender Systems. In RUE@ RecSys 2012 Sep 9 (pp. 30-32).

[29] Meyer F, Fessant F, Clérot F, Gaussier E. Toward a new protocol to evaluate recommender systems. arXiv preprint arXiv:1209.1983. 2012 Sep 10.

[30] March ST, Smith GF. Design and natural science research on information technology. Decision support systems. 1995 Dec 1;15(4):251-266.

[31] Hevner AR, March ST, Park J, Ram S. Design science in information systems research. Management Information Systems Quarterly. 2008;28(1):6.

[32] Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. Knowledge-based systems. 2013 Jul 1;46:109-32.

[33] Isinkaye FO, Folajimi YO, Ojokoh BA. Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal. 2015 Nov 1;16(3):261-73.

[34] Burke R. Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction. 2002 Nov;12(4):331-70.

[35] Mobasher B, Jin X, Zhou Y. Semantically enhanced collaborative filtering on the web. InEuropean Web Mining Forum 2003 Sep 22 (pp. 57-76). Springer, Berlin, Heidelberg.

[36] McSherry D. Explaining the Pros and Cons of Conclusions in CBR. InEuropean Conference on Case-Based Reasoning 2004 Aug 30 (pp. 317-330). Springer, Berlin, Heidelberg.

[37] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. InProceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 161-168).

[38] Larose DT, Larose CD. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons; 2014 Jul 8.

[39] Schafer JB, Frankowski D, Herlocker J, Sen S. Collaborative filtering recommender systems. InThe adaptive web 2007 (pp. 291-324). Springer, Berlin, Heidelberg.

[40] Burke R. Hybrid web recommender systems. The adaptive web. 2007:377-408.

[41] Park DH, Kim HK, Choi IY, Kim JK. A literature review and classification of recommender systems research. Expert systems with applications. 2012 Sep 1;39(11):10059-72.

[42] Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth". InProceedings of the SIGCHI conference on Human factors in computing systems 1995 May 1 (pp. 210-217).

[43] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine learning. 1997 Nov;29(2):131-63.

[44] Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973 Jan.

[45] Bishop CM. Pattern recognition and machine learning, vol. 4, no.4. Springer, New York; 2006.

[46] Billsus D, Pazzani MJ. User modeling for adaptive news access. User modeling and user-adapted interaction. 2000 Jun;10(2):147-80.

[47] Mooney RJ, Roy L. Content-based book recommending using learning for text categorization. InProceedings of the fifth ACM conference on Digital libraries 2000 Jun 1 (pp. 195-204).

[48] Bridge D, Göker MH, McGinty L, Smyth B. Case-based recommender systems. The Knowledge Engineering Review. 2005 Sep 1;20(3):315.

[49] Ricci F, Cavada D, Mirzadeh N, Venturini A. Case-based travel recommendations. Destination recommendation systems: behavioural foundations and applications. 2006:67-93.

[50] Miranda T, Claypool M, Gokhale A, Mir T, Murnikov P, Netes D, Sartin M. Combining content-based and collaborative filters in an online newspaper. InIn Proceedings of ACM SIGIR Workshop on Recommender Systems 1999.

[51] Billsus D, Pazzani MJ. A hybrid user model for news story classification. In: Kay J, editor. In: Proceedings of the seventh international conference on user modeling, Banff, Canada. Springer-Verlag, New York; 1999. p. 99–108.

[52] O'keefe DJ. Persuasion: Theory and research. Thousand Oaks: Sage Publications; 2015 Feb 18.

[53] Michener HA, DeLamater JD, Myers DJ. Social Psychology. Wadsworth: Thomson Learning, Inc. 2004.

[54] John OP, Robins RW, Pervin LA, editors. Handbook of personality: Theory and research. Guilford Press; 2010 Nov 24.

[55] Ewen RB. An introduction to theories of personality. Psychology Press; 2014 Jan 21.

[56] Sullivan HS, editor. The interpersonal theory of psychiatry. Routledge; 2013 Nov.

[57] Mairesse F, Walker MA, Mehl MR, Moore RK. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research. 2007 Nov 28; 30:457-500

[58] Uher J. Comparative personality research: methodological approaches. European Journal of Personality. 2008 Aug;22(5):427-55

[59] Goldberg LR, Johnson JA, Eber HW, Hogan R, Ashton MC, Cloninger CR, Gough HG. The international personality item pool and the future of public-domain personality measures. Journal of Research in personality. 2006 Feb 1;40(1):84-96

[60] Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist. 2000 Jan;55(1):68

[61] Marczewski A. Even Ninja Monkeys like to play. CreateSpace Independent. Publish Platform, Charleston, Chapter User Types. 2015:69-84.

[62] McCrae RR, John OP. An introduction to the five-factor model and its applications. Journal of personality. 1992 Jun;60(2):175-215

[63] Tupes EC, Christal RE. Recurrent personality factors based on trait ratings. Journal of personality. 1992 Jun;60(2):225-51

[64] Dunn G, Wiersema J, Ham J, Aroyo L. Evaluating interface variants on personality acquisition for recommender systems. In International Conference on User Modeling, Adaptation, and Personalization 2009 Jun 22 (pp. 259-270). Springer, Berlin, Heidelberg.

[65] Lambiotte R, Kosinski M. Tracking the digital footprints of personality. Proceedings of the IEEE. 2014 Oct 29;102(12):1934-9

[66] Soto CJ, John OP. Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. Journal of research in personality. 2009 Feb 1;43(1):84-90

[67] John OP, Srivastava S. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research. 1999 Mar;2(1999):102-38.

[68] Gkika S, Skiada M, Lekakos G, Kourouthanasis P. Investigating the role of personality traits and influence strategies on the persuasive effect of personalized recommendations. In 4th Workshop on Emotions and Personality in Personalized Systems (EMPIRE) 2016 Aug (p. 9)

[69] Lim A. The Big Five Personality Traits. Simply Psychology. Retrieved from https://www.simplypsychology.org/. (June, 15, 2020).

[70] Dennis M, Masthoff J, Mellish C. The quest for validated personality trait stories. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces 2012 Feb 14 (pp. 273-276).

