KU LEUVEN

FACULTY OF SCIENCES

Let's watch a movie together

Achieving a final decision for group movie recommendations

Ethan Van den Bleeken

Supervisor: *Prof. dr. K. Verbert*
Tutor: *Oscar Luis Alvarado Rodriguez*

Master thesis submitted in fulfillment
of the requirements for the degree in
Master of Science in Applied Informatics: A.I.

Academic year 2020-2021

Preface

What you are about to read is the result of the accumulation of research, brainstorming, endless indecisiveness, late nights, and a lot of being stuck at home instead of on-campus (sadly). As this was by no means a solo piece and I was accompanied by a great orchestra, acknowledgments are in place.

I want to start by thanking the supervisor of this thesis, *Prof. dr. Katrien Verbert*, for giving me the opportunity to work on this amazing and interesting topic.

Then I want to the conductor, my mentor *Oscar Luis Alvarado Rodriguez*, for without his guidance, time, patience, and advice, the result would not nearly be what it is today. Thank you!

I also want to thank my parents for being the backbone and driving force not only throughout my studies but throughout my entire life. Always there when needed the most.

Then my girlfriend, *Jilka*, the one who has endured the most the past year, all my ups and downs. No matter if it were complaints or euphoria, you were there when I needed you to support me and to help me make the right decisions.

And finally I want to thank all the people who have participated during my studies, as for them none of this would have been possible.

Ethan Van den Bleeken

Abstract

Every year more people are using online movie streaming platforms like Netflix. They often offer advanced recommendation systems but mainly towards a single user. Recommending movies gets more complex if a pair, instead of an individual, of people want to watch a movie together. This thesis explores a new and final phase to group movie recommendations systems, a final decision phase in which users reduce an initial list of recommended movies to a final decision, in terms of design solutions, user experience, and manipulation behavior. This research conducted three different studies in an iterative and incremental methodology to explore each research question. The results of the first study indicate that users want a quick approach and they suggested a layered approach, that combines a fast mechanism to filter and a complex mechanism to add depth to the final decision. The second study showed that this layered approach provides a better user experience compared to a negative elicitation approach. And from the analysis of the final study, it seems that participants accept possible manipulation enabled by the layered approach. This thesis proposes a new phase, the final decision phase, to the current process of a group recommendation system (GRS), as well as design guidelines for this phase. A system for this phase should allow enough control over the final decision but also be fast and leave room for discussion over the final discussion This thesis presents an implementation, MovieNight. Manipulation was accepted in this group context as participants expected users not to manipulate as this defeats the purpose of the application and taking away control to counteract it might be too restrictive.

Acronyms

CBF content-based filtering.

CF collaborative filtering.

GRS group recommendation system.

MS musical sophistication.

RS recommendation system.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Every year more and more people are turning to online movie streaming services that also offer movie recommendations. Just take a look at the data of Netflix, the platform now has 203,66 million subscribers[1], this has almost increased tenfold compared to 2011. With all these movies available at the touch of a finger, people still ask the question *"What movie to watch"*, admittedly platforms like Netflix contain quite sophisticated recommendation systems but these are mainly based on a single user.
Consider a pair of people who pose the same question, the situation gets a lot more complex. There are a series of techniques based on a group model that do recommend movies to groups of users, but how to effectively and efficiently reduce the list of recommended movies to a final decision remains an open problem.
This part of the process has also never been mentioned in the existing literature. Isinkaye et al.[2] mention the current process for group movie recommendations, but this doesn't include the phase I describe, the last phase in this framework is the recommendations phase in which movies get recommended to the group. That's why this research will propose an addition to this process, a *"final decision phase"* where users get the opportunity to reduce a list of recommendations to a final decision that is optimal for them.
This research will try to fill this gap and propose a possible solution for this problem as well as to shine some light on this issue and encourage further research. This thesis achieves this by splitting up the problem into three research questions:

- What are design solutions expected by users to achieve the final decision of group movie recommendations?
- In what ways do different mobile features, to achieve a final decision, present an improvement in user experience of the recommender?
- In what ways do different mobile features, to achieve a final decision, impact manipulation behaviour in a group movie recommender?

This thesis will explore each research question iteratively and incrementally. First, this research explored what users themselves expect from such a system by letting them design a possible solution during a co-design session. They will also explore two existing prototypes to spark their imagination more. Second, the research uses the proposed designs of the first iteration to explore which design provides a better user experience. Finally, this thesis explores manipulation behavior in the proposed design that provides the best user experience.

It will explore whether or not the design enables it and if/how the design should be modified to possibly counteract it. The proposed solution will be a mobile one, as almost all users own a smartphone, this makes it easy to use daily. The focus mainly lies in proposing a system that best aids the users in achieving the best possible final decision.
This thesis proposes the final decision phase, to the current process of group recommendations systems, as explained at the beginning of this chapter. It also presents design guidelines to implement this phase into a GRS.
Such an implementation should aim to strike a balance between the speed of the process and the control over it. It should also leave room for discussion, between the users, over the final decision.
An interface design example of this phase is the mobile app that this thesis presents, MovieNight. This app follows all the design guidelines and will be explained in detail in the following chapters.
A mobile solution provides privacy over the process, as users use the app on their mobile device. This privacy was valued by the participants and is not feasible in a tv-app solution.
In general participants accepted manipulation in this group context as they expected users to not manipulate if they knew each other well. Trying to counteract manipulation might also lead to less control over the process and participants would find this too restrictive. Future research can focus on other group contexts, where more manipulation behavior is exhibited, because some participants expected to have negative reactions to manipulation behavior if it occurred.
This thesis is split up into different chapters. The following chapter will focus on the research that has already been done to provide the needed background for this thesis. This literature study will go over the research of different topics related to the research questions: recommendation systems, group recommendation systems, and manipulation behavior.
The next three chapters, chapters 3, 4, and 5, will explain and lay out the three different user studies that were conducted to explore each research question respectively. The chapters will include methodology, results, and discussion.
The results of each user study will be discussed in chapter 6 to reach a final conclusion in chapter 7 and go over the most important takeaway points from the research done and point out possible further research aspects that seem most interesting based on the results of this thesis.

Chapter 2

Literature Review

This chapter contains a brief overview of the work already done in the field of recommendation systems. Followed by going more in-depth about the topic of this master thesis, namely group recommendation systems, highlighting the most important differences and the main challenges that GRS bring compared to regular recommendation system (RS). As well as explaining the gaps in the current literature, that will be brought to light in this research.

2.1 Recommendation systems

Recommendation systems are a heavily researched topic as well as being heavily used in a lot of online platforms (e.g. entertainment streaming platforms, webshops, recipes, ...). According to the meta-analysis by Jannach et al.[3] the main application domains were movies and general e-commerce products. This makes it clear that RS is a valid topic to research as it is so widely used nowadays and can be used to enhance the user experience of the system.
The research of RS can be split up into two main parts, the research about RS algorithms, which will be used to provide the user with a recommendation, and the user interface of the system itself, how the user will interact with the system. This will be the main focus of this thesis, but it is important to understand the underlying techniques used to recommend an item or certain items to the user.

Recommendation system algorithms

In this section, the most used algorithms will be highlighted and explained on a high level to provide a general understanding of the workings of the algorithms. Isinkaye et al.[2] explains three different recommendation filtering techniques; content-based filtering, collaborative filtering, and hybrid filtering. The following sections will go more in-depth about these techniques.

Content-Based filtering

Isinkaye et al.[2] explains content-based filtering in the following way:

> In content-based filtering technique, recommendation is made based on the user profiles using features extracted from the content of the items the user has evaluated in the past.

This means that this technique requires prior data of the user before it can make a recommendation for that user. It will then recommend the items most similar to the items rated as positive by the user.
One of the benefits of this approach is that it does not need information about other users and how they rated items and user preferences are quickly updated when new items are rated and added to the user profile.
However the major downfall of this approach is that the system will need an in-depth description and knowledge about the items themselves to measure the similarity. Meaning that the effectiveness of content-based filtering (CBF) depends heavily on the available data of the items.

Collaborative filtering

According to the findings of Jannach et al.[3] collaborative filtering has been the most used and researched recommendation technique in a period spanning from 2011 until 2016. And is explained by Isinkaye et al.[2] like:

> collaborative filtering (CF) technique works by building a database of preferences for items by users. It then matches users by calculating similarities between their profiles to make recommendations.

Meaning that similar users will impact the recommendation. So the user will get recommendations of items he has not rated before when similar users have rated this item positively.
The CF technique has a major advantage over the aforementioned CBF technique in that it does not require in-depth data about the items to recommend items to the user. It can also recommend relevant items to the user without them being in the users' profile.
Still the CF technique has some potential problems with the main ones being:

- Cold-start problem: When the profile of the user is (nearly) empty so the system does not know its taste, this decreases the quality of recommendations significantly.

- Data-sparsity problem: When only a few items are rated, the system has a hard time finding similar users and this results in weak recommendations.

Hybrid filtering

As the name suggests hybrid filtering techniques combine different techniques to get a better performing technique as a result. Isinkaye et al.[2] explains:

> Hybrid filtering technique combines different recommendation techniques in order to gain better system optimization to avoid some limitations and problems of pure recommendation systems.

The goal of this technique is then to create a superior technique by combining different ones. By doing this one technique can overcome the weaknesses of another and thus suppressing those weaknesses.
The combining can be done in the following ways, by implementing different techniques and combining the result or by creating a unified system that combines different techniques.

Recommendation system user experience

In this section the user experience of RS will be explored. When giving features to the user to use to get recommendations it is also important to investigate the impact of those features on the overall user experience of the system, only then the feature will add value to the system. As Konstan and Riedl[4] put it:

> The sweet spot is recommenders that balance serving users effectively while ensuring that the users have the control they desire.

An important aspect to keep in mind when recommending items is that users might not understand why a certain item was recommended to them, this can be solved by providing explanations about the recommendations. Konstan and Riedl[4] give three motivations for explanations in RS.

1. Transparency: Shows how the recommendation was formed, so the user knows how much to trust it.
2. Trust: Which encourages the user to take the recommendation, independent of how accurate it is.
3. Scrutability: Which enables to let the system know of mistakes in the data used for the recommendations, so future recommendations can be improved.

Millecamp et al.[5] also explored explanations in RS, specifically in music RS. They conducted a study on whether the system should explain the recommendations or not based on personal characteristics. Their findings were that explanations did not raise the confidence of all users, users with a high need for cognition preferred no explanations because explanations lowered their confidence. They mention that more research needs to be done to conclude a reasoning behind why confidence dropped in some users. The main takeaway from this study is that:

> Explanations, much like recommendations themselves, should be personalised for different end-users.

Another interesting point to take into account when designing the interface of a RS is the overall complexity of the interface, some users might prefer a more simple interface so they are not overwhelmed by all the features and options provided, but other users might actually enjoy the abundance of features to explore the system. This is also researched by Jin et al.[6] where two interfaces, both containing bubble charts, were used to recommend music. The complex interface presented the option of labeling the X- and Y-axis with musical attributes to get a deeper understanding of the recommended songs, while the simple one did not present such an option, to keep the chart easy to understand. One of the results of this research was that users with a higher musical sophistication preferred the complex interface and the opposite was true for users with a lower musical sophistication (MS). Although

according to their results 54% of the participants perceived the complex system as more informative which is to be expected as it contained more information but not all users will be able to process the information to experience a gain in information compared to the simpler interface.

