

# User Perceived Qualities and Acceptance of Recommender Systems: The Role of Diversity

THÈSE N° 4680 (2010)

PRÉSENTÉE LE 5 JUILLET 2010

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

GROUPE DE SCIENTIFIQUES IC

UNITE DU DR. P. PU FALTINGS

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Nicolas JONES

acceptée sur proposition du jury:

Prof. A. Schiper, président du jury  
Dr P. Pu Faltings, directrice de thèse  
Prof. A. Boyer, rapporteur  
Prof. P. Dillenbourg, rapporteur  
Dr A. Jaimes, rapporteur



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

Suisse  
2010



# Acknowledgements

The four years of work on my thesis have been memorable in many ways. I was given the opportunity to explore and to immerse myself in one topic; this work has taught me a lot about myself. It gave me confidence in my intellectual and scientific aptitudes and it forced me to push back limits. All this would not have been possible without the help and support of the people named hereafter.

Firstly I would like to present my utmost gratitude to my thesis supervisor, Dr. Pearl Pu. Her support and guidance were invaluable and have allowed me to develop my research skills. Dr. Pu has provided me with insightful feedback to help me stay on track and to surpass myself. We went through some challenging times together and I learned a great deal from Dr. Pu. I would also like to thank the members of the jury, Prof. Anne Boyer, Prof. Pierre Dillenbourg and Dr. Alejandro Jaimes, and the president of the jury, Prof. Andr Schiper, for the time they gave to evaluate my work and to make constructive suggestions for this dissertation.

I am grateful to the members of the HCI group with whom I had the chance to work during these years, in particular Jiyong Zhang and Li Chen who readily shared their experience and time with me and who gave me valuable comments on my research. I would also like to thank Rong Hu for her kindness and encouragement and Sylvain Castagnos for his enjoyable collaboration in several experiments. My thanks go to Dr. James Reilly for the contribution and collaboration on the CritiqueShop system and to Dr. Rosta Farzan for her valuable assistance.

I would especially like to thank my colleagues Olivier Crameri and Gilles Dubochet for their friendship and daily support. Animated discussions with them kept me focussed and motivated. My thanks also go to Yannick Burri and my friends outside EPFL who provided me with an escape window, much needed at times.

Last but not least, I would like to thank my parents and my friend Virginie for believing in me, and who showed me their understanding, patience and encouragement at all times.



# Abstract

Recommender systems have become important, as users are faced with an ever-increasing amount of information available on internet. Much of the research work on the topic has been focused on recommendation techniques, aiming at improving the accuracy of recommended items. Today, researchers use accuracy-metrics for evaluating goodness, when in fact these do not capture users' expectations and criteria for evaluating recommendation usefulness. We must ask ourselves whether a less accurate recommendation is necessarily a less valuable one for the user. To support this, we centre our investigations in this thesis on users, and explore their acceptance behaviours when using recommendations, and their perceived qualities. We present results in four areas.

First, we study users' perceptions leading to the acceptance of recommendations and the possible long-term adoption of the system. We run two user studies using two online music recommenders relying on different recommendation techniques. Our results show that the perceived usefulness in terms of quality, and the perceived ease of use in terms of effort, are directly correlated with the users' acceptance of the recommendations. The results also show the necessity for low-involvement recommenders to be highly reactive, helping to take the users' search context into account.

Secondly, we evaluate a behavioural recommender, where recommendations are made from implicitly expressed user preferences. We take profile sizes into account and compare such recommendations to an explicit search & browse interface. Our experiment reveals that users perceive the smaller effort required to use a behavioural recommender, but find the explicit solution to yield more diverse suggestions and gives them more control. Overall, users perceive both approaches as being satisfactory, providing the profile size is big enough.

Thirdly, we analyse the impact on users' perceptions of a visual rendering. We designed an iconised representation of compound critiques, usually textual, and observed the differences in users' appreciation. Our results reveal that users prefer the visual interface, that it reduces their interaction efforts, and that users are attracted to apply the critiques more frequently in complex product domains, which have more product-features.

In a fourth area, we examine the role of diversity of recommendations in users' acceptance. A first study shows that diversity is the dimension which most influences users' satisfaction. We also highlight that users have more confidence in their choice using an organised layout interface for the same perceived ease of use as with a list view, even though the organised layout creates longer interactions. For the first time in a study, we show that diversity correlates with the trust of users. In a second study, we use an eye-tracker to carry out an in-depth study of users' decision

process. We show how the influence of a recommender increases throughout a user's purchase decision process until the decision is close to being taken. At this moment, we observed that users rely on the recommender to enhance their confidence in the purchase decision, and that they need diversity to prioritise the suggestions.

To end our work, we propose a theoretical diversity-model for maximising users' overall satisfaction by balancing users' needs for recommendation accuracy and diversity throughout the decision process. In addition, we derive a set of design guidelines from all of the experimental results. They are elaborated around four primary axes: user effort, purchase intentions, complex systems and diversity.

### **Keywords**

Recommender systems, interaction design, usability evaluation, acceptance issues, critiquing, user preferences, diversity, entertainment e-commerce.

# Résumé

Les systèmes de recommandations sont devenus importants pour les utilisateurs qui sont confrontés à des quantités toujours croissantes d'informations disponibles sur Internet. Une grande partie du travail de recherche sur ce sujet a été dirigée vers les techniques de recommandation, le but étant d'améliorer la justesse des objets recommandés. Aujourd'hui, les chercheurs utilisent des métriques de précision pour l'évaluation des bons choix, alors que celles-ci ne captent pas les attentes et les critères des utilisateurs pour évaluer l'utilité de recommandation. A ce stade, il serait donc judicieux de s'interroger sur le rapport entre recommandations produites et perceptions de l'utilisateur. Une recommandation avec moins de justesse a-t-elle nécessairement une valeur inférieure pour l'utilisateur? Pour explorer cette hypothèse, nous avons concentré nos investigations sur des utilisateurs et leurs comportements d'acceptation lorsqu'ils reçoivent les recommandations, ainsi que sur leurs qualités perçues. Nous présentons des résultats dans quatre domaines.

En premier lieu, nous avons étudié les perceptions des utilisateurs qui mènent à l'acceptation des recommandations et la possible adoption du système à long terme. Nous avons élaboré deux études d'utilisateurs avec deux systèmes recommandeurs de musique en ligne, qui se basent sur des techniques de recommandations différentes. Nos résultats démontrent que l'utilité ressentie en termes de qualité, ainsi que la facilité perçue par l'utilisateur en termes d'effort, sont directement en corrélation avec l'acceptation des recommandations par l'utilisateur. Les résultats montrent également qu'il est nécessaire pour les systèmes recommandeurs à faible implication d'être fortement réactionnels afin de favoriser la prise en considération du contexte des recherches de l'utilisateur.

Deuxièmement, nous avons évalué des recommandations établies grâce à des préférences implicitement exprimées par les utilisateurs. En prenant en considération la taille des profils, nous les avons comparés à une interface explicite de type chercher & consulter. Notre expérience montre que les utilisateurs perçoivent un recommandeur implicite comme nécessitant moins d'effort à utiliser, mais en même temps ils trouvent que la solution explicite donne une plus grande diversité et un contrôle accru. Dans l'ensemble, les utilisateurs sont satisfaits avec les deux approches, pourvu que la dimension du profil implicite soit suffisamment grande.

Troisièmement, nous avons analysé l'impact des effets visuels sur les perceptions des utilisateurs. Nous avons conçu une représentation iconisée des critiques composées, traditionnellement textuelles, ce qui nous a permis d'observer les différences d'appréciations par les utilisateurs. Nos résultats montrent que les utilisateurs préfèrent l'interface visuelle, que leurs efforts d'interaction sont réduits, et qu'ils sont amenés à appliquer des critiques plus fréquemment dans

des domaines complexes de produits.

Finalement, nous avons également examiné le rôle de la diversité dans l'acceptation par l'utilisateur. Une première étude montre que la diversité est le paramètre qui influence le plus la satisfaction de l'utilisateur. Nous mettons l'accent sur le fait que les utilisateurs sont plus confiants dans leurs choix quand ils utilisent une interface organisée, qui leur donne la même facilité d'utilisation qu'une présentation par listes, même si un schéma organisé crée des interactions plus longues. Pour la première fois dans une étude, il est démontré que la diversité est en corrélation avec la confiance des utilisateurs dans le recommandeur. Dans une deuxième étude, nous avons utilisé un oculomètre pour faire une analyse approfondie du processus de décision de l'utilisateur. Nous expliquons pourquoi l'influence d'un système de recommandation augmente d'une manière interactive à travers le processus de décision d'achat d'un utilisateur jusqu'à ce que la décision soit proche d'être prise. C'est dans cette phase que nous avons remarqué que les utilisateurs dépendent du système recommandeur pour améliorer leur confiance de décision d'achat et qu'ils ont besoin de diversité pour établir une priorité parmi les propositions reçues.

A la fin de nos travaux, nous proposons un modèle théorique de diversité pour maximiser la satisfaction générale des utilisateurs en équilibrant les besoins pour la précision des recommandations et pour la diversité tout au long du processus de décision. En conclusion, nous déduisons de nos résultats une série de directives élaborées autour de quatre axes primaires: l'effort de l'utilisateur, les intentions d'achat, la complexité des systèmes et la diversité.

### **Mots-clés**

Systèmes de recommandations, design d'interaction, évaluation d'ergonomie, problématiques d'acceptation, critique, préférences des utilisateurs, diversité, e-commerce, divertissement.



# Contents

|  |            |
|--|------------|
| <b>Acknowledgements</b>  | <b>i</b>   |
| <b>Abstract</b>  | <b>iii</b> |
| <b>Résumé</b>  | <b>v</b>   |
| <b>1 Introduction</b>  | <b>1</b>   |
| 1.1 Research Motivation . . . . .  | 3          |
| 1.2 Problem Definition . . . . .   | 5          |
| 1.2.1 Attraction, Acceptance and Adoption in Recommendations . . . . .                   | 5          |
| 1.2.2 Explicitly and Implicitly Stated User Preferences in Recommender Systems . . . . . | 7          |
| 1.2.3 Diversity’s Role in Recommendations . . . . .                                      | 8          |
| 1.3 Main Contributions . . . . .   | 9          |
| 1.4 Overview of the Dissertation . . . . .   | 11         |
| <b>2 State of the Art</b>  | <b>15</b>  |
| 2.1 Recommendation Techniques . . . . .  | 15         |
| 2.1.1 Collaborative Filtering Technology . . . . .                                       | 15         |
| 2.1.2 Content-Based Recommendations . . . . .  | 17         |
| 2.1.3 Hybrid Recommender Systems . . . . .   | 18         |
| 2.2 Critiquing-based Recommenders . . . . .  | 19         |
| 2.2.1 Example Critiquing . . . . .   | 20         |
| 2.2.2 Unit Critiques and Compound Critiques . . . . .                                    | 21         |
| 2.2.3 Dynamic Critiquing . . . . .   | 23         |
| 2.3 User-Centred Evaluations of Recommenders . . . . .                                   | 25         |
| 2.3.1 Taxonomy of Recommender Systems in E-Commerce . . . . .                            | 25         |
| 2.3.2 Interaction Design for Recommender Systems . . . . .                               | 25         |
| <b>3 User Attraction and Recommender Acceptance</b>                                      | <b>27</b>  |
| 3.1 Introduction . . . . .   | 27         |
| 3.2 Motivation for Research on Acceptance of Recommendations . . . . .                   | 29         |
| 3.2.1 Technology Acceptance . . . . .  | 29         |
| 3.2.2 Research Model . . . . .   | 31         |

|          |  |           |
|----------|--|-----------|
| 3.3      | Experiment Framework: Online Music Recommenders . . . . .                      | 32        |
| 3.3.1    | Why Music Recommenders? . . . . .  | 33        |
| 3.3.2    | Pandora.com . . . . .  | 34        |
| 3.3.3    | Last.fm . . . . .  | 36        |
| 3.4      | Experiment 1: User Acceptance in <i>Pandora</i> and <i>Last.fm</i> . . . . .   | 37        |
| 3.4.1    | Setup and Procedure . . . . .  | 37        |
| 3.4.2    | Analysis of Results . . . . .  | 39        |
| 3.4.3    | Discussion . . . . .   | 48        |
| 3.5      | Experiment 2: User Adoption in <i>Pandora</i> and <i>Last.fm</i> . . . . .     | 49        |
| 3.5.1    | Setup and Procedure . . . . .  | 49        |
| 3.5.2    | Analysis of Results . . . . .  | 52        |
| 3.5.3    | Discussion . . . . .   | 56        |
| 3.6      | Conclusions . . . . .  | 58        |
| <b>4</b> | <b>Losing Control: Are Preferences Best Revealed Explicitly or Implicitly?</b> | <b>61</b> |
| 4.1      | Introduction . . . . .   | 61        |
| 4.2      | Background and Related Work . . . . .  | 62        |
| 4.3      | Hypotheses . . . . .   | 63        |
| 4.4      | Experiment 3: Search & Browse vs. Implicit Recommendations . . . . .           | 65        |
| 4.4.1    | Setup and Procedure . . . . .  | 65        |
| 4.4.2    | Analysis of Results . . . . .  | 67        |
| 4.4.3    | Discussion . . . . .   | 70        |
| 4.5      | Conclusions . . . . .  | 70        |
| <b>5</b> | <b>The Effects of Layout in Critiquing-Based Recommenders</b>                  | <b>73</b> |
| 5.1      | Introduction . . . . .   | 73        |
| 5.2      | Related Work . . . . .   | 74        |
| 5.3      | Experiment Framework: CritiqueShop . . . . .                                   | 75        |
| 5.3.1    | Textual Interface . . . . .  | 75        |
| 5.3.2    | Visual Interface . . . . .   | 77        |
| 5.4      | Experiment 4: Textual vs. Visual Compound Critiques . . . . .                  | 78        |
| 5.4.1    | Setup and Procedure . . . . .  | 78        |
| 5.4.2    | Analysis of Results . . . . .  | 82        |
| 5.4.3    | Discussion . . . . .   | 85        |
| 5.5      | Conclusions . . . . .  | 86        |
| <b>6</b> | <b>How Diversity Leads to Confidence</b>                                       | <b>89</b> |
| 6.1      | Introduction . . . . .   | 89        |
| 6.1.1    | Can Layout be a Vector of Diversity? . . . . .                                 | 89        |
| 6.1.2    | Recommenders' Influence on Buyer's Decision Process . . . . .                  | 90        |
| 6.2      | Related Work . . . . .   | 92        |
| 6.2.1    | Diversity in Recommendations . . . . .   | 93        |
| 6.3      | Experiment Framework: Perfume Recommender . . . . .                            | 94        |
| 6.3.1    | Diversity Metrics . . . . .  | 97        |

|          |  |            |
|----------|--|------------|
| 6.4      | Experiment 5: Diversity in a Content vs. Layout Approach . . . . .         | 99         |
| 6.4.1    | Experiment Setup . . . . .   | 99         |
| 6.4.2    | Experiment Procedure . . . . .   | 102        |
| 6.4.3    | Analysis of Results . . . . .  | 104        |
| 6.4.4    | Discussion . . . . .   | 111        |
| 6.5      | Experiment 6: Diversity in Buyers' Decision Process . . . . .              | 113        |
| 6.5.1    | Hypotheses . . . . .   | 114        |
| 6.5.2    | Experiment Setup . . . . .   | 115        |
| 6.5.3    | Experiment Procedure . . . . .   | 116        |
| 6.5.4    | Analysis of Results . . . . .  | 120        |
| 6.5.5    | Discussion . . . . .   | 128        |
| 6.6      | Conclusions . . . . .  | 129        |
| <b>7</b> | <b>Design Guidelines and Diversity Model</b>                               | <b>133</b> |
| 7.1      | User Effort . . . . .  | 133        |
| 7.2      | Purchase Intentions . . . . .  | 135        |
| 7.3      | Complex Systems . . . . .  | 137        |
| 7.4      | Diversity . . . . .  | 139        |
| 7.4.1    | Generalisation of Guidelines . . . . .                                     | 140        |
| 7.5      | Revised Research Model . . . . .   | 143        |
| 7.5.1    | Time dependent Accuracy vs. Diversity model . . . . .                      | 146        |
| <b>8</b> | <b>Conclusion</b>  | <b>149</b> |
| 8.1      | Contributions . . . . .  | 149        |
| 8.1.1    | Attracting New Users to Recommender Systems . . . . .                      | 149        |
| 8.1.2    | Evaluating Implicitly Expressed User Preferences . . . . .                 | 150        |
| 8.1.3    | The Impact of Visual Renderings in Critiquing-Based Recommenders . . . . . | 150        |
| 8.1.4    | The Role of Diversity in Recommendations . . . . .                         | 151        |
| 8.1.5    | Guidelines and Diversity Model . . . . .                                   | 152        |
| 8.2      | Future Research Directions . . . . .                                       | 153        |
| 8.2.1    | Adoption after Acceptance . . . . .  | 153        |
| 8.2.2    | Explicit and Implicit User Preferences . . . . .                           | 154        |
| 8.2.3    | Refining Diversity Explorations . . . . .                                  | 154        |
| 8.3      | Take Home Message . . . . .  | 155        |
| <b>A</b> | <b>Appendix: Experiment 1</b>  | <b>157</b> |
| A.1      | Description of the User Study . . . . .                                    | 157        |
| A.2      | Quick Presentation of Last.fm . . . . .                                    | 158        |
| A.3      | Quick Presentation of Pandora . . . . .                                    | 159        |
| A.4      | Questions . . . . .  | 160        |
| A.5      | Template . . . . .   | 161        |

## CONTENTS

---

|  |            |
|--|------------|
| <b>B Appendix: Experiment 2</b>              | <b>163</b> |
| B.1 Description of the User Study . . . . .  | 163        |
| B.2 Post Study Interview . . . . .           | 165        |
| <b>C Appendix: Experiment 3</b>              | <b>167</b> |
| C.1 Description of the User Study . . . . .  | 167        |
| C.2 Template . . . . .                       | 169        |
| <b>D Appendix: Experiment 4</b>              | <b>171</b> |
| D.1 Introduction to the User Study . . . . . | 171        |
| <b>E Appendix: Experiment 5</b>              | <b>175</b> |
| E.1 Description of the User Study . . . . .  | 175        |
| <b>F Appendix: Experiment 6</b>              | <b>177</b> |
| F.1 Description of the User Study . . . . .  | 177        |

# List of Figures

- 1.1 Overview of the organisation of the thesis. . . . . 13
- 2.1 A snapshot of Entrée’s main GUI. . . . . 22
- 2.2 Snapshot of a prototype dynamic critiquing shop (Qwikshop). Highlighted are unit and compound critiques. . . . . 24
- 3.1 Technology Acceptance Model . . . . . 30
- 3.2 Illustration of Kamis and Stohr’s research model. . . . . 30
- 3.3 Model of experiment setup . . . . . 32
- 3.4 Snapshot of Pandora . . . . . 35
- 3.5 Snapshot of Last.fm . . . . . 37
- 3.6 Graph of reasons for not discovering more. . . . . 42
- 3.7 Graph of enjoyability . . . . . 44
- 3.8 Graph of discovery and appreciation . . . . . 44
- 3.9 Distribution of system vs. friends’ recommendations . . . . . 45
- 3.10 Agreement levels to statements of the final questionnaire. . . . . 51
- 4.1 Background questions . . . . . 67
- 4.2 Detailed graph of preferences of users. . . . . 68
- 5.1 Screenshot of the interface for textual compound critiques (with laptop dataset). 76
- 5.2 Example of a compound critique from both the textual and visual interface. . . 77
- 5.3 Average sessions lengths (left) and average application frequency of compound critiques (right) for both user interfaces. . . . . 82
- 5.4 Average recommendation accuracy for both user interfaces. . . . . 84
- 5.5 Results from the post-stage assessment questionnaire. . . . . 84
- 5.6 Results from the final preference questionnaire. . . . . 85
- 5.7 Screenshot of the visual interface for the online shopping system (with laptop dataset). . . . . 88
- 6.1 The six stages of purchase decision, as proposed by Maes *et al.* . . . . . 92
- 6.2 The *search page* of the perfume recommender. . . . . 95
- 6.3 The *detail page* of the perfume recommender. . . . . 97
- 6.4 Comparison of List and Organised view. . . . . 100

## LIST OF FIGURES

---

|      |   |     |
|------|---|-----|
| 6.5  | Four possible combinations of content and layout of Experiment 5. . . . .   | 102 |
| 6.6  | Background profile of users. . . . .  | 105 |
| 6.7  | Average number of pages viewed. . . . .   | 106 |
| 6.8  | Average session length. . . . .   | 107 |
| 6.9  | Results from the post-stage assessment questionnaires (sessions 1 & 2). . . . .                                   | 108 |
| 6.10 | Results from the assessment questionnaires divided into List and Organised view.                                  | 109 |
| 6.11 | Results from the assessment questionnaires divided into Amazon and Editorial<br>Picked Critiques content. . . . . | 109 |
| 6.12 | Hotspots recorded by an eye tracker on a page of the perfume framework. . . . .                                   | 113 |
| 6.13 | AOIs of the <i>search page</i> and <i>detail page</i> in the perfume framework. . . . .                           | 117 |
| 6.14 | Users' background knowledge about perfumes (Exp 6). . . . .   | 119 |
| 6.15 | Bubble view of the usage of tools provided in perfume framework. . . . .  | 121 |
| 6.16 | Cumulative usage-times of RS and MCF. . . . .   | 122 |
| 6.17 | Usage of MCF and RS over time, with purchase decision peaks. . . . .  | 123 |
| 6.18 | Proportion of time spent looking at RS categories. . . . .  | 126 |
| 6.19 | Answers to the second experiment's assessment questionnaire . . . . .   | 127 |
|      |   |     |
| 7.1  | Dimensions of the original research model. . . . .  | 142 |
| 7.2  | Key correlations of Experiment 2 modelled according to the TAM. . . . .   | 143 |
| 7.3  | Proposed research model focused on experience and decision variables. . . . .                                     | 144 |
| 7.4  | Factors and correlations from implicit profiles. . . . .  | 145 |
| 7.5  | Time dependent Accuracy vs. Diversity model. . . . .  | 147 |
|      |   |     |
| A.1  | Snapshot from Last.fm . . . . .   | 158 |
| A.2  | Snapshot from Pandora . . . . .   | 159 |
| A.3  | Template provided. . . . .  | 162 |
|      |   |     |
| C.1  | Template provided. . . . .  | 169 |
|      |   |     |
| D.1  | Screenshot of the system with Unit and Compound Critiques. . . . .  | 172 |
| D.2  | Textual vs. Visual Compound Critiques. . . . .  | 172 |
| D.3  | Iconic representation of weight variations. . . . .   | 173 |

# List of Tables

- 3.1 Interface quality results. . . . . 41
- 3.2 Subjective variables results. . . . . 43
- 3.3 Objective measures from templates. . . . . 45
- 3.4 Users’ preference in systems. . . . . 47
- 3.5 Prediction quality of recommendations, based on correlations  $r$ . . . . . 48
- 3.6 List of questions addressing Perceived Usefulness - Quality. . . . . 52
- 3.7 List of questions addressing Perceived Ease of Use - Effort. . . . . 53
- 3.8 List of questions addressing Acceptance. . . . . 54
- 3.9 Correlation: PEOU (effort) with PU (quality). . . . . 55
- 3.10 Correlation: PU (quality) with acceptance. . . . . 56
- 3.11 Correlation: PEOU (effort) with acceptance. . . . . 56
  
- 4.1 Post-stage assessment (S) and template (T) questions. . . . . 66
- 4.2 Summary of main correlations. . . . . 69
  
- 5.1 CritiqueShop post-stage assessment questionnaire. . . . . 79
- 5.2 Final preference questionnaire. . . . . 80
- 5.3 Demographic characteristics of participants. . . . . 81
- 5.4 Design of the real-user evaluation. . . . . 81
  
- 6.1 Average Intra-List Similarity for six recommendations (ILS), Average Similarity between a recommendation and the perfume of the current detail page (Sim), and Relative Diversity of a recommendation relative to the current perfume (RD). . . . . 99
- 6.2 Design of the real-user evaluation. . . . . 102
- 6.3 Post-stage assessment questionnaire. . . . . 104
- 6.4 Demographic characteristics of participants. . . . . 105
- 6.5 Diversity scores for each configuration. (diversity =  $1 - ILS$ ) . . . . . 107
- 6.6 Which system did users prefer between both sessions. . . . . 111
- 6.7 Average Relative Diversity between two perfumes coming from MCF tool. . . . . 115
- 6.8 Post-stage assessment questionnaire. . . . . 118
- 6.9 Demographic characteristics of participants. . . . . 119
- 6.10 Statistics of sessions for the overall set of users. . . . . 120
- 6.11 Average numbers of cycles (all users). . . . . 124

## LIST OF TABLES

---

|      |   |     |
|------|---|-----|
| 6.12 | Average numbers of clicks (all users). . . . .                            | 125 |
| 6.13 | Average effects of recommendation categories. . . . .                     | 126 |
| 6.14 | Number of products added to the basket that came from RS vs. MCF. . . . . | 126 |



# Chapter 1

## Introduction

“It’s not what you look at that matters, it’s what you see.”  
– *Henry David Thoreau.*

“The Internet is about usability. ... The computer industry has been able to ship difficult-to-use products because you buy first, and then you try to use it. With the Web, usability comes first, then you click to buy or become a return visitor.”  
– *Jakob Nielsen.*

The Web is growing dramatically fast. In May 2009, The Guardian reported on IDC’s latest numbers, showing that “at 487bn gigabytes, if the world’s rapidly expanding digital content were printed and bound into books it would form a stack that would stretch from Earth to Pluto ten times”<sup>1</sup>. Such a sizeable growth has changed our daily lives, and continues to do so everyday. E-commerce services have thrived into a huge business market. According to the Census Bureau of the United States, the U.S. retail e-commerce sales for the third quarter of the year 2009 was estimated to have reached \$34.0 billion, despite being in the middle of the 2007-2010 Financial Crisis<sup>2</sup>. With so much information available, it becomes difficult for users to make decisions: the Web has a real information overload problem. Henry David Thoreau’s quotation above is very fitting. Artist and naturalist, he pronounced these wise words back in 1859, a long time before the Web was invented. He had understood that even when available, anything could only become useful if people saw it. Regarding today’s Web, the analogy is simple: it is easy to know how many people look at an e-commerce website, but it is harder to know how many actually see the products that they are looking for. What Jacob Nielsen says in the second introductory quote, is even more revealing about today’s challenges on the Web. He pinpoints how important usability has become because of the Web’s access structure, which is especially true when we consider the information overload problem.

One approach which has emerged to help users deal with the amount of information is Recommender Systems. The goal of recommenders is to quickly and efficiently sort through vast quantities of information and bring the (hopefully) relevant pieces of information to users atten-

---

<sup>1</sup><http://www.guardian.co.uk/business/2009/may/18/digital-content-expansion>

<sup>2</sup>Data source: <http://www.census.gov/retail/mrts/www/data/pdf/09Q3.pdf>

tion. In doing so, recommender systems help us navigate through complex information spaces, allowing us to be more efficient and helping us to cope with the overload of information.

Recommender Systems (hereafter also RS) were originally proposed by Goldberg *et al.* in 1992 [50], and have been supporting users for the last eighteen years, by providing interesting items according to users' profiles and items' attributes. A recommender system typically compares a user's profile to some reference characteristics, and seeks to predict which items the user would have rated highly. The characteristics may come from the information about items in a content-based approach, or from the user's social environment in a collaborative filtering approach. We cover the main recommendation techniques which are used in this thesis in Chapter 2. In the last few years, the expansion of recommenders in mainstream commerce websites has started to become significant. Their positive effects on sales have been shown in several marketing studies such as [33, 34, 35]. However, as pointed out at the end of Nielsen's quote, even if systems can recommend users a product they like, the real challenge is to bring them to purchase or to come back to the site. We call these two concepts the *acceptance* of recommendations, and the *adoption* of the system, and define them in Chapter 3.

It is tempting to question whether one really understands users and their preferences? Understanding users' preferences is an important and necessary issue, in order to make sure that the system suggests relevant items to the user. It is however very easy for a recommendation to be horribly wrong, whether coming from a system or a human. Imagine you received a tip from a friend at work, about a new rock song. You love rock, and decide to play it to your guests who came round for supper. Even though you might love the song, it could be a hit or total miss for your friends. The context in which users' preferences are used is crucial in making a good recommendation. In 2002, Jeffery Zaslow wrote an article in the Wall Street Journal about recommender systems entitled, "If TiVo Thinks You Are Gay, Here's How to Set It Straight", which highlighted another issue with preferences: controlling them. The article relates the mishaps of many users, in trying to correct the recommendations of their TiVo digital video recorder<sup>3</sup>.

Mr. Iwanyk, 32 years old, first suspected that his TiVo thought he was gay, since it inexplicably kept recording programs with gay themes. A film studio executive in Los Angeles and the self-described "straightest guy on earth", he tried to tame TiVo's gay fixation by recording war movies and other "guy stuff".

"The problem was, I overcompensated", he says. "It started giving me documentaries on Joseph Goebbels and Adolf Eichmann. It stopped thinking I was gay and decided I was a crazy guy reminiscing about the Third Reich".

Such issues of context and control in users' preferences, are central to this thesis. Much of the research work on recommenders has been focused on the recommendation techniques, aiming at improving the accuracy of recommended items. Today, researchers use accuracy-metrics for evaluating goodness, when in fact these do not capture users' expectations and criteria for evaluating recommendation usefulness. We must ask ourselves whether a less accurate recommendation is necessarily a less valuable one for the user? Could users, for instance, benefit from having more diversity among the suggested items? To support this, we centre our investigations on users, and explore their behaviours when encountering and using recommender systems.

---

<sup>3</sup><http://www.tivo.com/>

The first goal of this thesis was to study how users perceived the different qualities of recommender systems, leading them to accepting the proposed recommendations and ultimately driving them to the adoption of the system. Throughout our investigations, we managed to single out one essential quality: the diversity of among items recommended to users. When an e-commerce site presents a set of recommendations to a user, the later is readily able to perceive the degree of diversity offered. We then studied this effect in greater detail, allowing us to understand the role of diversity in the process leading users to accept the recommendations. We designed six experiments where we evaluated the reactions of 306 users to different systems, collecting a large amount of data covering users' perceptions of recommendations. We used the Technology Acceptance Model [42] as a baseline model, and refined it throughout our transversal investigations. We revealed how accuracy and diversity of the recommendations were interleaved, introducing *time* as a new element in a user-satisfaction model. Moreover, based on the experimental results, we derived a set of design guidelines.

## 1.1 Research Motivation

Internet is open to all and is constantly changing, evolving and growing. What was a trend only a few months ago is probably already being pushed aside by a new upcoming concept. At the time when we were preparing this thesis, we had written in a first research-plan proposal the following positioning about recommender systems:

Recommender systems continue to play a critical role in helping users overcome the information overload problem on internet. Currently, recommendation technology is not only used in recommendation-*giving* sites where the system suggests an item which may interest users based on their history, but also in recommendation-*seeking* sites where users actively seek advice and recommendations as first-time users. The main difference between the two kinds of recommendation contexts lies in the actor who initiates the recommendation process, i.e. the system in the first context and the user in the second. In the earliest rating-based recommender systems, users were given recommendations as a result of items that they had rated or bought (purchase was used as an indication of preference). Such systems observe users and learn about their interests and tastes in the background. They then propose items that may interest their users. At present, an increasingly large number of users go to websites to *seek* advice and suggestions for electronic products, vacation destinations, music, books, etc. Users are often first-time visitors to the websites and they are looking for recommendations without necessarily having established a history with the store. The users therefore actively seek advice and recommendations as first-time users.