[71] Filho JR, Freire EO. On the equalization of keystroke timing histograms. Pattern Recognition Letters. 2006 Oct 1;27(13):1440-6.

[72] Khan IA, Brinkman WP, Fine N, Hierons RM. Measuring personality from keyboard and mouse use. In Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction 2008 Jan 16 (pp. 1-8).

[73] Nunes MA, Bezerra JS, de Oliveira AA. Personalityml: a markup language to standardize the user personality in recommender systems. Revista GEINTEC-gestão, inovação e tecnologias. 2012 Sep 22;2(3):255-73.

[74] Nunes MA. Recommender systems based on personality traits (Doctoral dissertation). Université Montpellier 2, 2008.

[75] Rammstedt B, John OP. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality. 2007 Feb 1;41(1):203-12

[76] John OP, Naumann LP, Soto CJ. Paradigm shift to the integrative big five trait taxonomy. Handbook of personality: Theory and research. 2008 Aug 5;3(2):114-58

[77] Costa Jr PT, McCrae RR. The Revised NEO Personality Inventory (NEO-PI-R). Sage Publications, Inc; 2008.

[78] Tkalcic M, Chen L. Personality and recommender systems. In Recommender systems handbook 2015 (pp. 715-739). Springer, Boston, MA.

[79] Jones JG, Simons HW. Persuasion in society. Taylor & Francis; 2017 Apr 7

[80] Oinas-Kukkonen H, Harjumaa M. A systematic framework for designing and evaluating persuasive systems. In International conference on persuasive technology 2008 Jun 4 (pp. 164-176). Springer, Berlin, Heidelberg.

[81] Cialdini RB. Harnessing the science of persuasion. Harvard business review. 2001 Oct 1;79(9):72-81

[82] Kellermann K, Cole T. Classifying compliance gaining messages: Taxonomic disorder and strategic confusion. Communication Theory. 1994 Feb;4(1):3-60.

[83] Vargheese JP, Collinson M, Masthoff J. Exploring susceptibility measures to persuasion. In International Conference on Persuasive Technology 2020 Apr 20 (pp. 16-29). Springer, Cham.

[84] Cialdini R. Pre-suasion: A revolutionary way to influence and persuade. Simon and Schuster; 2016 Sep 6.

[85] Hu R, Pu P. Enhancing collaborative filtering systems with personality information. In Proceedings of the fifth ACM conference on Recommender systems 2011 Oct 23 (pp. 197-204)

[86] Burger, J.M. Personality. Wadsworth Publishing, Belmont, CA. 2010.

[87] Hu R, Pu P. Acceptance issues of personality-based recommender systems. In Proceedings of the third ACM conference on Recommender systems 2009 Oct 23 (pp. 221-224).

[88] Rentfrow PJ, Gosling SD. The do re mi's of everyday life: the structure and personality correlates of music preferences. Journal of personality and social psychology. 2003 Jun;84(6):1236.

[89] Rentfrow PJ, Goldberg LR, Zilca R. Listening, watching, and reading: The structure and correlates of entertainment preferences. Journal of personality. 2011 Apr;79(2):223-58.

[90] Hu R, Pu P. A study on user perception of personality-based recommender systems. In International conference on user modeling, adaptation, and personalization 2010 Jun 20 (pp. 291-302). Springer, Berlin, Heidelberg

[91] Odić A, Tkalčič M, Tasič JF, Košir A, A.: Personality and Social Context: Impact on Emotion Induction from Movies. UMAP 2013 Extended Proceedings (2013)

[92] Odić A, Tkalčič M, Tasič JF, Košir A. Predicting and detecting the relevant contextual information in a movie-recommender system. Interacting with Computers. 2013 Jan 1;25(1):74-90

[93] Chen L, Wu W, He L. How personality influences users' needs for recommendation diversity? InCHI'13 extended abstracts on human factors in computing systems 2013 Apr 27 (pp. 829-834)

[94] Wu W, Chen L, He L. Using personality to adjust diversity in recommender systems. Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13 (May), 225–229 (2013). DOI 10.1145/2481492.2481521

[95] Kompan, M., Bielikova, M.: Social Structure and Personality Enhanced Group Recommendation. UMAP 2014 Extended Proceedings (2014).

[96] Gkika, S., and Lekakos, G. "Investigating the effectiveness of persuasion strategies on recommender systems." Semantic and Social Media Adaptation and Personalization (SMAP), 2014 9th International Workshop on. IEEE, 2014

[97] Bothos E, Apostolou D, and Mentzas G. A recommender for persuasive messages in route planning applications. Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on. IEEE, 2016.

[98] Sengvong S, Bai Q. Persuasive public-friendly route recommendation with flexible rewards. In 2017 IEEE International Conference on Agents (ICA) 2017 Jul 6 (pp. 109-114). IEEE.

[99] Wu S, Bai Q, Sengvong S. GreenCommute: an influence-aware persuasive recommendation approach for public-friendly commute options. Journal of Systems Science and Systems Engineering. 2018 Apr 1;27(2):250-64

[100] Herse S, Vitale J, Ebrahimian D, Tonkin M, Ojha S, Sidra S, Johnston B, Phillips S, Gudi SL, Clark J, Judge W. Bon appetit! robot persuasion for food recommendation. In Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction 2018 Mar 1 (pp. 125-126)

[101] Pilloni P, Piras L, Carta S, Fenu G, Mulas F, Boratto L. Recommender system lets coaches identify and help athletes who begin losing motivation. Computer. 2018 Mar 20;51(3):36-42.

[102] Adaji I, Kiron N, Vassileva J. Evaluating the Susceptibility of E-commerce Shoppers to Persuasive Strategies. A Game-Based Approach. In International Conference on Persuasive Technology 2020 Apr 20. pp. 58-72. Springer, Cham.

[103] Ünal P, Taskaya-Temizel T, Eren P. An Exploratory Study on the Outcomes of Influence Strategies in Mobile Application Recommendations. In BCSS@ PERSUASIVE 2014 May 22 (pp. 27-40)

[104] Heras S, Rodríguez P, Palanca J, Duque N, Julián V. Using argumentation to persuade students in an educational recommender system. In International Conference on Persuasive Technology 2017 Apr 4 (pp. 227-239). Springer, Cham.