Conclusion

This section covered the mainly used recommendation techniques and also touched upon the Human-Computer Interaction side of this by talking about the importance of the user interface in RS.
A lot of research about the techniques has already been done to create a collection of very accurate recommendation techniques. On the human-computer interaction side of things, there are still a lot of challenges to be solved as there are far more factors to keep in mind when designing a user interface because of the inherent diversity of a user base. The main method for researching those problems will be conducted on a population of real users and can't be done theoretically which makes it harder to research in general.

2.2 Group recommendation systems

This section will cover one of the problems of RS, namely the more specific GRS, where recommendations need to be made for a group of users instead of individual users. One of the difficulties of GRS is that the system will need to take all the group members' preferences into account when making recommendations. Often the group profile is a union of every members' individual user profile.

Group model

Kompan and Bielikova[7] describe the process of user or group modeling in three steps:

1. Data collection from various sources.
2. User model inference, process users data into higher levels.
3. Adaptation and personalization, use the constructed model to provide content.

As well as laying out two approaches for group modeling: merging single-user profiles and group profile construction. The latter of the two is not so widely used because of its shortcomings in terms of not being a dynamic approach that is suited for changing groups and it is not possible to keep personal preferences and characteristics into account when the system does not hold information about every user. Thus the merging of single-user profiles is the most widely used approach. This approach was also used in a proposed system for tv program recommendations for multiple viewers by Yu et al.[8].

Recommendation generation

Kompan and Bielikova[7] lay out three different ways of generating recommendations for a group model, one is based on a single group profile in which the systems recommends as the

group profile were to be an individual user profile, as discussed earlier this approach is not optimal.
Two other approaches are to merge the single-user profile and make recommendations for the aggregated group model or to recommend for every single-user model and merge the recommendations of every single one. The most used approach is the one based on the aggregated group model.
This means that after the aggregation of the single-user profile, recommendations can be generated as if they were a single-user profile with the techniques mentioned in the section above, collaborative, content-based or hybrid filtering.

Manipulation in group recommendation systems

The general idea of how GRS works, in terms of recommending items to a group of people, has been explained in the previous two sections. Now it is important to keep in mind that those were not the only challenges that GRS brought compared to individual RS. When recommending to a group of users some users will be more inclined to push their preferences even if this means that the overall satisfaction of the group decreases. Manipulation is a factor that has to be taken into account when designing a GRS and should be limited as much as possible if not eliminated completely.

Aggregation mechanisms

Jameson[9] presents two aggregation mechanisms that prevent manipulation. The first method he proposes is the median method, this makes sure that no group member can distort the outcome by specifying an extreme (either low or high) preference. This is what he calls a hand-crafted mechanism, which can yield suboptimal results.
For this problem Conitzer and Sandholm[10][11] introduce *automated mechanism design*. These are designs that are generated to fit a given setting, this can take various preferences into account and is non-manipulable and optimal in terms of a given objective function such as utility or equity.

User interfaces

Another way of counteracting decision manipulation in GRS is to design the interface in such a way manipulation behavior is discouraged or completely prevented. The work of Tran et al.[12] provides a possible answer to this problem, after comparing two user interfaces differing in level of transparency of the preferences of other group members, they confirmed the Hawthorne effect, which means if users know that their preferences could be seen by other users they will tend to avoid decision manipulation.
In short if users know they are being watched they will tend to avoid bad behavior, in this context manipulating the decision. This does bring the problem of privacy to the surface which could prevent the system from showing every user's preferences but this heavily depends upon the context of the group.

Conclusion

This section covered the technical background of GRS as well as some of the challenges faced in the field of Human-Computer Interaction. As groups can be much more dynamic, diverse,

and complex compared to an individual user, more challenges arrive in terms of group dynamic and different personalities of the group members. As this topic is not yet heavily researched this leaves room for future research about GRS. Another gap found in the current literature will be explained in detail in the next section and will be the topic of this thesis.

2.3 Gaps

There exists a lot of research about giving recommendations to a group of users instead of an individual user, if these recommendations are a list of items the group will need tools to make a final decision.
Isinkaye et al.[2] describe a framework for splitting up the whole process that the system/user will go through to reach a recommendation. It consists out of three phases:

1. Information collection phase
2. Learning phase
3. Recommendation phase

As mentioned earlier after receiving a list of recommendations, the group needs to decide on a final decision, this phase is not accounted for in the framework above. The absence of this phase is the gap to be filled in this thesis, let's call this phase the final decision phase.
Resulting in the new framework:

1. Information collection phase
2. Learning phase
3. Recommendation phase
4. Final decision phase

In this fourth and final phase, the system will present the list of recommended items to the group of users and will provide a set of tools based on different methods to accommodate the process of achieving a final decision that is satisfactory for all the group members.
This topic has not been completely ignored in the literature; there has been research conducted about consensus negotiation techniques but it has never been identified as a phase of GRS and the existing research does not go in-depth about these methods.
Salamó et al.[13] talk about different aggregation methods to find recommendations that maximally satisfy all members of the group but don't talk about features to accommodate this process.
Jameson[9] also advocates the need for the phase described above. He presents three different ways this problem has been solved in the past:

- The system returns the highest rated item without consent of any users.
- It is assumed one group member is responsible for making the final decision.
- It is assumed the group will arrive at the final decision through face-to-face discussion.

The methods all leave a lot to be desired and will not be optimal in a lot of different cases. This leads to the need for further research about this topic.

Chapter 3

Iteration 1

3.1 Methodology

The first iteration concerned itself with research question 1, *What are design solutions expected by users to achieve the final decision of group movie recommendations?*. In other words, the goal of this iteration was to explore the design space for mobile features that users deemed necessary and how they would design such a system to fit their needs. Not only how they would design it was important, but also why they would design it in such a way, to generalize their design solutions to different criteria that are important in the system. To achieve this, this research conducted a co-design session, a session focussed on collaboration and the ideas of the participants, which is a perfect fit considering that the goal of this study is to find out what users expect. The co-design session consisted out of three main parts. In the first part, the participants design their solution and explain it and in the second part two prototypes were presented for the participants to test, and finally, they reflected on their solution and created one final design. Two different prototypes were designed beforehand based on existing research on ways to handle the final decision in group recommendation systems.
From these proposals, the most popular design or design criteria was used to create two high fidelity prototypes to research in the second iteration.

Design rationale

Two prototypes were created to let the participants test during this iteration. The two prototypes each provide a different approach to come to a final movie to watch together starting from an initial list of recommended movies. The different approaches use different mechanisms for the users to elicit their preferences and to merge them to come to the best possible decision. These two mechanisms were picked from existing research based on three factors: simplicity, effectiveness, amount of research done. From the current literature popular mechanisms are negative elicitation[14], rating[15][16][17][18][19] and veto[20][21][22][23][24][25], these approaches are simple as well. The veto approach was left out because this would defeat the purpose of the system as then only one person would decide upon which movie to watch. Another approach, third-party[26][9][27][28][29][30] is more used in the current literature compared to the negative elicitation approach, but it is way more complex. Given these reasons, the negative elicitation and the rating approach were built into prototypes.

Negative Elicitation

With the negative elicitation approach, both users have the opportunity to iteratively remove movies they don't like from the list of recommended movies. This can lead to two different scenarios:

- A single movie is left: in this case that movie is the final decision.
- No movies are left: if all movies get removed from the list, the system will recommend new movies to the users and the cycle continues.

In the prototype these two scenarios are explored, the second scenario will be explored first and after that, the system will recommend new movies and then a single movie will be left. Below the flow of the prototype is shown with screenshots.

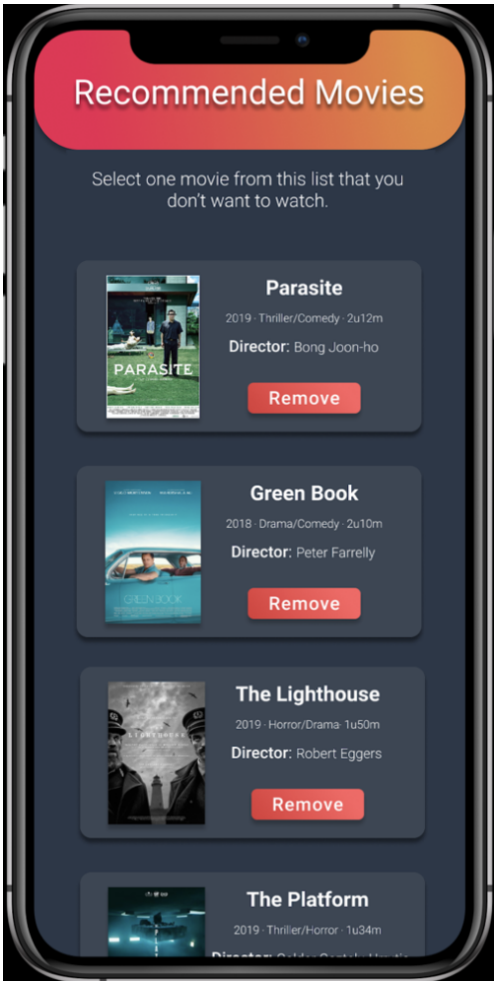

Figure 3.1: Initial list of movies

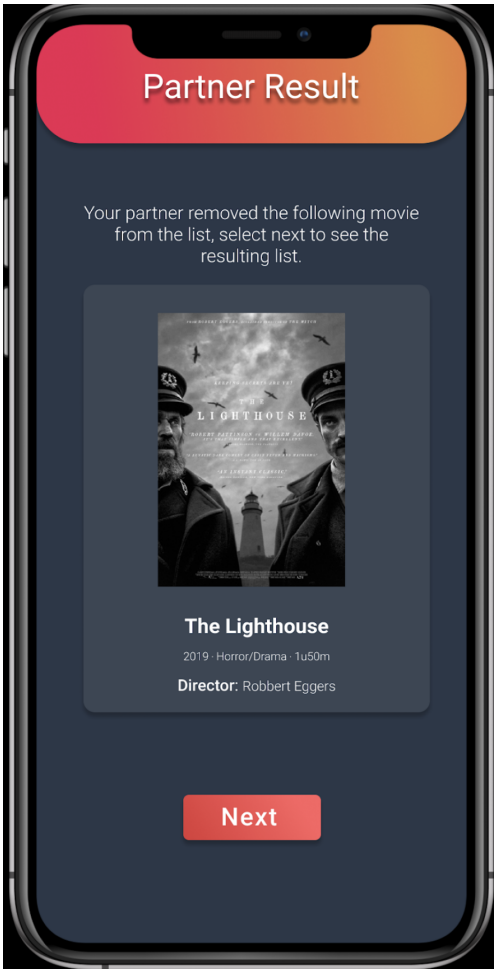

Figure 3.2: Result of partner

Users will receive an initial list of recommended movies as shown in figure 3.1 and after removing a movie, the system will show the movie that the partner has removed as in figure 3.2

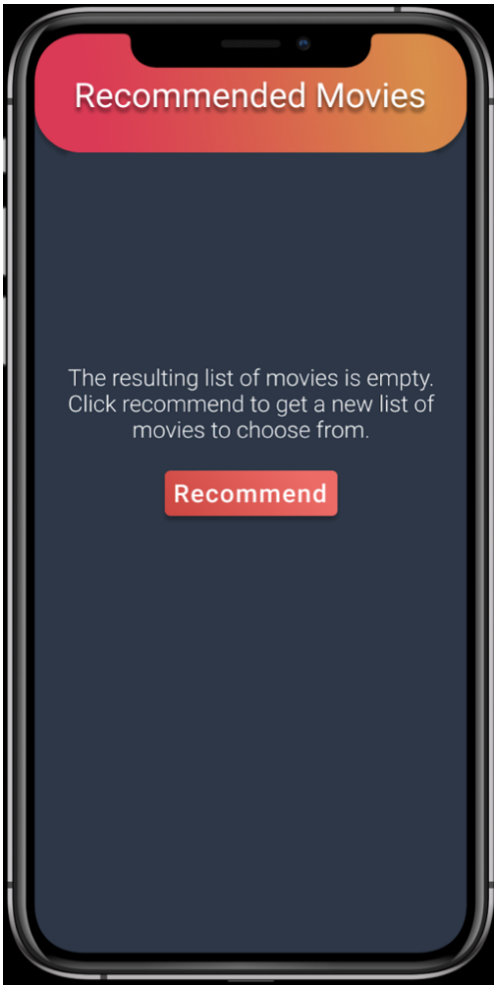

Figure 3.3: No movies left

Figure 3.4: Final decision

This process continues until there is either no movie left as showcased in figure 3.3 or there is one movie left, which then serves as the final decision as shown in figure 3.4.