When written, this observation was a reflection of a fresh analysis on how recommenders were being spread and adopted across the Web. However, as said in preamble, the Web is rapidly changing and our painting of the recommender landscape is already outdated. Recommenders have spread seeing the majority of e-commerce sites propose some suggestive mechanism. The

frontier between recommendation-giving and recommendation-seeking has often started to disappear. On the one hand, sites using the first technique have been increasing the space dedicated to the recommendations, thus bringing them closer to a seeking system. On the other hand, dedicated recommenders such as *Deezer*<sup>4</sup> have rapidly evolved to include cross-linking mechanisms, bringing it closer to giving schemes. New paradigms relying more strongly on implicit preference elicitation have started to appear already. At the same time, the general layouts of websites have globally changed and the way in which information is displayed has become more graphical, seeing a strong progression of data visualisations. Visualisations, as opposed to raw numbers and data, have even started to be used as a mechanism to compare more abstract elements such as songs<sup>5</sup>, books<sup>6</sup> or clothes [121]. Furthermore, internet was then seeing the rise of *Web 2.0* and the emergence of social interactions, where content is created and shared by users, leading to new contexts where recommender systems were used. Today, *Web 3.0* is already being discussed, defining that tomorrow's internet will be more trustful and relying on a solid layer of semantic web, according to Tim O'Reilly, the man who popularised the 2.0 term. Despite the fact that our initial analysis is not as true anymore, we feel it is important to this thesis, as it gives a first insight into parts of the motivation behind this thesis.

Our motivation to work on recommendation-seeking systems goes beyond these first impressions. Some prior research had investigated the implications of different recommendation technologies on users' interaction experience. For historical reasons, the usage context had always been assumed to be the recommendation-giving environment. As the scope and context of recommendations extended to recommendation-seeking, user interaction issues also changed. In the seeking mode, a user's initial interaction with a system becomes an especially prominent factor influencing the user's reaction and subsequent behaviour towards such systems. If the first step is not intuitive, and if the results obtained are not a minimum convincing, users are very likely to feel frustrated and leave the website. Because of our interest in understanding factors which lead users to accepting recommendations, these requirements of recommendation-seeking sites on users further persuaded us to choose them to start our enquiries. Furthermore, even though much research has emphasised developing and improving the underlying algorithms that fulfil the recommendation objectives for both contexts, very little research has addressed the contextual differences and how users may interact with different types of recommender systems. This was another leading factor motivating our research.

As stated by the title, the core of our drive is user perceived qualities. We are interested in understanding how users truthfully feel about recommenders, and what are the qualities which users distinguish, which are those that go unnoticed and which are unnerving. We wanted to compare user experience with different technologies and aimed to understand the key design features that make an effective recommender, first in the seeking context, and later in the giving context, whilst taking into account social interactions and awareness. We are motivated to find out what is the link between a technology and the way it is implemented in a specific environment. This will help understand well established problems of current popular recommendation algorithms such as the new-user (cold start) problem, and its implications for website design.

---

<sup>4</sup><http://www.deezer.com/>

<sup>5</sup><http://www.musicoverly.com/>

<sup>6</sup><http://www.zoomii.com/>

## 1.2 Problem Definition

Historically and by nature, scientific research has always had a strong bond with algorithmic discoveries and improvements. However beautiful and captivating these advances can be, algorithms are only one part of the game, especially when considering recommender systems. Through the inherent trends of our society, recommenders are currently mainly being implemented within websites (as opposed to within standalone applications), creating for a complex frame of work where the attention span of users is very short, where the slightest problem encountered can change a user's mind for good, and where the users first test before buying, as highlighted by Nielsen's quotation.

This thesis encapsulates a range of issues. Beyond algorithms, we investigate what are users' perceptions of the qualities of a recommender? Obtaining a recommendation involves to produce a certain effort and requires a minimal *involvement* from users, which we believe could be an important element for them. Furthermore, we are interested to find out how do users perceive *novelty* and *enjoyment* among a selection of items, and we ask whether they are seen as a measure of quality? Similar interrogations occur in our work about users' attitudes towards the adoption of technology and their behaviour in the recommendation process are also motivational to this work. Today's Web 2.0 model highly encourages the so-called social tools where users interact and give opinion in all sorts of ways. But what are users' real advantages, motivating them to contribute to social computations, and how do their *emotions & moods* affect the whole recommendation process remain open debates. Such are the questions which lead the reflection to this work. In order to present the problem definition of this thesis clearly, we define three main topics in the following subsections.

### 1.2.1 Attraction, Acceptance and Adoption in Recommendations

Technologies such as recommender systems are complex. Because they are new, and because an element of uncertainty exists in the minds of decision makers with respect to the successful adoption of them, people form attitudes and intentions toward trying to learn to use the new technology prior to initiating efforts directed at using it [11]. Attitudes towards usage and intentions to use may be ill-formed or lacking in conviction. Some usage attitudes may even occur only after preliminary strivings to learn the system have been conducted. Thus, actual usage may not be a direct or immediate consequence of exposition to the system.

This problematic of how users behave and feel during an encounter with a new online system is one of the challenges explored by this thesis. The value of a system where users rarely get past the entrance barrier, is strongly diminished. Furthermore, as explained in preamble, web services are changing rapidly, seeing popular systems emerge for a few months, only to disappear again before long. Such elements are the basis of a major challenge: getting users to come to a service, and making sure they accept it and stay. Our research question here is to understand: how do systems manage to *attract* new users, get them to *accept* the recommendations, and motivate them to *adopt* the system on the long term? These three words, attraction, acceptance and adoption are the essence of this research question.

Technology acceptance research investigates how and when users come to accept and use a website, a software system, or a technology. The history of this subject can be traced to as

early as the 1970s relating to the adoption of new technological innovations. With the arrival of computers, software, and lately websites, there has been a growing interest to apply the general framework of technology acceptance to specific domains. In 1989, Davis *et al.* proposed a *Technology Acceptance Model* (TAM) as an adaptation of Fishbein's theory of reasoned action [41]. This model has already been applied to certain websites but not to recommenders. We decided to use it in devising our research model, to measure and understand users' experiences with recommender systems.

As explained previously, recommendation technology was first used in recommendation-giving sites where the system would observe users' behaviours and learn about their interests and tastes. The users would select articles to read [114], or items to purchase, and the RS would then propose items that may potentially interest them based on the observed history. Therefore, users were *given* recommendations as a result of items that they had rated or bought (purchase was used as an indication of preference). In this regard, recommendations were offered as a value-added service for users to discover new items and as a way for the site to interest users in buying items that they did not look for initially. If the context of use of recommender systems had remained unchanged, understanding how users may *accept* recommender systems and recommendation results would not have been such a crucial field of study.

However, now that recommenders have emerged where users go to actively *seek* advice and suggestions for electronic products, vacation destinations, music, books, users interact with such systems as first-time customers and may not get the usual benefit of receiving recommendations "automatically" [118] due to the lack of a personal preference history. Compared to receiving unsolicited recommendations, users who specifically seek recommendations are likely to have a higher level of expectations on the results they obtain and the ease of use of the system. In the seeking context, a user makes a conscious decision to use a system for a specific goal. If they do not find what they are looking for, or the system is too hard to use, they may quickly leave. We therefore argue that as recommender systems are broadening their scope of use, the acceptance process is becoming more complex and pertinent to study. Site designers must somehow identify the right balance between the benefits they offer relative to the effort they require from users in order to increase their ability to attract new users and provide a high level of staying power.

We wish to explore and address the following questions:

- What are the key design features of a recommender system that attract new users and motivate them to accept this technology? We question if there are strong differences perceived by users between recommendation-giving and recommendation-seeking sites, where users actively seek advice, often as first-time users. We aim to identify key factors that motivate users and influence them to adopt a regular usage of a recommender website.
- Do key design features in recommenders exist that motivate users to stay and to adopt the system? This investigation should allow to understand better what makes users return to a recommendation website, and what motivates them to purchase. We believe that knowing a user's history helps the site to motivate the user to stay on. Users' purchase histories and the items that they have rated can be combined to provide persistent and personalised recommendations.

- Can other tools such as visualisations help users to accept the site? When recommendations based on a group of users are suggested to a user, this one may not be prepared to accept them due to a low level of system transparency. Bonhard *et al.* showed ways to improve collaborative filtering based recommender systems by including information on the profile similarity and rating overlap of a given user [17]. Herlocker *et al.* have investigated visualisation techniques that explain the neighbour ratings and help users accept the results [59]. We feel there is a strong chance that visualisations can play a larger role in recommendations. By extension, we also believe that other higher-level abstractions which help to reduce system transparency could change users' perceived system qualities.

### 1.2.2 Explicitly and Implicitly Stated User Preferences in Recommender Systems

There are at least three principle ways to elicit preferences from users: by observing and recording what they did in the past [145], based on items that they have liked in the past (rating information) [72, 13, 76], and based on the users' demographics [116, 75].

Pure collaborative filtering methods [62, 72, 115, 120, 21] base their recommendations on community preferences. These can have been explicitly expressed such as user ratings, or they can be implicitly collected such as purchase histories, ignoring user and item attributes like demographics and product descriptions. For example, a music recommender might combine explicit ratings data (e.g. John rates "So What" by Miles Davis a 4 out of 5) and implicit purchase data (e.g. John purchased the CD "Rainbow Children" by Prince) to make recommendations on songs and new CD releases to John. Collaborative filtering techniques compute correlations among similar users, or "nearest neighbours". Prediction of the attractiveness of an unseen item for a given user is computed based on a combination of the rating scores derived from the nearest neighbours, hence explicitly expressed preferences. However, the philosophy behind such systems is to recommend items from "like-minded" people. These could very easily be inferred from other information, such as implicitly acquired preferences. Some argue that originally, collaborative filtering technology was developed to function in the background of an information provider, and thus should also recommend items based on implicitly gathered profile data.

As this thesis is interested in understanding users' perceived qualities in a recommender system, and as highlighted by the introductory TiVo example, the question of implicitly and explicitly stated user preferences is highly stimulating. The following questions motivate our research:

- Do users perceive differences between systems operating on preferences being expressed explicitly and implicitly? There are many factors which can change users' perceived qualities. One issue is: how reliable is the information when it is explicitly acquired vs. when it is implicitly generated? Another quality issue is whether the past is a good predictor of the future.
- How to choose whether to propose explicit or implicit preference elicitation? It is unclear today under what circumstances the users prefer one approach rather than the other. Similarly, we question possible trade-offs which users might feel strongly about. Experiments such as [119] suggest that when users implicitly give feedback, the performance of the RS

can be close to the more traditional ones using explicit feedback, but much of the research is incremental and there are no studies directly comparing both extremes.

- How does knowing users' histories, help to make them return to the site? This problem is linked to the previous section on adoption. How can one extract the maximum from a user's preference, whether stated explicitly or implicitly, in order to maximise the chances that the user will return? This problem can also be seen as the trade-off between ephemeral persistent recommendations.

### 1.2.3 Diversity's Role in Recommendations

Another research dimension which has been investigated in this thesis on user perceived qualities is the diversity of recommended items. This dimension was not originally foreseen in our research plan, but it rapidly emerged in several of our studies, as an element which was influencing users' perceived overall appreciation. Diversity thus became a key topic of interest.

Diversity is a sought-after property. Early studies such as [43, 90, 122, 127] had already proposed a certain number of mechanisms which help to introduce some form of diversity. Since, it became clear that diversity had a role to play in recommendations. There are mainly three questions about diversity which motivate our work: *how* to add diversity, *why* add diversity and *when* to add diversity. *How* has already been studied by a small number of papers. Traditional methods for diversifying search results rely on attribute-based diversification, and focus on re-ranking the top  $N$  results, making the documents most likely to be preferred by the user, appear higher. Bradley and Smyth were among the first to propose a bounded greedy algorithm for retrieving the set of cases most similar to a user's query, but at the same time most diverse among themselves [20]. The questions of *why* and *when* to add diversity have less been studied. Our questioning includes:

- How can diversity be introduced more easily? As explained earlier, most methods rely on filtering and re-ranking the top suggested items. However, our research drive is to find what really impacts on users. We intend to explore if alternative approaches can yield more diversity among results, and how users react when faced with such diversity. This includes changing the layout and measuring how users perceive the change.
- Do users feel the diversity? This is an important question because it is closely linked to the *why* dimension: why do we need diversity? Fleder *et al.* covered the discussion of whether recommenders really helped users discover new products, or if they rather pushed forward the already popular ones [47]. Inversely, we might want to ask whether we really need recommendation accuracy? Indeed, McNee *et al.* raised concerns about how accuracy metrics had not only misguided but actually harmed the field of recommenders, questioning whether a probabilistically less accurate recommendation is necessarily less valuable [88]. We wish to further explore these questions.
- When should diversity optimally be introduced? McGinty and Smyth, in the context of conversational recommender systems, attempted to clarify the role of diversity. They



showed in [86] that generally, introducing diversity has the potential to significantly enhance the efficiency of recommendations. It may, however, lead to new challenges. One of these was that diversity might not be desirable during every recommendation cycle. We wish to repeat this observation, clarifying what role *time* has in suggesting diverse recommendations.

### 1.3 Main Contributions

The objective of this work was less the in-depth exploration of a low-level phenomena, but the more general understanding of factors that lead users to accepting recommendations and adopting the whole system. By *general understanding* we do not mean that the work remains superficial, but rather that we have kept our topics of investigation as closely related to each other as possible, seeking to achieve a coherent transversal investigation.

More precisely, after considering some large-scale differences between two online music recommender systems, we looked at how perceptions of control through explicitly and implicitly revealed preferences could influence the satisfaction of users. Differences between layout and content were also explored and tested, before focusing on the need for diversity in the different stages of the purchase process. These investigations have allowed us to gain a detailed understanding of how users perceive and use recommender systems, from the first time they are confronted with them, to the more regular usage of recommendations in online shopping. The main contributions of this thesis can be briefly summarised as follows.

**Attraction, Acceptance and Adoption** We set up two user studies in Chapter 3 where we address issues raised in Section 1.2.1. In both cases we compare two music recommender websites, *Pandora* (a content-based recommender) and *Last.fm* (a rating-based social recommender). In the first experiment, we compare them side-by-side in a within-subject user study involving over sixty participants. Results show that a simple interface design, the requirement of less initial effort, and the quality of recommended items in terms of accuracy, novelty and enjoyability, are some of the important design features that such websites rely on to break the initial entrance barrier in becoming a popular website.

We analyse these findings more deeply in the second experiment, where we run an in-depth between-subject lab study. We show that perceived usefulness (quality) and perceived ease of use (effort) are the key dimensions which are sufficient to incite users to accept recommendations. We adapt Davis *et al.*'s Technology Acceptance Model and show that this model is suitable for entertainment recommenders. Measures of quality such as accuracy, enjoyment, satisfaction and having music tailored to a user's taste are directly correlated with acceptance, and measures of effort like the initial time to reach interesting recommendations and the ease of use for discovering music are strongly linked to acceptance. Finally, the results highlight the necessity for low-involvement recommenders to be highly reactive.

**Explicit or Implicit Preference Elicitation** We address several issues raised in Section 1.2.2 in the experiment of Chapter 4. We report an in-depth user study comparing Amazon's

implicit book recommender with a baseline model of explicit search and browse. We set up a comparative between-group user study. Results show both approaches can yield similar overall results, as perceived by users. Specifically, implicit recommenders were perceived as being trust-worthy. Moreover people felt that these interfaces required less effort for finding items. At the same time however, they were just as satisfied as with more traditionally controlled interfaces, notably because diversity was found to be higher. Results further suggested how measures such as confidence, trust, control and intentions to buy evolve as the users' purchase profile grows.

**Layout or Content** In Chapter 5 we apply a Layout vs. Content approach to try and stimulate greater reactions from users. We propose a new *visual* design of a critiquing interface, by representing compound critiques via a selection of value-augmented icons. We set up a user evaluation based on the *CritiqueShop* system, an online shopping prototype developed earlier in our research group. Our study compares the performance of our visual design with traditional textual approach. Results from our evaluation show that the visual interface can improve the performance of critique-based recommenders by attracting users to apply the compound critiques more frequently and reducing users' interaction effort substantially when the product domain is complex. Users' subjective feedback also shows that the visual interface is highly promising in enhancing users' shopping experiences.

**Diversity in Recommendations** We set up two users-studies in Chapter 6 where we address issues raised in Section 1.2.3 of diversity in recommendations. In the first experiment, we run a between-group user study with over sixty users on a simulated perfume e-commerce website. Our results show that diversity is the dimension which most influences users' overall satisfaction and experience with the system. We also show that users have more confidence in their choice using an organised layout interface for the same perceived ease of use as with a list view, even though the organised layout creates longer interactions. Furthermore, when exposed to both, participants strongly prefer the organised layout.

In the second experiment, we employ an eye tracker and run an in-depth user study with eighteen users. Through our online perfume simulation website, we collect over 48,000 fixation data points and 7,720 areas of interest. This second study shows that recommender systems provide users with decision confidence and that users grab new opportunities when adding a product to the basket by satisfying their need for diversity. Most importantly, the influence of recommendations increases as the purchase decision is about to be made, and sometimes even surpasses the preferences expressed through the fixed search criteria. Our results also show a need to find a good compromise between accuracy and diversity in order to increase quality of recommendations perceived by users.

**User-Studies** A strong part of this dissertation relies on experiments where we carried out user studies. We refer to these as *user studies* or *experiments*. All of the studies were taken by *real-users*. We refer to participants as being "real" users as they took part in an online evaluation, as opposed to offline studies where corpora of data are used, such as MovieLens or Netflix.<sup>7</sup> The fact that this work rests so strongly on user studies anchors it in

---

<sup>7</sup>The reader should keep in mind that these users were in all cases simulating a purchase process; no money was

users' perceptions and allows to draw very practical lessons. Such an approach had to be chosen because of the user-centric goals of this work. In order to truthfully understand users' perceptions, and to see what are the qualities that people see in a system, we could only rely on real-user studies, either with online websites or as in-depth lab-studies.

The experiments were carried out on three product domains: music or books or perfumes. Despite these domains being apparently different, they share three common characteristics which support our choice in using them. Firstly, they are all everyday consumer products, also called public taste products, which users are accustomed to. Secondly, they are low risk or low involvement products. Through their low price range, they can very easily be bought by users, without spending a lot of time thinking about the purchase, contrary to the acquisition of a car or an apartment [129]. Thirdly, these product domains are complex in features. They all three can be classified according to a high number of individual features.

**Design Guidelines** Through our numerous user-evaluations, we were able to derive a set of design guidelines which should be helpful for researchers wanting to design and develop their preference-based recommender systems. The guidelines cover crucial dimensions including user effort, purchase intentions, complex systems and usage of diversity in recommender systems.

## 1.4 Overview of the Dissertation

Following the presentation of the motivation behind our research, we hereafter introduce the primary elements that we have worked on throughout the thesis. Two main questions have driven our work. First of all we are interested in exploring the dimensions which users feel strongly about when discovering a recommender system, and leading them to accept recommendations, with a possible long-term adoption of the system. Secondly we are eager to show how diversity can be perceived by users as being just as important as accuracy in suggestions. The different chapters are summarised in the list hereafter and the schema of Figure 1.1 presents their organisation.

**Chapter 2** discusses in detail several research fields that exist in the current domain of recommender systems. We expose the state of the art of recommendation techniques, presenting the main differences between collaborative filtering and content-based recommendations. Both approaches are used on several occasions throughout the thesis. Time is also taken for introducing main implications of critiquing-based recommenders. Remaining references, and more technical references are directly presented at the beginning of chapters when needed.

**Chapter 3** presents the first two experiments of this thesis. We arrange two user studies where we compare two music recommender systems. Both experiments focus on user issues which attract them, and lead them to accept (to listen to) the suggested recommendations.

---

spent for real. In order to maximise truthfulness of behaviours, incentives were always proposed, as is usually done in other studies in the field.

The adoption of the system in a long-term perspective is briefly surveyed in this part. The chapter also presents our research model on acceptance, and compares the obtained results with Davis *et al.*'s well known Technology Acceptance Model, revealing that it can be successfully adapted and applied to such recommender systems.

**Chapter 4** introduces a comparison of a traditional user-controlled interface with a more recent personalised system using recommendations. We establish a direct comparison for users between a behavioural recommender, which has implicitly gathered user profiles, and between a common search & browse mechanism.

**Chapter 5** describes an experiment where we confront a content-driven approach with a layout-enhanced one. It presents a slightly different strategy than in previous chapters, seeking to generate broader reactions from users. The experiment is run on a dynamic-critiquing platform and presents our new visual critiquing, an iconised representation of critique features.

**Chapter 6** reports the two last studies of the thesis. Both are oriented around the importance of diversity in recommendations. The chapter describes the perfume recommender platform used, and how the Editorial Picked Critique method was employed. In the first experiment users encountered a layout vs. content experiment, like in the previous chapter. In the second study, users' interactions were recorded through an eye-tracker and their decision processes were analysed.

**Chapter 7** summarises all of the experimental results and derives a catalogue of design guidelines, associated with user effort, purchase intentions, dealing with complexity of systems and above all diversity. The guidelines concern critiquing recommenders, and online entertainment e-commerce systems. The chapter also discusses the more global repercussions of our results, synthesising the different research models considered throughout our experiments. We propose a diversity-model for maximising users' overall satisfaction, which expresses how accuracy and diversity should ideally be interleaved during a user's interaction with a system.

**Chapter 8** concludes with the main contributions of this thesis, and indicates the limitations of our work and on-going researches with the aim to further enhance our recommender technologies.

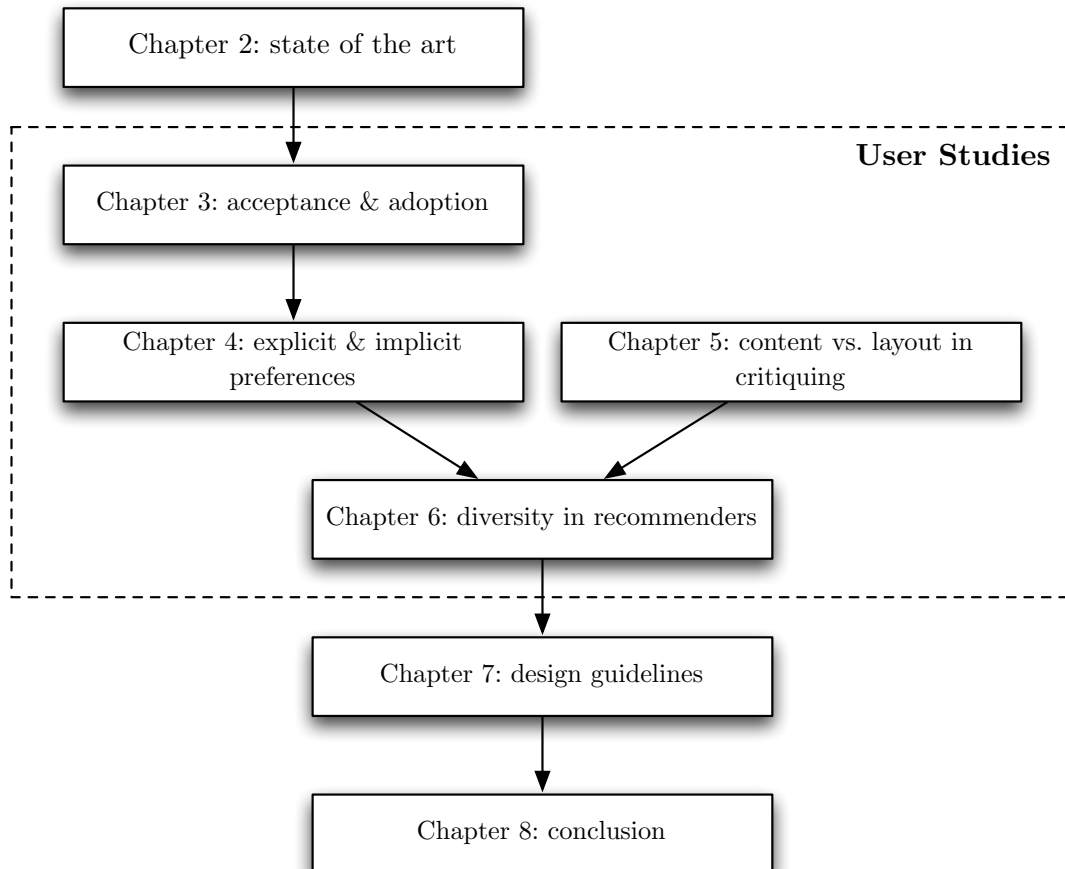


Figure 1.1: Overview of the organisation of the thesis.



## Chapter 2

# State of the Art

The following chapter covers general topics linked to recommender systems, and related to the experiments of this thesis. Specific technicalities are directly covered by related work sections in each chapter. Below, we first present the two most popular recommendation techniques, Collaborative Filtering and Content-Based Recommendations. We then detail the current knowledge regarding Critiquing, a specific extension of content-based recommenders, upon which we rely in several of our experiments. Finally, we present some work which has already compared multiple recommender systems from a user-centric perspective, in order to justify our approach in the user studies of this thesis.

### 2.1 Recommendation Techniques

There has been a great deal of research and literature produced about recommendation technologies. Much of which has focused on comparing them, especially regarding their technical and algorithmic performances. In this section, we present what we believe are the two main recommendation techniques: collaborative filtering and content-based recommendations. We chose these two techniques as they appear repeatedly in the experiments described later in the thesis. The descriptions first present the general concept of the recommendation techniques, before introducing some of the classical problems encountered. As a complement, we also present works on hybrid approaches which often alleviate the disadvantages of one approach with the advantages of the other.

#### 2.1.1 Collaborative Filtering Technology

The idea behind collaborative filtering (CF) can be summarised into one sentence: similar users like similar things. If we know for example, that *Billy* and *Jean* have similar tastes in life, and we also know that *Billy* just discovered a new song that he likes, then it is reasonable to imagine that *Jean* will also like this song. In this approach, user profiles are necessary. The usual method is to require from users to state their preferences by rating a set of items (e.g. products), which are then stored in a user-item rating matrix. This matrix is  $R = \{r_{i,j}\}$  and a couple  $r_{i,j}$  represents the rating value given by user  $i$  to item  $j$ . By this method, the similarity between two users

can be determined by their rating values on items. For each user, a set of neighbours can be determined, which have expressed similar preferences on similar items.

One of the earliest collaborative filtering recommender systems was implemented as an email filtering system called *Tapestry* [50]. Later on this technique was extended in several directions and was applied in various domains such as music recommendation [120] and video recommendation [62]. Possibly the most well known example of collaborative filtering can be found at Amazon<sup>1</sup>, where on detail pages, a “people who bought this item also bought” selection of recommendations can be found. It was one of the earliest commercial adoptions of this technique, and we use it in several experiments of this thesis.

It is usually accepted that collaborative filtering approaches can be classified into two kinds: those which are *model-based* and those which are *memory-based* as detailed by Breese *et al.* [21]. Memory-based collaborative filtering tries to foresee what vote the current user would give to an item based on the votes from some other neighbour users. Such algorithms operate over the entire user voting database to make predictions on the fly [114]. The most frequently used approach in this category is nearest-neighbour collaborative filtering: the prediction is calculated based on the set of nearest-neighbour users for the current user (*user-based* CF approach) or, nearest-neighbour items of the given item (*item-based* CF approach). The second category of collaborative filtering algorithms is memory-based. Such approach relies on the users’ voting database to estimate or learn a probabilistic model (such as cluster models, or Bayesian network models, etc.), and then uses the model for predicting what a user’s vote might be for a given item. Hybrid approaches between the memory and model based CF also exist such as Aggarwal *et al.*’s Horting collaborative filtering [2] and more recently Gong *et al.*’s combining approach [52].

User-based CF is the approach most often encountered in our work [114] and works as follows. The general prediction process aims at selecting a set of nearest-neighbour users for the active user based on a certain similarity criterion (such as the Pearson correlation [120]), and then to aggregate their rating information to generate the prediction for the given item. Memory-based approaches present the advantage that they are easy and evolve dynamically, since a profile update immediately changes the prediction calculation. However, they were rapidly found to scale difficultly [21] because of the algorithmic complexity and the high amount of memory needed for the databases. Item-based CF approach had originally been proposed to improve the system scalability [81, 117]. Item-based CF explores the correlations or similarities between items. Since the relationships between items are relatively static, the item-based collaborative filtering approach may be able to decrease the online computational cost without reducing the recommendation quality.

As users’ profiles grow, as their preferences become sufficient, collaborative filtering can yield good prediction accuracy as numerous studies have shown. Despite much effort to improve the accuracy of collaborative filtering methods [21, 97], several problems still remain unsolved. A classic problem in traditional collaborative filtering recommendation is what is often labelled the *cold start* problem. It represents the difficulty of making a recommendation to a new element in the system. When new users come to a recommendation website, the system is unlikely to recommend interesting items because it knows nothing about them. This is also sometimes

---

<sup>1</sup><http://www.amazon.com/>



called the new-user problem. There have been attempts to reduce this problem such as [108] where six techniques are studied which can help collaborative filtering algorithms to learn about new users. Another part of the cold start problem occurs with new items: when a new item becomes available in a database, the system will not be able to find users having rated this item and thus will not be able to recommend it. One more issue is the data sparsity problem. When there are too many items for users to rate, the user-item rating matrix is very empty and only a small number of ratings can be used during the prediction process. A promising techniques for reducing this problem was described by Zhang and Pu, who proposed a recursive prediction CF algorithm which allows those nearest neighbour users to join the prediction process even if they have not rated the given item. Lastly, collaborative filtering gives users little autonomy in making choices. When users deviate from the interests and tastes of their neighbours, they have little chance of seeing items that they may actually prefer among recommendations. We believe that this issue is important because it is related to the issue of *accepting* recommendations, one of the key points of this thesis. When recommendations based on a group of users are suggested, the user may not be prepared to accept them due to a low level of system transparency. Herlocker *et al.* have investigated visualisation techniques that explain the neighbour ratings and help users to better accept the results [59]. Later, Bonhard *et al.* [17] showed ways to improve collaborative filtering based recommender systems by including information on the profile similarity and rating overlap of a given user.

Beyond the technical questions such as cold start or accuracy of the approach, are ethical dimensions such as a user's privacy and preservation of intimacy. Despite these questions being crucial, as they are linked to the human rights topic and will consequently influence the wide-scale implementation and generalisation of recommenders in our everyday life, we will not directly address them in this thesis. An extensive overview of privacy issues in collaborative filtering techniques is discussed by Castagnos and Boyer in [25].

### 2.1.2 Content-Based Recommendations

Content-based (CB) recommendation technology determines an item to suggest to a user based upon a description of the item and a profile of the user's interests. This technology has its roots in information retrieval and information filtering, and rests strongly on textual information about items. For example, when making a recommendation, a content-based music recommender will try to recognise what aspects of music a user has liked (or disliked) in the past (e.g. what genre of music, if the songs are melodic, with heavy syncopation or prominent percussion) and recommend music that best matches those aspects. Generally speaking, content-based recommenders must address two challenges:

- how to represent items textually
- how to construct a profile that accurately represents user preferences

Depending on the domain, item descriptions can be structured, unstructured or semi-structured. Structured items are usually stored in a database where each item is described in terms of a finite number of features (also called attributes) and there is a known set of values that each feature

may have. Machine learning algorithms can be employed to learn a user profile from item selections by analysing which features and values the user prefers. To the opposite, unstructured items are described by plain textual information. Typically in this case the unstructured items are converted into structured ones before the recommendation process. For example, an item could have a list of boolean features indicating whether some particular keywords are included or not. Semi-structured items are in between structured and unstructured data. For instance, a mp3 music file is a semi-structured item: it has some header fields containing the basic information such as the title or singer, and some unstructured music data. In this case most likely we still need to convert the unstructured data into some kind of structured features before recommendation process. In the case of *Pandora*, a content-based music recommender system, professional musicians are paid to classify music into a large set of features. This system is used and detailed in Chapter 3.