[105] Zanker M, Schoberegger M. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In Joint Workshop on Interfaces and Human Decision Making in Recommender Systems 2014 Sep 6 (Vol. 1253, pp. 33-36)

[106] Yoo KH, Gretzel U. Creating more credible and persuasive recommender systems: The influence of source characteristics on recommender system evaluations. In Recommender systems handbook 2011 (pp. 455-477). Springer, Boston, MA.

[107] McNee S, Lam S, Konstan J, Riedl J. Interfaces for eliciting new user preferences in recommender systems. In International Conference on User Modeling 2003 Jun 22 (pp. 178-187). Springer, Berlin, Heidelberg.

[108] Cosle D, Lam S, Albert I, Riedl J. Is seeing believing? How recommender systems influence users' opinions. In Proceedings of the computer-human interaction 2003 conference on human factors in computing systems. Fort Lauderdale, FL 2003 (pp. 585-592).

[109] Pu P, Chen L. Trust-inspiring explanation interfaces for recommender systems. Knowledge-Based Systems. 2007 Aug 1;20(6):542-56.

[110] Petty RE, Cacioppo JT. Attitudes and persuasion: Classic and contemporary approaches. Routledge; 2018 Feb 20.

[111] Berkovsky S, Freyne J, Oinas-Kukkonen H, editors. Influencing individually: fusing personalization and persuasion. ACM Transactions on Interactive Intelligent Systems, 2(2) (2012)

[112] Oyibo K, Adaji I, Orji R, Olabenjo B, Vassileva J. Susceptibility to persuasive strategies: a comparative analysis of Nigerians vs. Canadians. In Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization 2018 Jul 3 (pp. 229-238).

[113] Oyibo K, Adaji I, Orji R, Vassileva J. The Susceptibility of Africans to Persuasive Strategies: A Case Study of Nigeria. PPT@ PERSUASIVE. 2018;2018: 8-21.

[114] Oyibo K, Orji R, Vassileva J. Investigation of the Influence of Personality Traits on Cialdini's Persuasive Strategies. PPT@ PERSUASIVE. 2017; 2017:8-20.

[115] Alkış N, Temizel TT. The impact of individual differences on influence strategies. Personality and Individual Differences. 2015 Dec 1; 87:147-52

[116] Kaptein, MC: Adaptive persuasive messages in an e-commerce setting: the use of persuasion profiles. In: European Conference on Information Systems, p. 183 (2011)

[117] Kaptein M, Van Halteren A. Adaptive persuasive messaging to increase service retention: using persuasion profiles to increase the effectiveness of email reminders. Personal and Ubiquitous Computing. 2013 Aug 1;17(6):1173-85.

[118] Kaptein MC. Personalized persuasion in ambient intelligence. Technische Universiteit Eindhoven, 2012. 226 p. https://doi.org/10.6100/IR729200.

[119] Kaptein, MC., Eckles, D. Selecting effective means to any end: Futures and ethics of persuasion profiling. In T. Ploug, P. Hasle, & H. Oinas-Kukkonen (Eds.), Persuasive technology. 2010. pp. 82—93. Springer Berlin / Heidelberg.

[120] Alslaity A, Tran T. The Effect of Personality Traits on Persuading Recommender System Users. IntRS '20 - Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 26, 2020. pp. 48-56.

[121] Alslaity A, Tran T. On the Impact of the Application Domain on Users' Susceptibility to the Six Weapons of Influence. InInternational Conference on Persuasive Technology 2020 Apr 20 (pp. 3-15). Springer, Cham.

[122] Basili VR. Software modeling and measurement: the Goal/Question/Metric paradigm. University of Maryland. CS-TR-2956, UMIACS-TR-92-96; 1992 Sep.

[123] Kaptein M, Eckles D. Heterogeneity in the effects of online persuasion. Journal of Interactive Marketing. 2012 Aug 1;26(3): pp.176-88.

[124] JASP team. (2019). JASP (version 0.10. 2)[computer software]. See https://jasp-stats.org.

[125] Gelman A. Analysis of variance—why it is more important than ever. The annals of statistics. 2005;33(1):1-53.

[126] Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in psychology. 2013 Nov 26;4:863.

[127] Cohen J. Statistical power analysis for the behavioral sciences. Academic press; 2013 Sep 3.

[128] Curran PG. Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology. 2016 Sep 1;66:4-19.

[129] Abdullahi A, Orji R, Oyibo K. Personalizing persuasive technologies: Do gender and age affect susceptibility to persuasive strategies? In Adjunct publication of the 26th conference on user modeling, adaptation and personalization 2018 Jul 2 (pp. 329-334).

[130] Oyibo K, Orji R, Vassileva J. The influence of culture in the effect of age and gender on social influence in persuasive technology. In Adjunct publication of the 25th

conference on user modeling, adaptation and personalization 2017 Jul 9 (pp. 47-52).

[131] Soto CJ, John OP. Development of Big Five Domains and Facets in Adulthood: Mean-Level Age Trends and Broadly Versus Narrowly Acting Mechanisms. Journal of Personality. 2012 Aug;80(4):881-914.

[132] Rothmann S, Coetzer E. The big five personality dimensions and job performance. SA Journal of Industrial Psychology. 2003 Jan 1;29(1):68-74.

[133] Alslaity A, Tran T. Towards persuasive recommender systems. In2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT) 2019 Mar 14. pp. 143-148. IEEE.

[134] Núñez-Valdéz ER, Lovelle JM, Martínez OS, García-Díaz V, De Pablos PO, Marín CE. Implicit feedback techniques on recommender systems applied to electronic books. Computers in Human Behavior. 2012 Jul 1;28(4):1186-93.

[135] Isinkaye FO, Folajimi YO, Ojokoh BA. Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal. 2015 Nov 1;16(3):261-73.

[136] Thathachar MA, Sastry PS. Networks of learning automata: Techniques for online stochastic optimization. Springer Science & Business Media; 2011 Jun 27

[137] Narendra KS, Thathachar MA. Learning automata-a survey. IEEE Transactions on systems, man, and cybernetics. 1974 Jul (4):323-34

[138] Thathachar MA, Sastry PS. Varieties of learning automata: an overview. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2002 Dec 10;32(6):711-22.