Rating

With the rating approach, both users will have the opportunity to rate each movie with a score ranging from 1 to 5, the movie with the highest rating will be selected as the final decision. This can again lead to two different scenarios:

- A single movie has the highest rating: in this case, that movie is the final decision.
- More movies are the highest rating: in this case, the users can rate those movies again.

In this prototype the two scenarios are also explored similarly, after rating there will be two movies with the highest rating, then the user can rate the remaining movies again to reach a final decision. In this prototype the input rating of the user will have no impact on the flow and decision because it is a prototype, therefore the user will only have the possibility to input the rating of a single movie to feel how the interaction will go, not to impact the decision.

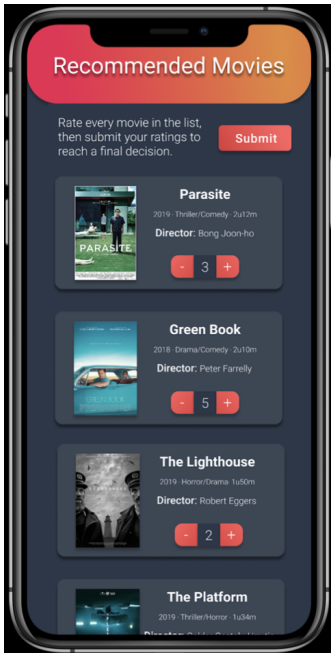

Figure 3.5: Initial list

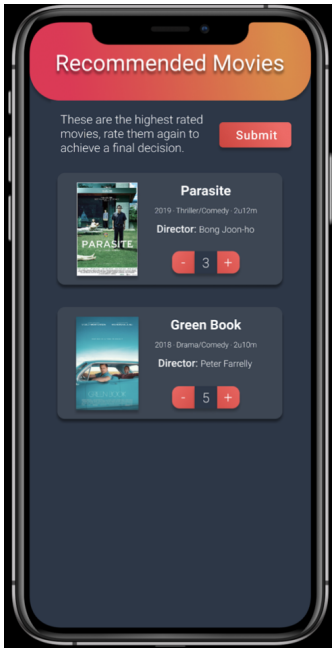

Figure 3.6: Tied ratings

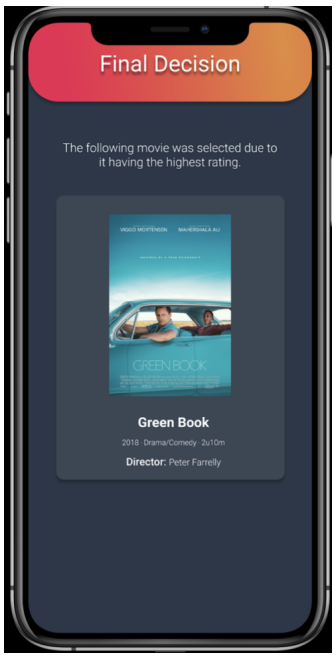

Figure 3.7: Final decision

Users will again receive an initial list of movies in this prototype, they then need to give a rating ranging from 1 to 5 for each movie in the list before submitting as shown in figure 3.5. This can lead to either a tie as shown in figure 3.6. The users then have to give a rating again to all the tied movies or it can result in a final decision, the movie with the single highest rating, as shown in figure 3.7

Methods

Because I conducted a co-design session to find out ideas, I used a qualitative data collection approach. The data was collected with audio and video capture software as well as a whiteboard tool for sketches made by the participants.

Participants

In total 12 participants took part in the co-design session, meaning 6 pairs of people. The pool of participants consisted of 8 male participants and 4 female participants, with 75% of them belonging to the 18-24 age group and 25% of them to the 45-54 age group. Half of all participants reported having an intermediate level of knowledge about technology with 33% a novice level and 17% reported a high level of knowledge. A majority of 83% reported using streaming platforms between 1 and 10 hours each week and only 17% reported more, with half of those 17% reporting 30+ hours each week.

Friends, colleagues, family, and friends of friends were recruited for this iteration. There were some criteria required from the participants:

- Users of Netflix, Amazon Prime, or any other movie streaming platform with a recommender system.
- Users with at least one year using a subscription on the platform.
- Users that commonly watch movies together.

These requirements reflect the target audience of the system, so they will be the best audience to propose design solutions as they need such a system the most.

Co-design setup

The session was split up into a sensitizing activity, to stimulate the participants to think about movie recommendations in general, and three main parts, one to design, one to test, and one to reflect. The average time in which the sessions were conducted was around 45 minutes each.

Sensitizing activity The sensitizing activity was performed by the participants before the co-design session started. They were asked to fill in a short questionnaire consisting of the following questions:

- Did you know your movie streaming platform recommends movies to you?
- Did you know these recommendations are personalized?
- Do you often watch movies the system recommended to you?
- To which extent do you think you understand why the platform recommends you those movies?
- To which extent do you think you have control over the recommended movies?

This sensitizing activity is meant to enhance the participants' ideas and creativity about the general topic of movie recommendation systems by making them think about the topic in advance of starting the co-design session. Alvarado et al.[31] argue that a sensitizing activity combines users' situated experiences and a general understanding of the presence of the hidden, more technical aspects of computing and thus preparing the participants for the design aspects of the co-design session.

Part 1 After a brief explanation of the context of the research, participants were invited to present ideas for possible design solutions based on their preferences, needs, experience, and understanding of the topic. This was done by using an online whiteboard tool where participants could sketch their ideas. During this, the participants were presented with a series of questions to spark their imagination or to go deeper into why they choose to include a certain aspect in their design. These questions were in the lines of the following:

- Given a list of recommended movies, how would you want to reach a final decision?
- Can you sketch a possible solution?
- Why would you prefer such an approach?
- What things did you keep in mind when you figured out your preferred approach?

This part of the session lasted about 15 to 20 minutes for each group.

Part 2 After the design part, the participants were introduced to the two existing lo-fi prototypes containing the rating and negative elicitation approach. Then they were invited to test and interact with the prototypes and ask questions to get a deeper understanding of how these approaches worked. To avoid bias a Latin square distribution, as proposed by James V. Bradley[32], was used so that every other group would use the prototypes in a different order as shown in Table 3.1. The reasoning behind this is that after using a prototype they are more familiar with it so they would understand the second prototype faster, which results in a bias towards the second prototype, this is mitigated by using the Latin square. After each prototype the participants were asked the following questions:

- What did you like most about this approach? Why?
- What did you like the least about this approach?
- Where do you feel this approach would not aid you in achieving a final decision? Why?
- Did the given approaches match your expectations built during the first part of this session? How would you change the proposed design? Why?
- Which one of the proposed approaches would you prefer? Why?

This part of the session also lasted 15 to 20 minutes for each group.

even group number	prototype 1	prototype 2
odd group number	prototype 2	prototype 1

Table 3.1: Latin square.

Part 3 This final part lasted around 5 to 10 minutes for most groups and consisted of a discussion about the prototypes the participants had just tested and a reflection on their designs. Participants were invited to talk about how they would compare their designs to the prototypes and if and how they would change their designs given the new information and ideas gained from using the two approaches the prototypes offer. The results of this part were the final designs for each group, in other words, their best possible solution.

Analysis

As mentioned above, the collected data was mainly audio and video footage of the conducted interviews. To analyze this, the thematic analysis, as introduced by Braun and Clarke[33][34], was used. This consisted of the following steps, transcribing the data, coding the data and forming themes from the codes to represent the main ideas that were discussed by the participants. 5 themes were formed, 4 concerning the design solution made by the participants as well as the most important criteria for the system, and one theme about general design improvements for the next iterations about the existing prototypes

3.2 Results

The data were analyzed using thematic analysis, thus the results are presented in 5 different themes each representing an important aspect that was discussed during the co-design sessions.

The relevance of control regardless of the approach

A lot of participants mentioned the need for control regardless of the used approach to achieve the final decision as their needs might vary over time.
Participants mentioned that they wanted control over the recommendations in a group movie recommender application. P9 suggested a way of taking one's current state of mind into account: *"And also maybe a way of taking your current mood into account"*. And P8 suggested a way of saying what types of movies they wanted to see so the list would be adjusted to that: *"And maybe at the beginning both of us can enter the genres of movies that we would like to watch."*. This was also mentioned by P11 and P12 but they went a step further to also enter subdivisions after entering the preferred genre to filter even more. These remarks all point to the fact that users want control even before movies are recommended regardless of the approach. This way the user feels in control and feels that the application is personalized to the users' current feelings.

Start watching a movie as quickly as possible

The efficiency of the process to find a movie to watch together was discussed a lot during the different workshops. Almost all participants wanted a fast process so they could start watching the movie.
When the participants were being asked what they liked the most/least about the two prototypes I designed, most of them said that they preferred the negative elicitation prototype over the rating one because it was more efficient. P7 said: *"I like the simplicity of it."* and P9 mentioned: *"Yes I liked this approach, it was very nice to use and went pretty quickly"*. Most participants also mentioned that the rating prototype was too complex so that they would lose interest in the application. For example, P3 said: *"Giving ratings might take too long if you want to watch a movie."*, P10 expressed: *"I think giving ratings for all the movies takes way too long because then you need to also compare it to the rest of the list."* and according to P12: *"This takes too long, I'd get put off by this"*.
Some participants did describe the rating prototype to be more *"accurate/detailed"*. But they still did not prefer it over the other one as they would trade off "accuracy" for

efficiency. P5 said: *"Giving a rating for all movies will take too long for me but it could be more accurate"*, P1 mentioned: *"I would not use this one because it's too much effort to rate the movies, but the system would be accurate."* and P3 stated: *"It's better to get a good result very fast instead of a perfect one after some time."*.
P1 did propose a design that is supposedly faster than the negative elicitation approach. In figure 3.8 the sketched design of P1 is shown, a binary choice instead of negative elicitation is proposed. So both users enter what movies they like and what movies they dislike, the movies they both liked are matched and the system will pick a random one out of the matched movies as a final decision. In this way, there should only be one step before the final decision, so it should be faster than the negative elicitation prototype that I proposed and thus making it a preferred approach.
In conclusion participants would like an approach that is efficient, one of them even proposed an approach that was not contemplated before that could even be faster.