It is often considered that the *FindMe* restaurant recommender was the first content-based recommender system [23]. Since, content-based recommender systems have been successfully developed to recommend items in various domains such as news articles [16], television programs [126], computers [112, 113, 139] or even web-graphics [133].

One requirement of this type of recommender system is that all items must first be encoded into a set of features. In most electronic catalogues used in e-commerce environments, products are encoded by the physical features such as the processor speed or the screen size in the case of portable PCs. This has significantly alleviated the time-consuming task of encoding the item profiles. Due to the difficulty of such tasks, the general belief is that CB recommender systems are not feasible for domains such as music or perfume, where items are not easily amenable to meaningful feature extraction. Another shortcoming of such systems is overspecialisation: users tend to receive recommendation that are limited to the preferences that they have specified. However, several researchers have developed techniques to overcome this limitation by considering proposing attractive items that users did not specify (called suggestion techniques) [105] or how to add diversity among the suggested items such as [86, 144]. The later, diversity, is discussed and studied at length in Chapter 6.

### 2.1.3 Hybrid Recommender Systems

Both collaborative filtering and content-based recommender systems have their respective advantages and disadvantages. There are not equally suitable for every domain or recommendation scenario. Often the strengths of one technique are offset by its weaknesses or limitations, and the research community rapidly started considering mixing both approaches. Hybrid recommender systems attempt to leverage the power of multiple techniques in order to improve the overall prediction accuracy of recommendations made to users.

Balabanovic and Shoham described the *Fab* digital library project at the Stanford University [12]. It is a content-based collaborative recommender that maintains user profiles based on content analysis, but uses these profiles to determine similar users for collaborative recommendation. Good *et al.* described a hybrid recommendation framework in [53] which combines information filtering and collaborative filtering. They differentiate these approaches by considering that information retrieval focuses on tasks involving fulfilling ephemeral interest queries,

information filtering deals with tasks involving classifying new content into categories, and collaborative filtering pays attention to finding which items should be suggested, determining how much a users will like them. Based on this classification, they propose a hybrid framework where results from information filtering agents based on content analysis can be combined with the opinions of a community of users to produce better recommendations. Another kind of hybrids can be found with demographic recommenders. They have attempted to address some of the deficiencies of the content and collaborative approaches by avoiding the cold-start problem by assuming a set of preferences based on demographic data [92].

Adomavicius and Tuzhilin present an overview of the field of recommender systems where they include hybrid recommendation approaches [1]. They present a survey of the recommendation techniques that were usual in 2005 and include scenarios for combining these methods. Since, many more hybrid approaches have been invented or refined. One recent example was proposed by Al-Shamri *et al.* who presented a “fuzzy-genetic” approach to recommender systems based on a novel hybrid user model. Thanks to hybrid features, a new kind of user model was built, helping to achieve significant reduction in system complexity and sparsity whilst making the neighbour transitivity relationship hold. In their approach, the user model is employed to find a set of like-minded users within which a memory-based search is carried out. This set is much smaller than the entire set, thus improving system’s scalability.

## 2.2 Critiquing-based Recommenders

In this section we present the general concepts behind critiquing-based recommenders. Critiquing-based recommenders form a sub-category of content-based recommenders, and focus on users’ interactions with the system. We introduce this technique as several experiments rely on modified forms of critiquing. The experiment of Chapter 5 relies on dynamic critiquing, and both studies of Chapter 6 on Editorial Picked Critique, an extended form of example-critiquing.

In order to understand the evolution of critiquing techniques, we must look at its origins in preference elicitation. User preferences are a challenging topic. It would be very easy to think that users’ preferences could be elicited by simply asking users to state them. Many online search tools use a form-filling type of graphical user interface or a natural language dialogue system to collect such information. Until recently, users were asked to state their preferences on every aspect. For instance, if we consider the first kind of online plane booking systems, users were often asked to state the departure and arrival date and time, the airline company, the intermediate airports, and were given the impression that all fields had to be filled. Such approaches are called non-incremental, since all preferences must be obtained up-front. In order to understand why this simple-minded approach does not work, we turn to behaviour decision theory literature to grasp the nature of user preference expression. According to the adaptive decision theory [95], user preferences are inherently adaptive and constructive depending on the current decision task and environment. Due to this nature, users may lack the motivation to answer demanding initial elicitation questions prior to any perceived benefits [129], and they may not have the domain knowledge to answer the questions correctly. In other words, if a system imposes a heavy elicitation process in the beginning, the preferences obtained in this way are likely to be uncertain and erroneous.

Similar literature reveals that users' preferences are context-dependent and are constructed gradually as users are exposed to more domain information regarding their desired product [95, 96]. There is a well known example reported by Tversky *et al.* They asked subjects of a user study to buy a microwave oven [135]. Participants were divided into two groups of 60 users. In the first group, each user was asked to choose between an Emerson priced at \$110 and a Panasonic priced at \$180. Both items were on sale, and these prices represented a discount of one-third off the regular price. In this case only 43% of the users chose the more expensive Panasonic at \$180. A second group was presented with the same choices except with an even more expensive item: a \$200 Panasonic, which had a 10% discount from its original \$220 price. In this context, 60% of the users chose the Panasonic priced at \$180. In other words, more subjects preferred the same item just because the context had changed. This finding demonstrates that people are not likely to reveal their preferences as if they were innate to them, but construct them in an incremental and adaptive way based on contextual information.

### 2.2.1 Example Critiquing

Critiquing is a technique whereby users can very easily indicate and refine their preferences over one or several attributes of products in an e-commerce catalogue. Conceptually, it is a mechanism allowing users to transmit a feedback, which serves to tell the system how to refine the displayed selection of items. Hence the analogy with a spoken *critique*. This intuitive method allows users to convey a sufficient amount of preference information.

If users are often unable to input all their preferences precisely at one time, it has been observed in behaviour theory that people find it easier to construct a model of their preferences when considering examples of actual options [96]. These observations lead to the first known form of critiquing, called *example critiquing*. Example critiquing was first mentioned by Williams and Tou [137] in a new interface paradigm for database access, especially destined to novice users. They described it as follows: "To make a query, ... the user interactively refines partial descriptions of his target item(s) by criticising successive example (and counterexample) instances that satisfy the current partial description". This method is called example critiquing since users build their preferences by critiquing the example products that are shown to them. The general process requires users initially to state preferences on any number of attribute values that they determine to be relevant. From that point on, the system engages the user in successive cycles of "examples and critique". Typically, the system returns a set of example products to which the user replies by indicating feedback in the form of critiques such as "I like this laptop computer, but with more disk space". The critiques determine the set of examples to display next. Such interaction terminates when users are able to identify their preferred product(s).

Users can quickly build their preferences by critiquing the example products shown to them. As users only have to state critiques rather than preferences, the model requires little effort from users. Most importantly, the example critiquing paradigm appears to satisfy both the goal of educating users with available options and the goal of stimulating them to construct their preferences in the context of given examples.

Over the years, example critiquing has been used in two principal approaches by several researchers: those supporting product recommendation based on an explicit preference model, and those supporting product catalogue navigation.

In the first type of example-critiquing systems, each user feedback in the form of a critique is added to the model to refine the original preference model. An explicit preference model is thus maintained. An example of a system with explicit preference models is the *SmartClient* system used for travel planning [102, 130]. It shows up to 30 examples of travel itineraries as soon as a set of initial preferences have been established. By critiquing the examples, users state additional preferences. These preferences are accumulated in a model that is visible to the user through the interface and can be revised at any time. ATA [80], ExpertClerk [122], and the Adaptive Place Advisor [49] function similarly. The advantage of maintaining an explicit model is to avoid recommending products which have already been ruled out by the users.

In the second type of systems, the system first retrieves and displays the best matching product from the database based on a user's initial query. It then retrieves other products based on the user's critiques of the current best item. The technique is sometimes called *tweaking*, since it allows users to express preferences with respect to a current example, such as "look for an apartment similar to this, but with a better ambiance". Such critiquing was first introduced as a form of feedback for recommender interfaces as part of the *FindMe* recommender systems [22, 23] and is, as explained earlier, a content-based recommender. FindMe is perhaps best known for the role it played in the *Entrée* restaurant recommender. During each cycle *Entrée* presents users with a fixed set of critiques to accompany a suggested restaurant case, allowing users to *tweak* or critique this case in a variety of directions. For example, the user may request another restaurant that is *cheaper* in price or *more formal* in style or even *quieter* in noise. FindMe was actually a larger scope system applying critiques in multiple domains: *RentMe* for apartments, *PickAFlick* for movies and *CarNavigator* for automobiles. *Entrée*, the part dedicated to restaurants, was destined to the Chicago region and available as an online website. Users could start to navigate the site by either specifying a restaurant directly, or they could express their preferences through a set of restaurant features (i.e. cuisine, price, style, atmosphere and occasion). Once this was set, the system would return the best matching restaurant, which is *Legal See Foods* in the example of Figure 2.1. Below is displayed a recommendation which can be tweaked through the critics displayed at the bottom, allowing to search for a less expensive place for example. This two-stage process, revealing one's preferences at first before critiquing suggested items remains the commonly adopted paradigm.

### 2.2.2 Unit Critiques and Compound Critiques

The basic and simplest form of critiques is a *unit critique* which allows users to give feedback (i.e. increase or decrease) on a single attribute or feature of the products at a time [23]. It is a mechanism that gives direct control to each individual dimension. A unit critique is easy to represent as a button alongside the associated product feature value and it can be easily selected by the user. In addition, it can be used by users who have only limited understanding of the product domain. This simplicity can be put into perspective by considering how challenging value elicitation approaches can be: these must accommodate text entries for a specific feature value from a potentially large set of possibilities, via drop-down list for example. Furthermore, critiquing can be used by users who may only have a limited (or vague) understanding of the product domain.



Figure 2.1: A snapshot of Entrée’s main GUI.

These significant usability benefits are unfortunately offset by several issues. The main problem is that unit critiques are not very efficient because the feedback provided by users is rarely sufficiently detailed to sharply focus the next recommendation cycle: if a user wants to express preferences on two or more attributes, multiple interaction cycles between the user and the system are required and big jumps in the data space are not possible in one operation. For example, by specifying that they are interested in a digital camera with a *greater zoom* than the current suggestion, users are helping the system to narrow its search but this may still lead to a large number of available products to choose from. In contrast, if the users could tell the system that they are looking for at least a *200mm* zoom, this would most likely reduce the number of potential products more effectively.

To make the critiquing process more efficient, an alternative strategy is to consider the use of what we call *compound critiques* [22, 109, 128]. Compound critiques are collections of individual feature critiques and allow the user to indicate a richer form of feedback, but limited to the presented selection. For example, the users might indicate that they are interested in a digital camera with a higher resolution *and* a lower price than the current recommendation in one single compound critique. The compound critique would typically be *lower price, higher resolution* in

this case. The concept of multiple-feature tweaks is not fully novel in itself, as the seminal work of Burke *et al.* [22] introduced it in a two-step process. In the CarNavigator system, the *sportier* critique was actually increasing both the *acceleration* and *horsepower* features, while allowing for a *greater price*. Similarly if we consider our digital camera example, a *more professional* critique would certainly lead to increased *price*, *zoom*, *screen size* and *thickness* properties. Obviously, compound critiques have the potential to improve recommendation efficiency because they allow users to focus on multiple feature constraints within a single cycle, thus navigating in bigger and more precise hops across the product space. However, compound critiques were initially hard-coded by the system designer resulting in the users being presented with a fixed set of compound critiques in each recommendation cycle. These compound critiques may, or may not, be relevant depending on the products that remain at a given point in time.

### 2.2.3 Dynamic Critiquing

McCarthy *et al.* [84, 85] proposed a method of discovering the compound critiques dynamically through the *Apriori* algorithm [3, 4]. It considers each critique pattern as the shopping basket for a single customer, and the compound critiques are the popular shopping combinations that the consumers would like to purchase together. Based on this idea, Reilly *et al.* [109, 128, 110] have developed an approach called *dynamic critiquing* to generate compound critiques. As an improved version, the incremental critiquing [111] approach has also been proposed to determine the new reference product based on the user's critique history. We show in Figure 2.2 a snapshot of their prototype system based on this approach, with both some unit and compound critiques highlighted.

The Apriori algorithm is a data mining approach which is used in the market-basket analysis method [4]. It treats each critique pattern as the shopping basket for a single customer, and the compound critiques are the popular shopping combinations that consumers often purchase together. The Apriori algorithm is efficient in discovering compound critiques from a given data set. However, selecting compound critiques according to their frequency in the data set may lead to some problems. This approach can for instance reveal “what the system would provide”, but does not tell “what the user likes”. For example, in the digital camera domain if 90 percent of the products have a larger screen size than the current reference product, it is still unknown whether the current user likes cameras with large screens or not. If the users find that the compound critiques cannot help them find better products within several interaction cycles, they may be frustrated and give up the interaction process.

Zhang *et al.* [140, 142] proposed a new algorithm to generate compound critiques for online product search with a preference model based on the multi-attribute utility theory *MAUT* [70]. In each interaction cycle, their approach first determines a list of products via the model of preferences of the user, and then generates compound critiques by comparing them with the current reference product. In this approach, the user's preference model is maintained adaptively based on the user's critique actions during the interaction process, and the compound critiques are determined according to the utilities they gain instead of the frequency of their occurrences in the data set.

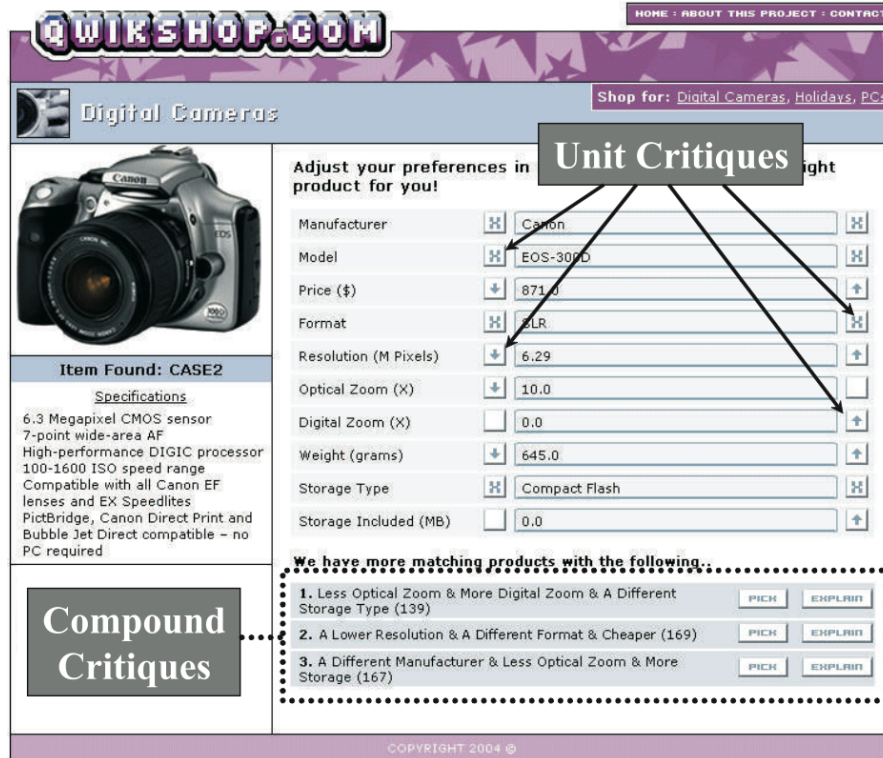


Figure 2.2: Snapshot of a prototype dynamic critiquing shop (Qwikshop). Highlighted are unit and compound critiques.

For a given user, this approach uses the weighted additive utility function to calculate the utility of a given product  $\langle x_1, x_2, \dots, x_n \rangle$  as follows:

$$U(\langle x_1, \dots, x_n \rangle) = \sum_{i=1}^n w_i V_i(x_i) \quad (2.1)$$

where  $n$  is the number of attributes that the products may have, the weight  $w_i (1 \leq i \leq n)$  is the importance of the attribute  $i$ , and  $V_i$  is a value function of the attribute  $x_i$  which can be given according to the domain knowledge during the design time.

The system constructs a preference model which contains the weights and the preferred values for the product attributes to represent the user's preferences. When the user selects a compound critique, the corresponding product is assigned as the new reference product, and the user's preference model is updated based on this critique selection. For each attribute, the attribute value of the new reference product is assigned as the preference value, and the weight of each attribute is adaptively adjusted according to the difference between the old preference value and the new preference value. Based on the new reference product and the updated preference model, the system is able to recommend another set of compound critiques. As the topic of this



this thesis is not directly concerned with the algorithmic implications of such approaches, we will not go into a higher level of detail about this approach. A more in-depth explanation of this approach to generating compound critiques dynamically is contained in [141].

## 2.3 User-Centred Evaluations of Recommenders

Recommender systems have not been the subject of many studies which take a user-centred approach, considering what qualities users perceive from such systems. This is even more true when considering comparisons of multiple recommenders. Yet it is more than ever important to not only consider giving good, accurate recommendations to users, but also to look at the situation from the user's point of view. Hereafter we briefly present the few works which were carried out on parts of the problem, often realised in similar experimental setups as those of our upcoming studies, and we explain how we position our research with respect to the current findings.

### 2.3.1 Taxonomy of Recommender Systems in E-Commerce

Schafer *et al.* [118] examined six e-commerce websites employing one or more variations of recommendation technologies to increase the website's revenue. They created a taxonomy of RS through the methodical reporting and analysis of the six websites' applications, interfaces, recommendation technology, and how users find recommendations for all of the example applications. Their classification of technologies was proposed along two main criteria: the degree of automation and the degree of persistence. The former refers to the amount of user effort required to generate the recommendations (manual vs. automatic). The level of persistence measures whether the recommendations are generated based on a user's current session only (ephemeral) or on the current session together with the user's history (persistent). It is a positive study as it is the first (to the best of our knowledge) to compare several recommenders, and does so extensively. Even though it was run in 1999, the main analyses are still sound today. It singled out the context of use (recommendation giving vs. seeking) as a critical dimension to characterise recommendation technologies, thanks to its automatic vs. manual classification, and it also compared content- vs. rating-based technologies to closely examine issues, including the new-user problem. Several works have extended proposed concepts, such as Cranor who included the ephemeral vs. persistent dimension as one of four axes of personalisation in a study of systems and their impacts on privacy [38]. Unfortunately, Schafer's study [118] is more of a technical comparison of recommenders, than a user-centric inspection. The research did not address design issues from a user's motivation-to-join or to-accept perspective.

### 2.3.2 Interaction Design for Recommender Systems

A few years later, Swearingen and Sinha [131] examined system-user interaction issues in recommender systems in terms of the types of user input required, information displayed with recommendations, and the system's user interface design qualities such as layout, navigation, colour, graphics, and user instructions. We believe this is the first work to directly compare

recommenders' design issues from the user's perspective. Six collaborative filtering based RS were compared in a user study involving 19 users in order to determine the factors that relate to the effective design of recommender systems beyond the algorithms' level. At the same time, the performance of these six online recommendation systems was compared with that of recommendations from friends of the study participants.

The main results were that an effective recommender system inspires trust in a system which has a transparent system logic, points users to new and not-yet-experienced items, and provides details about recommended items and ways to refine recommendations by including or excluding particular genres. Moreover, they indicated that navigation and layout seemed to be strongly correlated with the ease of use and perceived usefulness of a system. While the focus of this thesis is also on system-user interaction issues, we decided to split our work. First, we chose to focus on design simplicity, users' initial effort requirement and the time it takes for them to receive quality recommendations, in the first two experiments on acceptance and adoption (Chapter 3). Second, we explored more possible implications of layout and its importance in users' perceived ease of use and usefulness in Chapter 5.

More recently (after our Experiments 1 & 2 had been ran), McNee presented a thesis in which he underlined the importance of having a user-centric approach [87]. He explained how, in order to build relevant, useful, and effective recommender systems, researchers needed to understand why users come to these systems and how users judge recommendation lists. One of his key contributions was in showing how accuracy-based metrics cannot capture users' criteria for judging recommendation usefulness. He argues that we not only need to know about a user (i.e. about the user's profile), we above all need to know what the user is looking for. McNee explores how to tailor recommendation lists not just to a user, but to the user's current information seeking task, and proposes a new set of recommender metrics. In the process, he proposes a Human-Recommender-Interaction theory, which we discuss in more detail in Section 3.2, positioning it in relation to our work.

## Chapter 3

# User Attraction and Recommender Acceptance

### 3.1 Introduction

As explained in introduction to this thesis, recommender systems have become a popular solution to help users deal with the information overload. Unfortunately, what makes a website successful while others fail is currently more an art than a science. To begin the explorations of this thesis, we need to define two concepts, *acceptance* and *adoption*. These have been studied in several fields (as explained later) but have not been defined within research on recommender systems.

**DEFINITION: ACCEPTANCE** *The acceptance of a recommendation is the action or event whereby a user shows a willingness to consider a proposed recommendation. This most often takes the form of a click on the recommended item.*

**DEFINITION: ADOPTION** *The adoption of a recommender system is characterised by a user's decision to employ the system and the user's intention to return to the system in the future.*

Technology acceptance research investigates the conditions whereby users come to accept and use a website, a software system, or a technology. The history of this subject can be traced back to the 1970s and work relating to behaviours and attitudes towards new technological innovations [46]. With the arrival of computers, software, and lately websites, there has been a growing interest to develop a general framework of technology acceptance and apply it to specific domains. We decided to test the original *Technology Acceptance Model* (TAM) proposed by Davis *et al.* in 1989 [41]. Relying on it for devising our research model, we aimed at measuring and understanding users' experience with recommender systems. We detail the TAM later in Section 3.2. In short, this model suggests that when users are presented with a technology, a number of factors influence their decision about how and when they will use it, notably the perceived usefulness (PU) and the perceived ease-of-use (PEOU).

The understanding of usability issues of RS begins with an analysis of why users come to such systems for recommendations. As explained in Section 1.1, recommendation technology was historically used in recommendation-*giving* sites such as Amazon where the system observes users' behaviours and learns about their interests and tastes in the background. The system then proposes items that may interest a potential buyer based on the observed history. Concretely, users were *given* recommendations as a result of items that they had rated or bought. In this regard, recommendations are offered as a value-added service to increase the site's ability to attract new users and more importantly obtain their loyalty. In the last five years however, an increasingly large number of users have started to go to websites in order to *seek* advice and suggestions for electronic products, vacation destinations, music or books. They interact with such systems as first-time customers, without necessarily having established a history. Because the users come mainly for the recommender service, such RS must be more dedicated at making sure users accept and ultimately adopt their system. For this reason, we decided to work above all on the seeking paradigm in this first chapter. In order to explore the implications of the TAM in the field of seeking recommenders, we decided to rely on two music recommender websites, *Pandora* (a content-based recommender) and *Last.fm* (a collaborative-filtering social recommender). They are both music seeking recommenders and are presented in Section 3.3, where we also justify this choice. The two underlying technologies being central to recommender systems in general (collaborative-filtering and content-based recommendations), we presented them in detail in Chapter 2. We ran two user studies, where the music websites were compared side-by-side.

The contribution of this chapter lies first of all in a detailed understanding of user experience issues with recommender systems. We study users' initial attraction to a system, and elements shaping users' attitudes towards the initial acceptance of recommendations, & the eventual adoption of the technology. From this perspective, the outcome of these two studies is a broad range of observations about building effective RS which help achieving the aim of attracting new users. As some of the first works comparing two recommender systems which use different technologies in the field of recommendations, our studies aim to evaluate these systems as a whole and not solely reduce them to their background algorithmic nature.

Secondly, this research points out two important results. We show that overall user satisfaction in music recommender systems is not just a question of accuracy; it can be reached through several dimensions such as novelty. However the system's recommendation accuracy remains a crucial component as it is the only one which correlates with the intention to buy the proposed songs. Whilst highlighting new control dimensions for music recommender systems, the study's second results is that we show the necessity for low-involvement recommenders to be highly reactive. In this context, content-based recommenders appear to be most appropriate. These observations have stirred many decisions in this thesis and are central to the model proposed at the end of Chapter 7.

This chapter is organised as follows. We introduce our motivation, our research model and how it relates to the TAM. We then present the experimental framework: the two music recommenders, *Pandora* and *Last.fm*, which we compare in both of our user studies. Follow the descriptions of both experiments, where we include the user study settings, the analysis of results, and related discussions.

## 3.2 Motivation for Research on Acceptance of Recommendations

The Technology Acceptance Model, hereafter TAM, invented by Davis in 1989 [41], is an important model, which has become very influential over time, in several fields. As explained in the introduction, with the growing popularity that RS are getting, it is important to work on obtaining a better understanding of how users accept recommendations. In this perspective, testing the TAM on recommender systems is important. We selected this model as our baseline for two main reasons. First of all, as Section 3.2.1 explains, numerous improvements and variations of the original model have been proposed and tested, but each-time the same core elements remain essential, making it a strong model. Secondly the TAM is extremely simple, thus fitting a high number of situations. The combination of these two reasons led us to select it. The next section will highlight past work on elements leading to *acceptance* and will start by presenting Davis *et al.*'s Technology Acceptance Model.

### 3.2.1 Technology Acceptance

Computer technology acceptance by users is a topic which Davis *et al.* tackled already back in 1986, forging research works which have become very influential in the field today [41, 42, 136]. They introduced the Technology Acceptance Model, shown in Figure 3.1, as an adaptation of the Ajzen and Fishbein's Theory of Reasoned Action [7], which combines behavioural intention, attitude and subjective norm. The TAM hypothesises that perceived usefulness and perceived ease of use influence users' intention to use a system and eventually how they will use it. The factors are defined as follows.

**DEFINITION: PERCEIVED USEFULNESS (PU)** *Perceived usefulness is defined as the degree to which a person believes that using a particular system would enhance his or her job performance.*

**DEFINITION: PERCEIVED EASE-OF-USE (PEOU)** *Perceived ease-of-use is defined as the degree to which a person believes that using a particular system would be free from effort.*

These two dimensions, PEOU and PU, have since then been at the heart of the research on user related issues that lead to acceptance and eventually adoption. In 2004, Hassanein and Head wrote about building trust through socially rich web interfaces [58] (i.e. e-commerce websites) in a study where the TAM was used. Indeed, trust in an online shopping context is a complex issue which deals with a broad range of aspects. In a push to explore trust and its determinants, the research coupled the TAM with "social presence" and "enjoyment" as an interconnected network leading to trust. Previously, Koufaris and Hampton-Sosa had set the groundwork for similar research by examining the role of the experience with the website in customer trust online [73]. The TAM was also used as a baseline but augmented with the influence of "enjoyment" and "perceived control" on PU and PEOU, and linking the whole model to effects on "intention to return" and "intention to purchase". In both studies, results reinforced the importance of the TAM, while highlighting the fact that other dimensions can belong to the model, often as a catalysers of either PU or PEOU or both. Others like Ahn *et al.* studied the impact of quality

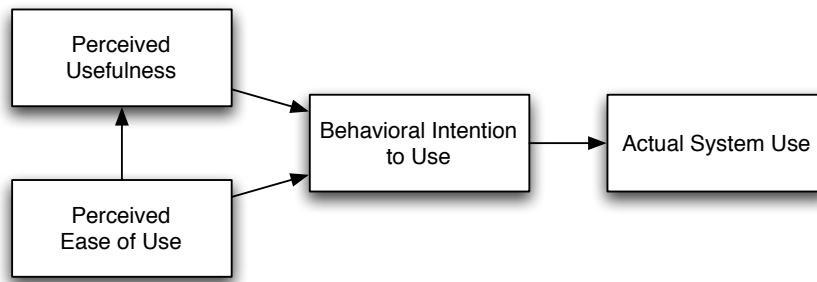


Figure 3.1: Davis's Technology Acceptance Model (TAM).

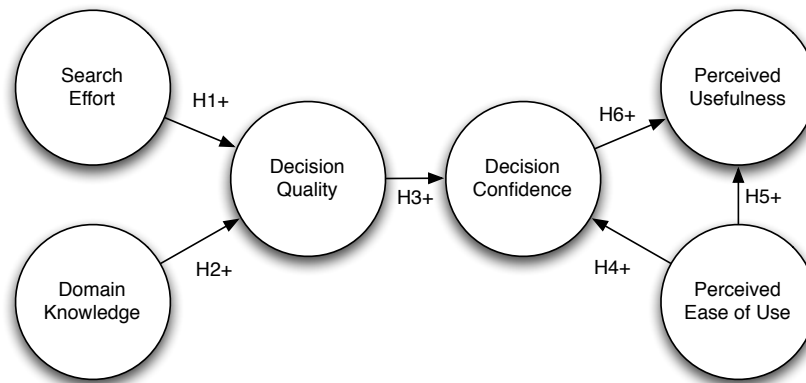


Figure 3.2: Illustration of Kamis and Stohr's research model.

and playfulness on user acceptance in online retailing, by making additions to the TAM [6].

A nice example of work on these potential components is [69], where Kamis and Stohr studied parametric search engines. The goal was to model the effectiveness of four parametric search engines in using search effort and domain knowledge to increase decision quality, decision confidence, PEOU and PU. Based on their previous work, they came up with a two-dimensional classification where subjective and objective elements were placed with respect to decision inputs and decision outcomes. They formed a research framework which encompassed decision process inputs (like search effort, domain knowledge) and decision outcomes (such as decision quality) and their respective evaluations. On this base they created a model where search effort and domain knowledge influence decision quality, which in return impacts decision confidence, itself being directly linked to PU and PEOU. In order to help visualising what such a model might look like, we present this model in Figure 3.2. Their research showed that search efforts and domain knowledge were mediated through decision quality and decision confidence, and that these impacted both the perceived ease of use, and the perceived usefulness.

Not all models are as succinct as the TAM. At the CHI 2006 conference, McNee, Riedl and Konstan proposed an analytic model for RS, which they named Human-Recommender Interaction (HRI) [89]. They proposed this framework and methodology, as part of [87], because they felt the need to obtain a deeper understanding of users and their information seeking tasks. Their model is not based on the TAM, it rests on three pillars of what they call the interaction process: the recommendation dialogue, the recommender's personality (perceived by the user over time) and the user's information seeking task. For each pillar, they come up with a range of dimensions that are thought to influence users' behaviour. In the case of our music RS studies, we will mainly be observing the user's interaction, part which in the HRI model is classified as the Recommendation Dialogue. The later is divided into eight elements such as correctness, quantity or saliency. Two elements, usefulness or usability, appear to us to be equivalent (or at least very similar) to our definition of PU and PEOU. This HRI model suggests that these two dimensions are far from being enough to model the main user-interaction dialogue and that they stand parallel to six other dimensions: serendipity, correctness, transparency, saliency, quantity and spread. We did not select this model as our research baseline, for the simple reason that it had not yet been published at the time when our first two experiments were designed.

The main observation about the results of all these studies on users' acceptance of technology and recommendations is the importance of very domain specific dimensions. However, and with the exception of the HRI model, the perceived ease of use and perceived usefulness stick out as key features. It was thus quite logically that we decided to use the TAM as our baseline research model in our experiments, whilst trying to find new domain specific elements. PEOU and PU are our two main pillars, then decomposed into smaller dimensions as explained hereafter.

### 3.2.2 Research Model

As highlighted in the previous section, many dimensions which influence a user's perception of a website and the potential acceptance of recommendations, are associated with the perceived usefulness and perceived ease of use initially proposed by the TAM. However the problematic of a user's interaction is very task or context specific where several different dimensions are essential. Instead of including PEOU and PU within other categories such as presented in other models (for example the HRI model of [89]), we propose to consider them as key categorisers and include several other dimensions within them, just like the TAM initially did.

In order to define a set of domain specific criteria, we reasoned around the meaning of PEOU and PU in the case of our two systems. We considered that a music RS is *useful* when the songs proposed are of good *quality*. We also believe that the *ease of use* of a RS website must essentially be a question of *effort*. It is this core classification that we use as model, in order to infer hypothetical domain specific dimensions, as shown in Figure 3.3.