[139] Charness G, Gneezy U, Kuhn MA. Experimental methods: Between-subject and within-subject design. Journal of Economic Behavior & Organization. 2012 Jan 1;81(1):1-8.

[140] Alslaity A, Tran T. Towards a Comprehensive Evaluation of Recommenders: A Cognition-Based Approach. InCanadian Conference on Artificial Intelligence 2018 May 8. pp. 310-315. Springer, Cham.

[141] Alslaity A. A Unified Evaluation Framework for Recommenders. In Canadian Conference on Artificial Intelligence 2018 May 8. pp. 351-354. Springer, Cham.

[142] Alslaity A, Tran T. ComPer: A Comprehensive Performance Evaluation Method for Recommender Systems. International Journal of Information Technology and Computer Science (IJITCS), Vol.11, No.12, pp.1-18, 2019. DOI: 10.5815/ijitcs. 2019.12.01.

[143] Sohail SS, Siddiqui J, Ali R. A comprehensive approach for the evaluation of recommender systems using implicit feedback. International Journal of Information Technology. 2019 Sep 1;11(3):549-67.

[144] Moody J, Glass DH. A novel classification framework for evaluating individual and aggregate diversity in top-n recommendations. ACM Transactions on Intelligent Systems and Technology (TIST). 2016 Feb 1;7(3):1-21.

[145] Olmo FH, Gaudioso E. Evaluation of recommender systems: A new approach. Expert Systems with Applications. 2008 Oct 1;35(3):790-804.

[146] Krathwohl DR. A revision of Bloom's taxonomy: An overview. Theory into practice. 2002 Nov 1;41(4):212-8

[147] Adomavicius G, Tuzhilin A. 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. 2005 Apr 25;17(6):734-49.

[148] Campos PG, Díez F, Cantador I. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. User Modeling and User-Adapted Interaction. 2014 Feb 1;24(1-2):67-119

[149] Erdt M, Jomrich F, Schüler K, Rensing C. Investigating crowdsourcing as an evaluation method for TEL recommender systems. ECTEL meets ECSCW. 2013 Sep 21:25-9.

[150] Najjar NA, Wilson DC. Evaluating group recommendation strategies in memory-based collaborative filtering. In Proceedings of the ACM recommender systems conference workshop on human decision making in recommender systems 2011 (pp. 43-51). New York, NY: ACM.

[151] Trattner C, Said A, Boratto L, Felfernig A. Evaluating group recommender systems. In Group Recommender Systems 2018 (pp. 59-71). Springer, Cham.

[152] Monti D, Palumbo E, Rizzo G, Morisio M. Sequeval: An offline evaluation framework for sequence-based recommender systems. Information. 2019 May;10(5):174.

[153] Ludewig, M., Jannach, D. Evaluation of session-based recommendation algorithms. In User Model User-Adaptation International Conference. 2018. 28, 331–390. https://doi.org/10.1007/s11257-018-9209-6

[154] Said A, Bellogín A, De Vries A. A top-n recommender system evaluation protocol inspired by deployed systems. In LSRS Workshop at ACM RecSys 2013 Oct 1.

[155] Dooms S, Bellogín A, Pessemier TD, Martens L. A framework for dataset benchmarking and its application to a new movie rating dataset. ACM Transactions on Intelligent Systems and Technology (TIST). 2016 Feb 11;7(3):1-28.

[156] Pu P, Chen L, Hu R. A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems 2011 Oct 23 (pp. 157-164).

[157] Böhmer M, Ganev L, Krüger A. *Appfunnel*: A framework for usage-centric evaluation of recommender systems that suggest mobile applications. In Proceedings of

the 2013 international conference on Intelligent user interfaces 2013 Mar 19 (pp. 267-276).

[158] Schröder G, Thiele M, Lehner W. Setting goals and choosing metrics for recommender system evaluations. In UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA 2011 Oct 23 (Vol. 23, p. 53).

[159] Arana-Llanes JY, Rendón-Miranda JC, González-Serna JG, Alejandres-Sánchez HO. Design and user-centered evaluation of recommender systems for mobile devices-Methodology for user-centered evaluation of context-aware recommender systems. In 2014 International Conference on Computational Science and Computational Intelligence 2014 Mar 10 (Vol. 2, pp. 277-280). IEEE.

[160] Duwairi RM. Machine learning for Arabic text categorization. Journal of the American Society for Information Science and Technology. 2006 Jun;57(8):1005-10.

[161] Triantaphyllou E, Shu B, Sanchez SN, Ray T. Multi-criteria decision making: an operations research approach. Encyclopedia of electrical and electronics engineering. 1998 Feb;15(1998):175-86.

[162] Guo G, Zhang J, Sun Z, Yorke-Smith N. LibRec: A Java Library for Recommender Systems. In Proceedings of the 23rd Conference on User Modelling UMAP 2015 Jun 29 (Vol. 4).

[163] Zheng Y, Mobasher B, Burke R. Carskit: A java-based context-aware recommendation engine. In2015 IEEE International Conference on Data Mining Workshop (ICDMW) 2015 Nov 14 (pp. 1668-1671). IEEE.

[164] Garimella K, Morales G, Gionis A, Mathioudakis M. Reducing controversy by connecting opposing views. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining 2017 Feb 2 (pp. 81-90).

[165] Christakopoulou E, Karypis G. Local item-item models for top-n recommendation. In Proceedings of the 10th ACM Conference on Recommender Systems 2016 Sep 7 (pp. 67-74).

[166] Cheng W, Yin G, Dong Y, Dong H, Zhang W. Collaborative filtering recommendation on users' interest sequences. PloS one. 2016 May 19;11(5): e0155739.

[167] Hofmann T, Puzicha J. Latent class models for collaborative filtering. In IJCAI 1999 Jul 31 (Vol. 99, No. 1999).

[168] Hofmann T. Latent semantic models for collaborative filtering. ACM Transactions on Information Systems (TOIS). 2004 Jan 1;22(1):89-115.