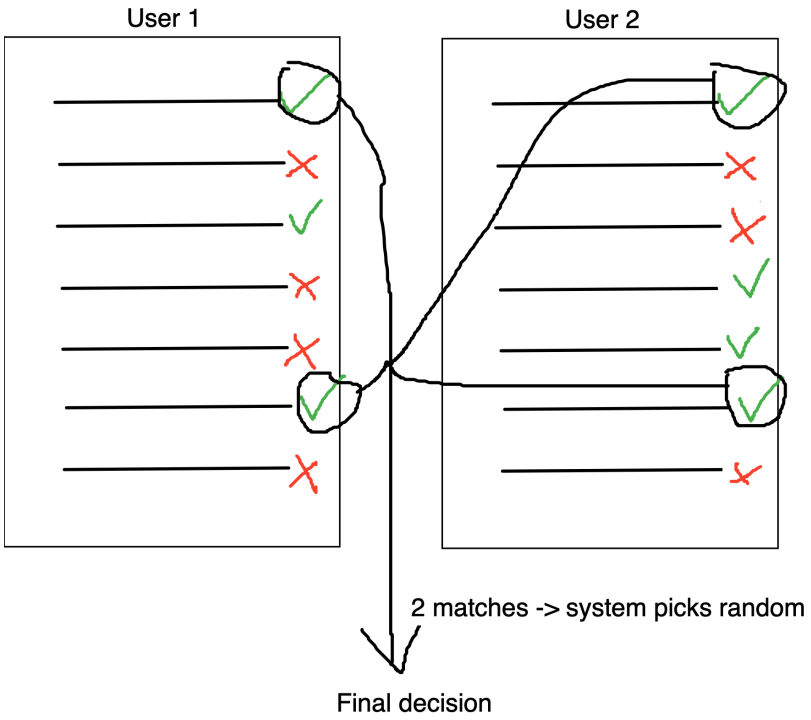

Figure 3.8: P1 design sketch

Filtering for faster preference elicitation

As shown in theme one participants wanted control over the process but in theme two it was shown that participants wanted an efficient approach above all. These two needs are contradictory, that's why some participants proposed a combination of the two.
These proposed designs also used the same binary choice mechanism as shown in theme two but they added another step for more *"detail/accuracy"*. P3 proposed the following: *"Maybe a more detailed way of filtering instead of yes/no, in the later stages, when the list has already been reduced."*. P8 said the following about their own design: *"The yes/no*

mechanism makes the list a lot shorter and afterward you can be more precise to get the best movie out of the resulting list."
P7 proposed such a mixed design, the sketch is shown in figure 3.9. The first step of the mixed design would be the binary choice (to filter movies) and the second step would be a scoring approach with the matched movies from the first step. The movie that has the highest combined score after the second step would be the final decision. By only adding one extra step to the process, the mixed design adds *"detail/accuracy"* without losing a lot of efficiency.
This idea of filtering was also prevalent in other proposed designs, but in a different way. P3, P4, P5, and P6 all wanted the system to use the characteristics of their liked/disliked movies to create a *"better"* list in case no matches were found.
In conclusion participants, rather than choosing only one approach, prefer a filtering mechanism that provides some control but still remains simple enough so they don't sacrifice efficiency.

Figure 3.9: P7 design sketch

Making a final decision over movies

Participants also considered how to achieve a final decision in the case that there is a tie between matched movies as a result of the preference elicitation step. Several different ways of handling this problem were proposed.
P3 proposed to pick a random movie when several movies both have the same highest combined score: *"And if there is a tie in scores maybe both 8/10 then the system will pick a random movie as a result and final decision."*. P6 proposed that it could be based on external information (the most popular, the highest rated on IMDb, ...).
These ways will lead to a final decision but participants also expressed the need for control even after the final decision. This was mentioned by participants in different workshops. P2 said: *"You could show a new list so you can choose again (with a button)."* and P9 mentioned something similar: *"Yes it's also good that you get a new list of movies in case the first list of movies is not that great, but maybe that should also be an option after you achieve a final decision."*.

The users will also be physically present with each other so they can solve the issue without the system providing the solution to them, this was also mentioned by some participants in different workshops, P1 said: *"Yes but this could be done between the users themselves."*, P10 mentioned: *"The users can discuss this between themselves and then choose what they want after the final decision."* and according to P11: *"And you're next to each other so you don't have to go through the whole process, you can still decide on your own."*.
Because of this the application should give control to the users so that they can use the application the way they want so that they will achieve the final decision in the best way possible for them.

Design improvements for the next iterations

Participants also mentioned parts of the design that could be improved and parts of the design that they liked, during the workshops.
A common trend among the participants was that they disliked the number of movies in the presented prototypes (six movies each). P2, P4, and P12 all suggested a longer list of movies, the latter two wanted at least ten movies and the former eight, P5 also said that six is not enough but did not suggest a number of movies. P7 and P8 both suggested a max amount of movies, they suggested a maximum of fifteen movies. And P5 said the following: *"The movies can be listed in two columns instead of one so I don't have to scroll that much to see everything."*
The overall look and feel of the prototypes was received positively among most participants, for example, P9 said: *"The prototype was nice to look at"* and P5 mentioned: *"I liked the prototype and it looked nice as well"*.
In conclusion, a balance should be found between more movies and keeping the list concise, and presenting the information in a way that does not require excessive scrolling. Because the prototypes "looked good" for most participants the design will be used in the application as well.

Chapter 4

Iteration 2

4.1 Methodology

In the second iteration, the user experience of the recommendation system was explored. More specific the aspects of user experience that applied the most to the system.
During the study, the participants were asked to use and test each hi-fi prototype and to fill in a questionnaire after using each prototype. Afterward, the participants were be asked a series of qualitative questions on why they made certain choices and to explore possible design improvements for the next iteration.
Based on these results the prototype for the third and final iteration was created.

Design rationale

Two prototypes were created to let the participants test during this iteration. The two prototypes each provide a different approach to come to a final movie to watch together starting from an initial list of recommended movies (powered by TMDb[35]). The different approaches use different mechanisms for the users to elicit their preferences and to merge them to come to the best possible decision. In this iteration, the two approaches used were picked from the results of the previous iteration.
The co-design session resulted in a final design from the participants. A layered approach filters the initial, longer, list of movies with a fast mechanism and uses a more in-depth and complex mechanism to find a final decision from the, shorter, filtered list.
During the previous iteration participants also tested two existing prototypes of which the negative elicitation approach was the clear favorite due to it being a faster and less complex way of finding a final decision.
This is why these two mechanisms were chosen in this iteration to explore which of the two has the better user experience.

Negative elicitation

The negative elicitation mechanism has been updated from the previous iteration. In this prototype, both participants have to remove all movies they don't want to watch from the initial list of movies. The final decision is now a list of movies. The list consists of all movies that were not removed by at least one participant. From this list, the participants can choose

a final decision themselves. In case there is no movie left, the system will generate a new list. The option to generate a new list if the list is not empty is also given to provide more control.

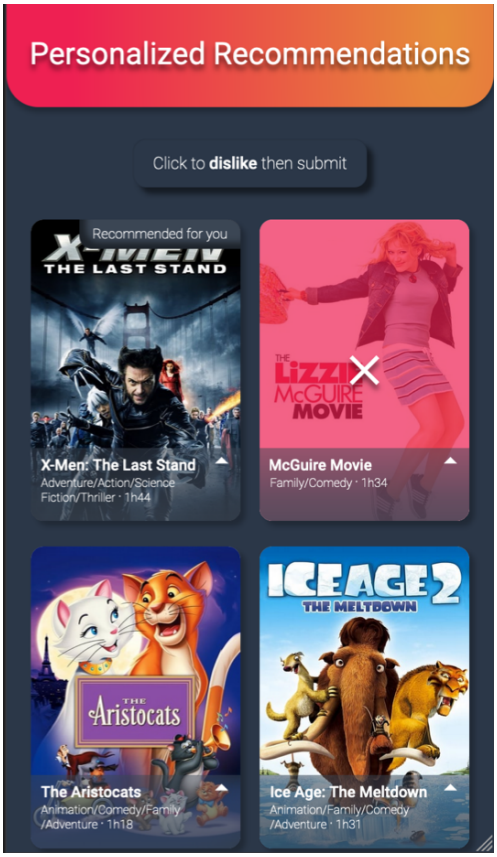

Figure 4.1: Initial list

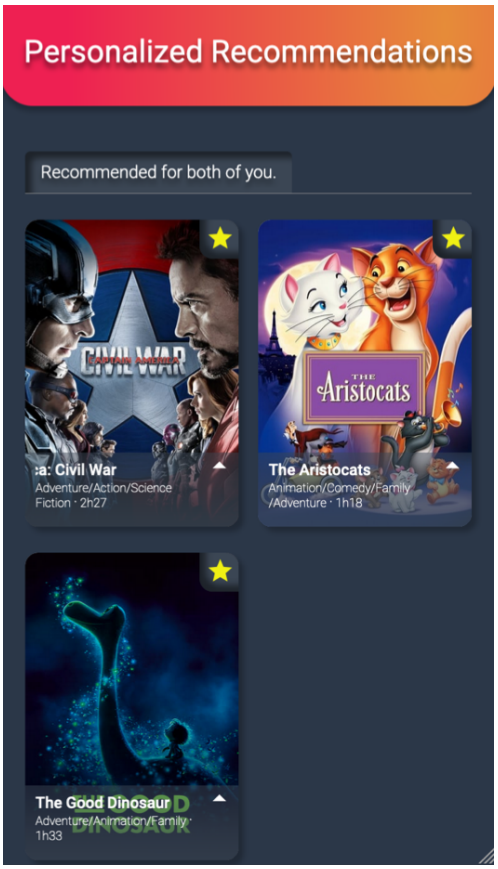

Figure 4.2: Final decision

The system will present an initial list of movies, each user can then dislike the movies that they don't want to watch by clicking on a movie as shown in figure 4.1. When both users have submitted their disliked movies the system will present their list of final decisions as presented in figure 4.2, this list contains all movies that were disliked by neither of them.

Layered approach

The layered approach uses a fast mechanism to quickly reduce the initial list of movies to a shorter list and then continues with a more complex mechanism to come to a more detailed final decision from the shorter list.
In this prototype the fast mechanism used is positive elicitation. In this mechanism, the participants have to select all the movies that they do want to watch. The resulting list consists of all the movies that both participants have liked.
The more complex mechanism is a rating approach, in which all participants have to give a score ranging from 0 to 3. The option to score movies that were liked by only one participant is included but is not mandatory, this leaves room for compromise.
After the rating step, the system will present the participants with the final decision list. In the layered approach, this list will consist of all movies with the highest combined rating. This list will never be empty so the system will not have to generate a new list, but again this option is presented to provide more control.

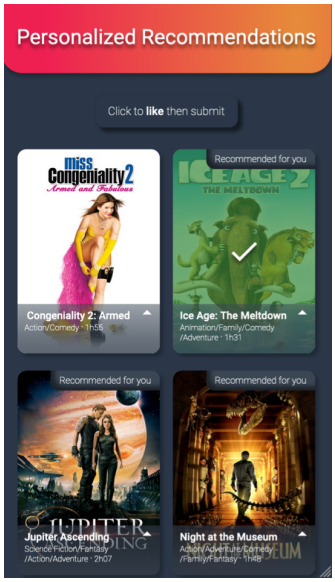

Figure 4.3: Initial list

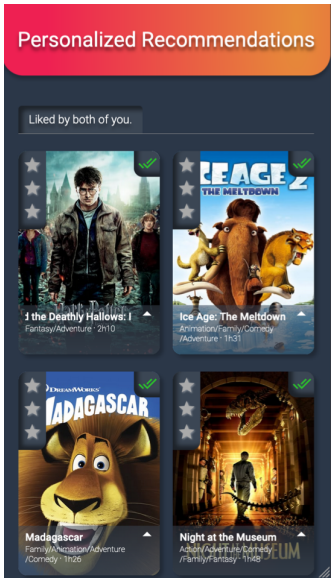

Figure 4.4: Rating step

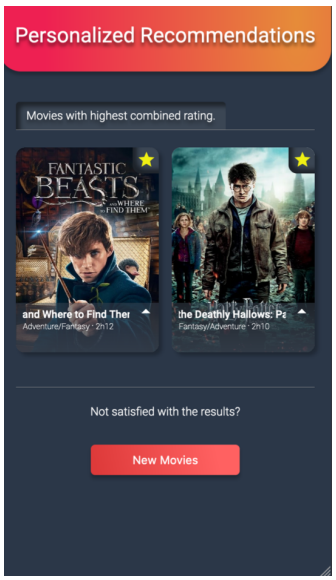

Figure 4.5: Final decision

The system will provide the users with an initial list of recommended movies, the users have to click on the movies that they want to watch and then submit as shown in figure 4.3. Afterward, the system will show the movies that were liked by both users as in figure 4.4. Below that list, the system also shows the movies like by only one user, to leave room for compromise. Finally, the system will present the final decision list as shown in figure 4.5, this list contains the movies with the highest combined score.