The perception of quality in a suggested song is first of all influenced by the *accuracy* of the recommendation. We define the term accuracy as the adequacy of a recommendation with respect to a users' theoretical preferences. In return, we expect this accuracy of recommendations to impact elements such as users' enjoyment or satisfaction, as part of what defines a RS's quality. We also considered other dimensions like how diverse the songs were, if they were new to users and if they approved of the heard ones they already knew. We believe that such dimensions help the users in building their trust in the system. Less conventional dimensions

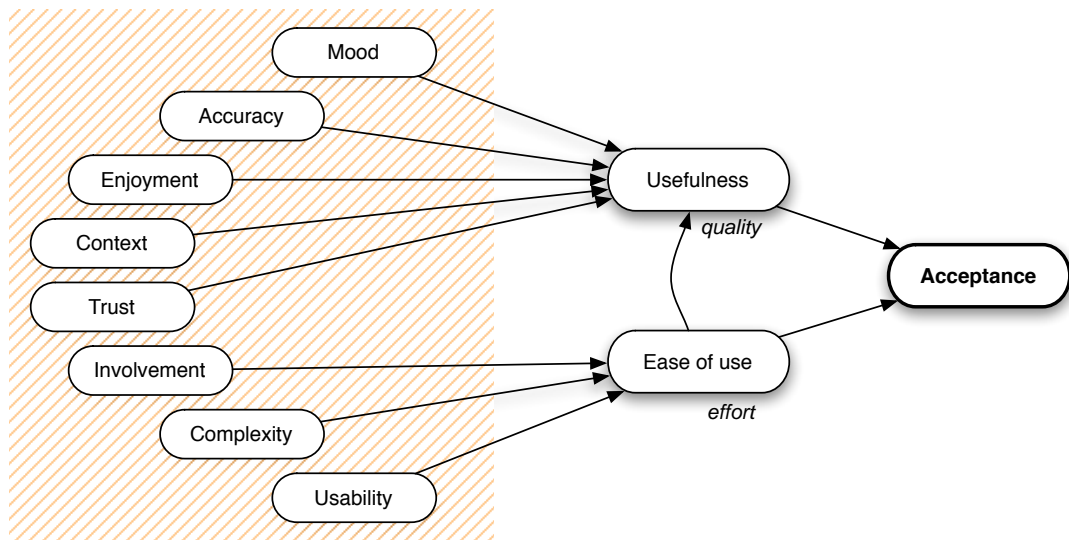


Figure 3.3: Dimensions of our research model.

might include mood: if songs are tailored to users' tastes and suit their mood, this could also clearly impact their perception of the RS's quality, and thus usefulness. Similarly, the context in which we listen to music is important in guiding our perception of quality. Finally, more global elements such as enjoyment and satisfaction are obviously part of what defines a RS's quality.

The effort required to operate in such a website can also be separated into several dimensions. The complexity of the system will impact the user's perception of effort and ease of use. Complexity in itself is not a sufficient measure, as a simple website can still suffer from usability issues which can make it very challenging to operate. Finally, the amount of involvement necessary to obtain the desired songs is possibly another influential factor.

We included all of these element in a research model highlighted in Figure 3.3. We are of the opinion that these features are all foundational characteristics that support PU and PEOU in music recommenders, and that they lead to the user's acceptance of a RS's recommendations. The exact questions that we asked to evaluate these multiple dimensions are detailed throughout Section 3.5.2.

### 3.3 Experiment Framework: Online Music Recommenders

We specifically chose music recommender systems to conduct our research based on the ease of accessibility to such systems. The following sections explains why we chose this product domain, and the advantages which it offers.



### 3.3.1 Why Music Recommenders?

Radio stations have existed for a long time on Internet. Back in the early 1993, three years after Tim Berners Lee invented the web [14, 15], Carl Malamud developed the first *internet radio* station called “Internet Talk Radio”. Rapidly more stations appeared and 1995 saw the first full-time only internet-radio appear, “Radio HK”. The worldwide broadcasting possibilities that internet offered attracted many people and completely independent radios appeared, differentiating themselves from traditional *terrestrial* radios. Today there are so many of these radios that you can hardly count them. But more interestingly, new paradigms have emerged such that the end users are starting to be able to personalise radio stations or even create their own. Online community services have started providing users with collective tools for putting music on their personal pages and it has proved a big hit. The social networking website *MySpace*<sup>1</sup> is an obvious example and was, according to *Compete*, the most popular social networking site (February 2008) [48].

Music was arguably the first type of rich media to really go *Web 2.0* in the sense that web-based communities rapidly emerged, and that they led to the creation of music folksonomies<sup>2</sup>. Music has become a very popular playground. As a result, there are some great Rich Internet Applications built around social music. In 2006, there were already more than eighty documented and well established social-music services such as MusicStrands<sup>3</sup>, Liveplasma<sup>4</sup>, Audiri<sup>5</sup>, Pandora<sup>6</sup>, Yahoo LAUNCHcast<sup>7</sup> or Last.fm<sup>8</sup>. This number just has not stopped growing since. Two of these already have several million users, proving how popular this field is. However, for the exception of a few singular systems, the majority still have very much in common in terms of functionalities, limitations and problems. Regrettably, a classification by genres of music is still dominant and popular bands are pushed forward whilst more obscure artists are still hard to discover[9]. These issues are resulting in some tangible frustration across the communities of music lovers and seekers.

The emergence of RS as a common tool has produced a real shift in the way we listen to music (and more generally perceive online entertainment, whether music, films, books, etc.). While highly captivating, this change is at the same time worrisome, as fundamental user-experience issues seem unresolved. Annoyance and disappointment are frequent feelings in online music threads, new services keep appearing and disappearing just as rapidly, and people continue to switch from one system to another, as if they had not yet found the system that fitted their needs. Beyond the points raised in Section 1.3 about our global framework of study for this thesis, the arguments listed above are the motivating factors for choosing to work on fundamental user issues in *music* recommender systems.

---

<sup>1</sup><http://www.myspace.com/>

<sup>2</sup>A folksonomy can be defined as a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorise content.

<sup>3</sup><http://www.musicstrands.com/>

<sup>4</sup><http://www.liveplasma.com/>

<sup>5</sup><http://www.audiri.com/>

<sup>6</sup><http://www.pandora.com/>

<sup>7</sup><http://music.yahoo.com/>

<sup>8</sup><http://www.last.fm/>

We see many advantages in our choice of the musical domain. These include:

- Music is an attractive and highly inspiring topic where it is easy to motivate users to be involved in a trial compared to other domains such as news articles.
- Music has a relatively short validation process to determine the quality of recommendation results. Compared to books and movies, a user can more quickly and easily determine if a recommended song is enjoyable, novel, pleasant, etc.
- Music items are available in large quantities and in a wide range of varieties.

By deciding to initially focus on low-involvement entertainment products, we gain an advantage since products carry a smaller financial commitment compared to most of the other entertainment-related commodities, such as electronic and travel products. They do however propose other challenges: because they are relatively easy to acquire, users are unlikely to spend much time choosing them, hence the name low-involvement products.

In choosing the systems to evaluate, we were encouraged in our choice by what was at that time, a recent blog publication by Steve Krause[74]. He had entitled his article “Pandora and Last.fm: Nature vs. Nurture in Music Recommenders”, and compared the features of *Pandora* and *Last.fm*, two music recommender systems that employ rather different technologies: *Pandora* is a content-based recommender and *Last.fm*, on the other hand, is a collaborative filtering recommender. The article started by opposing CB and CF recommenders, similarly to the nature versus nurture debate. The point was that algorithmically, *Pandora*’s recommendations are based on the inherent qualities of the music (its “genes” as *Pandora* call them) hence the analogy with nature. On the nurture side (where by nurture we mean what people around us have influenced), *Last.fm* is a social recommender, which knows little about the songs’ inherent qualities. However it knows who listens to the same music as you, and can thereby suggest songs. Beyond this nice analogy, Krause then considered how these differences might influence several interesting perspectives. He conversed about surfacing new artists, locked loops where people keep receiving the same kind of recommendations, or *Last.fm*’s delivery versus *Pandora*’s promise and finally *Pandora*’s unique features and possibilities. His inspired article finalised our choice, and we adopted the use of these two systems for our experiments. They serve our purpose of comparing recommender systems and evaluating their ability to attract new users, while employing two different technologies. We are in no way affiliated with either of the companies providing these systems. Both were contacted without success, in view of establishing a collaboration.

### 3.3.2 Pandora.com

*Pandora* is a content-based music recommender. When a new user first visits *Pandora*, a flash-based radio station is launched within 10-20 seconds. Without any requirement on registration, you are prompted to enter the name of an artist or a song that you like, and the radio station starts playing an audio stream of songs. For each song played, you can give thumbs up or down to refine what the system is recommending to you next. A snapshot is shown in Figure 3.4. You can start as many stations as you like with a seed that is either the name of an artist or



Figure 3.4: A snapshot of Pandora’s main GUI with the embedded flash music player.

a song. One can sign in immediately, but the system will automatically prompt all new users to sign in after the first fifteen minutes, whilst continuing to provide music. As a recognised user, the system remembers your stations and is able to recommend more personalised music to you in subsequent visits. From interacting with *Pandora* and in accordance with indications on its website, it appears that this is an example critiquing-based recommender, based on users’ explicitly stated preferences. As announced on their website, *Pandora* employs hundreds of professional musicians to encode each song in their database into a vector of hundred features. The system is powered by the *Music Genome Project*, a wide-ranging analysis of music started in 2000 by a group of musicians and music-loving technologists. The concept is to try and encapsulate the essence of music through hundreds of musical attributes (hence the analogy with *genes*). As an example, the first twelve features of the song “Billy Jean” by Michael Jackson, are (according to the website):

- pop rock qualities
- r & b influences
- disco influences
- danceable grooves
- a subtle use of vocal harmony
- extensive vamping
- a clear focus on recording studio production
- groove based composition

- minor key tonality
- prominent bass riffs
- an emotional male lead vocal performance
- subtle use of strings

The focus behind such a concept is on properties of each individual song such as harmony, instrumentation or rhythm, and not so much about a genre to which an artist presumably belongs. At the time of the studies, the system included songs from more than 10'000 artists and had created more than 13 million music stations.

It is conceivable that *Pandora* uses both content- and rating-based approaches. However, in the initial phase of using *Pandora.com*, we believe that the system operates in the content based mode, according to our experience and to several blog posts about the radio.

### 3.3.3 Last.fm

*Last.fm* is a music recommender engine based on a massive collection of music profiles. Each music profile belongs to one person and describes their taste in music. *Last.fm* uses these music profiles to make personalised recommendations by matching users with people who like similar music, and generate personalised radio stations, also called recommendation radios, for each person. While it is hard to know the exact technology that powers *Last.fm*, we believe that it uses user-to-user collaborative filtering technology from the ways *Last.fm* behaves and based on information on the website (and forums). In January 2006, *Last.fm* described itself as follows:

Join the social music revolution at Last.fm. It's fun, it's free, it's all about the music. You get your own online music profile that you can fill up with the music you like. This information is used to create a personal radio station and to find users who are similar to you. Last.fm can even play you new artists and songs you might like.

This description further supports our belief. It is a social recommender and knows little about songs' inherent qualities. It functions based on users' rating of items.

With *Last.fm*, a user interacts by first downloading and installing a small application, i.e. the music player. *Last.fm* also provides a plug-in for recording your music profile through a classic music player like *iTunes*<sup>9</sup>, but could not take feedback into account (at the time of the studies). After the download, the user needs to create a user profile which in return must be indicated to the player. You can then specify an artist's name, such as "Miles Davis". A list of artists that *Last.fm* believes to be neighbours from the same group as Miles Davis will then appear, and you can start to listen to an audio stream of songs that belong to that group. For each song, you may press a "I like" or "I don't like" button. It is also possible to specify a tag or a set of tags, such as "Indie pop", in the player's interface in order to listen to another suggested stream of audio. Additional features are proposed on the website, as shown on Figure 3.5, including detailed fact-sheets on each artist, popularity charts and discussion forums.

---

<sup>9</sup><http://www.itunes.com/>

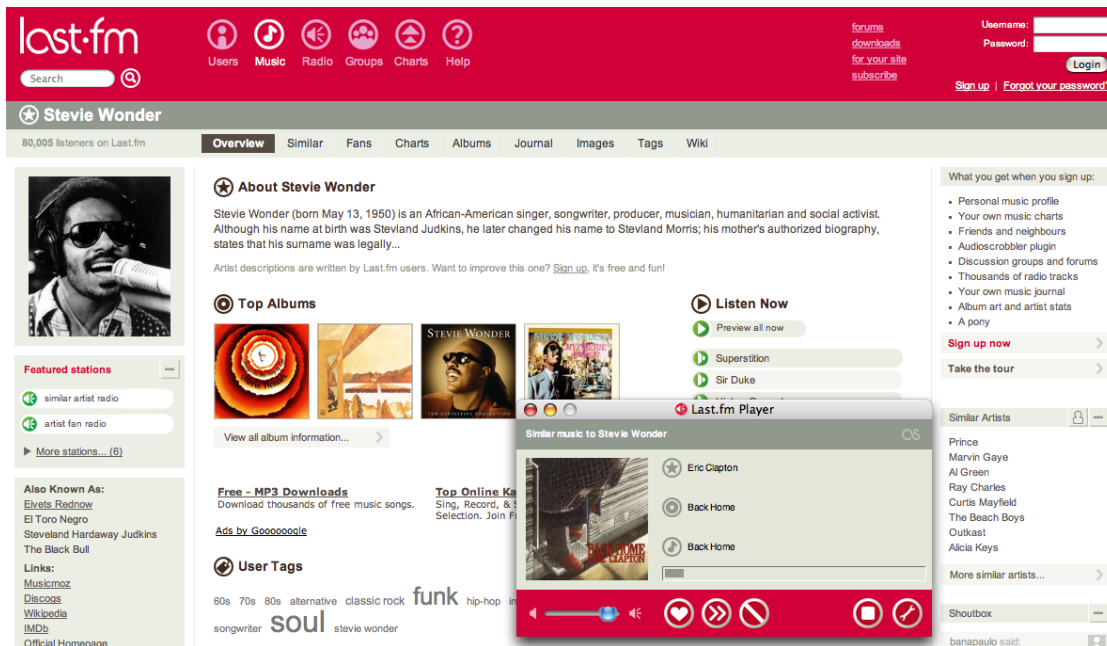


Figure 3.5: A snapshot of Last.fm’s main GUI, with the music player application in foreground.

Information from the *Last.fm* website indicates that after a few days, new users get a personalised recommendation radio based on their music profile. There is no precise timeframe indicated, but from our experiments we believe that the standard update-duration is around five days. According to most of our participants, the recommendations in the beginning were not very relevant to their input. However, the songs already became more interesting and closer to what they liked after several hours of listening to the radio stations created in *Last.fm*. It is possible that *Last.fm* uses some hybrid techniques to bootstrap the system in the beginning.

### 3.4 Experiment 1: User Acceptance in *Pandora* and *Last.fm*

For our first experiment, the two music recommender websites, *Pandora* and *Last.fm*, were compared side-by-side in a within-subject comparative user study involving 64 participants. We investigated users initial adoption of recommender technology and their subjective perception of the respective systems.

#### 3.4.1 Setup and Procedure

Precise instructions were given throughout the user study and are summarised as follows. In order to complete the experiment, students followed steps provided by a website. The main task was to setup and listen to one of the radio systems for one hour and then answer an online questionnaire of thirty-one questions. Background information was obtained through an initial

questionnaire of nine questions. One week later, students tested the other system with the same main questionnaire. At the end, seven preference questions were asked. These steps are detailed hereafter. The detailed material given to participants is detailed in Appendix A.

The same time limit was given to evaluations to ensure that results would be comparable. In order to maximise the chances that each student would not try to directly compare the two tested systems, *Pandora* and *Last.fm*, they were not informed that they would be testing both and were randomly assigned one system as their first assignment. They were further instructed not to share evaluation results with others. Subjects were first given a summary of what they would be doing (system names were kept hidden) before being directed to a website where the detailed instructions were presented and a system to test was automatically selected. The website was designed to accompany the students through the whole experiment, step by step. Since both systems functioned in different ways, we defined a common base of functionalities in which we would be interested, and decided to give testers detailed information in order to maximise possibilities of comparing the two music recommenders. We decided to focus on the music discovery process; participants were instructed not to use *Last.fm*'s social features, nor *Pandora*'s profile or backstage features. Here are the steps users encountered.

**Step 1** In order to make sure that students understood the goals of the experiment, they were provided with a summary of the tasks to complete and informed of technical requirements. The opportunity was taken to remind them to behave normally, with the intention of reducing outliers.

**Step 2** The users were then asked some initial questions about their background (gender, age group). Familiarity with internet & computers was assessed before targeting two aspects: the subjects' inherent attitudes towards music and recommendations.

**Step 3** Participants were presented with detailed instructions on the system they were about to test. The page included the list of tasks they should accomplish, timing indications and a checklist. In addition, a summary was created for both RS (summary based on the official texts provided online by *Pandora* and *Last.fm*), including precise details on how to get the application running and a short explanation on how to give feedback within the system.

**Step 4** Subjects were then told to start. They were expected to execute the necessary actions to get their designated system up and running, according to the instructions, before being left to listen to the suggested music during the remaining time. Finally, at the end of the hour, the website automatically redirected the students to the main online questionnaire.

**Step 5** The user is asked to fill in a post-stage assessment questionnaire to evaluate the system that was just tested. The questions are detailed in tables of Section 3.4.2.

**Step 6** One week later, participants were asked to evaluate the system they had not yet tested. Subjects were not asked to re-state their background, but otherwise the testing procedure was precisely the same.

Once finished, seven *preference* questions were added at the end of the main questionnaire. These were aimed at summarising the subjects' opinions and giving us their preferences on seven selected aspects of the tested radios. All questions in this study use five-point Likert scale.

In order to complement users' subjective opinions (recorded in the questionnaires), we tried to obtain some objective data. As we had no way of logging user's actions on both remote website, we handed participants a paper template, designed to help us log and analyse their experience with the system. Users could write down each song with its title and the name of the artist. Above all, they could indicate if a song was *new* to them, if they *liked it* or *hated it*, and if they would be prepared to *buy it online* given the opportunity (hereafter new, love, hate and buy).

### Participants' Background

Participants were mainly computer and communication science students in their third year at university. There was no financial incentive, but course credits were offered to ensure that the participants were serious about the experiment. We acknowledge that having only computer science students in the study possibly introduced a bias into the data. This was a voluntary choice; we believe that usability problems met by such qualified users can only be more challenging for less skilled computer users.

The participants' background was made-up of 62% from the 18-24 age group, 34% 25-30 and 4% were above. Subjects' preferred pastime seems to be "sport" for 32% of the cases, "reading" for 15% and "music" for 47% of them. Since both music RS allow users to buy music, students were initially questioned about their experience with buying songs through internet. Surprisingly, 88% of them had never bought a song or album through an online shop and 6% had only ever bought one song online. A few other questions aimed at determining users' affinity for music were asked, revealing that 44% of the students play an instrument and 30% of those consider themselves musicians (i.e. 13% of all subjects). This is encouraging, as the subjects will certainly be discerning in their testing thanks to their strong interest in music. We also tested subject's predispositions towards recommenders. When asked if they had any confidence in computers accurately predicting songs they would like, the subjects were surprisingly positive. 40% of subjects answered "maybe", 35% "cautiously yes" and 12% were "definitely" convinced that computers would be able to recommend songs with precision, leaving only 13% of users unconvinced. Before the study, only one person had heard of *Pandora.com*, and none of *Last.fm*. This assures that the results do not carry much prior bias towards these two systems.

### 3.4.2 Analysis of Results

We relied on three questionnaires for this study: one for the background profile of users, one for the main assessment after testing each system, and one for users' final preferences. As this was our first study, we chose a broad selection of questions (for the main questionnaire) in order to get a feel for users' opinions. The domains of questions covered the following themes: the effectiveness of RS in terms of its interface quality, the perceived quality of recommended items (accuracy, novelty, enjoyability) relative to its requirements on the users (time to register and download software, time to recommendation) and users' attitudes in adopting the underlying recommender technologies. Additionally, initial effort was investigated through a small set of mixed questions.

### Initial effort

A subset of questions were designed to measure users' task time in setting up the respective recommender systems (download and registration time) and the time it takes for a user to receive useful recommendations (time to recommendation). For *Pandora*, the time to get the flash plug-in and to register is around 2-5 minutes, although you may start listening to music without immediately registering. As for *Last.fm*, the time to download, install the audio player application and to register is 5-15 minutes. However, to get a personalised recommendation radio, a new user has to build up a profile and wait for an average of five days after the registration for that profile to be updated. To conclude, the initial effort required by *Pandora* is only a few minutes, whereas *Last.fm* requires more than few days for users to get started.

**Bias due to registration time?** It is clear that this *update time* of a few days means that testers in our study did not have the complete opportunities to enjoy *Last.fm*'s personalised radio recommendations. In this sense, it might have introduced a certain bias in the results. However, this time-limitation was voluntarily kept as it reflects the computational complexity of the underlying algorithm. More importantly, the experiment intended to evaluate acceptance and adoption mechanisms. If we had bypassed this aspect, we would have reduced the experiment's interest as it would not have fully represented a real-life scenario.

### Interface quality

Several questions were asked about users' experience with the interfaces of both systems. To the first question Q1, "how satisfied with the interaction are you", subjects were clearly more at ease with *Pandora* as indicated by the difference in means in Table 3.1 [*Pandora*: median=4 mode=4 | *Last.fm*: median=4 mode=4]. A T-test reveals that the difference in means is significant ( $p < 0.01$ ). Globally, users were satisfied with both systems as in both cases more than half of the subjects expressed a preference that was above the average score of the five-point Likert scale. However a solid 22.7% more users found *Pandora* excellent and in total 30% more users found its interaction above the average mark. Strikingly, only 5% found it bellow average, against 17% for *Last.fm*.

Subjects were questioned on what had worsened their satisfaction. *Pandora* users indicated two main reasons that were that feedback options were not detailed enough (3 users), and that there was no way of having multiple artists for one radio channel<sup>10</sup> (4 users). For *Last.fm*, the two main problems were installation difficulties (initial effort), and interface difficulties (not intuitive, not clear or not comfortable). The difference between the two systems is striking. *Last.fm* users mentioned fundamental usability problems which have complicated the usage of the system, whereas *Pandora* users talked about some secondary issues, not fundamental in making the radio work. Furthermore, for the first system only 7 people mentioned these issues, against 16 in the second case.

A certain number of other elements were considered for providing further explanation to this satisfaction difference, the first being *feedback*. The question "did you find it easy to provide

---

<sup>10</sup>This is not true, but clearly users did not find how to do it.



|                                |  | Mean (Std. Dev.)   |                |
|--------------------------------|--|--------------------|----------------|
|                                |  | <i>Pandora.com</i> | <i>Last.fm</i> |
| <i>Significant results</i>     |  |                    |                |
| Q1                             | How satisfied with the interaction are you? ( $p < 0.05$ ) | 4.1 (1.0)          | 3.4 (1.1)      |
| <i>Not significant results</i> |  |                    |                |
| Q2                             | Did you find it easy to provide feedback? ( $p > 0.05$ )   | 3.7 (1.0)          | 3.5 (1.0)      |

Table 3.1: Interface quality results.

feedback” does not help explain this difference. Indeed, users find both systems equally easy to operate for this topic [*Pandora*: median=4 mode=4 | *Last.fm*: median=4 mode=4]. The results are not significantly different ( $p = 0.228$ ). The same is true from the evaluation results of the first exposure, such that there cannot be any influence of the order in which the systems are tested ( $p = 0.464$ ). These results are not very surprising as both radio interfaces use similar and simple systems for providing feedback, whereby users can make a single click to show that they love or hate a song.

More feedback issues were investigated through a proposed selection of three reasons for feeling that the user could have discovered more music. Users were asked to select a label for each of the three suggested causes. The labels were “small”, “medium” and “big” problem for the three reasons: 1) feedback options being too limited, 2) hearing certain songs twice and 3) having enough time to listen. Graph on Figure 3.6 shows the results for both systems. Clearly there is some similarity between the results. The most visible difference appears to be that more *Last.fm* users found that limitations in feedback options were a “big problem”, and less so for *Pandora*. However this difference is relatively small. We believe that the feedback contrast might be an indication that *Last.fm* users did not feel that their feedback had sufficiently changed the music proposed. This is supported by the fact that both systems propose very similar actions for providing feedback. We therefore see no “a priori” reason for such a difference. Another small difference is that less *Pandora* users found the time limit to be a problem: we believe this small effect might be due to the ease of use of the interface, which just starts in a couple of seconds, even for a novice user. More results later in this chapter explore this concept as well.

The interface quality questions seem to indicate that *Pandora*’s ease of use makes it a more satisfying interface than *Last.fm*. The tools proposed for giving feedback are clearly easy to use, but some users obviously would like to give more detailed indications than just “I like” or “I don’t like”, possibly because they felt that their feedback did not influence the proposed music sufficiently.

### Subjective attitudes

The results on subjective questions are shown in Table 3.2. The first subjective question was “How enjoyable were the recommended songs?”. As the distributions on graph of Figure 3.7 highlights, a strong number of subjects gave *Pandora* a score of 4 out of 5, whereas the distribution for *Last.fm* is more centred, a bit like a Gaussian distribution [*Pandora*: median=4 mode=4 | *Last.fm*: median=3 mode=3]. Results are only marginally significant ( $p = 0.08$ ). It is interesting to observe that 67% of users gave *Pandora* an above average score (4 or 5) against only

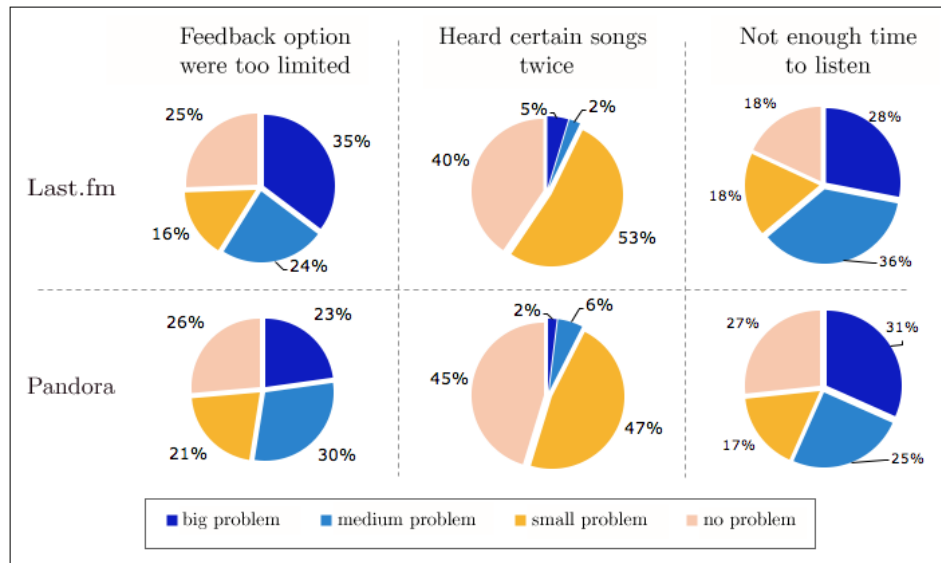


Figure 3.6: Pie-charts of users' reasons for feeling that they could have discovered more.

45.3% for *Last.fm*. In an attempt to gain more understanding about this observed difference, we normalised results by reclassifying it into three categories (instead of five), *above average* (4 or 5), *average* (3) and *below average* (2 or 1). This revealed that more *Pandora* testers enjoyed songs above average (43 vs. 29,  $p = 0.01$ ). We therefore believe that there is a higher level of *global* enjoyability for *Pandora*.

The questionnaire gave the subjects the possibility to explain what hampered their listening experience in terms of enjoyability. For both systems, the main reason mentioned was the poor quality of recommended songs as eleven users of *Pandora*, respectively nineteen of *Last.fm*, reported this as the main enjoyability problem. It seems obvious that a RS cannot always be “right”, and this 11-19 ratio seems reasonable. However the difference is significant ( $p < 0.01$ ) and tends to indicate that the first system provides better recommendation quality (under the constraints of the experiment setup); this result will be addressed further in this chapter. Another reason indicated by *Pandora* users, was that upon entering marginal artists or songs as a starting point, the system did not seem to have any such “data” on which to make recommendations. This problem was reported by four users and is a known issue with such systems, especially since each song has to be analysed and classified according to multiple attributes. Finally, a third main concern was expressed about proposed music sometimes being too similar, although all those who mentioned this point added that it was probably normal since it was the goal of the system. Including diversity in recommendations is a well know issue that many papers study [86]. *Last.fm* users also felt that marginal and obscure artists were a problem for the system (five subjects). Many other issues were mentioned for this system, each time only by one or two subjects, but nothing significant. Critiques go from “choice is too wide”, to “the lyrics are missing”, and mention the cold start problem “initial songs proposed were bad” or

| <i>Significant results</i>    |   | Mean (Std. Dev.)   |                |
|-------------------------------|---|--------------------|----------------|
|                               |   | <i>Pandora.com</i> | <i>Last.fm</i> |
| Q3                            | Was the system good compared to recommendations you may receive from a friend? ( $p < 0.05$ ) | 3.4 (0.8)          | 2.9 (1.0)      |
| <i>Marginally significant</i> |   |                    |                |
| Q4                            | How enjoyable were the recommended songs? ( $p < 0.1$ )                                       | 3.6 (0.9)          | 3.3 (1.1)      |
| <i>Not significant</i>        |   |                    |                |
| Q5                            | The system gave more personalised recommendations based on my feedback. ( $p > 0.05$ )        | 2.8 (1.0)          | 2.7 (0.9)      |

Table 3.2: Subjective variables results.

even “recommendations got worse over time”, for example. *Pandora* subjects also referred to the these last two defaults, and some hinted at “too much commercial influence” or “feedback options were too limited”.

These results give us a first indication that under this precise setup, *Pandora*’s recommendations might be better than *Last.fm*’s. They also tend to indicate that *Pandora*’s interface is easier to use and corresponds better to the users’ mental models, thus making their musical experience better. However, the enjoyability measure is not so clear-cut between the two systems in terms of satisfaction: we question whether this is not simply inherent to the music domain where any song, even randomly chosen, has a reasonable chance of being pleasing to the ears of the average listener. The following paragraph reinforces this statement.

The next subjective interrogation challenged participants to decide if they appreciated or really discovered music. The four possible answers were: “neither”, “appreciate”, “discover” and “both”. Figure 3.8 shows the distribution of users’ answers. It is striking to see that many more users both discover and appreciate songs suggested by *Pandora* rather than *Last.fm* and that this difference is significant ( $p = 0.049$ ). When considering the total number of subjects who selected “discover”, it appears that *Pandora* is significantly better ( $p = 0.058$ ) than *Last.fm*. In other words, *Pandora* seems to not only provide more new songs, but also new songs that people like. We believe this to be an important result.