[169] Natarajan S, Vairavasundaram S, Natarajan S, Gandomi AH. Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data. Expert Systems with Applications. 2020 Jul 1;149:113248.

[170] Herlocker JL, Konstan JA, Terveen LG, Riedl JT. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS). 2004 Jan 1;22(1):5-53.

[171] Zhang M, Hurley N. Avoiding monotony: improving the diversity of recommendation lists. In Proceedings of the 2008 ACM conference on Recommender systems 2008 Oct 23 (pp. 123-130).

[172] Bellogin A, Castells P, Cantador I. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In Proceedings of the fifth ACM conference on Recommender systems 2011 Oct 23 (pp. 333-336).

[173] Abdollahpouri H, Adomavicius G, Burke R, Guy I, Jannach D, Kamishima T, Krasnodebski J, Pizzato L. Multistakeholder recommendation: Survey and research directions. User Modeling and User-Adapted Interaction. 2020 Mar;30(1):127-58.

[174] Abdollahpouri H, Burke R, Mobasher B. Recommender systems as multistakeholder environments. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (pp. 347-348). ACM. (2017).

[175] Burke R, Abdollahpouri H, Mobasher B, Gupta T. Towards Multi-Stakeholder Utility Evaluation of Recommender Systems. In UMAP. 2016 Jul 13.

[176] Sohrabi B, Toloo M, Moeini A, Nalchigar S. Evaluation of recommender systems: A multi-criteria decision making approach. Iranian Journal of Management Studies (IJMS). Vol. 8, No. 4. pp: 589-605 (2015).

[177] Said A, Tikk D, Stumpf K, Shi Y, Larson MA, Cremonesi P. Recommender Systems Evaluation: A 3D Benchmark. In RUE@ RecSys 2012 Sep 9 (pp. 21-23)

[178] Alwidian S, Amyot D, Babin G. Evaluating the potential of technology in justice systems using goal modeling. In International Conference on E-Technologies 2017 May 17 (pp. 185-202). Springer, Cham.

[179] Yu E. Towards modelling and reasoning support for early-phase requirements engineering. In Proceedings of ISRE'97: 3rd IEEE International Symposium on Requirements Engineering 1997 Jan 6 (pp. 226-235). IEEE.

[180] Chung L, Nixon BA, Yu E, Mylopoulos J. Non-functional requirements in software engineering. Springer Science & Business Media; 2012 Dec 6.

[181] Giorgini P, Mylopoulos J, Sebastiani R. Goal-oriented requirements analysis and reasoning in the tropos methodology. Engineering Applications of Artificial Intelligence. 2005 Mar 1;18(2):159-71.

[182] Lamsweerde A. Requirements engineering: From system goals to UML models to software. Chichester, UK: John Wiley & Sons; 2009.

[183] ITU-T, Recommendation Z.151 (10/12): User Requirements Notation (URN) – Language Definition, Geneva, Switzerland (2012). http://www.itu.int/rec/T-REC-Z.151/en

[184] Freeman R. Strategic management: A stakeholder approach. Cambridge university press; 2010 Mar 11.

[185] Akhigbe O, Alhaj M, Amyot D, Badreddin O, Braun E, Cartwright N, Richards G, Mussbacher G. Creating quantitative goal models: governmental experience. In International Conference on Conceptual Modeling 2014 Oct 27 (pp. 466-473). Springer, Cham.

[186] Mussbacher G, Ghanavati S, Amyot D. Modeling and Analysis of URN Goals and Scenarios with jUCMNav. In2009 17th IEEE International Requirements Engineering Conference 2009 Aug 31 (pp. 383-384). IEEE. http://softwareengineering.ca/jucmnav/

[187] Perry DE, Porter AA, Votta LG. Empirical studies of software engineering: a roadmap. In Proceedings of the conference on the future of Software engineering 2000 May 1 (pp. 345-355).

Appendix A: Big Five Inventory (BFI-44)

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others*? Please write a number next to each statement to indicate the extent to which **you agree or disagree with that statement.**

1 Disagree Strongly	2 Disagree a little	3 Neither agree nor disagree	4 Agree a little	5 Agree strongly

I am someone who…

1. _____ Is talkative
2. _____ Tends to find fault with others
3. _____ Does a thorough job
4. _____ Is depressed, blue
5. _____ Is original, comes up with new ideas
6. _____ Is reserved
7. _____ Is helpful and unselfish with others
8. _____ Can be somewhat careless
9. _____ Is relaxed, handles stress well.
10. _____ Is curious about many different things
11. _____ Is full of energy
12. _____ Starts quarrels with others
13. _____ Is a reliable worker
14. _____ Can be tense
15. _____ Is ingenious, a deep thinker
16. _____ Generates a lot of enthusiasm
17. _____ Has a forgiving nature
18. _____ Tends to be disorganized
19. _____ Worries a lot
20. _____ Has an active imagination
21. _____ Tends to be quiet
22. _____ Is generally trusting
23. _____ Tends to be lazy
24. _____ Is emotionally stable, not easily upset
25. _____ Is inventive
26. _____ Has an assertive personality
27. _____ Can be cold and aloof
28. _____ Perseveres until the task is finished
29. _____ Can be moody
30. _____ Values artistic, aesthetic experiences
31. _____ Is sometimes shy, inhibited
32. _____ Is considerate & kind to almost everyone
33. _____ Does things efficiently
34. _____ Remains calm in tense situations
35. _____ Prefers work that is routine
36. _____ Is outgoing, sociable
37. _____ Is sometimes rude to others
38. _____ Makes plans and follows through with them
39. _____ Gets nervous easily
40. _____ Likes to reflect, play with ideas
41. _____ Has few artistic interests
42. _____ Likes to cooperate with others
43. _____ Is easily distracted
44. _____ Is sophisticated in art, music, or literature

Appendix B: Persuasion Questionnaire

Part 1: On a scale of -1 to 5, to what extent may the website influence (by encouraging or discouraging) you to buy items recommended to you by providing each of the following options?

- **(-1) Negative Effect**: indicate that the sentence may discourage you from buying the item.
- **(0) Neutral:** presenting or removing the sentence will not affect your decision.
- **(1 to 5) Positive Effect.** You are very low (1) to very high (5) agree that the sentence may persuade you to buy the recommended item.