Methods

Data was collected in a mixed manner. The main data about the user experience of the system was collected using a big part of the ResQue[36] questionnaire, a User-Centric evaluation framework for recommender systems. The following constructs were included, resulting in a questionnaire of 16 questions:

- Interaction adequacy
- Perceived ease of use
- Control
- Transparency
- Perceived usefulness
- Overall satisfaction
- Use intentions

This data is qualitative data on a 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5). At the end of the study, some quantitative questions were also asked to the participants to get a deeper understanding of their opinions and leave room for design improvement suggestions. This data was collected using audio recording software.

Participants

In total 40 participants took part in the co-design session, meaning 20 pairs of people. The pool of participants consisted of 19 male participants and 21 female participants, with 72.5% of them belonging to the 18-24 age group, 10% of them to the 25-34 age group, 12.5% was between 45-54 years old, and 5% between 55-64. 42.5% of all participants reported having an intermediate level of knowledge about technology with 32.5% a novice level and 25% reported a high level of knowledge. A majority of 95% reported using streaming platforms between 1 and 10 hours each week and only 5% reported more between 11 and 20 hours each week.
Friends, colleagues, family, and friends of friends were recruited for this iteration. All participants were not yet recruited before as this would influence how they would perceive the user experience of the system as they would be familiar with it and thus understand it quicker. The same criteria as in iteration 1 were required from the participants:

- Users of Netflix, Amazon Prime, or any other movie streaming platform with a recommendation system.
- Users with at least one year using a subscription on the platform.
- Users that commonly watch movies together.

Study setup

The session was split up into three main parts and lasted around 1 hour in total. The first two parts each consisted of the participants using the prototypes and filling in the questionnaire afterward. And during the final part, the participants will be presented with a couple of questions about why they made certain choices.

Prototype and questionnaire The two first parts were almost identical, they both consisted of the participants using the system for around 15 minutes and afterward filling in the questionnaire. Which can be found in appendix A, this took around 10 minutes on average. The only difference between the two parts is that participants would test a different prototype each time. For this a Latin cross distribution was used again, like in iteration 1, to avoid bias.

Final part Afterward, the participants were asked a couple of questions. These questions were asked to further explore their reasoning on certain choices they made in the questionnaire, as well as to explore possible design improvements for the next iterations. The following questions were asked:

- What is your favourite prototype and why?
- Which are the best features about each prototype and why?
- Is there anything that you would change about a certain prototype?

This final part of the study lasted anywhere from 5 to 15 minutes depending on the group and on how extended their answers to these questions were.

Analysis

The collected data in this iteration was of a mixed nature, it consisted mainly of quantitative data but also some quantitative data. The quantitative data were analyzed using thematic analysis and were also used to strengthen patterns found in the qualitative data.
The results of the questionnaires were analyzed using hypothesis testing split up into each construct of the ResQue questionnaire that was used. The null hypothesis for this analysis stated that there was no significant difference in user experience between the two prototypes tested. When there was a significant difference and the null hypothesis could be rejected, descriptive data were used to explore which prototype was favored.

4.2 Results

As mentioned above, the data was mainly quantitative mixed with a few qualitative questions. This section will first handle the results of the questionnaires and afterward introduce some themes that were the result of the thematic analysis.

Quantitative analysis

To investigate the second research question, two prototypes were compared and tested with a large part of the ResQue questionnaire for user experience of the recommendation system. The goal was to find aspects of the recommender that were statistically different in the two prototypes. In other words, to disprove the null hypothesis being: $p1 = p2$.
To analyze if this was the case, the results of each question for both prototypes were compared using a paired samples t-test with the Wilcoxon hypothesis test, for the hypothesis $p1 \neq p2$, so that the null hypothesis could be rejected. The results are displayed in figure 4.6.

Paired Samples T-Test

			Statistic	p
1.1	2.1	Wilcoxon W	30.0 [a]	0.020
1.2	2.2	Wilcoxon W	45.0 [a]	0.116
1.3	2.3	Wilcoxon W	95.5 [b]	0.712
1.4	2.4	Wilcoxon W	50.0 [d]	0.378
1.5	2.5	Wilcoxon W	81.0 [a]	0.842
1.6	2.6	Wilcoxon W	46.0 [b]	0.018
1.7	2.7	Wilcoxon W	95.0 [b]	0.697
1.8	2.8	Wilcoxon W	88.0 [e]	0.032
1.9	2.9	Wilcoxon W	54.0 [f]	0.023
1.10	2.10	Wilcoxon W	27.0 [g]	0.003
1.11	2.11	Wilcoxon W	57.0 [h]	0.165
1.12	2.12	Wilcoxon W	75.0 [g]	0.404
1.13	2.13	Wilcoxon W	45.0 [h]	0.052
1.14	2.14	Wilcoxon W	34.0 [a]	0.027
1.15	2.15	Wilcoxon W	36.0 [h]	0.017
1.16	2.16	Wilcoxon W	59.0 [a]	0.395

Figure 4.6: Results of all constructs

To reject the null hypothesis a p-value of 0.05 or lower is needed, meaning that the two prototypes don't perform the same on that specific aspect with at least 95% confidence. As seen in the results above, certain aspects do in fact have a p-value of 0.05 or lower meaning that the two prototypes don't perform the same in all aspects.
To now conclude which of the two prototypes performs better than the other one, a more detailed analysis and a deeper look into the data are needed. For this, the questions will be grouped by the different constructs of the ResQue questionnaire that were used in the test.

Interaction adequacy It seems that the extra rating step in prototype 2 allows users to better tell the recommender what they (dis)like. This can be shown by the fact that the null hypothesis can be rejected for question 1 in figure 4.7 and the median, shown in figure 4.8 of the second prototype is the highest. This idea is also supported by the qualitative data as P22 said; *"Scoring is nice to give more depth to a movie that you've liked."* and P23 mentioned: *"With prototype 2 you can more specifically say what movie you want."*.

Paired Samples T-Test

			Statistic	p
1.1	2.1	Wilcoxon W	30.0 [a]	0.020
1.2	2.2	Wilcoxon W	45.0 [a]	0.116
1.3	2.3	Wilcoxon W	95.5 [b]	0.712

Figure 4.7: Q1-3 t-test results

Descriptives

	1.1	2.1
N	40	40
Median	4.00	5.00
Minimum	2	4
Maximum	5	5

Figure 4.8: Q1 descriptives

Question 2 and 3 don't show a significant difference according to figure 4.7. This can be explained by the fact that the way that a user tells the system what (s)he (dis)likes is more or less the same for both prototypes. When looking at the descriptive data in figure 4.9, the high median scores for questions 2 and 3 indicate that users liked this way of telling the system what they liked.

Descriptives

	1.2	2.2	1.3	2.3
N	40	40	40	40
Median	5.00	5.00	5.00	5.00
Minimum	2	3	3	2
Maximum	5	5	5	5

Figure 4.9: Q2,3 descriptives

In summary these results show that for interaction adequacy the second prototype performs better because of the extra step in the process and participants like the current way of eliciting their preferences.

Perceived ease of use In regards to perceived ease of use, no significant differences were found between the two prototypes, as illustrated in figure 4.10 below. This can be explained by the fact that both interfaces were almost identical except for prototype-specific design features.

Paired Samples T-Test

			Statistic	p
1.4	2.4	Wilcoxon W	50.0 [a]	0.378
1.5	2.5	Wilcoxon W	81.0 [b]	0.842

Figure 4.10: Q4,5 t-test results

The results, as shown in figure 4.11, indicate that the majority of the participants found that both prototypes were easy to use, this is indicated by the median scores being 5 for both prototypes. This view was also echoed by P26 who said that: *"Both are pretty similar to use in my opinion."*.

Descriptives

	1.4	2.4	1.5	2.5
N	40	40	40	40
Median	5.00	5.00	5.00	5.00
Minimum	3	3	2	1
Maximum	5	5	5	5

Figure 4.11: Q4,5 descriptives

Figure 4.11 does show a minimum score of 1 for 2.5, but from figure 4.12, it is clear that that low minimum in 2.5 is merely an outlier and does not represent a large part of the participants.

Frequencies of 2.5

Levels	Counts	% of Total	Cumulative %
1	1	2.5%	2.5%
2	1	2.5%	5.0%
3	2	5.0%	10.0%
4	12	30.0%	40.0%
5	24	60.0%	100.0%

Figure 4.12: Q5 P2 frequencies

Overall, these results seem somewhat counter-intuitive as the second, more complex, prototype was not perceived harder to use by the participants. But this might be in favor of the second prototype as it was not perceived harder to use but does provide more functionality.

Control It seems that the majority of the users felt that they have more control over the recommendation process and their taste profile with the second prototype due to the layered approach.
Figure 4.13 below shows a significant difference between the two prototypes for question 6 and 8. But the median scores, as shown in figure 4.14, are equal for both questions. So this can't be used to conclude which one is better.

Paired Samples T-Test

			Statistic	p
1.6	2.6	Wilcoxon W	46.0 [a]	0.018
1.7	2.7	Wilcoxon W	95.0 [a]	0.697
1.8	2.8	Wilcoxon W	88.0 [b]	0.032

Figure 4.13: Q6-8 t-test results

Descriptives

	1.6	2.6	1.8	2.8
N	40	40	40	40
Median	4.00	4.00	4.00	4.00
Minimum	2	2	1	3
Maximum	5	5	5	5

Figure 4.14: Q6-8 descriptives

Figure 4.15 and figure 4.16 below shows that for the second prototype the scores are more centered around the higher scores compared to the first prototype, indicating a tendency towards the second prototype.

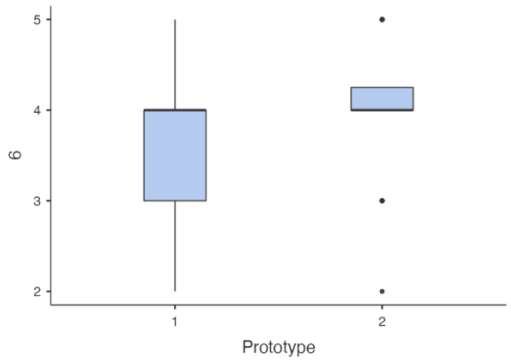

Figure 4.15: Q6 boxplot

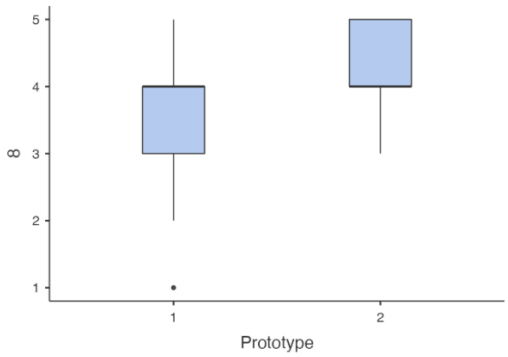

Figure 4.16: Q8 boxplot

Participants also mentioned that the second prototype allowed giving more detail about what movie they wanted to watch. P10 said: *"Variation in how much you like a movie is better."* and P19 mentioned that: *"Saying what you like is easier with scoring."*.
These results suggest that the majority of the participants prefer a layered approach, as implemented in prototype 2, as they felt that it gave them more control over their taste profile in the recommendation process.