One essential point of this study was to measure the quality in terms of perceived accuracy of the recommendations. In order to do so, the students were asked “Was the system good compared to recommendations you may receive from a friend?” Testers of *Last.fm* indicated a slightly negative emphasis, as their average was below the middle score on the five-point scale. On the contrary, users felt that *Pandora* was better than this middle score. [*Pandora*: mean=3.4 median=3 mode=3 stddev=0.8 | *Last.fm*: mean=2.9 median=3 mode=3 stddev=1.0]. The difference in means is significant ( $p < 0.01$ ) though both medians and modes are the same. The data is reported in Table 3.2 and shown on the graph of Figure 3.9, where a trend-line was added to facilitate the visualisation of the data distributions for both systems. *Last.fm* users are very centralised as in a Gaussian-like distribution, whereas *Pandora* users have a stronger concentration above the middle-score mark. We believe the significant separation in data frequency is an

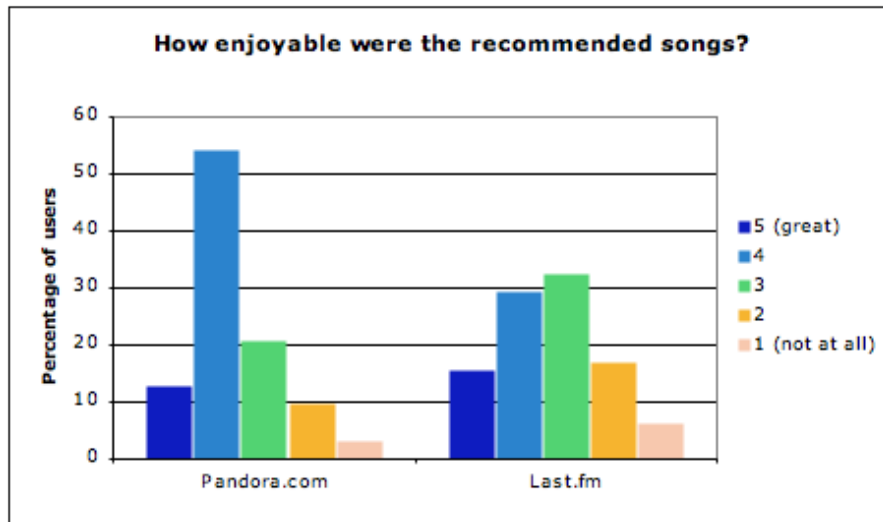


Figure 3.7: Graph of enjoyability of recommended songs.

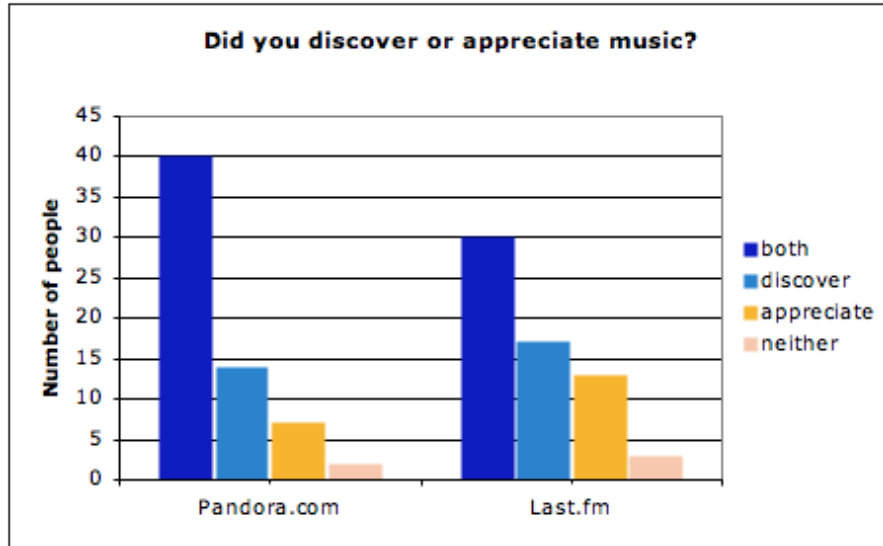


Figure 3.8: Graph comparing discovery and appreciation of music.

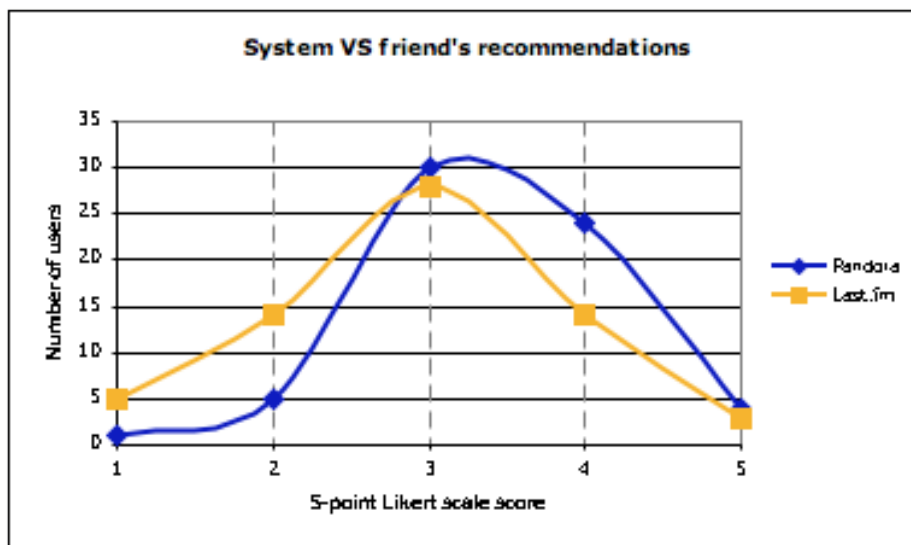


Figure 3.9: Distribution of appreciation of recommendations vs. friends' recommendations.

important measure because comparing system and user's (here a friend) recommendations is an indirect measure of the accuracy of the recommender system.

Average nb of songs, in 1 hour

| Nb of songs:           | Median (Mean)  |                |
|------------------------|----------------|----------------|
|                        | <i>Pandora</i> | <i>Last.fm</i> |
| people listened to     | 19 (20.1)      | 20 (20.4)      |
| people <i>love</i>     | 12 (12.5)      | 10 (9.8)       |
| people <i>hate</i>     | 5.3 (5.3)      | 6 (7.5)        |
| people find <i>new</i> | 13 (14.1)      | 12 (11.1)      |

Table 3.3: Objective measures from templates.

### Objective quality

The objective variables were aimed at obtaining impartial measures of what users listened to, and how good systems and their recommendations were. Users were asked how many songs they listened to, or how many songs they really loved. The templates that we collected gave precise indications on how many songs students listened to, for how long they listened to the music, which songs they loved and which they hated, if a song was new to them and if they were prepared to buy it. The duration of listening was used to adjust results on a linear scale so that all results were comparable on a one hour listening period.

Table 3.3 shows the results from the templates. The difference between the two systems for the number of songs a user is able to listen to in one hour is rather small as the averages of 20.1 and 20.4 demonstrate. The results are not significantly different ( $p > 0.1$ ) and the standard deviation for *Pandora* and *Last.fm* are respectively 4.0 and 7.8, which are high values. This result does not surprise us: as long as the songs played do not displease the users, they will listen to roughly the same number of songs in a fixed amount of time.

The next attribute considered was the number of songs subjects really *loved*. The results show a stark difference between the two systems, in favour of *Pandora*, where on median people liked two more songs per hour than with *Last.fm*. The results are significant ( $p = 0.021$ ) and even more evident since the mode for *Pandora* is 10, compared to 6 for the other system.

The next attribute was about novelty. Under the current setup, *Last.fm* seems not to be performing as well since its median value was one song fewer and the average number of songs played was three lower than *Pandora*. [*Pandora*: mode=10 stddev=5.04 | *Last.fm*: mode=6 stddev=3.65]. Again the results are significant ( $p = 0.022$ ).

The fourth main evaluated attribute was how many bad recommendations were made so that users came to *hate* a song. On average over one hour, *Last.fm* users hated two more songs than those proposed by *Pandora*. [*Pandora*: mode=6 stddev=3.17 | *Last.fm*: mode=4 stddev=5.60]. Results are marginally significant according to the p-value obtained ( $p = 0.061$ ). A deeper analysis of the results shows that *Last.fm* has more records of users hating a high number of songs and that these cases do not come from adjusted values, but are all from people who had listened for a full hour.

The templates were also used to evaluate how many songs the subjects were ultimately prepared to buy. The most frequent answer was “0”. However, a small number of people did show some interest in purchasing some of the music. If we only consider the first exposure to the two recommender systems, we can see that out of the 64 candidates, 29 (45%) said they would be ready to purchase a song online, with the median value of the non-zero answers being 2. The difference between the two radio systems when measured in this context is not significant.

### User preferences in systems

The final part of the study concentrated on obtaining the users’ final preference. This was done through a set of six related questions where the students had to choose between the two systems for each question. The results presented in Table 3.4 are significantly in favour of *Pandora*.

The first preference question asked subjects directly which system they preferred most, and the answer was very clear as 71% users preferred *Pandora*. A Chi-Square test of independence was computed (for all preference questions) to make sure that these results were not influenced by the order in which the students tested the systems, and the test is conclusive that there is no order-effect correlation since the p-value is high ( $p = 0.57$ ).

Further preference results were just as conclusive. When asked what interface users preferred in terms of music recommendations, 70% voted for *Pandora*. Again tests of independence show no order influence ( $p = 0.35$ ). And to the question “Which interface do you prefer to use as an internet radio?”, subjects had strong opinions again as only 38% voted for *Last.fm* ( $p = 0.16$ ). However, this result is not as clear-cut as the previous ones. Curiously, this happens on the question where a more social dimension is approached, an internet radio. Subjects

| Questions |  | Nb people      |                |
|-----------|--|----------------|----------------|
|           |  | <i>Pandora</i> | <i>Last.fm</i> |
| P1        | Which system do you prefer most?   | 71%            | 29%            |
| P2        | Which interface do you prefer for getting music recommendation?                            | 70%            | 30%            |
| P3        | Which interface do you prefer to use as an internet radio?                                 | 62%            | 38%            |
| P4        | Which interface inspires more confidence in you in terms of its recommendation technology? | 70%            | 30%            |
| P5        | If I want a recommendation in the future, I will be likely to use:                         | 66%            | 33%            |
| P6        | I felt comfortable using the following interface:  | 66%            | 33%            |

Table 3.4: Users' preference in systems.

previously indicated, very clearly, that *Pandora* had a better interface and was easier to use, two aspects clearly important for an internet radio. Despite that, they still seem to indicate that this is the best function for *Last.fm*: being an internet radio, in the more classical way. We believe this possibly comes from the vast amount of social manipulations that can be done on *Last.fm*'s website, such as writing blog entries, defining musical friends, leaving comments and many more similar actions.

One of the most important dimensions in these preference questions was to determine the quality of the recommendations. So we asked the students: "Which interface inspires you more confidence in terms of its recommendation technology?". 70% clearly designed *Pandora* as the most accurate system and ordering has no effect ( $p = 0.47$ ).

To the question "Which system would you use to get a recommendation in the future", participants designated *Pandora* in 66% of cases; ordering has no effect ( $p = 0.45$ ). The last preference question considers the interface design, one last time, through the word "comfort". Again, students vote for *Pandora* in 66% of cases; ordering has no effect ( $p = 0.32$ ).

We extended the analysis of the preference questions' results by computing the inter-rater-reliability amongst the answers. This is useful for determining how much homogeneity there is in the answers given by the users, therefore indicating if these six preference questions were perceived as representing different dimensions or not. The computed Intra-Class Correlation coefficient,  $ICC = 0.29$ , shows that there is no consensus across the questions.

Users' answers for the preference questions are highly in favour of *Pandora*. Whether considering the recommendation interface or simply the best system, the main trend of responses always points to *Pandora*. Furthermore this 30%-70% separation does not come from two groups of people voting exclusively for one system, but reflects user's diverse opinions on multiple criteria used to judge these two music RS.

### Correlation Analysis

Correlation analysis (Table 3.5) among the measured variables shows that enjoyability of songs, interface satisfaction, and the number of songs loved are the most important factors in predicting

| <i>Factors that predict RS quality</i>    | <i>Corr. r (sig.)</i> |
|---|-----------------------|
| Enjoyability of recommendations           | 0.760 (0.000)         |
| Interface satisfaction                    | 0.574 (0.000)         |
| No of songs subjects loved                | 0.315 (0.000)         |
| No of songs subjects were prepared to buy | 0.289 (0.001)         |

| <i>Factors that do not predict RS quality</i>      | <i>Corr. r (sig.)</i> |
|--|-----------------------|
| No of songs people listened to                     | -0.062 (0.484)        |
| Interrupted whilst listening                       | 0.019 (0.830)         |
| Trying other features whilst listening             | 0.035 (0.697)         |
| Would you have discovered this system on your own? | 0.127 (0.152)         |

Table 3.5: Prediction quality of recommendations, based on correlations  $r$ .

the relative quality of recommendations as being better than what the user may get from their friends. Interestingly but not so surprisingly, the number of songs subjects were prepared to buy correlates positively with recommendation quality. Analysis of users' detailed comments show that the main problems causing users dissatisfaction with *Last.fm*'s interface are the initial time required to set up the proper environment (initial effort), the time it takes to get useful recommendation (time to recommendation) and the fact that the interface is not intuitive and comfortable to use (simplicity). However it is not clear whether the high time to recommendation is a problem as such, or if it is above all linked to the other two dimensions as a kind of side effect. As for the shortcomings of *Pandora*, users wished that they could provide more refined feedback. Another deficient feature for *Pandora* was that users did not always find how to create a radio channel with more than one artist.

### 3.4.3 Discussion

This first study reveals that users strongly prefer *Pandora* to *Last.fm* as a general music recommender system. Beyond this general observation, the study points out that participants in the study are more likely to use *Pandora* again, prefer to use *Pandora*'s interface for getting music recommendations and as an internet radio, and perceive *Pandora*'s interface as more capable of inspiring confidence in terms of its recommendation technology. Moreover, users are generally more satisfied with *Pandora*'s interface, and found the songs that it suggested were significantly more enjoyable and perceivably better than their friends' suggestions. Finally, under this specific setup, users also loved more songs suggested by *Pandora* than *Last.fm*, found the songs more novel, and disliked fewer of the suggested songs from *Pandora*, albeit this might be partly linked to *Last.fm*'s inability to recalculate users' profiles in less than a few days.

Even though the direct intention to purchase music at both sites was not very significant, users expressed more intention to return to *Pandora*. An increased intention to return due to positive experiences gained on the initial visits is likely to bring revenue for the websites. Ac-



According to a 2006 marketing report by WebSideStory<sup>11</sup>, returning visitors are eight times more likely to purchase than first-time visitors.

This opening experiment provides a first understanding of how recommender websites attract new users as a result of the site design. Based on our result analysis, we are already able to distinguish three directions which seem promising in terms of users' perceived qualities. We could summarise them as follows:

1. The importance to minimise user effort, such as the time needed to register, to download and to get recommendations.
2. The possibility of maximising the quality of recommendations in terms of accuracy, but relative to a commonly shared measure such as friend's suggestions. This appears to help to increase enjoyability and novelty.
3. The need to maximise the interface's ease of use. According to our study, users clearly prefer such recommender systems and as a result are more convinced of its underlying technology.

As simple as they may appear, these initial findings show that focusing on the recommendation's technology alone is not enough to attract new users. An analysis of the website design and especially the human factor aspects are crucial in understanding users' technology adoption issues. Furthermore, the captured dimensions are highly similar to those highlighted by repeated studies by Forrester Research, such as [45], which stress that most important factors in user-web-interaction are ease of use (i.e. minimising user effort) and content quality (i.e. maximising recommendation quality). However, these findings do not yet tell us much about the acceptance of the recommendations themselves. For this matter we devise a more detailed study in Experiment 2, where we focus on the same dimensions as those highlighted in the TAM and our research model (see Section 3.2).

## 3.5 Experiment 2: User Adoption in *Pandora* and *Last.fm*

### 3.5.1 Setup and Procedure

We setup an in-depth between-subject lab study involving twenty participants where we, again, compared both music recommender systems. An in-depth study was favoured in order to first observe the users' interaction, and secondly to be able to discuss more fundamental issues that affected users. Due to the lengthy nature of the study, we decided not to perform a within subject study because of the potential fatigue risk that such an experiment would impose on our subjects.

The experiment was performed on a single machine, guaranteeing the same setup, conditions and material for each tester. Quality earphones were provided allowing each subject to feel immersed in the listening experience. The test was conducted in two distinct half an hour periods, separated by a fifteen day lapse. The interval was necessary since *Last.fm* requires a period of several days before it starts making personalised recommendations for new users. Although

---

<sup>11</sup><http://www.websidestory.com/>

*Pandora* does not require such an interval, we imposed it to all the users to make the experience with both systems comparable. At the end of each half hour, users had a quick oral interview to verify elements such as their first impression and any potential observations. To conclude the experiment, users had to answer a detailed online questionnaire of 28 questions (detailed in Tables 3.6, 3.7, 3.8) at the end of the second half hour. The detailed material given to participants is detailed in Appendix B.

In order to maximise users' comfort, each user interaction was observed remotely through a Virtual Network Computing (VNC) client, and directly encoded into text by a unique observer. The experiment machine was connected to a high-speed internet access and the VNC was adequately configured, to ensure that no bandwidth shortages occurred. The users were informed that their interaction was being observed.

The users taking part in this study received precise written instructions on the tasks they had to complete. The experiment was divided into the following steps:

**Step 1** Users are asked to fill in a six-questions questionnaire about their background profile (gender, age, etc.). An outline of the user study is provided, informing users about the topic of the experiment: music.

**Step 2** Users are then informed of which system they will be testing. They are given a scenario to follow, rather than a set of detailed tasks to complete.

**Step 3** The goal of this step is to create an account and to get used to the system. When the users feel comfortable, they may stop. A rapid interview wraps up this part.

**Step 4** Fifteen days later, the users are invited to come back. They are given another scenario where the goal is to get some recommendations for discovering new music in a half hour session.

**Step 5** To finish, each user is guided to the main questionnaire, before having a final short debriefing interview.

All 26 assessment questions in this study (except Q20 & Q22) are statements to which a users can indicate their level of agreement on a five-point Likert scale, ranging from 0 (strongly disagree) to 5 (strongly agree); 3 is neutral. Four questions used a reversed scale (Q21, Q23, Q26, Q27). The questions for *quality*, *effort* and *acceptance* were asked all at once, but are respectively shown in Tables 3.6, 3.7, 3.8 and detailed directly in the results' section for greater readability.

### **Participants' Profiles**

In order to have a balanced study which covered different types of interactions, two kinds of users were selected: the first half were computer and communication science Ph.D. students, and the second half were non-computer science people of university level and who used a computer regularly, making the range of users from normal to expert. Because of the slightly advanced nature of the topic, no beginner users were selected. A financial incentive was proposed to

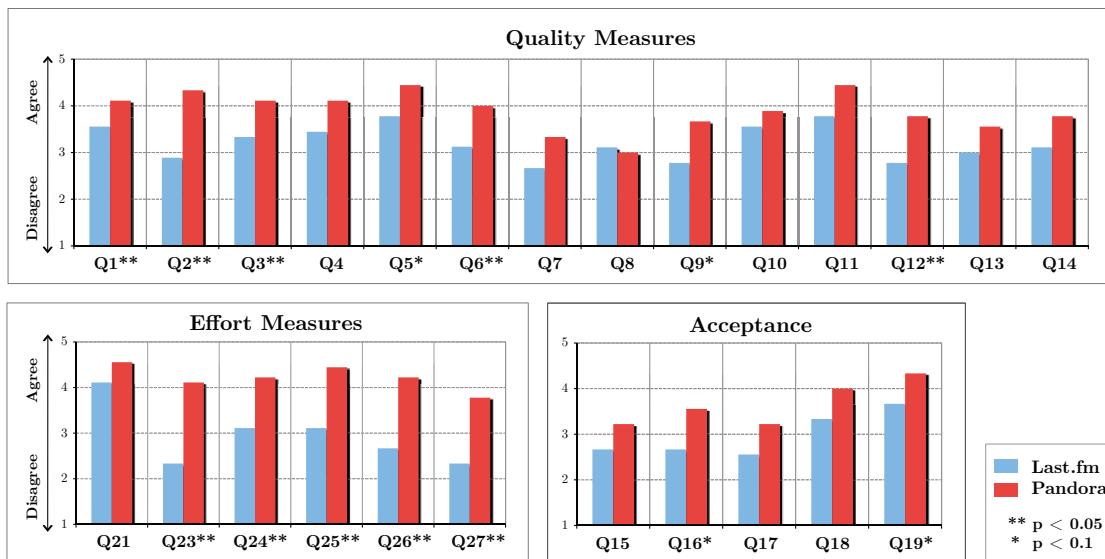


Figure 3.10: Agreement levels to statements of the final questionnaire.

ensure that the participants were serious about the experiment, and all subjects took part in a draw to win a high value present. With the exception of one user, all had never heard of or used *Last.fm* and *Pandora* ensuring that the selected participants represented sound first-time user experiences. People were randomly attributed which system they would be testing in order not to introduce any bias.

Seven of the twenty participants were female. Most users were in the 25-30 age group, with two in the 18-24 range, and three in the 30-40 range. As a base measure of the users' affinity for music in general, we chose to ask a subjective question to assess the users' own perception of their bond with music: we questioned them about the size of their personal music collection. Most of the subjects had an average collection, except for three who thought theirs was small and three who qualified theirs as being large. We operate under the assumption that there is no bias in terms of the users' connection to music, so the uniform and centred distribution of this measure is a positive factor for the quality of our results.

### Experiment Unfolding

As this study was carried out on systems that we did not control, we paid attention to a few issues. The experiment was carried out over one month. In this lapse of time, both services remained constant throughout the experiment, with respect to features tested by users. An administrative problem was encountered when *Pandora* changed its licensing terms, restricting access to U.S. citizens only (through the detection of the connecting IP address). At the time, five *Pandora* users had not yet finished the experiment. A temporary solution was found through the usage of a VPN access in America.

During the pilot studies, we noticed that users were at ease following our guidelines in both

systems. We thus decided to allow users to continue to use the music recommender system on their personal computers during the fifteen day interval. This gave us the opportunity to take into account some emotional mechanisms that occur when users discover a new service and become more familiar with it.

The musical profile of *Last.fm* users was examined after the first session to verify how many songs had been recorded. This was necessary because the system only considered that a song had been *listened to* if it was heard for the shorter of three quarters of the song length or 3 minutes. The natural behaviour of first time users was to skip through songs in order to better understand and explore the proposed features. This situation, detected in the pilot studies, led us to define a baseline before allowing users to come back for the second session: at least five songs must have been recorded to their profile. When this was not the case after the first session, we asked the involved user to listen to a few more songs at home, and ensure that at least five songs had been saved to the profile.

### 3.5.2 Analysis of Results

In this section we report the results from the questionnaire, through the three selected dimensions: *quality*, *effort* and *acceptance*. We then present the results from correlation analysis before discussing further observations. Statistical significance was computed for each questions with a T-Test, two-sample assuming equal variances. We report strongly significant ( $p < 0.05$ ) but also marginally significant ( $0.05 < p < 0.1$ ) results because this is a lab-study, and the post-interviews supported these values.

#### Perceived Usefulness - Quality

| Perceived Usefulness - Quality |   |
|--------------------------------|---|
| Q1                             | The songs recommended to me were enjoyable.   |
| Q2                             | The songs recommended to me suited my mood.   |
| Q3                             | The songs recommended to me were tailored to my taste.                                |
| Q4                             | The songs recommended to me were novel.   |
| Q5                             | The recommended songs that I already knew are songs I like.                           |
| Q6                             | In general, I am satisfied with the songs recommended to me.                          |
| Q7                             | The recommended songs are as good as those I would receive from my friends.           |
| Q8                             | Too many recommended songs were similar to each other.                                |
| Q9                             | The system's recommendation technology is accurate.                                   |
| Q10                            | The system has enough music to propose as recommendations.                            |
| Q11                            | I like the fact that the system elicits preferences from me.                          |
| Q12                            | I am able to influence the quality of recommendations through my preference feedback. |
| Q13                            | The system understands my musical taste and preferences.                              |
| Q14                            | I am able to determine how the system recommends music to me after using it.          |

Table 3.6: List of questions addressing Perceived Usefulness - Quality.

The questions assessing *quality* are listed in Table 3.6, and the results are shown in the first graph of Figure 3.10. The first surprising observation is that for each question where the

difference between the two systems is statistically significant, the distinction is in favour of *Pandora*. Still, overall the answers are above 3, the average value of the agreement scale, which is a positive indicator that quality is an important dimension for users.

Possibly the most striking significant difference can be noted for Q2 where *Pandora* seems to be much better at providing songs that “suit a user’s mood”. The two averages have quite a large difference of 1.4. During the interviews, several people spoke about their mood to justify a certain interaction, positive or negative, with the tested system. Another notable difference can be seen with Q12 which checked with participants if they were “able to influence the quality of recommendations through their preference feedback”. When asked in the post-study interview if they felt their feedback was taken into account, several *Last.fm* users mentioned that some artists they had banned were proposed again, sometimes even within the next five minutes. Furthermore, several of them simply said that they had not observed any difference from their feedback.

One more remarkable difference can be seen for Q9. Users perceived *Last.fm*’s recommendation technology as being less accurate than that of its counterpart. Although only marginally significant, this is supported by post-study interviews where *Last.fm* users often reflected negatively on the accuracy of the system, contrary to several *Pandora* subjects who were curious to know more about how the system was achieving such good results.

The differences for Q1 and Q6 were also found to be significant. *Pandora* participants thought that the recommended songs were more enjoyable and more satisfying. For both systems, users considered that songs were novel and above the middle value of the Likert scale, but the difference was not significant. We did not find this very surprising since playing songs randomly, with no personalisation, can also guarantee novelty. The quality dimensions which could help explain the users’ higher enjoyment and satisfaction with *Pandora* might come from Q3 and Q5. With Q3, users claimed that songs were better tailored to their taste, and in Q5 they stated that the recommended songs that were already known, were songs that they liked. For Q5, the difference was only marginally significant.

### Perceived Ease of Use - Effort

| Perceived Ease of use - Effort |  |
|--------------------------------|--|
| Q20                            | How long did it take you to register?  |
| Q21                            | The registration process required too much effort. ( <i>reverse</i> )                                    |
| Q22                            | How long did it take for the system to initially propose the first few enjoyable songs?                  |
| Q23                            | The initial time it takes for the system to recommend interesting music is too long. ( <i>reverse</i> )  |
| Q24                            | The website was easy to use as an internet radio for listening to music.                                 |
| Q25                            | The website was easy to use as a recommender system for discovering music.                               |
| Q26                            | The website offers too many features which are not relevant to music recommendations. ( <i>reverse</i> ) |
| Q27                            | There were too many navigable links which made the website confusing. ( <i>reverse</i> )                 |

Table 3.7: List of questions addressing Perceived Ease of Use - Effort.

The questions for *effort* are listed in Table 3.7 and the results for the *effort* assessment are

shown in the first graph of the second line in Figure 3.10. These questions are also measured on the Likert scale, but the scale was reversed for Q21, Q23, Q26 and Q27. The general observation for this graph is that *Pandora*'s measures are very high, whereas those from *Last.fm*, to the exception of Q21, are average, if not sub-average.

All questions have statistically significant differences, with the exception of Q21: users from both systems found that the registration process did not require too much effort, since both averages were above 4 out of 5 (reverse scale), with no significant difference. This is coherent with textual answers from Q20. Furthermore, we believe that there are two reasons why users did not find the registration as requiring much effort, contrary to results in Experiment 1. First of all no beginner users were selected for this experiment, meaning users had already been confronted to this step in the past. Secondly the nature of this study was such that there was very little pressure on people at the first stage of the study, since we wanted to focus on more long-term effects. The biggest difference comes from Q23: *Last.fm* users found that the initial time for the system to propose interesting music was too long contrary to those from *Pandora*. Their difference between the systems' averages is 1.8. Q22 textually questioned users about this dimension. *Last.fm* people report times between 15 minutes and 3 hours, with even two users saying that they had not yet received interesting music at the end of the experiment. On the contrary, *Pandora* users report times of 3 to 12 minutes, with the exception of one user who could not find her preferred style of music (oriental).

The users were also more in agreement that *Pandora* was easy to use as an "internet radio for listening to music" (Q24) and as a "recommender system for discovering music" (Q25), with a bigger difference for this second question. It is further supported by a difference in median values, unlike Q24. Finally it seems that contrary to *Pandora* users, *Last.fm* subjects were in agreement that "the website offered too many features which are not relevant to music recommendations" (Q26) and that "there were too many navigable links which made the website confusing". During the interviews, several mentioned being confused on the website and did not know where and when to click or not. This observation was made independently of the users' level, normal or expert.

### Acceptance

| Acceptance |  |
|------------|--|
| Q15        | If a similar technology existed for recommending other things to me (books, movies), I would use it. |
| Q16        | I would like to own the recommended songs.   |
| Q17        | I would purchase the recommended songs given the opportunity.  |
| Q18        | I found this website useful for listening to music that I like therefore I will use it again.        |
| Q19        | I found this website useful for discovering new music that I like therefore I will use it again.     |

Table 3.8: List of questions addressing Acceptance.

The questions for *acceptance* are listed in Table 3.8 and the results for the *acceptance* assessment are shown in the second graph in line two of Figure 3.10. Of the three graphs, it is the one where the values are the lowest on average.

The differences for Q19 are only marginally significant, indicating that users found *Pandora* a bit more useful for discovering music than *Last.fm* and were hence more inclined to use it again, but for Q18 the difference was not meaningful. The mean scores for both systems reveal that users find them useful for listening to and discovering music. The scores for the other questions are not as high. People only half agreed that “if similar technology existed for recommending other items (books, movies) then they would use it” (Q15), and similarly they were not really prepared to purchase the songs given the opportunity (Q17). We find these results interesting as they tend to show that people find these recommendation techniques useful (for listening or discovering) but are still fairly hesitant when it comes to going one step further, like buying.

### Correlation Analysis

Though we only had data from twenty participants, we chose to perform a correlation analysis of results, in order to gain some deeper insight into users’ perceptions and behaviours. We considered all results from both systems together. We found numerous correlations  $r$ , from which we report a selection in the three following subsections. Correlations between the *effort* and *quality* dimension are in Table 3.9), those between *acceptance* and *quality* in Table 3.10, and with *effort* in Table 3.11. For enhancing readability, only statistically significant results are reported in the tables. Readers are reminded that correlations don’t imply causality.

*Effort & quality:* The strongest link comes from the ease of use of the website as an internet radio (Q24) which strongly correlates with Q1, Q2, Q6 and Q9. The ease of use for discovering music is also highly linked with Q1 and Q6. The initial time a system takes to recommend good music (Q23) is correlated with enjoyability and satisfaction (a long initial time reduces enjoyability and satisfaction). Surprisingly the number of features (Q26) and navigable links (Q27) correlate inversely with having songs suited to a user’s mood.

| Correlation $r$ of effort with quality |        |         |        |        |
|--|--------|---------|--------|--------|
|  | Q1     | Q2      | Q6     | Q9     |
| Q23                                    | .600** | -       | .782** | -      |
| Q24                                    | .692** | .595**  | .678** | .632** |
| Q25                                    | .680** | -       | .678** | -      |
| Q26                                    | -      | -.609** | .608** | -      |
| Q27                                    | -      | -.616** | -      | -      |

\*\* Correlation is significant at the 0.01 level (2-tailed)

Table 3.9: Correlation: PEOU (effort) with PU (quality).

*Quality & acceptance:* The enjoyment of songs (Q1), and having them tailored to one’s taste (Q3) are clearly the two factors that most influence *acceptance*. We found that Q1 was correlated with the wish to own (Q16), the PU for listening (Q18) and the PU for discovering (Q19). Q3 was found to be correlated with exactly the same dimensions. In parallel with enjoyment, satisfaction (Q6) was also highly correlated with PU for listening (Q18) and discovering (Q19), and Q11, the fact that a system elicits users’ preferences is also correlated with Q18 and Q19. Finally, the perceived accuracy of the technology is important as it is the only quality element that correlates with the intention to purchase (Q17). We believe that all these results support the idea that the

quality of recommendations is a key issue in the recommendations process that leads to user acceptance, and this in particular through generating a perceived usefulness.

| Correlation of quality with acceptance |       |       |       |       |        |
|--|-------|-------|-------|-------|--------|
|  | Q15   | Q16   | Q17   | Q18   | Q19    |
| Q1                                     | -     | .470* | -     | .555* | .679** |
| Q3                                     | -     | .499* | -     | .576* | .593** |
| Q4                                     | .549* | .568* | -     | -     | -      |
| Q6                                     | -     | -     | -     | .485* | .635** |
| Q9                                     | -     | -     | .513* | -     | .489*  |
| Q11                                    | -     | -     | -     | .612* | .507*  |

\*\* Correlation is significant at the 0.01 level (2-tailed)

\* Correlation is significant at the 0.05 level (2-tailed)

Table 3.10: Correlation: PU (quality) with acceptance.