#		(-1)	(0)	(1)	(2)	(3)	(4)	(5)
1	Giving you something for free (e.g., samples, gift, or free delivery)	○	○	○	○	○	○	○
2	Display a countdown, beside an item, indicating the time remaining for an offer on that item	○	○	○	○	○	○	○
3	Presenting an image of an expert uses the recommended item (ex: a doctor uses particular medication for his patients, or a security guard uses the recommended security lock)	○	○	○	○	○	○	○
4	Presenting the best seller's items	○	○	○	○	○	○	○
5	Well designed (Fancy and professional) website's interface and product's presentation.	○	○	○	○	○	○	○
6	Using "add to wish list" option. *("wish list" contains items that you wish to consume (i.e., buy, watch, use, etc.) in the future*)	○	○	○	○	○	○	○

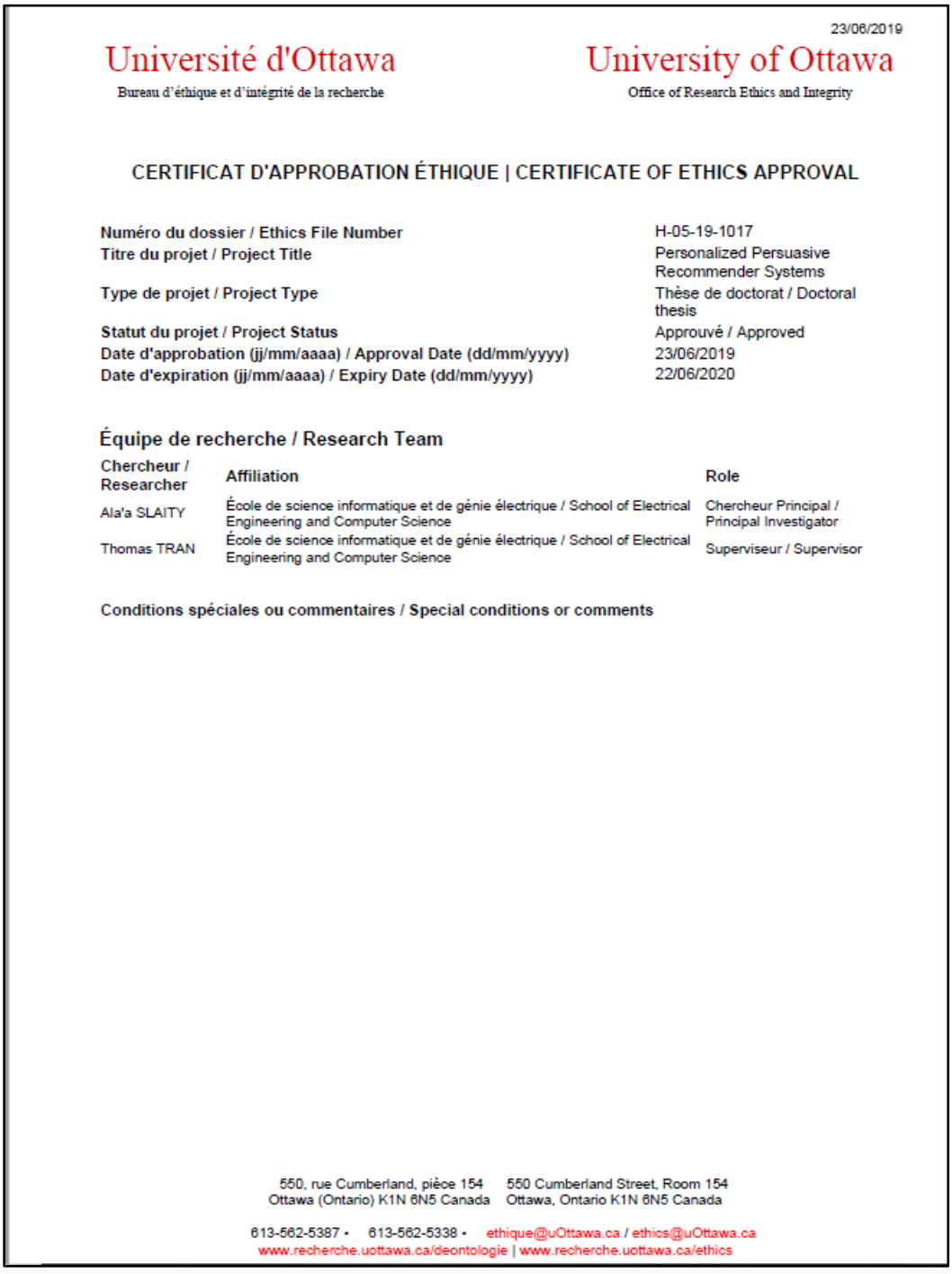
Part 2: Suppose that you are using an e-commerce system (ex: Amazon, ebay, etc). on a scale of -1 to 5, to what extent do you think that each of the following statements will influence your decision (by encouraging or discouraging you) to buy an item recommended to you based on your preferences?

#		(-1)	(0)	(1)	(2)	(3)	(4)	(5)
1	A friend of you, who bought the item that you suggested to him/her in the past, recommends you this item	○	○	○	○	○	○	○
2	The recommended item will be available for 2 months only	○	○	○	○	○	○	○
3	The recommended item won 3 prizes as the best manufactured product	○	○	○	○	○	○	○
4	87% of users rated the recommended item with 4 or 5 stars	○	○	○	○	○	○	○
5	Your Facebook friends bought this item	○	○	○	○	○	○	○
6	This item belongs in the kind of items you usually buy	○	○	○	○	○	○	○

Part 3: Suppose that you are using an online movie system (ex: Netflix, YouTube, etc.). on a scale of -1 to 5, to what extent do you think that each of the following statements will influence your decision (by encouraging or discouraging you) to watch a movie recommended to you based on your preferences?