Transparency In terms of transparency, it seems that most participants were in favor of the second prototype. This can be evidenced by the significant difference between the two prototypes as shown in figure 4.17 below and by the fact that the second prototype has a higher median and minimum score as presented in figure 4.18.

Paired Samples T-Test

			Statistic	p
1.9	2.9	Wilcoxon W	54.0 [a]	0.023

Figure 4.17: Q9 t-test results

Descriptives

	1.9	2.9
N	40	40
Median	4.00	5.00
Minimum	2	3
Maximum	5	5

Figure 4.18: Q9 descriptives

As the way handling transparency for the initial list of movies was the same for both prototypes, this result might be because the extra step gives more insight into the process of achieving a final decision as mentioned by P2 who said that: *"The final movies in prototype 2 are more clear for me."*.
This was not mentioned by the participants explicitly during the study but in general, they were quicker to understand their final recommendations with the second prototype.
In general it seems that participants prefer a system with a layered approach that offers more insight on how the final recommendations are achieved, like prototype 2.

Perceived usefulness In terms of perceived usefulness, it seems that the second prototype better helped the participants find the ideal item. Figure 4.19 below shows that the null hypothesis can be rejected for this question. And 4.20 shows that the favor is toward the second prototype with the mean and minimum scores being higher.

Paired Samples T-Test

			Statistic	p
1.10	2.10	Wilcoxon W	27.0 [a]	0.003
1.11	2.11	Wilcoxon W	57.0 [b]	0.165
1.12	2.12	Wilcoxon W	75.0 [a]	0.404

Figure 4.19: Q10-12 t-test results

Descriptives

	1.10	2.10
N	40	40
Median	4.00	5.00
Minimum	3	4
Maximum	5	5

Figure 4.20: Q10-12 descriptives

These results are also reflected in the qualitative data as P1 said: *"The second prototype gives better final decisions."*, P9 mentioned: *"Prototype gives the ultimate choice."* and P13 said that: *"Because of the scoring step you get closer to each other's taste."*.
For questions 11 and 12 the null hypothesis can't be rejected, so there is no significant difference between the two prototypes. For question 11 this is also prevalent in the qualitative data as some participants mentioned that they found liking more intuitive and others found disliking movies easier. And the results of the 12th question were to be expected because both recommenders gave the same suggestions.
These results show that the majority of the participants perceived the second prototype more useful in terms of finding the ideal movie to watch. As that is the main goal of the application, this forms a strong argument for prototype 2 being the better one.

Overall satisfaction It seems that most participants were overall more satisfied with the second prototype type than the first one. Although figure 4.21 does not show a significant difference between the two prototypes it does show a strong tendency because of the p-value being only slightly higher than 0.05.

Paired Samples T-Test

			Statistic	p
1.13	2.13	Wilcoxon W	45.0 [a]	0.052

Figure 4.21: Q13 t-test results

Figure 4.22 below indicates that the median for prototype 2 is higher as well as that the answers are more centered around the higher scores compared to the first prototype. This suggests that the majority of the users preferred the second prototype.

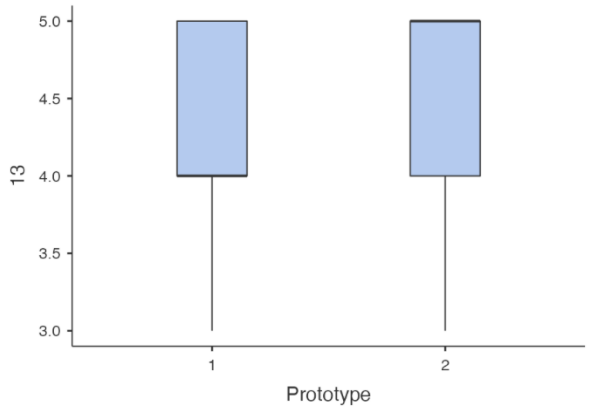

Figure 4.22: Q13 boxplot

The results of the qualitative data show the same tendency towards the second prototype. At the end of the study, the participants were asked which prototype they liked best and after analyzing the answers, 73% of the participants liked the second prototype better. Together these results provide arguments for the second prototype providing a better overall satisfaction for the participants even though the null hypothesis could not be rejected.

Use intentions It seems that the majority of the participants have the intention to use the second prototype again instead of the first prototype.
Figure 4.23 below shows a significant difference between the two prototypes for questions 14 and 15. However figure 4.24 does not show a difference in median scores for the two questions, so this cannot be used to decide which prototype is better.

Paired Samples T-Test

			Statistic	p
1.14	2.14	Wilcoxon W	34.0 [a]	0.027
1.15	2.15	Wilcoxon W	36.0 [b]	0.017
1.16	2.16	Wilcoxon W	59.0 [a]	0.395

Figure 4.23: Q14-16 t-test results

Descriptives

	1.14	2.14	1.15	2.15
N	40	40	40	40
Median	4.00	4.00	4.00	4.00
Minimum	3	3	2	3
Maximum	5	5	5	5

Figure 4.24: Q14-16 descriptives

A deeper look into the data in figure 4.25 below shows that for both questions the density is higher around a score of five for the second prototype showing a tendency towards the second prototype. The fact that 73% of the participants liked the second prototype better also adds to the tendency of the second prototype being better.
These results suggest that most participants will use the second prototype again and also frequently. If this is the case then it adds to the argument that in general, the second prototype is better in terms of user experience for the recommender.

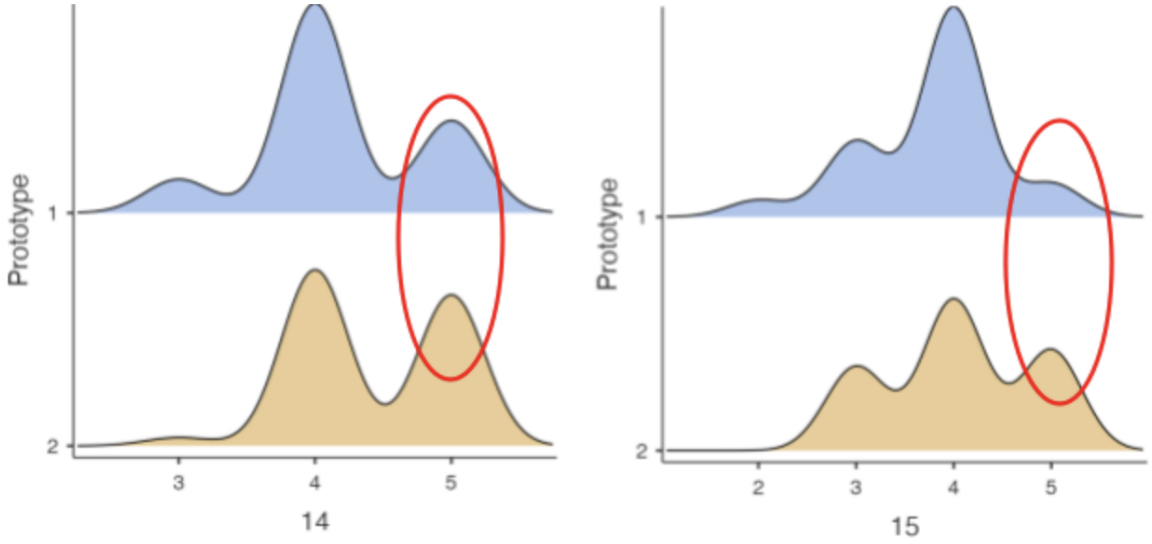

Figure 4.25: Q13 boxplot

Thematic analysis

More information about the movies Some participants expressed the need for more information about the movies so it will get easier to give an opinion about a movie that the participant had not watched yet. This could be done with trailers as expressed by P24: *"Having a trailer to watch will give more info about the movies."* and P18 said: *"I would like to watch a trailer of the movies."*. Or this could also be done with some more information beyond the title, genre, and description that was shown already. P19 said: *"I'd like some more information such as actors and directors about the movies."*.

Limited final list of movies The final list of recommended movies in the application now is not limited and can result in a long list of movies (6 for example) to choose from. This didn't make the choice a lot easier in the end and was something that participants mentioned as well during the study. Some participants only wanted one movie at the end, P39 said: *"One movie at the end would be better, so there is no discussion possible."* and P20 said: *"One movie at the end is less overwhelming."*. Some participants proposed a limited list of 2 or 3 movies, P29 said: *"I would like two final movies so there is still another option but not too much choice."* and P8 said: *"I like the choice at the end but not too much, maybe three movies."*. As pleasing every single user might be impossible, limiting the number of final recommendations seems to be a smart idea so the initial goal of making it easier to pick a movie is not contradicted. A good amount might be 2 or 3 movies but there is definitely room and opportunity for further research.

Chapter 5

Iteration 3

5.1 Methodology

In this last iteration manipulation behavior in the layered approach was explored. The goal of this iteration was to find whether or not the layered approach enabled, or even encouraged, manipulation behavior, as Tran et al.[12] confirmed the *Hawthorne effect*, meaning that if users know that others can see their actions, they will tend to manipulate less. This might not be the case for the layered approach as the manipulation that could occur in the second step is not as transparent as the first step.
This goal was achieved using a two-part quantitative study. During the first part, participants were invited to use and test the prototype multiple times until they felt that they understood the system well. Afterward, they were presented with a series of questions about manipulation in the layered approach, to find out if, why, how, and when they manipulated as well as their opinions on manipulation behavior in relation to the current system prototype. From these results the final suggestions were made on how to design a group movie recommendation system for the proposed final decision phase.

Design rationale

For this final iteration, no new prototype was created. From the results of the previous iteration, the layered approach was deemed superior in terms of user experience by the participants while also providing more control. That's the reason why the layered approach prototype is the one that was tested in this iteration. The main problem this system still faced was manipulation behavior, this is why this was explored in this iteration.
Because part of the same participant pool from the second iteration also tested no modifications were made in the prototype. Also no design improvements were yet taken into account because this could have introduced another learning curve for the participants which was best avoided as this iteration focused on manipulation behavior, behavior that is mainly exhibited when an understanding of the system and it's inner workings is in place.

Methods

Because this study was a quantitative one, all data was collected using audio and video recordings, to record the ideas of the participants and all topics discussed. A whiteboard tool

was also offered to the participants to sketch ideas or changes they wanted to make to the system, however, this was not used by any participant.

Participants

In total 10 participants took part in the study, meaning 5 pairs of people. The pool of participants consisted of 5 male participants and 5 female participants, with 80% of them belonging to the 18-24 age group and 20% of them to the 25-34 age group. Half of all participants reported having an intermediate level of knowledge about technology and the other half reported a high level of knowledge. A majority of 90% reported using streaming platforms between 1 and 10 hours each week and only 10% reported between 11 and 20 hours each week.
A random pool of participants that also took part in the previous study was selected. This was done because the participants needed to have a good knowledge and understanding of the workings of the layered approach so that they would have the ability to manipulate more.

Study setup

The study was split up into two parts. During the first part the participants were asked to use and explore the system and in the second part, they were asked a series of questions about the topic of manipulation behavior in the layered approach. Each session lasted around 30 minutes. No sensitizing activity was included as each participant had already used the prototype in the previous iteration.