*Effort & acceptance:* Not all of the *effort* questions correlated with *acceptance*, but those that did correlated strongly. A short initial time for generating interesting recommendations (Q23) correlates with PU for listening and PU for discovering (Q18 & Q19). It is even more impressive how Q25, the ease of use of the website as a recommender system for discovering music, correlates with all acceptance measures with the exception of Q15. We believe these results support our hypothesis that effort is a key issue in acceptance.

| Correlation of effort with acceptance |     |       |       |        |        |
|---------------------------------------|-----|-------|-------|--------|--------|
|                                       | Q15 | Q16   | Q17   | Q18    | Q19    |
| Q23                                   | -   | -     | -     | .486*  | .518*  |
| Q25                                   | -   | .582* | .500* | .768** | .721** |

\*\* Correlation is significant at the 0.01 level (2-tailed)

\* Correlation is significant at the 0.05 level (2-tailed)

Table 3.11: Correlation: PEOU (effort) with acceptance.

### 3.5.3 Discussion

Like in the first experiment, users strongly preferred *Pandora*: twenty-four of the twenty-five questions assessing users' preferences are in favour of *Pandora*, though some were not statistically significant. The following section takes a look at some of the reasons why *Last.fm* was outperformed, and discusses how this all relates to the TAM proposed at the beginning of the chapter.

The TAM postulates that the PEOU influences the PU, and that both influence the behavioural intentions to use a system. The results strongly support this linkage as the correlation between effort (PEOU) and quality (PU) is highly favourable since the two direct assessments of "ease of use" (Q24 & Q25) are very strongly related to respectively four and two main quality questions, including the two direct questions on "usefulness" (enjoyability and satisfaction). This result is in total accordance with the TAM in terms of the link between PEOU and PU.



The next links in the TAM, PEOU and PU with behavioural intentions to use the system (acceptance), are also clearly supported. Four PU questions (Q1, Q3, Q6 & Q11) and two PEOU questions (Q23 & Q25) present strong correlations with the acceptance questions Q18 and Q19. We believe that these results strongly indicate that the TAM is an excellent model for music RS. It captures in a simple way the core interaction dimensions in the acceptance process of such a system. Impressively, questions which directly assess the simple dimensions of this model, usefulness and ease of use, are systematically highly correlated. Although Davis *et al.*'s TAM is very basic and not recent, it still encapsulates the major and fundamental components leading to acceptance. This forms a very solid reason for relying on a model as close as possible to the TAM. One such as the HRI mentioned in Section 3.2 or our model proposing domain-specific dimensions, as highlighted in Figure 3.3, seem less obvious to support from the data acquired in these first two experiments. One possible matter of our model is that all specific dimensions are either linked to Ease of Use or Usefulness, rather than being directly interleaved, transversally. We indeed note many correlations between all these dimensions, supporting the idea that they often interleave on a one-to-one basis.

This having been said, and in accordance with points made in related work, there are some significant domain-specific elements which do not fit our model. In this study, a miss-fit can be seen with Q9: the perceived accuracy of the underlying algorithm is the only value which correlates with the intention to purchase the recommended song, given the opportunity (Q17). This is interesting because for both systems, the score of the perceived accuracy of the recommendation technology is below that of the average quality question. Yet, questions such as Q1 or Q5 and the post-study interviews reveal that overall users were satisfied with either system, and dimensions like novelty show good correlation with acceptance questions. Based on these results, we question whether in order to please users, the system only needs to have a “minimal” recommendation accuracy, which should of course take into account elements such as novelty (or diversity, as shown in [144, 86]). But in order to get users to go further in the acceptance process and actually buy songs, the system’s recommendation accuracy seems crucial, whilst maintaining an easy to use system. What is here only an inspired conclusion from some initial data, will be reinforced by results in Chapter 6, and used to produce the diversity model in Chapter 7.

Interestingly, *effort* results show that *Last.fm* users do not find the recommendations adapted to their mood (contrary to *Pandora* testers), and that the number of features (Q26) and navigable links (Q27) correlate inversely with having songs suited to a user’s mood. This is quite surprising as one could easily assume that by providing users with more tools to input their preferences, the system’s recommendations should get more precise thus closer to their current mood. HCI has always had to find the balance between control and ease of use. In the case of *Last.fm* it seems that they have gone over a “tipping point” where small features are actually hampering the end-user’s experience. *Last.fm* is clearly a successful website with more than ten million users. However, based on our results we believe that this does not primarily come from the recommender system, which clearly poses some problems, but possibly more from the website’s social features (which were not captured by this study).

One unexpected result is for Q2: *Pandora* seems to be much better at providing songs that “suit a user’s mood”. In today’s RS, the default mechanism for users to provide feedback is

based on providing a kind of *score* either on a rating scale, or as with these music RS as a positive / negative score (“I like this song” / “I don’t like this song”). However several studies in psychology, linked to music, have come up with different music classification schemes related to emotions. They propose new dimensions such as *arousal* and *valence* [77, 71]. The fact that the *mood* component is so prominent in our study supports the idea that feedback mechanisms should be more reactive, helping to take into account the context of recommendations. The default positive / negative feedback process does not seem optimal and other control dimensions could possibly be invented.

Beyond this potential future control mechanism, we believe that there are other reasons which explain why *Pandora* manages to “suit a user’s mood” so well. When asked in the post-study interview if they felt that their input was influencing the system, the users responded very differently: *Pandora* users answered “yes”, whereas most *Last.fm* testers said “no”. It seems that the responsiveness of algorithms might be playing an important role here. As highlighted in the State of the Art 2, recommender systems using collaborative filtering techniques are prone to encounter a computational challenge when having a large set of users, which is the case with *Last.fm* and its millions of users. Such an approach here leads to profile updates being calculated offline, apparently less frequently, whereas *Pandora*’s content-based RS seems to respond immediately to users’ input. Despite a longer experiment duration in our second study, both mainly tested the early adoption of such systems. It is therefore hard to differentiate between two possible explanations for this lack of reactivity. This might be due to scalability problems, but it is also highly likely linked to the cold-start problem. We believe that the important message here is that the reactivity of a RS is a key component in the users’ satisfaction.

Due to time related constraints, a deeper long-term analysis of adoptive mechanisms was not performed fully. Ideas of future work are addressed consequently in the last chapter of this thesis. We are confident that key dimensions highlighted in this Chapter are essential to the process which leads users to adopt a system. It is reasonable to imagine that adoption is a step which ensues quite naturally after a satisfying experience, especially if users accepted recommendations. We therefore believe that our results are applicable in the larger scope of adoption challenges.

### 3.6 Conclusions

In our first study, we compared the two music recommenders *Last.fm* and *Pandora* side-by-side in a within-subject user study involving 64 participants. Our initial goal was to investigate the attraction, acceptance and adoption of recommender technology and specifically users’ subjective perception of the respective systems as new users. The study highlights some of the reasons which led users to strongly prefer *Pandora* to *Last.fm* as a music recommender system. Users clearly prefer *Pandora* and at the same time are more convinced of its underlying technology. Through preference questionnaires, they told us that they were more likely to use *Pandora* again and prefer to use its interface for getting music recommendations and as an internet radio. They perceive *Pandora*’s interface as more capable of inspiring confidence in terms of its recommendation technology. Moreover, users are generally more satisfied with *Pandora*’s interface, and found the songs that it suggested were significantly more enjoyable and perceivably better than

their friends' suggestions. Finally, under this specific setup, users also loved more songs suggested by *Pandora* than *Last.fm*, found the songs more novel, and disliked fewer of the suggested songs from *Pandora*.

This opening experiment provides a first understanding of how recommender websites attract new users as a result of the site design. Based on our result analysis, we are already able to distinguish three directions which seem promising in terms of users' perceived qualities: minimising user effort, maximising recommendation quality in terms of accuracy relative to a common measure, and making easy to use interfaces. As simple as they may appear, these initial findings show that focusing on the recommendation's technology alone is not enough to attract new users. An analysis of the website design and especially the human factor aspects are crucial in understanding users' technology adoption issues.

In order to deepen our understanding of user experience aspects linked to attraction, our second experiment focused on domain specific elements leading to acceptance. Through an in-depth comparison, our second study allowed us to reveal key interaction features. Our results showed that users perceive *Pandora*'s recommendations as being more accurate, more suited to their mood, with songs more tailored to their taste, more enjoyable and more satisfying than those from *Last.fm*. In terms of effort, *Pandora* testers are equally satisfied with the initial time it takes for the system to recommend interesting music, and its ease of use for listening to and discovering music. In comparison, *Last.fm* users were less positive on several quality issues, and clearly unhappy with the initial time to reach good recommendations and the website's complexity, which proposed irrelevant features and was at times confusing.

Beyond this comparison, this study is one of the first to investigate acceptance issues in recommendation-seeking systems, for entertainment products. It reveals that F. Davis's initial Technology Acceptance Model [41] can be successfully adapted and applied to such recommender systems. Although the model is quite simple, we provide convincing data which supports that this straightforward model clearly suffice to capture the fundamental interaction mechanisms leading to acceptance. As a result, this study supports that the perceived usefulness of a RS, in terms of quality, and the perceived ease of use, in terms of effort, are directly correlated with the user's ultimate acceptance of the recommendations. Some of the domain specific dimensions we proposed in our model are backed up by results. Measures of quality such as accuracy, enjoyment, satisfaction and having music tailored to a user's taste are directly correlated with acceptance, and that measures of effort like the initial time to interesting recommendations and the ease of use for discovering music correlate with acceptance.

Our data also reinforced the idea that it is crucial to understand users' real preferences. For both systems, the participants strongly appreciated the fact that the system elicited preferences for them. However, *Pandora* was reported as being much better than its counterpart at tailoring songs to users' tastes, thus being more precise in recommending. This study also emphasised that customising songs in accordance with a user's mood is important. This link with the emotions is a new dimension in the recommendation process that only a handful of systems incorporate today.

Finally, this second experiment points out two important findings. First, the results show that overall user satisfaction in music recommender systems can be reached through several dimensions such as novelty. However the system's recommendation accuracy is, at this stage, the

crucial component as it is the only one which correlates with the intention to buy the proposed songs. Second, while highlighting new control dimensions for music recommender systems, the study shows the necessity for low-involvement recommenders to be highly reactive. In this context, content-based recommenders appear to have an advantage through their default implementation.

## Chapter 4

# Losing Control: Are Preferences Best Revealed Explicitly or Implicitly?

### 4.1 Introduction

In this chapter we compare traditional user-controlled interfaces with more recent personalised systems using recommendations. Previously in Chapter 3, two results caught our attention. First of all, the ambiguous importance of the accuracy in recommendations. Whilst an accurate suggestion is expected to be coherent with a user’s preferences, we showed that a “good” recommender (in the sense of user’s overall satisfaction) does not solely need to focus on improving its algorithmic accuracy. Users have broad needs that can also be satisfied, in parts, by adding for example novelty or diversity. Secondly, having less initial effort is important. This dimension is a relatively reasonable one, but it is surprising to see how strongly users perceived it. These two results led us to direct our research on comparing systems where users control their preferences versus those where their profile is gathered implicitly. Control should afford users a direct mechanism for ensuring that, within the system’s ability, recommendations are as accurate as possible. On the other hand, implicit schemes require, by nature, much less effort from users. This polarity highlights both of the dimensions revealed in our first two experiments, hence the great interest we show in it for extending our research here.

A lot of research has been done on ways for users to reveal their preferences, and experiments such as [119] suggest that when users implicitly give feedback, the performance of the RS can be close to the more traditional ones using explicit feedback. But the work is highly incremental and there are no studies directly comparing both extremes. This second reason pushed us further; we decided to evaluate how recommendations based on *implicit preference feedback* compare with results provided to users who explicitly reveal their preferences in a traditional *user controlled* way. We chose to conduct this study on Amazon, because it has a well-established RS, often cited in works of our community<sup>1</sup>. We set up a comparative between-group user-study where users were instructed to search for five books. One group of users tested Amazon without the benefit of the RS, by *searching and browsing*. This represented the baseline

---

<sup>1</sup>We have no affiliation with Amazon.

measure for the experiment. Two other groups tested Amazon’s recommendations which were based on implicit preference information: their past purchase history. One group had a small purchase history whereas the second group had a larger profile. The experiment was conducted online and users’ opinions were collected through a post-study assessment questionnaire, evaluating multiple dimensions from satisfaction to intention to return. Results show both approaches can yield similar overall results, as perceived by users. Specifically, implicit recommenders are perceived as being trust-worthy. Moreover people felt that these interfaces required less effort for finding items. At the same time however, they were just as satisfied as with more traditionally controlled interfaces, notably because diversity was found to be higher. Results further suggest how measures such as confidence, trust, control and intentions to buy evolve as the profile grows.

This chapter is organised as follows. We first provide a brief review of the related work about user preferences and feedback techniques in recommenders, followed by our hypotheses. Next we describe the set-up and procedure of the real-user study, before reporting the evaluation results. Finally we discuss our hypotheses and key elements highlighted by the study before giving our conclusions.

## 4.2 Background and Related Work

This chapter takes position in the trail of a longstanding debate of control and automation. Twenty years ago, the classical buying-scheme was that when a user entered a shop, a knowledgeable seller would be available to give advice and information on products. With the emergence of the Web, online shops started to appear, proposing interfaces where the users had a high level of control, and where actions triggered predictable results. Classical interfaces have allowed people to express their preferences by browsing along a set of well defined categories. For books these might be poems, romance or thriller. In addition search tools rapidly appeared allowing users to more quickly navigate to their target items. Later on, recommender systems were introduced, often relying on explicitly expressed ratings of items. More recently, there has been a lot of research on indirect ways for users to reveal their preferences (e.g. through their purchase history), paving the way to behavioural recommenders. This evolution from a search & browse interaction to today’s behavioural recommenders follows very well a more general and long standing debate, central to the User Modelling community, about automation and direct manipulation which was voiced in [123]: to what extent should users give up control of their interaction with interfaces in favour of depending on intelligent “agents” that learn the likes and dislikes of a user?

In a recommender system, we assume that users’ information needs can be satisfied by knowing their preferences. There are mainly two ways RS collect users’ preferences: by engaging users in an explicit preference elicitation process, or by inferring them implicitly through their interaction behaviours. In the first category, several mechanisms exist which are often linked to the underlying technologies presented in Chapter 2. Collaborative filtering traditionally relies on users being involved and rating items explicitly, such as in MovieLens<sup>2</sup>. Content-based recommenders rely on users specifying their needs in terms of content or features [101], and

---

<sup>2</sup><http://www.movielens.org/>

have also seen the emergence of rating mechanisms for allowing users to give feedback explicitly. The more recent unit or compound critiquing techniques also rely on explicit information. Rather than single valued ratings, the techniques allow users to indicate feedback explicitly over multiple dimensions, helping to improve accuracy [30]. Such explicit and direct feedback is the most common interest indicator, offering a fairly precise way to measure user's preferences, but suffers from several drawbacks [36]. These include the fact that a user must stop to enter explicit ratings, which alters browsing and reading patterns. Users may not be very motivated to provide ratings unless this effort is perceived to be beneficial [114], or because the users might not yet know their preferences as they just started to use the system, and often change them in different contexts [70, 101]. This first category is also often referred to as preference-based recommenders.

In the second category, implicit information like a user's purchase history or navigation pattern, can be used to gather and build user preference profiles. Such recommenders are often called behaviour based RS [93, 145]. Explicit ratings, though common and trusted, might not be as reliable as often presumed. In that sense, implicit ratings are a valuable alternative that must be considered, especially as they remove the cost from the user, associated with examining and rating items. In Nichols' seminal paper on implicit rating and filtering [94], he discusses the costs and benefits of such ratings for information filtering applications. In the process, he identifies several types of data that can implicitly capture a user's interest, including past purchases, repeated uses, and decisive actions (printing, marking, examining). Since then, several of these indicators have been used like in [119], where Shapira *et al.* showed that mouse movements normalised by reading time were a good preference indicator, or as in [36] where Claypool *et al.* illustrate that the time spent on a page is a potentially good interest indicator.

Since the technology was invented more than a decade ago [50], research work measuring the progress of RS, with few exceptions, has unfortunately concentrated on improving the accuracy of algorithms, the most common metric being the mean average error (MAE) [60]. The way users' can express their preferences has evolved, becoming more implicit. Much of the work has been incremental, comparing the latest approach with the previous. A 2007 marketing survey by ChoiceStream [34] compared site with and without recommenders. It reported in this personalisation survey, involving 811 respondents, that consumers strongly preferred sites that provide personalised product recommendations, with 45% claiming that they are more likely to shop at sites with personalised recommendations than at sites without them. The survey is unfortunately not detailed enough to examine much of the other factors that influence users, such as the way in which preference are acquired. In that sense, our work is the first significant in-depth user study that reports on the users' perceptions of today's behavioural recommender systems compared to classical search & browse patterns.

### 4.3 Hypotheses

As explained in the introduction, the goal of this experiment is to understand the differences in users' perceptions between traditional search & browse interfaces, and behavioural recommendations. We chose this comparison in order to oppose the two main results of the previous chapter, accuracy and effort. We believe that through search & browse, users would have a direct

mechanism for controlling the accuracy of their selections. On the other hand, recommendations based on implicit user preferences would require much less effort from users. We established four simple hypotheses for our experiment: two hypotheses about *how* the recommendation process would evolve, one about effort and one about performance with respect to the type of users.

**HYPOTHESIS 1** *For users with a small profile size, search & browse should provide higher recommendation accuracy than indirect feedback.*

First, we evaluate what might happen at the beginning of such interactions. We expect that when a person starts to use a website, a user-controlled solution would be more effective than an indirect one, at supporting the users' information needs. If a user has a small purchase history, for example, there is perhaps not enough information to infer this person's preferences, most certainly resulting in an inadequate recommendation. This issue belongs to the cold-start problem. We thus propose HYPOTHESIS 1.

**HYPOTHESIS 2** *There exists a profile size as of which indirect feedback should propose a better accuracy than the baseline explicit elicitation.*

Second, we consider how recommendation quality might evolve. When users control a search, they may only cover a specific subset of all their preferences, whereas information gathered over time gives a much broader view of these preferences. At the same time, it is unclear how much of this broad information is useful. Works by Bonnin *et al.* such as [19, 18] exploit the active user's navigation path, by considering long and short-distance events in the history with a tractable model, in order to generate suitable recommendations. We expect the amount of data collected implicitly to become greater through time, and we therefore believe that it should progressively generate better suggestions than the baseline. We highlight this with HYPOTHESIS 2 where we fix an arbitrary cut-off level of twenty books (in our experiment).

**HYPOTHESIS 3** *Finding products should require a higher effort with search & browse than through recommendations from indirect feedback.*

Third, we postulate that in the frame of our experiment, testers of the search & browse will perceive it as requiring a bigger amount of effort than those using recommendations computed from their Amazon profile. The effort observation of Chapter 3 concerned the requirements of initial effort and we believe that this will be extended here to the overall perception of effort.

**HYPOTHESIS 4** *Non-expert users are likely, overall, to significantly benefit from recommendations based on indirect feedback.*

Finally, we consider the potential overall benefits. Since an indirect profile should cover multiple aspects of a user's real profile (as the amount of collected information grows), we believe that it is likely to benefit users with lower experience in the field concerned. We thus propose HYPOTHESIS 4.



## 4.4 Experiment 3: Search & Browse vs. Implicit Recommendations

### 4.4.1 Setup and Procedure

We conducted an in-depth real-user evaluation in order to compare the performance of *Amazon*'s implicit recommendation interface to its normal search & browse interface. The experiment was limited to the domain of *books*. We designed a between-group experiment of three user groups, with 20 users in each: the baseline search & browse group, and two recommendation-receiving groups with small and big purchase profiles respectively. All users were told to find five books to purchase, similar to what they would do on the real website. This way, the study simulated a user's evolution with a website: a start with no behavioural history, then with a small profile, and gradually building a significant profile.

We implemented a user study with a wizard-like online web application containing all the instructions, interfaces and questionnaires so that subjects could remotely participate in the in-depth evaluation. The detailed material given to participants is detailed in Appendix C. The general experiment procedure consists of the following steps.

**Step 1** Based on how many books participants bought in the past on *Amazon* (profile size), they are oriented to the adequate experiment (baseline or recommendations).

**Step 2** Basic background information is collected (gender, age, etc.).

**Step 3** A brief explanation of the experiment-to-come is shown to the users and they are presented with a small scenario, so as to put them in the right frame of mind. They are then given some detailed instructions.

*baseline* The testers of the search & browse interface are instructed to go to *Amazon.com*, make sure they are not logged in, and then to browse through the available categories of literature, until they find a book which they like.

*recommendation* The testers of the recommender system are asked to head to *Amazon.com* before logging in to their account. They are then asked to go the "my recommendations" section and to navigate through the book section of the recommendations until they find a book that they like.

**Step 4** The users start the experiment. They are asked to select *five* books; for each one, they must fill in a template-questionnaire allowing them to rate the book on the spot.

**Step 5** To conclude the study, the users are asked to complete a nine questions assessment questionnaire to evaluate the system they have just tested.

### Measured Variables

All questions in this study are statements to which users can indicate their level of agreement on a five-point Likert scale, ranging from  $-2$  (strongly disagree) to  $+2$  (strongly agree);  $0$  is neutral. They are listed in Table 4.1. Not having access to *Amazon*'s interaction logs, we recorded users' opinion about the recommendation quality through a template, immediately after selecting each

Table 4.1: Post-stage assessment (S) and template (T) questions.

| ID | Statement (b:baseline / r:recommendations)   |
|----|--|
| S1 | b: My overall satisfaction with the books that I selected is high.<br>r: My overall satisfaction with the books that were recommended to me is high.                                 |
| S2 | b: Selecting books this way requires low effort.<br>r: Selecting books out of the recommended set requires low effort.   |
| S3 | b: I am confident that the books I just selected are the best choice for me.<br>r: I am confident that the books recommended to me are the best choice for me.                       |
| S4 | b: The Amazon search tool allows me to select a sufficiently diverse set of books.<br>r: The set of recommended books is sufficiently diverse.                                       |
| S5 | b: The search tool seems trustworthy because it shows me books that match my preferences.<br>r: The recs. seem trustworthy because the proposed books match my preferences.          |
| S6 | b: Books I selected were good compared to recs. I may receive from a friend.<br>r: The books I selected were good compared to recs. I may receive from a friend.                     |
| S7 | b: I feel in control in the selection process because my preferences are well respected.<br>r: I feel in control in the selection process because my preferences are well respected. |
| S8 | b: I intend to use this search tool in the future for getting books.<br>r: I intend to use this recommendation feature in the future for getting books.                              |
| S9 | b: I would introduce this search tool to a friend.<br>r: I would introduce this recommender to a friend.   |
| T1 | b & r: I have never heard of this book.  |
| T2 | b & r: I think I will like this book.  |
| T3 | b & r: I am willing to buy this book.  |

book (novelty, appreciation, intention to buy). Once five books had been selected, an overall appreciation was recorded through a set of nine questions, measuring *experience* (satisfaction, effort, trust, confidence, novelty, diversity) and *decision* (acceptance of a recommended book, future usage, sharing with friends). Because of the setup of the experiment, each question was adapted into two variants such as to differentiate between the baseline and recommendation experiments, but tested identical dimensions. An additional question S10 was only given to users of the implicit experiment in order to make sure we did not have any outliers: “I have already used this recommendation feature on Amazon before this experiment”.

### Participants’ Profiles

The user study was carried out over a period of three weeks and an incentive was proposed. The study was taken by off-campus users (half of the participants), students (7%) and academic researchers in Switzerland. The study collected 60 users, resulting in a sample size of twenty participants per group. There were 17 female and 43 male, with 66% being aged between 25 and 30; 18% were younger, and 15% older. As shown in Figure 4.1, the group of *baseline* users showed slightly less familiarity with Amazon as 25% more users disagreed that they “read a lot

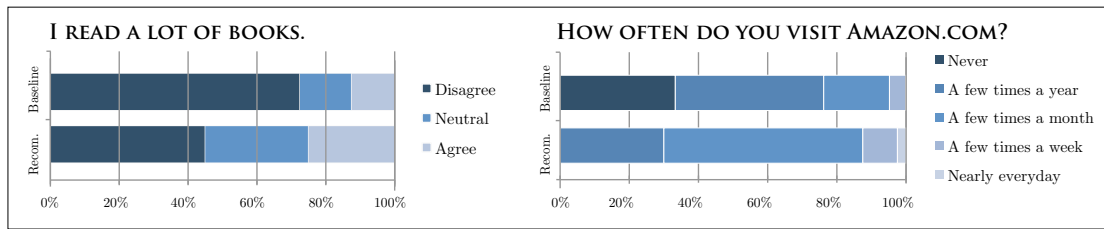


Figure 4.1: Background questions.

of books”, and 30% of them had never surfed on Amazon before. We accepted this potential bias: having less knowledgeable baseline testers possibly means they will demonstrate more first-time behaviours, central to the fundamental acceptance research of this thesis. Furthermore, such users have a fresh view of Amazon, less influenced by the evolution of the site.

## 4.4.2 Analysis of Results

### Post-Stage Assessment

We started by comparing results from the baseline group to all users from the recommendations (i.e. without taking into account profile sizes). This general analysis showed that most answers were the same, as only three questions presented statistically significant differences (S2, S3, S4). Following such an observation, we analysed more in detail this experiment by splitting the results in terms of three user-groups, *baseline*, *recommendation* with small and *recommendation* with big purchase profiles. Results are reported in Figure 4.2. An Anova analysis showed that five questions conveyed statistically significant different averages across all three groups of users. The question S2 shows an increase in results from *baseline* elicitation to *recommendation* users with a large profile, who found that the system required less effort (with an average of 1). The *recommendation* users with a small profile scored 0.6 on average. The difference between all three groups is significant ( $p = 0.02$ ). This result is further supported by objective data. We can indeed compare time-stamps between the background and assessment questionnaires for each user. They reveal that the median time for *baseline* is 31 minutes, against respectively 18 and 15 minutes for the *recommendations* small and big groups. S5, the question on trust, shows the same general tendency, albeit a smaller increase between the first two groups (significant,  $p = 0.05$ ). S3, the confidence about making the best choice, presents a *baseline* average around 0.5 and one of  $-0.5$  for the *recommendation* small group, with the *recommendation* big being somewhat in between them (significant,  $p = 0.02$ ). Diversity S4 shows a very similar pattern, but with an increased score from the *baseline* users, around 1 (significant,  $p < 0.01$ ). One of the template questions also shows a significant difference: T3, the intention to buy, where the *recommendation* small is much lower than both other groups ( $p = 0.04$ ).

For S1, satisfaction, the 0.5 difference between the first two groups is significant ( $p = 0.02$ ). T2, on perceived accuracy, gives much higher averages around 1.0, with a significant difference (t-test,  $p = 0.02$ ) between the two *recommendation* groups. Finally, the special question for *recommendation* users about them having “already used” this recommendation feature showed

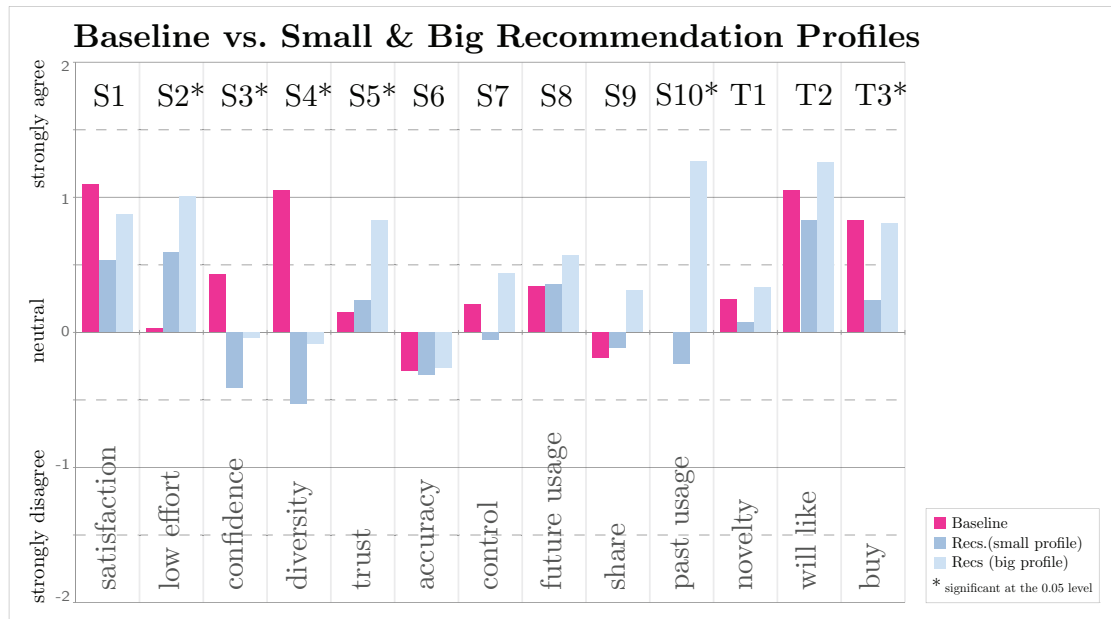


Figure 4.2: Detailed graph of preferences of users.

S10 the expected trend with a score close to 0 for the *recommendation* small group, and above 1 for the *recommendation* big group (significant,  $p < 0.01$ ). These results reveal that although a recommender interface provides users with an overall satisfaction and perceived benefits like a lower effort required, most users have to wait until their profile reaches a certain size to enjoy such benefits.

### Correlation Analysis

We computed a correlation analysis across all groups of users, and combinations of them, for an understanding of how factors of the two types of interface affect each other. For size reasons, we only report a selection of strongest or compelling correlations  $r$  in Table 4.2. We also computed a factor analysis among all *recommendation* data, allowing us to determine the most salient clusters of attributes. All factor nodes discussed in the text are valid (above 0.5 [54]) and tested for reliability (Cronbach's alpha  $\alpha > 0.7$ ).

The strongest overall correlation observed in this study is between the intention to use the feature in the future (S8) and the intention to introduce this feature to a friend (S9), with  $r = 0.658$  (significant,  $p < 0.01$ ). For each of the test groups, this relation is strong and significant. The intention to use in future also presents a favourable link with the fact that people have already used the recommendation feature in the past (S10). The factor analysis reveals that all these dimensions and S6, intention to buy, form a dominant factor for *recommendation* users ( $\alpha = .810$ ).

Table 4.2: Summary of main correlations.

| Overall<br>(60 users)          | Correlated measures                       | Baseline   | Recommendations        |                      |                    |
|--------------------------------|---|------------|------------------------|----------------------|--------------------|
|                                |   | (20 users) | (overall)<br>(40users) | (small)<br>(20users) | (big)<br>(20users) |
| 0.658**                        | Use in future & Introduce to friend       | 0.758**    | 0.693**                | 0.580**              | 0.712**            |
| 0.589**                        | Control & Trust                           | 0.574**    | 0.600**                | 0.771**              | 0.323              |
| 0.567**                        | Already used this feature & Use in future | NA         | 0.567**                | 0.533                | 0.701              |
| 0.501**                        | Satisfaction & Willing to buy             | 0.467*     | 0.497**                | 0.627**              | 0.224              |
| 0.394**                        | Control & Satisfaction                    | 0.115      | 0.542**                | 0.612**              | 0.416              |
| <i>Some other correlations</i> |   |            |                        |                      |                    |
| 0.563**                        | Will like & Willing to buy                | 0.330      | 0.641**                | 0.643**              | 0.401              |
| 0.459**                        | Diversity & Confidence                    | 0.174      | 0.432**                | 0.317                | 0.535*             |
| 0.252                          | Low effort & Satisfaction                 | -0.012     | 0.495**                | 0.345                | 0.604**            |

\*\* :  $r$  significant at the 0.01 level, 2 tailed.\* :  $r$  significant at the 0.05 level, 2 tailed.