#		(-1)	(0)	(1)	(2)	(3)	(4)	(5)
1	A Facebook friend, who saw the movie that you suggested him/her in past, recommends you this movie	○	○	○	○	○	○	○
2	The recommended movie will be available to view from 15/1/2014 to 31/1/2014 on cinemas	○	○	○	○	○	○	○
3	The recommended movie won 3 Oscars!	○	○	○	○	○	○	○
4	The 87% of users in this survey rated the recommended movie with 4 or 5 stars!	○	○	○	○	○	○	○
5	Your Facebook friends like this movie	○	○	○	○	○	○	○
6	This movie belongs in the kind of movies you enjoy watching	○	○	○	○	○	○	○

Appendix C: Certificate of Ethics Approval



23/06/2019

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

CERTIFICAT D'APPROBATION ÉTHIQUE | CERTIFICATE OF ETHICS APPROVAL

Numéro du dossier / Ethics File Number	H-05-19-1017
Titre du projet / Project Title	Personalized Persuasive Recommender Systems
Type de projet / Project Type	Thèse de doctorat / Doctoral thesis
Statut du projet / Project Status	Approuvé / Approved
Date d'approbation (jj/mm/aaaa) / Approval Date (dd/mm/yyyy)	23/06/2019
Date d'expiration (jj/mm/aaaa) / Expiry Date (dd/mm/yyyy)	22/06/2020

Équipe de recherche / Research Team

Chercheur / Researcher	Affiliation	Role
Ala'a SLAITY	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Chercheur Principal / Principal Investigator
Thomas TRAN	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Superviseur / Supervisor

Conditions spéciales ou commentaires / Special conditions or comments

550, rue Cumberland, pièce 154
Ottawa (Ontario) K1N 6N5 Canada

550 Cumberland Street, Room 154
Ottawa, Ontario K1N 6N5 Canada

613-562-5387 • 613-562-5338 • ethique@uOttawa.ca / ethics@uOttawa.ca
www.recherche.uottawa.ca/deontologie | www.recherche.uottawa.ca/ethics

Appendix D: Data Significance Analysis

Table 16 ANOVA significance test and Effect Size based on users' gender

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	0.019	0.108	9.095	1.086	1.911	1.397
P-value	0.892	0.742	**0.003**	0.298	0.168	0.238
Cohen's P	0.02	0.04	0.39	0.13	0.2	0.2

Table 17 ANOVA significance test based on users' ages

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	0.473	1.051	0.336	1.072	0.766	0.863
P-value	0.702	0.371	0.799	0.361	0.514	0.461

Table 18 ANOVA significance test based on users' culture

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	6.114	12.194	7.375	3.565	10.842	7.986
P-value	**0.003**	**<0.001**	**<0.001**	**0.03**	**<0.001**	**<0.001**

Table 19 Effect Size based on users' culture (A=Asia, E=Europe, N=North America)

Groups	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
A & E	0.15	0.13	0.4	0.29	0.3	0.2
A & N	0.01	0.15	0.16	0.22	0.13	0.19
E & N	0.14	0.18	0.44	0.16	0.27	0.13

Table 20 ANOVA significance test based on users' personality traits

Personality		Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
Extraversion	F	11.836	7.698	7.591	10.446	18.418	15.812
	P-value	**<0.001**	**0.006**	**0.006**	**0.001**	**<0.001**	**<0.001**
Agreeableness	F	7.235	3.154	1.012	2.476	2.443	0.27
	P-value	**0.008**	**0.077**	0.315	0.117	0.119	0.604
Conscientious-	F	1.333	6.317	1.174	0.003	0.192	0.192
ness	P-value	0.249	**0.013**	0.28	0.96	0.662	0.662
Neuroticism	F	1.343	4.198	1.012	1.239	0.043	4.198
	P-value	0.248	**0.041**	0.315	0.267	0.835	**0.041**
Openness	F	0.198	3.946	0.024	0.257	6.377	0.102
	P-value	0.657	**0.048**	0.877	0.613	**0.012**	0.75

Table 21 Effect Size based on users' personality

Personality	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
Extraversion	0.42	0.34	0.34	0.39	0.52	0.49
Agreeableness	0.33	0.22	0.12	0.19	0.19	0.16
Conscientiousness	0.14	0.30	0.13	0.11	0.15	0.11
Neuroticism	0.14	0.25	0.12	0.14	0.13	0.20
Openness	0.05	0.24	0.12	0.16	0.31	0.14

Table 22 RM-ANOVA significance test and Effect Size based on the application domain (the Whole sample)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	16.781	1.237	2.009	3.254	17.568	40.688
P-value	**<0.001**	0.267	0.158	0.072	**<0.001**	**<0.001**
Cohen's P	0.26	0.17	0.11	0.11	0.25	0.45

Table 23 ANOVA significance test based on the domain & gender (Female)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	5.746	0.566	-20.5841	2.055	19.823	34.521
P-value	0.019	0.454	1	0.155	<0.001	<0.001

Table 24 ANOVA significance test based on the domain & gender (Male)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	10.52	0.541	2.701	1.171	4.649	15.225
P-value	0.001	0.463	0.102	0.281	0.032	<0.001

Table 25 Effect Sizes based on the domain & gender

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
Female	0.26	0.19	0.12	0.14	0.43	0.59
Male	0.25	0.16	0.17	0.19	0.17	0.38

Table 26 ANOVA significance test based on the domain & age (16-25 years old)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	4.783	0.604	7.755	1.847	2.319	3.864
P-value	**0.034**	0.441	**0.008**	0.181	0.134	**0.055**

Table 27 ANOVA significance test based on the domain & age (26-35 years old)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	5.653	0.747	0.248	2.933	4.536	19.796
P-value	**0.019**	0.389	0.619	0.089	**0.035**	**<0.001**

Table 28 ANOVA significance test based on the domain & age (36-45 years old)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	11.174	1.277	0.748	0.044	1.475	7.476
P-value	**0.001**	0.263	0.391	0.835	0.23	**0.008**

Table 29 ANOVA significance test based on the domain & age (45 years and older)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	0.257	1.217	0.608	0.023	22.979	11.19
P-value	0.616	0.278	0.442	0.881	**<0.001**	**0.002**

Table 30 Effect Sizes based on the domain & age

Group	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
16-25	0.27	0.18	0.36	0.25	0.22	0.23
26-35	0.24	0.15	0.13	0.14	0.20	0.53
36-45	0.45	0.13	0.16	0.12	0.17	0.49
>45	0.13	0.12	0.11	0.11	0.72	0.64