Part 1 After a brief explanation of the context of the research, participants were presented with the prototype and instructed to use and explore the prototype thoroughly until they felt they had a deep understanding of the system. The participants were invited to talk during this process and share ideas in the form of a think-aloud study. This part of the session lasted about 10 minutes for each group.

Part 2 Following the exploration part, the participants were presented with a series of questions on the topic of manipulation behavior in the layered approach. These questions were posed to explore if, why, and how they manipulated the final decision while using the system. The following questions were asked:

- Did someone (try to) manipulate the final decision during the process? Why/why not?
- Did you notice that your partner manipulated the result? (or tried to
- Describe your reaction to this. Did you try to manipulate as well because of it?
- In what ways did you try to manipulate the final decision? In what step?
- In what ways did the system encourage/enable manipulation?
- Would you want the system to prevent this behaviour? Why/ why not? If yes, how?

This part of the study lasted about 15-20 minutes on average.

Analysis

As mentioned above, the collected data consisted of audio footage of the conducted interviews. To analyze this, thematic analysis was used. This consisted of the following steps, transcribing the data, coding the data and forming themes from the codes to represent the main ideas that were discussed by the participants. Four themes were formed, three concerning itself with the topic of manipulation behavior from the standpoint of the participants and one theme about general design improvements for the current system.

5.2 Results

The data were analyzed using thematic analysis, thus the results are presented in four different themes each representing an important aspect that was discussed during the conducted studies.

Noticed manipulation results in negative reactions

As explored before, the layered approach does make it possible to manipulate during the process, and some participants mentioned not only that this is noticeable but despite that participants felt that the system should not solve it, manipulation could lead to negative reactions.
For example P4 and P5 both mentioned that the second step is the most obvious for noticing manipulation as the systems shows who liked which movies. But by using the application frequently and understanding the system it also becomes clear in the final decision as mentioned by P1: *"I noticed that my partner manipulated the result because I understand how the system works."* and by P8: *"In the final step, it could be noticeable if no or few movies you liked are in there."*.
Both P3 and P8 said that if they noticed manipulation, that it would lead to negative reactions, P3 mentioned: *"I would be quite irritated and I would say to start over again but honestly this time. And if my partner would keep doing it, I would also start doing it to counteract it."* and P8 noted that: *"I would stop using the app and get frustrated I think, or in some cases, I would also start manipulating to counteract it out of frustration as well."*.
As the system is not too complex to understand after frequent use, manipulation becomes obvious to the users and this could lead to unwanted reactions. Some participants even mentioned that they would also start manipulating to counteract it, which further defeats the purpose of the application.

Possible manipulation is accepted

As expected the layered approach does enable manipulation behavior despite it also leading to a better final decision in general, as explained in the second iteration. This was discussed by most participants, but no obvious solution was proposed, nor needed in the eyes of the participants for multiple reasons.
P1 directly mentioned this fact: *"The system does enable us to manipulate, but I don't see how the system could counteract this, I think it's a necessary evil."*, P4 also did not see an obvious solution to this problem as P4 said that: *"I don't see how the app can solve it."*.

Even if the app could solve manipulation, no participant deemed it necessary, for multiple reasons. Both P2 and P8 deemed control more important than a possible solution to manipulation, as taking away control would be too restrictive. A lot of participants also discussed that the main reason for manipulation is personality, not the system itself. For example, P2 said: *"I think it's mainly the personality of the users that will be the reason for manipulation, not the system itself."* and P7 said: *"If the person wants their way, it's their fault. Not the systems' fault."*. Finally, some participants also said that the users should use the app with the mindset to find a consensus, P7: *"But I think if you use the app, you want to find a movie for both users and you won't manipulate because that is not the goal of the application."* and P4: *"I think if you use the app, you should be honest. Otherwise, it defeats the purpose of the app."*.
In conclusion even though the system enables manipulation, all participants deemed this as a problem not to be solved by the system itself, as it is mainly the personality and the mindset of the users that cause manipulation, not the system itself.

Manipulation is subjective

Only one out of the five groups interviewed did manipulate once during the process, meaning that eight participants did not. Their behavior was not deemed as manipulative by themselves but did resemble it.
When asked if they manipulated, participants mentioned that they did exhibit behavior that could be classified as manipulation, but not to manipulate the final decision, like P7 who said: *"Maybe in small amounts, but for the sake of eliciting our preferences, not trying to manipulate by doing it."*. Other participants also mentioned that they did not take their partners' tastes into account and picked movies quite selfishly, but with the expectation that if they both did that, the system would then find a movie that they both liked, which of course is the goal of the application. P5 mentioned that: *"I just tried to tell the system what movies I liked without thinking about my partner's taste and expected the system to find the best movie for us based on this."* and P4 said that: *"I gave high scores for my movies that I wanted to watch and expected the system to find a movie that we both would like based on this."*.
In summary, what gets classified as manipulation behavior is a subjective matter and seems to be less about the actions in the system themselves and more about the intent of the users behind those actions. And as discussed before, participants think that the intent behind using this app should be to find the best final decision possible, minimizing the need to find a solution to manipulation behavior.

Design improvements

In all studies done two main design improvements were suggested: a longer list of initial movies and the possibility to elicit genre preference before getting the initial list of movies. A longer list of movies was preferred because users want more options like P1 said: *"Longer list of movies would be nice for more options."*. And users did not think it would make the process too time-consuming as P4 mentioned: *"I would not mind a longer list, because you get through it pretty quickly."*. P2 even mentioned that a longer list would result in less chance of giving extreme scores and thus reducing manipulation.
Eliciting genre preference before getting the initial list was also suggested to get a broader

list of movies in the genre that they were in the mood to watch, as mentioned by P5: *"It would be nice if we could select genres at the beginning so the list of movies is broader."*. And P3 and P8 both mentioned that they would like this feature in general.
In conclusion users wants to have a longer list, in general, 16-20 movies, with an option to choose genres because they want a broader list of movies in a specific genre.

Chapter 6

Discussion

Every year streaming platforms have more users. Most of these platforms also offer movie recommendations, but they are mainly focused on individual users. There exist a set of techniques to recommend movies to groups of users, but not to aid users to reduce this list to a final decision. This research proposes a solution to achieve a final decision for groups.

6.1 Final decision phase

This thesis proposes a new phase called the final decision phase to be added to the current framework by Isinkaye et al.[2] that describes the process of GRS. The last phase of the current framework is the recommendation of items. In this new phase, groups can reduce the initial list of movies to a final decision. The final decision, in this case, is a list of 1 to 3 movies that satisfy both users, in other words, a concise list of movies that both users want to watch.

The goal of this phase is to speed up and improve the process of deciding which movie to watch out of the initial list of recommended movies.

This research also provides design guidelines to follow to implement this phase in a group recommendation system.

6.2 Design guidelines

From the results of the research, this thesis concludes that a design for this phase should respect three main criteria.

Fast process

First of all, the design should provide a fast process as discussed in section 3.2, users don't want to waste much time using the application as they expect it to speed up the regular process of choosing a movie to watch together. The goal of the application should always be to enhance the movie-watching experience, not to make it more lengthy and complex.

Control over final decision

Second, the design needs to give the users enough control such that they feel like they have a real impact on the process and on the final decision as mentioned in section 3.2. But this should not make the system too complex as this will make the process take too long and discourage users from using the application. The design should aim to strike a balance between speed and control.

Leave room for discussion

And finally, the design should keep in mind that users are next to each other and they want to discuss and communicate about the movies and the final decision as discussed in section 3.2. Therefore the system should not provide a hard final decision, but it should leave room for discussion over the final discussion, this can be achieved by presenting more than a single movie as a final decision and let the users start over when they are not satisfied with the result(s).

6.3 MovieNight as example

An example implementation of this phase that follows the proposed design guidelines is the app that this thesis presents, MovieNight. A mobile application that lets users achieve a final decision, each using their mobile device. MovieNight achieves this by using a layered approach, that uses a fast mechanism (positive elicitation) to filter the initial list of movies, to reduce the number of movies in the list. Followed by a more complex and time-consuming mechanism (rating from 0 - 3) to add control over the process and the final decision.

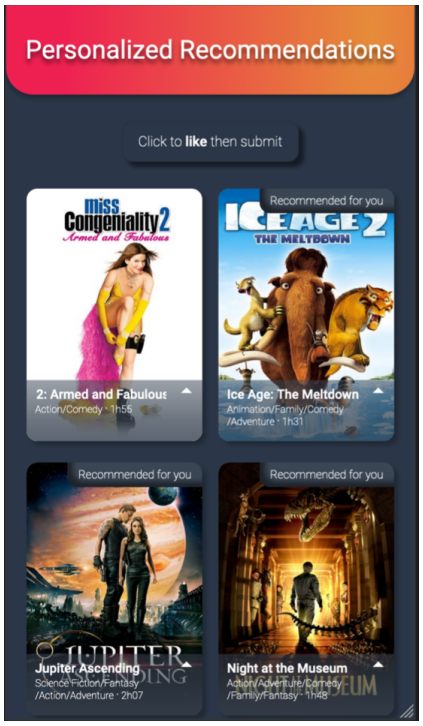

Figure 6.1: Initial list

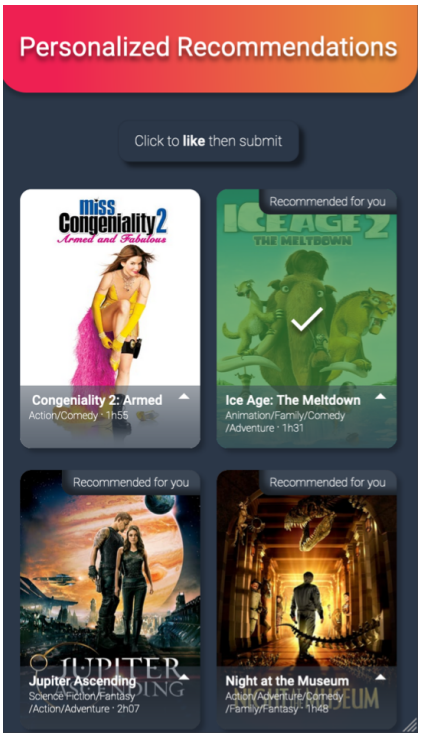

Figure 6.2: Movie liked

The initial list of movies is shown in figure 6.1. The users filter this list by liking the movies they want to watch (by clicking on the movies). Figure 6.2 shows that a movie is liked.