Another strong correlation from the overall pool is the tie-up between control (S7) and trust (S5) at  $r = 0.589$ , and control with satisfaction (S1). It seems that both for the *baseline* and the *recommendation* small group, the mentioned variables correlate strongly, but surprisingly not in the *recommendation* big group. Factor analysis shows a big cluster with control, trustworthiness, diversity, confidence, and recommendations being good compared to friends ( $\alpha = .800$ ). We also observe that satisfaction correlates with willingness to buy the selected books (T3), coherently with results from Chapter 3. This connection appears in all groups except the last one.

In order to complete the analysis, we ran a factor analysis on the data from the *recommendation* group (overall). We found four clusters of dimensions. The first cluster links confidence, diversity, trust and control (Cronbach Alpha  $\alpha = 0.800$ ). The second unites the intention to use in the future, the intention to introduce to a friends, the fact that the system has already been used in the past, and the impression that recommendations were good compared to friends ( $\alpha = 0.772$ ). The next cluster sees the fact that people have heard of a book, the fact that they will like it, and their willingness to purchase, correlate ( $\alpha = 0.767$ ). Finally, a last factor groups satisfaction and perception of low effort ( $\alpha = 0.733$ ). We discuss the specific implications of this factor analysis in Section 7.5 where we put these results into perspective with our research model. (The same analysis was run on the *baseline* data without success.)

### Users' Comments

Post-study feedback was collected. The main comment from the *baseline* group is that, when people search for books on Amazon with the categorised menu navigation (or search bar), they are mainly looking to find books that they have already heard of, or that friends would have recommended. (This behaviour leading to searching for a known-item first, was also seen in Experiment 6, and for all users.) Examined individually, such a statement is not surprising. But it is worthwhile to see that absolutely no user reported to have used the system to “browse” through the site like one browses through a furniture catalogue, for example. The comments from the *recommendation* groups highlight the diversity problem which was underlined in the results section. The quite classical problems of having “wrong” recommendations is also put

forward: several people report using Amazon for different purposes such as buying books for professional reasons at the same time as buying leisure books and then getting recommendations in those wrong domains.

### 4.4.3 Discussion

Through our HYPOTHESES 1 & 2, we predicted that at first a controlled search would be more accurate but that this would rapidly change, seeing the accuracy of recommendations increase with the profile size. The direct assessments of perceived accuracy, S6 and T2, are not strongly conclusive. This twist-and-turn between hypotheses and results is surprising. However, we would like to point out that if “accuracy” does *not* reveal itself as imagined, other dimensions *do* demonstrate some parallels with the predictions. Elements like confidence and diversity, show us that search & browse methods are more efficient at the beginning, but that larger recommendation profiles actually start to catch up as profiles grow. Diversity is actually the measure where the difference between the two approaches is the strongest. Nevertheless, and this brings us to HYPOTHESIS 4, there are not many measures where an implicit large profile strongly beats an explicit one (only trust S5 and low effort S2).

The summary of results points out that the two types of interface mechanisms being compared can provide quite similar overall satisfaction for the users. The difference in the amount of effort required to operate in both systems is highly noticed by users, and clearly in favour of the behavioural recommender system, as postulated in HYPOTHESIS 3. On the other hand, users clearly found the *baseline* as proposing a much more diverse set of books, which is problematic for the recommender engine. It is also disappointing to see such low scores for the novelty (T1) from the recommender. Measures of confidence show that users are more confident about their choices in the search & browse scheme. However, people are trusting the system’s implicitly generated recommendations, as soon as their profile reaches a certain size, which is encouraging. This was further reflected in users’ comments. When compared to books that friends might have proposed, neither methods were perceived as being very accurate; nevertheless users’ opinions were positive as in all groups they thought they would like the five selected books. Contrary to purchase intentions, decision variables about future usage of the system or introduction to a friend, were not very high on average, but all three showed good correlations with satisfaction.

## 4.5 Conclusions

In the frame of our experiment, we show that behaviour based recommendations such as proposed by Amazon can be similar in performance to controlled interfaces, for most common parameters. Specifically, our results show that although users may obtain limited benefits from recommenders, they perceive them as being trust-worthy and are willing to accept them. As expected, people especially felt that these interfaces required less effort for finding items. At the same time, they were just as satisfied as with more traditionally controlled interfaces. On the other hand, the study highlighted several areas to be improved. A notable such area was the diversity which was found to be higher with the search and browse interface. It was the question which presented the strongest difference between both systems. Furthermore the novelty

of items proposed by both interfaces was indistinguishable. Although this might be specific to Amazon, we believe that this is an important area to improve since one of the goals of recommender systems is to help users to discover new items. Finally, the study also shows very clearly how recommendations based on small user profiles are poor and need to be addressed. Results further suggest how measures such as confidence, trust, control and intentions to buy evolve as the profile grows.

A decade has passed since recommender technology was invented [50]. Today's systems based on this technology are in the mainstream practice of e-commerce and social websites. Even though some surveys demonstrate that acceptance and perception of this technology are showing good signs, we should not take them for granted. The work in this chapter demonstrates that investigating users issues pays off, and that several traditional problems remain unsolved. Additionally, the challenge of motivating initial users until they build a large profile (hence user loyalty) remains.





## Chapter 5

# The Effects of Layout in Critiquing-Based Recommenders

### 5.1 Introduction

At this stage in the thesis, some tendencies are starting to appear. Our first three experiments have highlighted the importance of keeping the user effort as low as possible. We have also shown how recommendation accuracy can be perceived as being crucial, whilst other dimensions like diversity actually seem to be just as important in users' overall satisfaction. At the same time, one of the observations of the previous chapter is that two different approaches for allowing users to discover new items are perceived as being equal in terms of satisfaction and accuracy.

This latest observation partly gave us the impression that our participants were not really experiencing main differences in the qualities of systems they were testing. Even in the first two studies, which both show a clear win in favour of *Pandora*, there is not much contrast among answers to the different preference-questions. As a result, we decided to try a slightly different approach hoping to trigger more discernible reactions from testers. One prime result from Chapter 3 was the importance of having simple-to-use interfaces, *Last.fm*'s being deemed as complex and proposing irrelevant features. This inspired us to change the visual representation of recommendations, and select a *layout* versus *content* approach for our fourth study, hoping to yield some stronger distinctions from users between two evaluated systems.

The nature of the selected comparison led us to use a system where we had full control. We indeed needed to be able to modify the appearance and layout of the interface, whilst being able to guarantee that the algorithm used to generate the recommendations was fixed. Ideally, we also wanted to have a system where had some data from past experiences in order to evaluate the global impact of our approach. We chose to rely on the CritiqueShop, a prototype e-commerce shopping website co-designed by our group. The system was first introduced by Reilly and Zhang in [112], and is directed at evaluating product search tools based on the critiquing technique. Critiquing-based recommenders, which we introduced in Chapter 2, are very fitting for the goal of this chapter. In critiquing, it is indeed important to encourage users to apply compound critiques frequently. Traditionally, compound critiques are represented textually with sentences [112, 113]. If the product domain is complex and has many features, it often requires

too much effort for users to read the whole sentence of each compound critique. We believe that such textual interfaces hamper the users' experience during the recommendation process.

We decided to propose a new *visual* design of the user interface, which represents compound critiques via a selection of value-augmented icons. We further developed the CritiqueShop framework, which is detailed later in Section 5.3, allowing us to setup a comparison between content and layout in a within-subject real-user evaluation. Results from this Experiment 4 show that the visual interface can improve the performance of critique-based recommenders by attracting users to apply the compound critiques more frequently and reducing users' interaction effort substantially when the product domain is complex. Users' subjective feedback also shows that the visual interface is highly promising in enhancing users' shopping experience.

The rest of this chapter is organised as follows. We first provide a brief review of some related work on visualisation techniques. Then the two interface designs for critiquing-based recommender systems are introduced. Next we describe the setup of the real-user study and report the evaluation results. Finally we present the discussion and conclusions of this chapter.

## 5.2 Related Work

We present the history and evolution of critiquing-based recommenders in detail in the state of the art, Section 2.2. In the experiment of this chapter, we rely on dynamic critiquing and the *MAUT* approach. *MAUT* was previously used in experiments such as [112, 113] on the same framework as the one selected for the experiment of this chapter (for details on the framework please refer to Section 5.3).

This section will be focused on interfaces and more specifically, visualisations. Information visualisation tools have been developed in past years, amongst other, to help users formulate their queries and understand the relationships between collection of information. In [5], Ahlberg and Shneiderman put the Starfield approach together with the dynamic query method, in order to allow users to explore information and data relationships in a large data collection. Users can manipulate attribute values using sliders, and once the values are changed, the display zooms in on a subspace, allowing information seeking at the detailed level. In [39], a Scatter/Gather approach is used by Cutting *et al.* to automatically cluster retrieved documents into categories and labels them with descriptive summaries. Kohonen maps cluster documents into regions of a 2-D map [79]. Recently in [104], Pu and Janecek implemented a visual interface using a semantic fish-eye view to expand search context and to allow users more opportunities to refine initial queries. Doody *et al.* combined product visualisation techniques and additions to the current methods of user preference extraction to recommend suitable eyeglasses to individual users [44].

Giving a detailed yet categorised view of how visualisations have progressed is not the purpose of this section. The drive behind visual representations in recent years has been so strong and widespread that it turns the descriptive task into a dissertation of its own, in nature and scale. Instead we would like to highlight how strongly information visualisations have been adopted in our current society. The most poignant example to us is the regular infographics that *The New York Times* releases. They have gone through the effort of setting up a dedicated team and

a *Visualisation Lab*<sup>1</sup>. Regularly, the journal illustrates news with some of these visualisations. A recent example saw a comparison, “A Year in Iraq and Afghanistan”, of the soldiers death rates. Several other journals have picked up this trend. The strength in these visualisation is that they sit at the cross frontier between information and aesthetics. Lev Manovich was the first to link these two concepts into one with the word “info-aesthetics”<sup>2</sup> and the term has since grown popular. For instance, a blog called *infosthetics* capitalised on this cross section and has become a reference. As we write this thesis, five years after the site’s launch, it has over 23’800 RSS feed subscribers, showing (if need be) how popular and powerful these visualisation can be<sup>3</sup>.

In this work we apply visualisation techniques on the critiquing-based product search tools. More specifically, we chose to use iconised representations. These have already been largely studied in works such as [83] and are particularly fitting, since unit critiquing already uses icon-sized tools to allow users to input their feedback. We present the compound critiques with various meaningful icons, instead of descriptions of plain text.

### 5.3 Experiment Framework: CritiqueShop

As explained in the introduction, in order to run the experiment we needed a platform where we had both the ability to control or modify it and some past experience in order to evaluate the real impact of our changes. We decided to rely on the CritiqueShop as first introduced in [112]. The CritiqueShop is an online platform which was co-designed by our group for designing and evaluating e-commerce product search tools based on the critiquing technique described in Section 2.2. It originally provided us with a unified user interface so that the performances of different recommendation algorithms could be evaluated under the same condition (in terms of interface and features). In this chapter we use it the other way round.

Having chosen an adequate platform, we designed an experiment where we would be able to make a direct comparison between two different layouts using the same content. By layout, we refer to visual design modifications which change parts of the interface whilst representing the same kind of information and content. In order to make the experiment comparable, we only modified one object of the interface: the representation of compound critiques. We hereafter describe the default textual interface of the CritiqueShop, followed by the new visual interface developed for this experiment. We are further motivated by the repeated observation in the previous studies that people find the compound critiques too complex and admit to not actually reading all the information provided.

#### 5.3.1 Textual Interface

The CritiqueShop is a critiquing-based recommender system we implemented in [112, 113] as an online shopping system on the product domains of both digital cameras and laptops. It is designed in a way that allows users to concentrate on the utilisation of both unit critiques and compound critiques as the main feedback mechanism.

---

<sup>1</sup><http://vizlab.nytimes.com/>

<sup>2</sup><http://http://www.manovich.net/>

<sup>3</sup><http://infosthetics.com/>

The interface layout is composed of three main elements, as shown in Figure 5.7:

- a product panel
- a unit critique panel
- a compound critique panel

The system functions as follows. On a first page, users can enter their initial preferences. Thanks to these, the system proposes the three main elements as second page. The product panel shows the current recommended product which best matches the user’s preferences expressed initially. On the left hand side, the unit critique panel allows users to make incremental changes: each feature (of a product) is surrounded by two small buttons, which allow users to increase or decrease a value in the case of numeric features, and to change a value in categorical features (such as the brand or processor type). On the bottom right hand side is the the compound critique panel. This panel shows a list of compound critiques. Users can perform a compound critique by clicking the button “I like this” on its right-hand side. The item then becomes the reference product, and is therefore shown in the product panel, the other two panels being adapted accordingly. These three elements make up the main shopping interface and are always visible to end-users.

By default, compound critiques are represented textually as was the case in [112, 113] and as shown in Figure 5.1. A typical compound critique might say that this product has “more memory, more disk space, but less battery” than the current best match item. A direct mapping is applied from the computed numerical values of the critique, to decide if there is more or less of each feature.

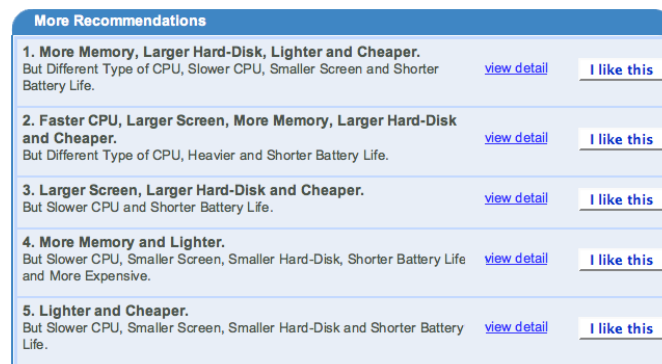


Figure 5.1: Screenshot of the interface for textual compound critiques (with laptop dataset).

In the experiment that follows, and in accordance with results from [113], we adopt the detailed interface where users are capable of seeing the product detail behind each compound critique. In addition, for each compound critique, the positive critiques are listed in bold on the first line, while the negative ones follow on the second line in a normal font-weight. Figure 5.1 presents a list of textual critiques, and Figure 5.2 compares an example of a single compound critique from the textual interface with one from visual solution, described hereafter.

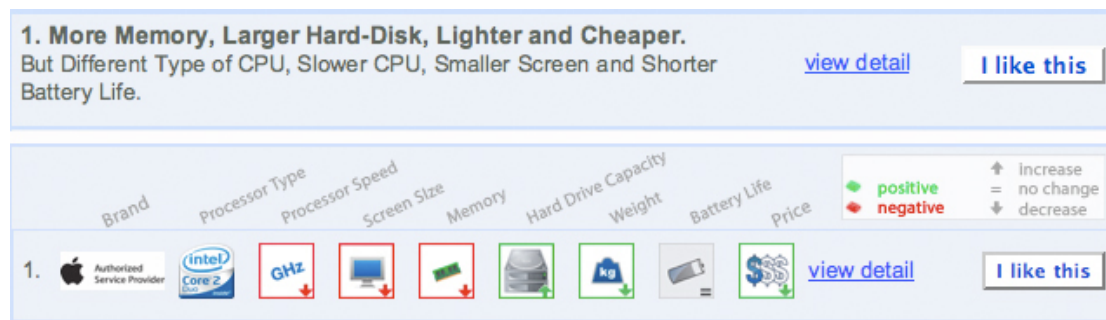


Figure 5.2: An example of a single compound critique from the textual interface (above) and from the visual interface (below).

### 5.3.2 Visual Interface

The visual interface used in this study was developed in several phases. The initial idea was to propose a graphical addition in order to complement the textual critiques, but this rapidly evolved into a complete alternative to a textual representation of the critiques. Three main solutions were considered: using icons, providing a graph of the different attributes or using text-effects such as tag-clouds. The amount of words displayed being already high, the tag-cloud idea was put aside. The first two solutions were kept and selected to build paper prototypes. The first test revealed that the icons were perceived as being closer in meaning to the textual representation, and they were hence chosen for this study.

Icons pose the known challenge that whilst being small they must be readable and sufficiently self-explanatory for users to be able to benefit from them [83]. One difficult task was to create a set of clear icons for both datasets that were used for the upcoming study (see 5.4). We refined them twice after small pilot-studies to make them uniform and understandable. They were then *augmented* such as to represent the critiques: the icon *size* was chosen as a mechanism to represent the change of value of the considered parameter (with respect to the currently selected ideal product). For each parameter of a compound critique, we know if the raw value is bigger, equal or smaller. We used this to adapt the size of the iconised object thus creating an immediate visual map. This method rapidly appeared to be insufficiently clear and even confusing at times. This made small icons unreadable and had to be adjusted. Furthermore, indicating an increase in value is not always a *positive* action: an increase in weight is a negative fact (for both cameras and laptops). This is a well known issue with icon design [83]. Secondly we rapidly understood that all the icons would have to be displayed for each compound critique, as a grid layout. The textual critiques only indicate the parameters that change, but doing so with the icons would have resulted in lines of different lengths, making them hard to compare through this alignment problem. These two potential issues led us to further extend the icons with additional labelling.

Consequently we decided to add a token to the corner of each icon: an up arrow, a down arrow or an equal sign, to further indicate if the critique was respectively an increase, a decrease or equal to the current best match. At the same time we gave colours to the border and the token

of each icon such as to indicate if the change in value was positive, negative or equal. Green was chosen for positive, red for negative and grey for the status quo. For those features without value change, the corresponding icons were shown in light grey. Thus all compound critiques had an equal number of icons and the potential alignment problem was avoided. More importantly, these lines of aligned icons form a comparison matrix and they are decision supportive: a user can quickly decide which compound critique to apply by counting the number of positive or negative icons.

During our pilot user study we found that the visual interface required a learning effort from users. Two measures were taken to tune down this effect. Firstly, a miniature legend of the icons was included at the top of the compound critique panel. Secondly, in our user study we provided an instructions page to users with explanations of the meaning of icons and some icon examples. A detailed example of these icons can be seen in Figure 5.2 which provides a quick comparison of a textual compound critiques and its visual equivalent. The final results with visual critiques inside the whole CritiqueShop can be seen in Figure 5.7. In earlier studies of this thesis, we made sure we did not tune down users' learning efforts. We indeed wanted to test the systems under their usual setup, taking into account potential effort effects. In this study however, we chose to make sure users understood the icons. There are two reasons motivating this choice. First, the visual representation being the tool itself, and not just an initial setup step, we wanted to make sure we were comparing two tools which users understood. Second, this experiment uses a new approach, not fully in line with that of previous chapters (as explained in the introduction). We hereby seek to observe more discernible reactions from testers. In order to do so, we chose to focus on the main interaction rather than the initial discovery of the system.

## 5.4 Experiment 4: Textual vs. Visual Compound Critiques

### 5.4.1 Setup and Procedure

We conducted a real-user evaluation to compare the performance of the two interfaces. There are two types of criteria for measuring the performance of a critiquing-based recommender system: the objective criteria from the interaction logs and the subjective criteria from users' opinions. In this real-user evaluation we mainly concentrated on the following objective criteria: the average interaction length, the application frequency of compound critiques, and the recommendation accuracy. Participants' subjective opinions included understandability, usability, confidence to choose, intention to purchase, etc. They were obtained through several questionnaires, which will be introduced later in this section.

For this user-study we extended the CritiqueShop with MAUT as explained in Section 5.3 and the material given to participants is detailed in Appendix D. We adopted a within-subjects design of the real-user evaluation where each participant is asked to evaluate the two different interfaces in sequence and finally compare them directly. The interface order was randomly assigned so as to equilibrate any potential bias. To eliminate the learning effect that may occur when evaluating the second interface, we adopted two different datasets (laptops and digital cameras) so that the user was facing different domains each time. As a result, we had four ( $2 \times 2$ ) conditions in the experiment, depending on interface order (visual first vs. textual first)

Table 5.1: CritiqueShop post-stage assessment questionnaire.

| ID  | Statement   |
|-----|---|
| S1  | I found the compound critiques easy to understand.  |
| S2  | I didn't like this recommender, and I would never use it again.                               |
| S3  | I did not find the compound critiques informative.  |
| S4  | I am confident that I have found the laptop (or digital camera) that I like.                  |
| S5  | Overall, it required too much effort to find my desired laptop (or digital camera).           |
| S6  | The compound critiques were relevant to my preferences.                                       |
| S7  | I am not satisfied with the laptop (or digital camera) I found using this system.             |
| S8  | I would buy the selected laptop (or digital camera), given the opportunity.                   |
| S9  | I found it easy to find my desired laptop (or digital camera).                                |
| S10 | I would use this recommender in the future to buy other products.                             |
| S11 | I did not find the compound critiques useful when searching for laptops (or digital cameras). |
| S12 | Overall, this system made me feel happy during the online shopping process.                   |

and product dataset order (digital camera first vs. laptop first). For all users, the second stage of evaluation was always the opposite of the first so that they would not take the same evaluation twice.

We implemented a wizard-like online web application containing all instructions, interfaces and questionnaires so that subjects could remotely participate in the evaluation. The general online evaluation procedure consisted of the following steps.

- Step 1** The participants are asked to input their background information.
- Step 2** A brief explanation of the critiquing interface and how the system works is shown to the users.
- Step 3** The users participate in the first stage of the evaluation. They are instructed to find a product (either laptop or camera, randomly determined) they would be willing to purchase if given the opportunity. The users are able to input their initial preferences to start the recommendation, and then can play with both unit critiques and compound critiques to find a desired product to select. Figure 5.7 illustrates the online shopping system with the visual interface and the laptop dataset.
- Step 4** The users are asked to fill in a post-stage assessment questionnaire to evaluate the system they have just tested. The statements are listed in Table 5.1.
- Step 5** Recommendation accuracy is estimated by asking the participant to compare the chosen product to the full list of products to determine whether or not they prefer another product.
- Step 6 – 8** These are steps for the second stage of evaluation which are almost identical to the steps 3 – 5, except that this time the users are facing the system with a different interface/dataset combination (i.e. for avoiding bias).

Table 5.2: Final preference questionnaire.

| ID  | Questions  |
|-----|--|
| Q1  | Which system did you prefer?   |
| Q2  | Which system did you find more informative?                                      |
| Q3  | Which system did you find more useful?   |
| Q4  | Which system had the better interface?   |
| Q5  | Which system was better at recommending products (laptops or cameras) you liked? |
| S13 | I understand the meaning of the different icons in the visual interface.         |

**Step 9** After completing both stages of evaluation, a final preference questionnaire is presented to the users to compare both systems they have evaluated. The users indicate which interface (textual or visual) is preferred in terms of several criteria (overall preference, informativeness, interface). The questions are listed in Table 5.2.

The final preference questionnaire contains an extra statement (*S13*) to evaluate if the icons that we designed were easy to understand. All post-stage assessment questions in this study are statements to which a user can indicate a level of agreement on a five-point Likert scale, ranging from  $-2$  (strongly disagree) to  $+2$  (strongly agree);  $0$  is neutral. We were careful to provide a balanced coverage of both positive and negative statements so that the answers were not biased by the expression style.

The datasets used in our experiment are relatively large (i.e. several hundreds of products each). Revealing all of these products to the user at once during the accuracy estimation of Step 5 would lead the user to confusion. To deal with this, we designed the accuracy test to only show 20 products in one page at a time, and we provided the function of allowing users to sort the products by different attributes. Such interfaces are called *Rankedlists* and have been used as baseline in earlier research such as [103].

### Datasets and Participants

The datasets used in this experiment were updated one week before the beginning of the experiment, resulting in them containing the most recent products currently available on the market. The laptop dataset contained 610 different items. Each laptop product had 9 features: *brand*, *processor type*, *processor speed*, *screen size*, *memory*, *hard drive*, *weight*, *battery life*, and *price*. The second one was the digital camera dataset consisting of 96 cases. Each camera was represented by 7 features: *brand*, *price*, *resolution*, *optical zoom*, *screen size*, *thickness* and *weight*. Besides, each product had a picture and a detailed description. In later discussions, we consider the laptop dataset as being more complex than that of cameras because of the number of features. This is further supported by some comments made by users in this study (and in [112, 113]).

To attract users to participate in our user study, we set an incentive of 100 EUR and users were informed that one of those who had completed the user study would take part in a draw to win it. The user study was carried out over two weeks. Users participated in the user study



Table 5.3: Demographic characteristics of participants.

| Characteristics            |                 | Users (83 in total) |
|----------------------------|-----------------|---------------------|
| Nationality                | Switzerland     | 36                  |
|                            | China           | 13                  |
|                            | France          | 12                  |
|                            | Ireland         | 6                   |
|                            | Italy           | 4                   |
|                            | Other Countries | 12                  |
| Age                        | <20             | 6                   |
|                            | 20-24           | 30                  |
|                            | 25-29           | 40                  |
|                            | $\geq 30$       | 7                   |
| Gender                     | female          | 15                  |
|                            | male            | 68                  |
| Online Shopping Experience | Never           | 2                   |
|                            | $\leq 5$ times  | 38                  |
|                            | $>5$ times      | 43                  |

remotely without any supervision. We recorded 83 users in total who completed the whole evaluation process. Their demographic information is shown in table 5.3. The participants were evenly assigned to one of the four experiment conditions, resulting in a sample size of roughly 20 subjects per condition cell. Table 5.4 shows the distribution of users with dataset and ordering they encountered in the user study.

Table 5.4: Design of the real-user evaluation.

| Group             | First stage |         | Second stage |         |
|-------------------|-------------|---------|--------------|---------|
|                   | Interface   | Dataset | Interface    | Dataset |
| I<br>(20 users)   | Textual     | Camera  | Visual       | Laptop  |
| II<br>(20 users)  | Textual     | Laptop  | Visual       | Camera  |
| III<br>(23 users) | Visual      | Camera  | Textual      | Laptop  |
| VI<br>(20 users)  | Textual     | Laptop  | Textual      | Camera  |

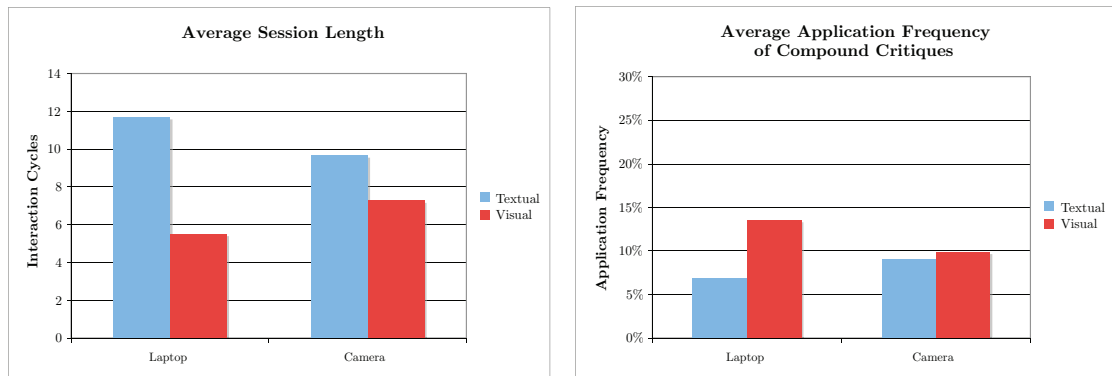


Figure 5.3: Average sessions lengths (left) and average application frequency of compound critiques (right) for both user interfaces.

## 5.4.2 Analysis of Results

### Recommendation Efficiency

One of the results highlighted in our previous studies is the importance of effort. In order to have an objective measure of effort in our evaluation, we measured the length of a session in terms of recommendation cycles, i.e. the number of products viewed by users before they accepted the system’s recommendation. For each recommendation interface and dataset combination we averaged the session lengths across all users. It is important to remember that any sequencing bias was eliminated by randomising the presentation order in terms of interface type and dataset. The first graph of Figure 5.3 presents the results of the average session lengths with different interfaces. The visual interface appears to be more efficient than the baseline textual interface. For the laptop dataset, the visual interface can reduce the interaction cycles substantially from 11.7 to 5.5, a reduction of 53%. The difference between these two results is significant ( $p = 0.03$ , with ANOVA test in this chapter). For the camera dataset, the visual interface can reduce the average interaction cycle from 9.7 to 7.3, a reduction of 25% (not significant,  $p = 0.31$ ).

We also looked into the detail of each interaction session to see how often the compound critiques had actually been applied, as shown in the second graph of Figure 5.3. For the system with textual interface, the average application frequencies are respectively 7.0% (for laptops) and 9.0% (for cameras). For the system with visual interface, the average application frequency is nearly doubled to 13.6% for the laptop dataset (significant different,  $p = 0.01$ ). For the camera dataset the application frequency is 9.9%, a 9.5% increase compared to the baseline textual interface (not significant,  $p = 0.70$ ). Since for both systems we are using exactly the same algorithm to generate the compound critiques, it appears from results that the visual interface can stimulate more users to choose compound critiques during their decision process. Also, compared to the two systems with different datasets, it seems to show that the visual interface can be more effective when the domain is more complex.

### Recommendation Accuracy

As this was the first study of this thesis where we had control over the recommendation generation process, we took the opportunity to measure the *quality* of the recommendations over the course of a session, as proposed in [91]. One factor for estimating recommendation quality is the recommendation accuracy, which can be measured by letting users review their final selection with reference to the full set of products (as in [103]). We chose to use such an accuracy-metric as it gives us the possibility to make a direct mapping between qualities which are perceived by users, and the accuracy they report. We defined recommendation accuracy formally hereafter. If users consistently select a different product, the recommender is judged to be not very accurate. The more people stick with their selected best-match product then the more accurate the recommender is considered to be.

**DEFINITION: RECOMMENDATION ACCURACY** *We define Recommendation Accuracy (for critiquing) as the percentage of times that users choose to stick with the product they selected in the critiquing interface, rather than changing to one from the full set of products.*

Figure 5.4 presents the average accuracy results for both interfaces on both datasets. The system with textual interface performs reasonably well, achieving an accuracy of 74.4% and 65.0% on the laptop and camera datasets respectively. By comparison, the system with visual interface achieves 82.5% accuracy on the laptop dataset and 70.0% on the camera dataset, which have been increased 10% and 7% respectively. It appears that the visual interface produces more accurate recommendations. However, these improvements are not significant ( $p = 0.378$  for laptop dataset, and  $p = 0.648$  for camera dataset).

### User Experience

In addition to the above objective evaluation results we were also interested in understanding the quality of the user experience afforded by the two interfaces. As we have mentioned earlier, a post-stage assessment questionnaire was given when each system had been evaluated. The twelve statements are listed in table 5.1. A summary of the average responses from all users is shown in figure 5.5.

From the results we can see that both systems with different interfaces received positive feedback from users in terms of their ease of understanding, usability and interfacing characteristics. Users were generally satisfied with both systems (see S2 and S7) and did not find that they required too much effort (see S5). We also noticed that overall, the visual interface has received higher absolute values than the baseline textual interface on all these statements. Three statements are clearly in favour of the visual interface, and present statistically significant differences: S4 ( $p < 0.001$ ), S5 ( $p < 0.01$ ) and S9 ( $p = 0.014$ ). These results show that the visual interface is significantly better than the textual interface in the criteria of effort, ease of use and leading to a more confident shopping experience.