Table 31 ANOVA significance test based on the domain & culture (Asia)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	0.973	2.638	4.239	1.801	0.392	0.298
P-value	0.327	0.109	0.043	0.184	0.533	0.587

Table 32 ANOVA significance test based on the domain & culture (Europe)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	1.158	6.76	0.022	0.288	1.669	10.353
P-value	0.305	0.025	0.884	0.604	0.223	0.008

Table 33 ANOVA significance test based on the domain & culture (North America)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	17.494	1.152	0.361	1.742	25.122	45.007
P-value	<0.001	0.285	0.549	0.189	<0.001	<0.001

Table 34 Effect Sizes based on the domain & culture

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
Asia	0.13	0.21	0.21	0.14	0.18	0.18
Europe	0.46	0.66	0.14	0.15	0.51	1.13
North America	0.31	0.19	0.17	0.19	0.39	0.57

Table 35 ANOVA significance test based on the domain & Personality trait (Extraversion)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	5.209	1.677	3.132	0.445	6.404	30.704
P-value	0.024	0.197	0.079	0.506	0.012	<0.001

Table 36 ANOVA significance test based on the domain & Personality trait (Agreeableness)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	10.041	1.509	4.217	1.595	18.547	33.654
P-value	0.002	0.001	0.042	0.209	<0.001	<0.001

Table 37 ANOVA significance test based on the domain & Personality trait (Conscientiousness)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	11.453	1.044	2.343	1.535	11.192	35.362
P-value	<0.001	0.309	0.128	0.218	0.001	<0.001

Table 38 ANOVA significance test based on the domain & Personality trait (Neuroticism)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	8.714	0.277	0.023	0.666	5.821	10.961
P-value	0.004	0.6	0.88	0.416	0.017	0.001

Table 39 ANOVA significance test based on the domain & Personality trait (Openness)

	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
F	16.605	0.665	0.477	3.183	17.391	24.659
P-value	<0.001	0.416	0.491	0.077	<0.001	<0.001

Table 40 Effect Sizes based on the domain & culture

Personality	Reciprocity	Scarcity	Authority	Social proof	Liking	Commitment
Extraversion	0.20	0.11	0.17	0.16	0.21	0.51
Agreeableness	0.32	0.11	0.21	0.14	1.38	0.47
Conscientiousness	0.34	0.18	0.15	0.14	0.25	0.57
Neuroticism	0.23	0.19	0.12	0.16	0.23	0.44
Openness	0.35	0.19	0.17	0.15	0.27	0.46

Appendix E: An Example of the Text used to infer the correlation between Bloom and Recommender Systems.

Scalability Dimension

- **The text:**

scale up to real at huge set. help users navigate in large collections of items. providing rapid results even for huge data sets consisting of millions of items. With the growth of the data set, many algorithms are either slowed down or require additional resources such as computation power or memory to evaluate the computational complexity of an algorithm in terms of time or space requirements to understand the scalability of the system it is also useful to report the consumption of system resources over large data sets. Showing how the speed and resource consumption behave as the task scales up. As recommender systems are expected in many cases to provide rapid recommendations online, it is also important to measure how fast does the system provides recommendations the throughput of the system, i.e., the number of recommendations that the system can provide per second. We could also measure the latency (also called response time) — the required time for making a recommendation online. How scalable is the system with respect to number of users, underlying data size, and algorithm performance? One of the most important goals of a recommendation system is to provide online recommendations for users to navigate through a collection of items. When the system scales up to the point where there are thousands of components to be recommended, the system must be able to process and make each recommendation within a reasonable amount of time. The scalability problem can be divided into two parts: (1) the training time of the recommendation algorithm and (2) the performance of the system or throughput when working with a large item database. The time that is required to train the algorithm can be evaluated by training different algorithms with the same dataset or by training them until they reach the same level of prediction correctness.

The performance of the system can be evaluated in terms of throughput— the number of recommendations that the system can generate per second. Performance (in terms of number of recommendations) can also adversely impact the usability of the recommendation system as response time may become too slow to be effective for its users. As recommender systems are designed to help users navigate in large collections. of items, one of the goals of the designers of such systems is to scale up to real data sets. As such, it is often the case that algorithms trade other properties, such as accuracy or coverage, for providing rapid results even for huge data sets consisting of millions of items. With the growth of the data set, many algorithms are either slowed down or require additional resources such as computation power or memory. One standard approach in computer science research is to evaluate the computational complexity of an algorithm in terms of time or space requirements. In many cases, however, the complexity of two algorithms is either identical, or could be reduced by changing some parameters, such as the complexity of the model, or the sample size. Therefore, to understand the scalability of the system it is also useful to report the consumption of system resources over large data sets. Scalability is typically measured by experimenting with growing data sets, showing how the speed and resource consumption behave as the task scales up. It is important to measure the compromises that scalability dictates. For example, if the accuracy of the algorithm is lower than other candidates that only operate on relatively small data sets, one must show over small data sets the difference in accuracy. Such measurements can provide valuable information both on the potential performance of recommender systems in general for the specific task, and on future directions to explore. As recommender systems are expected in many cases to provide rapid recommendations online, it is also important to measure how fast does the system provides recommendations. One such measurement is the throughput of the system, i.e., the number of recommendations that the system can provide per second. We could also measure the latency (also called response time) the required time for making a recommendation online.

- **The Vector:**

Performance Scalability Showing able accuracy additional adversely algorithm amount approach become behave called candidates case changing collection collections complexity

component compromises computation computational computer consisting consumption correctness coverage data database dataset design dictates different direction divided effective evaluate even example expected experimenting explore fast future general generate goals growth help huge identical impact important information item large latency level lower make many measure measurement memory millions model navigate number one online only operate over parameter part per performance point potential power prediction problem process property provide rapid reach real reasonable recommendation recommend reduced relatively report require requirement research resource respect response result same sample scalable scale science second set show size slow small some space specific speed standard system task terms thousand throughput time trade train training typically underlying understand usability useful users valuable when where with within working