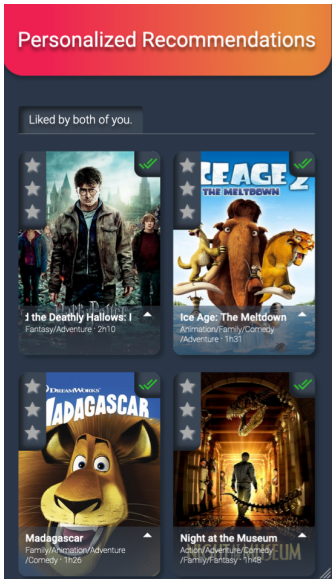

Figure 6.3: Filtered list

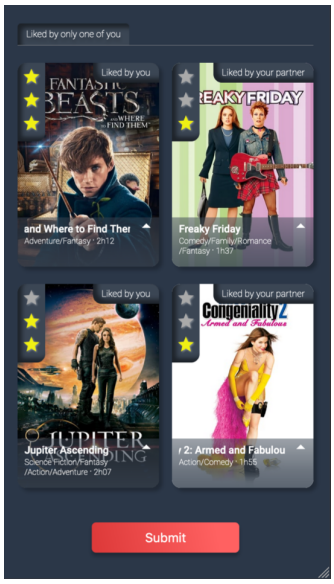

Figure 6.4: List below

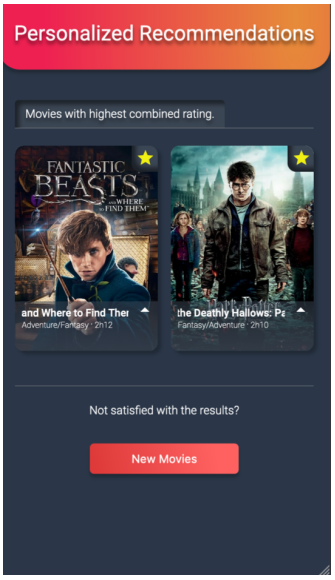

Figure 6.5: Final decision

The filtered list as shown in figure 6.3 contains all the movies that were liked by *both* participants the users both rate these movies, on a scale ranging from 0 to 3, to achieve a final decision.
Below the filtered list of movies the system displays a list of movies that were liked by only a single user as shown in figure 6.4. By doing this, the system leaves room for compromise. But movies don't need to be rated in this step (no rating is equal to a zero rating) the extra list does not add time and complexity.
Figure 6.5 shows the final decision, in this case, there were two movies that had the highest combined rating. Users can now discuss themselves which movie they want to watch and in case they are not satisfied they can choose to start over by pressing the button for new movies.

6.4 A mobile approach

For the group context of this research, pairs that commonly watch movies together, the mobile approach is the best solution. Not only because users have easy access to a mobile device. But it makes it easy for users to communicate with each other while maintaining a level of privacy that users want while going through the process. They perform the actions on their own device and can choose what they want to discuss. This privacy is not feasible is a tv-app solution.

6.5 Manipulation

From the results of the final iteration, manipulation seems to be accepted for the most part by, as mentioned in section 5.2 groups in the context of this research. As they know each other well and commonly watch movies together they expect their partners to have the mindset to find a good consensus and view the usage of the application as part of the movie-watching experience. As manipulation defeats the purpose of the application and ruins the full movie-watching experience, they expect groups in the same context to not manipulate.
Also mentioned in section 5.2 is that participants expected a system that counteracts manipulation to be too restrictive as it would take away control over the process, they would rather have this control with the possibility of manipulation behavior.
Even though manipulation is accepted, some participants did mention that they would react in a negative way to possible manipulation as discussed in section 5.2. Future research can focus on groups with different levels of familiarity as these groups may manipulate more and this could lead to more negative reactions and in general a more negative user experience of the system. In this case, it could be beneficial to try and counteract manipulation behavior, but as these are speculations, future research needs to be conducted to achieve a conclusive answer.

Chapter 7

Conclusion

This research aimed to explore and identify how to implement mobile features for the final decision phase, in group movie recommendation systems for pairs of people that commonly watch movies together, with regards to the users' needs, the user experience of the system, and manipulation behavior. From the two first studies done, it can be concluded that a possible solution is a layered approach that combines a simple mechanism (in this case positive elicitation) to filter the initial list of movies with a more complex mechanism (in this case rating from 0 to 3) to make a final decision with more detail from the filtered (shorter) list. Based on the final study it seems that it's not the job of the system to counteract or solve manipulation behavior. Mostly because participants expected that the system will be used with the mindset to find a consensus as this is the goal of the system.

Further research is still needed to determine different approaches to the final decision phase as not all possible ones were explored or suggested during this research. Different approaches could be possible and might perform better on user experience. For future research can also use a co-design session with more participants to explore new ideas and possibilities. This future research can add to design guidelines to keep in mind when designing a GRS that uses the final decision phase so that the final decision phase can be implemented in systems to create a better user experience.

To further understand the implications of the final study about manipulation behavior further studies could be conducted using participants with differing levels of familiarity to explore how this would impact manipulation behavior and the reactions to this. Differing levels of familiarity might change the dynamics of the group and thus users might be less or more inclined to manipulate and they might also react differently. Not only this but from this research, it can be concluded that people who commonly watch movies together have the mindset to reach a good consensus and will thus not manipulate and also accept possible manipulation. The mindset of people that don't know each other well might be different and the need for the system to solve or counteract manipulation might be different as well.

A design improvement that was suggested during all three studies by multiple groups is a mechanism to filter the initial list of movies on their current preferences of movie genres before receiving the list. This makes the process more dynamic and suited for different moods that users could be in. Despite this being an interesting topic, it was not explored during this research as it would steer away from the research questions. Further research is needed to determine the impact of such a mechanism on user experience, the complexity of the system, and possibly also on manipulation behavior.

This research explores a final phase to the current process of group recommendation

systems, the final decision phase, which aims to aid groups in reducing the initial list to a final movie to watch together. A possible, but not yet conclusive, solution is the layered approach that provides the needed control without adding too much complexity, which was not only suggested by the participants but also seems to provide a superior user experience over a less complex approach such as negative elicitation.

Bibliography

[1] B. D. Brian Dean. Netflix subscribers, 02 2021.

[2] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3):261–273, 2015.

[3] Dietmar Jannach, Markus Zanker, Mouzhi Ge, and Marian Gröning. Recommender Systems in Computer Science and Information Systems – A Landscape of Research. *Lecture Notes in Business Information Processing*, pages 76–87, 2012.

[4] Joseph A. Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.

[5] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[6] Yucheng Jin, Nava Tintarev, and Katrien Verbert. Effects of Individual Traits on Diversity-Aware Music Recommender User Interfaces. *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 2018.

[7] Michal KOMPAN and Mária BIELIKOVÁ. Personalized Recommendation for Individual Users Based on the Group Recommendation Principles. *Studies in Informatics and Control*, 22(3), 2013.

[8] Zhiwen Yu, Xingshe Zhou, Yanbin Hao, and Jianhua Gu. TV Program Recommendation for Multiple Viewers Based on user Profile Merging. *User Modeling and User-Adapted Interaction*, 16(1):63–82, 2006.

[9] Anthony Jameson. More than the sum of its members. *Proceedings of the working conference on Advanced visual interfaces - AVI '04*, 2004.

[10] Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. *UAI*, 06 2002.

[11] Vincent Conitzer and Tuomas Sandholm. Applications of automated mechanism design. 05 2004.

[12] Thi Ngoc Trang Tran, Alexander Felfernig, Viet Man Le, Müslüm Atas, Martin Stettinger, and Ralph Samer. User Interfaces for Counteracting Decision Manipulation in Group Recommender Systems. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP'19 Adjunct*, 2019.

[13] Maria Salamó, Kevin McCarthy, and Barry Smyth. Generating recommendations for consensus negotiation in group personalization services. *Personal and Ubiquitous Computing*, 16(5):597–610, 2011.

[14] Dennis Chao, Justin Balthrop, and Stephanie Forrest. Adaptive radio: Achieving consensus using negative preferences. *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*, 04 2004.

[15] Müslüm Atas, Alexander Felfernig, Martin Stettinger, and Thi Ngoc Trang Tran. Beyond Item Recommendation: Using Recommendations to Stimulate Knowledge Sharing in Group Decisions. *Lecture Notes in Computer Science*, pages 368–377, 2017.

[16] Christine Bauer and Bruce Ferwerda. Conformity Behavior in Group Playlist Creation. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[17] Şaban Düzgün and Aysenur Birturk. A web-based book recommendation tool for reading groups. *AAAI Workshop - Technical Report*, pages 10–17, 01 2012.

[18] George Popescu. Group Recommender Systems as a Voting Problem. *Online Communities and Social Computing*, pages 412–421, 2013.

[19] Wolfgang Wörndl and P. Saelim. Voting operations for a group recommender system in a distributed user interface environment. In *RecSys Posters*, 2014.

[20] L. Ardissono, A. Goy, G. Petrone, M. Segnan, and P. Torasso. Tailoring the Recommendation of Tourist Information to Heterogeneous User Groups. *Hypermedia: Openness, Structural Awareness, and Adaptivity*, pages 280–295, 2002.

[21] Stratis Ioannidis, Senthilmurugan Muthukrishnan, and Jinyun Yan. A consensus-focused group recommender system. 12 2013.

[22] Martin Stettinger. Choicla. *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14*, 2014.

[23] Martin Stettinger, Alexander Felfernig, Gerhard Leitner, and Stefan Reiterer. Counteracting Anchoring Effects in Group Decision Making. *Lecture Notes in Computer Science*, pages 118–130, 2015.

[24] Martin Stettinger, Alexander Felfernig, Gerhard Leitner, Stefan Reiterer, and Michael Jeran. Counteracting serial position effects in the choicla group decision support environment. volume 2015, 03 2015.

[25] Martin Stettinger, Gerald Ninaus, Michael Jeran, Florian Reinfrank, and Stefan Reiterer. WE-DECIDE: A Decision Support Environment for Groups of Users. *Recent Trends in Applied Artificial Intelligence*, pages 382–391, 2013.

[26] Berardina Carolis, Stefano Ferilli, and Nicola Orio. Recommending music to groups in fitness classes. 01 2014.

[27] Anthony Jameson, Stephan Baldes, and Thomas Kleinbauer. Enhancing mutual awareness in group recommender systems. 08 2003.

[28] Anthony Jameson, Stephan Baldes, and Thomas Kleinbauer. Two methods for enhancing mutual awareness in a group recommender system. *Proceedings of the working conference on Advanced visual interfaces - AVI '04*, 2004.

[29] Silvia Rossi, C. Napoli, F. Barile, and Luca Liguori. Conflict resolution profiles and agent negotiation for group recommendations. In *WOA*, 2016.

[30] Christian Villavicencio, Silvia Schiaffino, Andres Diaz-Pace, and Ariel Monteserin. *PUMAS-GR: A Negotiation-Based Group Recommendation System for Movies*, pages 294–298. 01 2016.

[31] Oscar Alvarado, Elias Storms, David Geerts, and Katrien Verbert. Foregrounding Algorithms: Preparing Users for Co-design with Sensitizing Activities. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 2020.

[32] James V. Bradley. Corrigenda: Complete Counterbalancing of Immediate Sequential Effects in a Latin Square Design. *Journal of the American Statistical Association*, 53(284):1030, 1958.

[33] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[34] Virginia Braun and Victoria Clarke. Thematic analysis. *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, pages 57–71, 2012.

[35] The Movie Database (TMDb).

[36] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11*, 2011.

Appendices

Appendix A

Questionnaire iteration 2

In this appendix, you find the questionnaire that was used in iteration two to measure the user experience of each prototype.

Interaction adequacy

1. The recommender allows me to tell what I like/dislike.
2. I found it easy to tell the system what I like/dislike.
3. I found it easy to inform the system if I dislike/like the recommended item.

Perceived ease of use

4. I became familiar with the recommender system very quickly.
5. I easily found the recommended items.

Control

6. I feel in control of modifying my taste profile.
7. The recommender allows me to modify my taste profile.
8. I found it easy to modify my taste profile in the recommender.

Transparency

9. I understood why the items were recommended to me.

Perceived usefulness

10. The recommender helped me find the ideal item.
11. Using the recommender to find what I like is easy.
12. The recommender gave me good suggestions.

Overall satisfaction

13. Overall, I am satisfied with the recommender.

Use intentions

14. I will use this recommender again.
15. I will use this recommender frequently.
16. I will tell my friends about this recommender.

AFDEL
Straat nr bus
3000 LEUVEN, BE
tel. + 32 16 00 0
fax + 32 16 00 0
www.kuleuve