With the final preference questionnaire we asked each user to vote on which interface (tex-

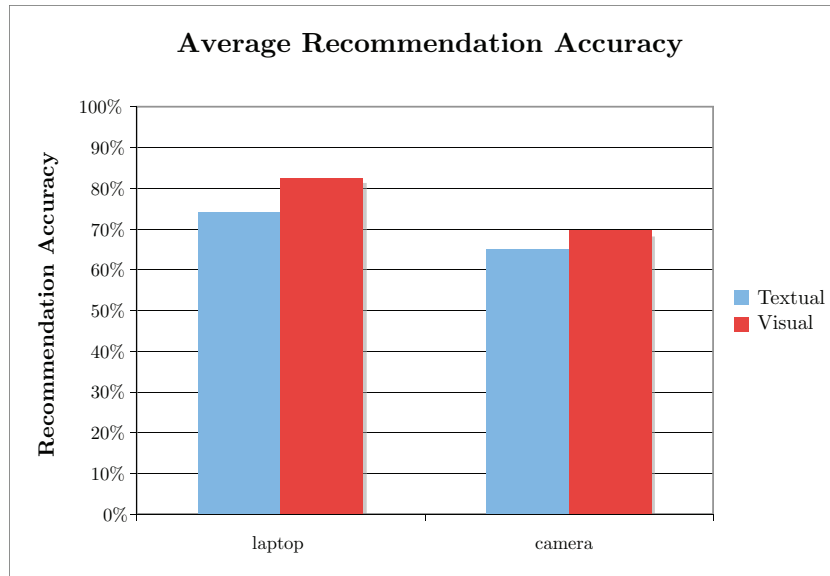


Figure 5.4: Average recommendation accuracy for both user interfaces.

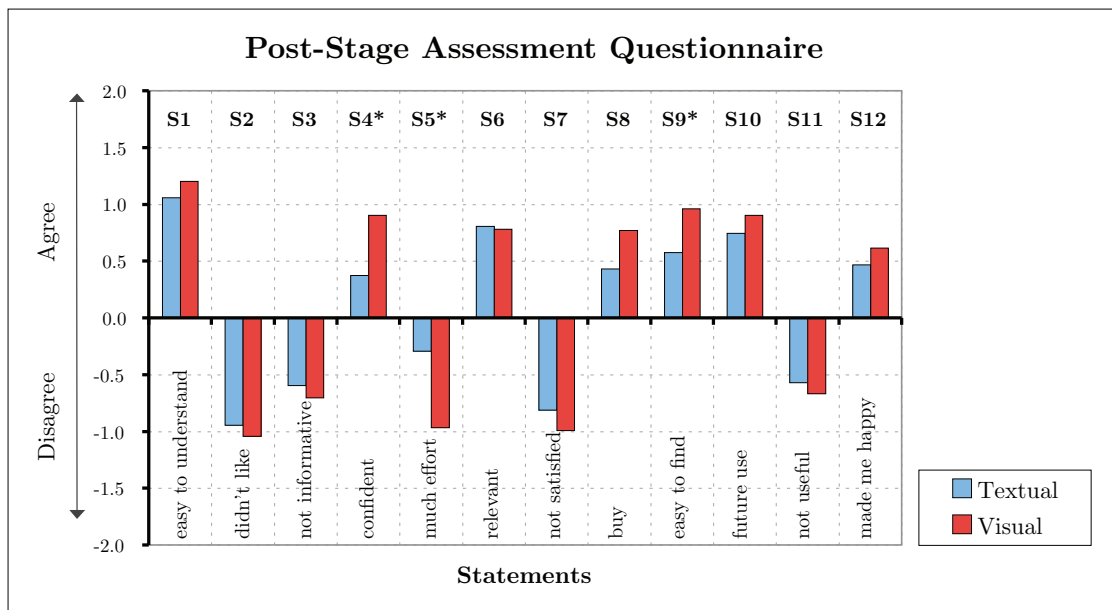


Figure 5.5: Results from the post-stage assessment questionnaire.

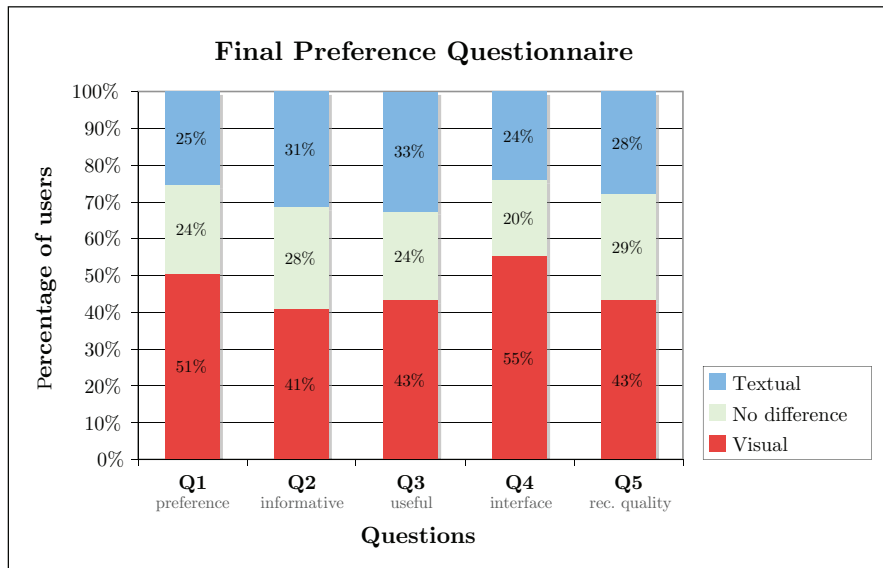


Figure 5.6: Results from the final preference questionnaire.

tual or visual) had performed better. The questions are listed in Table 5.2, and the results are presented in Figure 5.6. The results show that overall users feel that the visual interface is better than the textual interface in all given criteria. For instance, 51% of all users prefer the visual interface compared to 25% of whom prefer the textual interface (see Q1). Also, more than 55% of users think the visual interface is better (see Q4). Furthermore, although the two systems have exactly the same algorithm to generate compound critiques, the visual interface can enhance users' perception on the recommendation quality (see Q5). These results show that the visual interface has gained a much stronger support from end-users during the online shopping process.

The final questionnaire provided an extra statement (S13) to evaluate if the icons in the visual interface were understandable. The overall average score was 1.23 on the Likert scale, which shows that users had understood them well and that we had designed them adequately for the experiment. We re-examined the scores of Q1-Q5 of only people who stated that they had understood the meaning of the different icons. We found that among this subset of users, an even larger percentage of them prefer the visual interface. For example in Q1, 61% of those users voted for the visual system (13% for the textual one) against respectively 51% (and 25%) overall. These results suggest that if users understand the meaning of different icons, they are even more likely to prefer the visual interface.

### 5.4.3 Discussion

There are several aspects which are worth pointing out with these results. To start, it is interesting to notice that in the user study results, while the visual interface performed better than the textual interface with both laptop and camera datasets, the visual interface has achieved higher

performance improvements with the laptop dataset than with the camera dataset. The main difference between the two datasets is that the laptop assortment is more complex. It contains more products and each have more features than the cameras. When the product domain is rich, the textual interface will generate very long strings of text to describe the compound critiques, which are not easy for users to read. By comparison, the visual interface could provide an intuitive and effective way for users to make decisions (for example by simply counting the number of positive and negative icons). We believe that as the system or product space gets more complex, long textual descriptions become a burden, and the synthetic nature of a visual solution can become a real tool with real advantages.

Our results show that a large proportion of users preferred the visual interface for the critiquing-based recommender system. However we also noticed that there is still a small number of users who insist on the textual interface. As pointed out in Section 5.3.2, the pilot studies we ran highlighted that the visual interface required some additional learning effort to understand the meaning of the icons at first. This result is therefore not surprising. Potential solutions for helping users could include adding some detailed instructions and illustrative examples to educate new users, or the system could provide both textual and visual interfaces and let the users choose the preferred interfaces adaptively by themselves. Ideally, if the system had profile information about the users, it could guess whether a user is a novice or regular visitor, and chose which interface to propose by default. We believe that this observation is not unique to this experiment but was possibly highlighted through the multiple preference questions.

Finally, we compared our results with those obtained in [113] by Reilly & Zhang, a very similar study run on the CritiqueShop, but comparing two algorithmic approaches. They report that on average, both their approaches led users to performing 10 interaction cycles. The textual representation in our study also averages around 10.5 interaction cycles, which is expected since the algorithm we used is the same as one of the two in [113]. However, our visual critiques average around 6.5 interaction cycles, which is a strong improvement. Another insightful comparison can be done by looking at average recommendation accuracy. In Reilly and Zhang's study, the camera dataset produces a 60% to 80% accuracy which is similar to results of our study, where textual and visual average is respectively 65% and 70%. The results for the laptop dataset are much more in favour of our comparison. Whereas the first study showed 65% to 70% accuracy measures, our textual critiques gave 75% and above all, the visual solution was over 80% accuracy on average. Based on these observations, we believe that such a visual approach is highly promising. We choose to rely on it again for our next study, Experiment 5, detailed in Section 6.4.

## 5.5 Conclusions

In this Chapter, we chose to use a new approach for exploring users' perceived system qualities, hoping to trigger more discernible reactions from our participants. Despite the fact that accuracy metrics are not statistically different between the textual and visual representations, several dimensions participating in user's overall satisfaction show a strong preference for the visual critiquing. Users perceive the visual interface as requiring less effort, and the measured sessions lengths in the visual case are shortened. Finally, when confronted with complex product-

domains, users apply the visual critiquing more frequently. Beyond the fact that participants seem to rely strongly on the recommender, the visual interface clearly appears to play a large role in users' interactions.

User interface design is an important issue for critiquing-based recommender systems. Traditionally the interface is *textual*, which shows compound critiques as sentences in plain text. In this chapter we propose a new *visual* interface which represents various critiques by a set of meaningful icons. We developed an online web application to evaluate this new interface using a mixture of objective criteria and subjective criteria. Our comparative real-user study showed that the visual interface is more effective than the textual interface. It can significantly reduce users' interaction efforts and attract users to apply the compound critiques more frequently in complex product domains. Users' subjective feedback also showed that the visual interface is highly promising in enhancing users' shopping experience.

**critique**shop BETA

**Instructions:** Please use this **visualization** interface to find the laptop that you want to buy. You can either click the button for each attribute on the left panel, or select one of the recommended products below. [Click here for more instructions...](#)

**Refine Features**

Brand:

ProcessorType:

ProcessorSpeed(GHz):

ScreenSize(inches):

Memory(MB):

HardDriveCapacity(GB):

Weight(lbs):

BatteryLife(hours):

Price(\$):

**Product History**

- Lenovo  
Lenovo ThinkPad X60
- Apple  
Apple MacBook

**Our Recommendation**

Authorized Service Provider  
**Apple MacBook Pro**

**Price: 1999.0 USD**  
1599.2 EUR  
2498.75 CHF

[Buy now!](#)

---

**Main Features:**

- ProcessorType: **Core 2 Duo**
- ProcessorSpeed(GHz): **2.2**
- ScreenSize(inches): **15.4**
- Memory(MB): **2048.0**
- HardDriveCapacity(GB): **120.0**
- Weight: **5.5lbs (2.5kg)**
- BatteryLife(hours): **6.0**

**Product Description:**

Powered by the most advanced mobile processors from Intel, the new Core 2 Duo MacBook Pro is over 50% faster than the original Core Duo MacBook Pro and now supports up to 4GB of RAM. The NVIDIA GeForce 8600M GT delivers exceptional graphics processing power. Featuring 802.11n wireless technology, the MacBook Pro delivers up to five times the performance and up to twice the range of previous-generation technologies. Quickly set up a videoconference with the built-in iSight camera. Control presentations and media from up to 30 feet away with the included Apple Remote. Connect to high-bandwidth peripherals with FireWire 800 and DVI. Innovations such as a magnetic power connection and an illuminated keyboard with ambient light sensor put the MacBook Pro in a class by itself.

---

**More Recommendations**

|    | Brand  | Processor Type | Processor Speed | Screen Size | Memory | Hard Drive Capacity | Weight | Battery Life | Price |   |
|----|--------|----------------|-----------------|-------------|--------|---------------------|--------|--------------|-------|---|
| 1. | Apple  | Intel Core 2   | GHz             |             |        |                     |        |              |       | <a href="#">view detail</a>   <a href="#">I like this</a> |
| 2. | lenovo | Intel Core 2   | GHz             |             |        |                     |        |              |       | <a href="#">view detail</a>   <a href="#">I like this</a> |
| 3. | lenovo | Intel Core 2   | GHz             |             |        |                     |        |              |       | <a href="#">view detail</a>   <a href="#">I like this</a> |
| 4. | SONY   | Intel Core 2   | GHz             |             |        |                     |        |              |       | <a href="#">view detail</a>   <a href="#">I like this</a> |
| 5. | SONY   | Intel Core 2   | GHz             |             |        |                     |        |              |       | <a href="#">view detail</a>   <a href="#">I like this</a> |

[Next Step](#)

Figure 5.7: Screenshot of the visual interface for the online shopping system (with laptop dataset).



## Chapter 6

# How Diversity Leads to Confidence

### 6.1 Introduction

In the previous chapter, we considered and evaluated how interface design could change user's perceptions of recommendations in contrast to algorithmic changes determining the content displayed. The user evaluation showed some promising results in the domain of compound critiques, where a visual representation helped to generate a more frequent application of critiques and a lower perception of required effort. In this chapter, we decided to first apply a similar comparison in order to explore users' perceptions of *diversity*.

Diversity is a key dimension which was first put forward in Chapter 4. Much of the research on recommenders has focused on improving recommendation accuracy, often at the cost of diversity. This tends to lead to a lack of novelty that users will experience with suggested items, and ultimately to their dissatisfaction. In this chapter we run two experiments. We first consider how both layout and content fixes can influence the perceived diversity of a recommender, and how users' overall preferences are affected. Second, we decided to refine our analysis, by using an eye-tracker to record users' interactions. We study users' browsing and purchasing behaviours when they interact with a website that has a personal recommendation system, focusing on how diversity can change online search behaviours. The question addressed hereafter is less to know *what* to suggest, but more *how*, *when* and *why* do users need diversity.

#### 6.1.1 Can Layout be a Vector of Diversity?

The first question that this chapter specifically targeted was *how* do users perceive diversity? It can be said that diversity is not specific issue of recommender systems; it can be found in several domains like for instance search engines. When one considers these, one can see that search engines have tried to reduce this problem by proposing diversity through various post-search tools, designed to help users refine their initial search query. Such tools include query suggestions or refinements [37], result clustering and cluster naming, and mapping of results against a predetermined taxonomy such as ODP <sup>1</sup>. Contrary to recommender engines, where many of them try to perform another filtering of the recommendation set (ending up by taking out items that are too

---

<sup>1</sup>Open Directory Project <http://www.dmoz.org/>

similar to items already suggested), search engines propose alternate approaches such as relying on layout-driven tools to provide diversity.

This observation about search-engines using layout to introduce diversity is coherent with the approach we used in the experiment of Chapter 5. We believe that there is more to users' perceptions than accuracy-driven improvements. We therefore decided to apply the same kind of "layout vs. content" approach to the topic of diversity for the first of our two experiments. As said, our initial goal was to understand *how* users perceived diversity in recommendations systems. We were interested in bringing some answers to the following questions:

1. If the content of a recommendation set is more diverse, will the user experience the difference?
2. If we just change the layout of the recommendation set, will the user perceive a difference in diversity?
3. If we change both the content and the layout, would the users perceive the system as providing even better results?
4. Does diversity change users' perceptions of trust or confidence in the recommender?

In order to answer these questions, we decided to work on *organisation interfaces*. Organisation interfaces, as first introduced in Chen and Pu's work [31], have been shown to be effective in helping users reduce effort and errors in reviewing recommended items. In [100], it was even demonstrated that organisation-based interfaces display a diverse set of results, thereby more effectively enabling user's trust formation compared to traditional k-best interfaces. As a result, users are more likely to establish trust relationships with the recommender agent and are eager to conserve effort when they revisit the sites. The results look surprisingly encouraging from the perspective of diversity. In the frame of this work, we define organisation interfaces as follows:

|   |
|---|
| <p><b>DEFINITION: ORGANISATION INTERFACE</b> <i>An organisation interface is an interface-design element where recommendations are organised into different categories according to their similar trade-off properties.</i></p> |
|---|

In order to explore how both layout and content fixes were noticed by users, and how they influenced the perceived diversity of a recommender, we extended an online perfume shopping prototype system [26], and ran a between-subject real-user study with four groups of users. We used two approaches for selecting the content: we crawled the recommendation set of Amazon, which we compared to a second set that we generated through Editorial Picked Critiques [106]. For the layout we opposed items listed in a linear fashion (which we called List view) and items grouped into five categories (Organised view). This fifth experiment of the thesis is detailed in Section 6.4.

### 6.1.2 Recommenders' Influence on Buyer's Decision Process

The second part of our investigations around diversity was a more in-depth and in-detail focused study. It was aimed at understanding the role of diversity, more specifically *when* and *why* users

need diversity in their buying decision process. The decision process leading to purchase has been shown to be separated into six steps [82], but one does not yet understand where and how the recommender's role is so influential. This is not much of a surprise since *user preferences* are a vast and complex topic in recommenders, as highlighted in Section 4.2. Users are generally unable to accurately state their preferences up front [99], especially when confronted with an unfamiliar product domain or a complex decision situation with overwhelming information. More recently, Häubl and Trifts [56] suggested that there are two main steps during the decision process in an online product search environment. In the first step, the active users identify a subset of products that they want to compare (called the consideration set or the basket). We will refer to this step as *product brokering* to be consistent with concepts used in [82]. During the second step, users compare the different features and details of these products in order to make a decision. We will refer to this step as *product comparison*. In their work, two interaction decision aides were investigated for their roles in helping users make better decisions. The first tool, a recommender agent, assists users in the initial screening of the alternatives and establishing the basket set. The second one, a comparison matrix, helps users make an in-depth feature-by-feature comparison of the items in the basket. Their empirical studies showed that the use of a recommender agent leads to a significant reduction in the number of alternatives seriously considered for purchase, and increases the quality of consumers' decisions. The constructive influence of recommenders was consequently established.

There are reasons to believe that these findings established several years ago may not be correct anymore considering today's recommenders. At the time, the recommender system was indeed conceptualised in the study but implemented by personallogic<sup>2</sup>, a company that does not exist anymore. In the more current literature of recommenders, personallogic is considered closer to a multi-criteria product filtering tool rather than a recommender. The main difference is that in the former, users obtain a set of items after they have actively specified their preferences, whereas in the latter, users get suggested items without asking for them (in a recommendation-giving context of e-commerces). We used Häubl and Trifts findings as our baseline for characterising users' buying decision process, and conceived an experiment where a multi-criteria product filtering tool and a recommender would both be available. We wanted to analyse users' behaviours when confronted to a website providing a Multi-criteria Filtering Tool (MCF) and a Recommender System (RS). The MCF search tool highlights the interactions through a classical interface that does not supply personalisation, but simply helps users to reduce the number of displayed products from a set of constraints. The RS system aims at presenting relevant alternatives to a given product with several levels of diversity, thus assisting users in their choices. This was possible by using the same online perfume shopping system as for the other experiment of this chapter. We setup an in-depth lab-study with an eye tracker, and observed in detail during a one-hour experiment the behaviour of eighteen users when confronted with a perfume e-commerce website. Details about Experiment 6 are given in Section 6.5.

The experimental framework used for both of these studies is detailed Section 6.3.

---

<sup>2</sup><http://www.personallogic.com/>

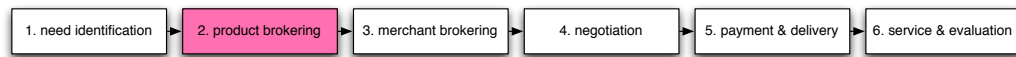


Figure 6.1: The six stages of purchase decision, as proposed by Maes *et al.*

## 6.2 Related Work

When confronted with an overload of alternatives, everybody is affected by limited rationality [124]. This means that every human decision is rational within limits such as time and cognitive capabilities. This limitation has strong economic repercussions, since a wealth of items creates a poverty of attention. E-commerce applications are no exception, and it has become essential to study the human-computer interactions leading to a decision, in order to improve filtering strategies related to attention economy [40] in recommender systems.

Several descriptive models seek to capture buying behaviours. However, everyone agrees to say there are six fundamental stages within the purchase decision making process, as proposed by Maes *et al.* [82]. We illustrate these stages in Figure 6.1. First, users become aware of a new need. Sometimes, this realisation results from companies' prospecting campaigns. Then, they have to determine who to buy from. Everything customers experience on a website feeds into the building of rapport between the buyers and the seller. Users search elements that promote trust, such as the ease of navigation or the relevance of answers and recommendations. In the meantime, they evaluate product alternatives to make a choice. At this stage called *product brokering*, interactions help recommender systems to iteratively characterise their needs and present options. Stages 4 and 5 consist in negotiation and purchasing. The seller has to provide security and confidence in order to close the sale. Finally, users' satisfaction in relation to the overall buying experience can be measured in a sixth stage, if post-purchase product service is involved.

The product brokering stage, highlighted in Figure 6.1, can be decomposed into two steps according to Häubl and Trifts [55]. During the first step, active users identify a subset of products that they want to compare. During the second step, they compare the different features and details of these products, in order to make a decision. In [56], it was proven that the use of a recommender system leads to a reduction in the number of alternatives considered seriously for purchase. They also showed that a recommendation agent increases the number of non-dominated alternatives (i.e. not objectively inferior to any alternative [95]) in the set of alternatives seriously considered for purchase. The influence of recommenders is consequently no longer to be demonstrated, assuming the fact that they provide items relevant to users' needs. However, the goal of personalisation is not only to provide the right item to the right person, but also at the right *time*. The time constraint has been overlooked by researchers for years. In this chapter, we aim to analyse both the impact of recommender systems over time at the product brokering stage, which is a transverse problematic of [55], and the factors that influence the users. The question is less to know *what* to suggest, but more *when* and *why*.

The work outlined by Ho in [63] represents a pioneering effort to study the impact of personalisation at different decision making stages. Through an experiment involving a ringtone

personalisation service, they highlight the decreasing probability of a tailored item to be selected at later stages of decision making. Nevertheless, the absence of selection does not mean that the recommender system does not play a role in the decision process. The mere fact of looking at a recommendation can affect the user's decision. In order to answer the *when* interrogation, we intend to show in this chapter that the influence of a recommender in comparison with a multi-criteria filtering tool constantly increases as time goes on. We will prove that the influence of RS is maximal when the active user is close to making a decision and adding a product to the basket. This influence was observable within the two decision steps identified by Häubl *et al.* [55], and we will measure that it has an effective impact on these steps. In order to do so, we use an eye-tracker to measure the usage of recommendations by coupling action logs with an analysis of eye movements. This way, we are able to superpose the search cycles and the function of influence with the goal of confirming or invalidating a potential link between them.

The conclusions in [63] can also underline an inappropriate combination between accuracy and diversity at later decision stages. This brings us back to the question of *why* to suggest items. Ho *et al.* [65, 64] showed the influence of need for cognition<sup>3</sup> and the size of the recommendation set on the decision making. In the context of our work, we rather decided to investigate the need for diversity in order to reach a decision.

### 6.2.1 Diversity in Recommendations

Diversity has a role to play in making good recommendations and is thus a sought-after property, but *why* and *how* to improve diversity remains an open subject of study. Although a number of diversity enhancing mechanisms had already been proposed [43, 90, 122, 127], the first major paper where diversity was not introduced in a static or uniform way was [86], where McGinty and Smyth attempted to clarify the role of diversity in the context of conversational recommender systems. They showed that, in general, introducing diversity has the potential to significantly enhance the efficiency of recommendations. It may however lead to new challenges. One of these was that diversity might not be desirable during every recommendation cycle, an observation repeated and confirmed in studies such as [26]. With more people starting to be aware of the importance of diversity, numerous debates have emerged.

Fleder *et al.* covered the discussion of whether recommenders really helped users discover new products, or if they rather pushed forward the already popular ones [47]. Later, McNee *et al.* raised concerns about how accuracy metrics had not only misguided but actually harmed the field of recommenders, questioning whether a probabilistically less accurate recommendation is necessarily less valuable [88]. Their work shed light on the possible roles of diversity, such as considering what items should be proposed to a user returning to a site compared to those for a new user.

Traditional methods for diversifying search results rely on attribute-based diversification, especially in personalised web search where different users look for different information even when entering the same search query. Bradley and Smyth were among the first to propose a bounded greedy algorithm for retrieving the set of cases most similar to a user's query, but at the

---

<sup>3</sup>The need for cognition is a personality variable reflecting the extent to which people engage in and enjoy effortful cognitive activities.

same time most diverse among themselves [20]. Common client-side approaches focus on re-ranking the top  $N$  search results, making the documents likely to be preferred by the user appear higher [134]. Anagnostopoulos *et al.* describe an algorithm for sampling search-results whereby they can reduce homogeneity [8]. Later, Radlinski *et al.* propose three alternative methods that increase the diversity of top results, which rely on analysing query-query reformulations in order to add interesting diversity within the result set [107].

More recently Zhang and Hurley suggested maximising diversity while maintaining adequate similarity with a binary optimisation problem. They evaluated their approach on the MovieLens dataset with reasonable success [143]. At the same time, Celma *et al.* not only proposed a new approach for evaluation novelty in recommendations (item- and user-centric) [28], but they also considered the diversity question from another perspective in [27] where they asked: how much can popularity bias a (music) recommender? In 2009, Hijikata *et al.* proposed a discovery-oriented version of a collaborative filtering algorithm, relying on a profile of acquaintance to predict a user's unknown items [61]. Despite these debates and years of research, even very recent user studies such as [67] by Jones *et al.* continue to point out that recommender engines used in industry still suffer from this over-specialisation phenomenon, drawing attention to how critical and necessary diversity can be.

Methods for increasing diversification are not only limited to the ranking and selection of results. In [144], Ziegler and McNee introduced a method for designing and diversifying personalised recommendation lists, especially in item-based collaborative filtering where they showed that even if the average accuracy level decreases, topic diversification can improve user satisfaction. This draws parallels with research on explanation interfaces which present a novel and alternative approach to the over-specialisation problem. Rather than relying on item attributes, explanations convey the reason for which a particular item is being recommended in order to diversify the results. Explanation interfaces have long been accepted in fields such as medical decision support [10] or data exploration systems [24]. Their usage in recommenders was first discussed by Herlocker *et al.* in [59] and tested by Pu and Chen in [31] which showed how they helped to build users' trust while introducing classification and diversity. More recently, they demonstrated that compared to a list view, an organisation-based layout tends to perform significantly more effectively in improving users' perceptions of recommendation quality, increasing their system-acceptance levels, such as the perceived ease of use and usefulness, and finally global satisfaction [32]. The perspectives with regard diversity are highly promising, and other recent works continue to explore this topic of diversity through explanation [138].

### 6.3 Experiment Framework: Perfume Recommender

We conducted two experiments to address three questions: how, why, when. They were both realised using the same experimental framework: an online perfume recommender. This online shopping website was previously developed in [106] and reproduced the layout and catalogue of a real e-commerce business. The dataset of perfumes was crawled from Amazon and contained recent and popular fragrances.


The website was composed of two main pages. The first page of the website served as an entry point. We refer to it as the *search page*. It was divided into two parts, composed of a multi-

**2363 total results for PERFUME (WOMEN)      European perfumes sold in retails stores worldwide**


| brand  | price   | quantity   | category   |
|--|---|--|--|
| <a href="#">Chanel</a> (42)<br><a href="#">Estee Lauder</a> (33)<br><a href="#">Calvin Klein</a> (15)<br><a href="#">Gucci</a> (6)<br><a href="#">Givenchy</a> (11)<br><a href="#">More...</a> | <a href="#">Less than 30 USD</a> (565)<br><a href="#">30-50 USD</a> (843)<br><a href="#">50-80 USD</a> (527)<br><a href="#">80-110 USD</a> (185)<br><a href="#">More than 110 USD</a> (243) | <a href="#">Less than 40 ml</a> (262)<br><a href="#">40-80 ml</a> (733)<br><a href="#">80-120 ml</a> (752)<br><a href="#">120-160 ml</a> (48)<br><a href="#">More than 160 ml</a> (22) | <a href="#">Eau de Parfum</a> (1071)<br><a href="#">Eau de Toilette</a> (1188) |

Total results: 2363


currency:  sort list by:  show:  1 | 2 | 3 ... >




[Givenchy Very Irresistible Eau de Parfum, 2.5 oz.](#)  
**GIVENCHY**  
 Around 77 USD  
 73ml  
 A red carpet of roses illuminated by star anise and the ...



[Kai Perfume Kai Perfume Oil](#)  
**KooKai**  
 Around 45 USD  
 An intoxicating blend of tropical gardenia and white exotic ...



[Kai Perfume Kai Eau de Parfum Spray](#)  
**KooKai**  
 Around 65 USD  
 Shipping March 5th..... What you've asked for ...



[Angel by Thierry Mugler for Women 0.8 oz Eau de Parfum Spray Refillable \(Decoded Box\)](#)  
**THIERRY MUGLER**  
 Around 47.98 USD  
 23ml  
 Angel by Thierry Mugler fragrance for women is a unique ...

Figure 6.2: Snapshot of the *search page* of the perfume recommender.

criteria filtering tool (MCF) at the top and a listed presentation below, as shown in Figure 6.2. The MCF was designed to allow testers to search for a perfume by the brand, by the price range and the type of perfume (Eau de Toilette, Eau de Parfum, Cologne, Aftershave). Below this, a double column, lexicographically ordered item-list of perfumes was available. This part displayed the perfumes respecting the selected criteria of the upper multi-criteria search tool. In this double choice-list, each perfume was laid out with a picture of the bottle, its exact name, the brand, price and quantity. As a complement, the first two lines of its description were shown. In addition to the MCF tool, we proposed a classical re-ordering tool allowing users to sort the list of results by brand, price (low to high, or high to low), and popularity. The search page also included a possibility to set the number of results displayed on one page, and the currency. By default, results were displayed in US dollars, sorted by popularity, sixteen at a time.

The second main window that users encountered was that presenting the detail of any perfume. We refer it as the *detail page*, is captured in Figure 6.3. In complement to presenting the same information as in the listed presentation described earlier, specific data was here given including: a full description, a big-sized picture, a best-selling rate, average user ratings, the gender, the source website, and the possibility to rate. The page had an “Add to shopping cart”

button. At the same time, on the right of this detail, a column for displaying recommendations was reserved. By default, these were proposed in five classified boxes, all displaying their classification label. This method of selecting and displaying the recommendations was developed in [106] and is called *Editorial Picked Critiques* and is hereafter described more in detail.

### **Editorial Picked Critiques**

Editorial Picked Critiques (EPC) is an extended form of example critiquing. The approach of editorial picked critiques was originally inspired from editorial picks as information aggregated from public opinions and at the same time, editors' own experience and knowledge in producing the final recommendations, and was adapted from the preference based critiquing method [31]. Experimental results from Smith *et al.* [125] revealed that editorial recommendations are largely preferred by users in many fields. Similarly, we showed in Chapter 3 of this thesis how *Pandora's* recommendations generated through manual classification by musical experts could beat recommendations made by *Last.fm's* collaborative-filtering. These promising results stimulated the creation of EPC as detailed in [106]. We decided to use this form of critiquing for mainly two reasons. First, it is an extension of critiquing. This allows our experiments to remain coherent since we used critiquing in Experiment 4. Second, as noted in the related work on diversity, explanation interfaces are promising in terms of diversity.

Editorial Picked Critiques are generated by combining editorial opinions, popularity information and preference based critiques. The EPC algorithm first establishes a tradeoff table reflecting attribute compromises among products. Relying on a list of critique categories, editorially selected and ranked, the approach seeks to find products for each category. Each item in the set has attributes that have a different importance level depending on the current category. The composition of these attributes allows generating compound critiques, which are then compromised in order to retrieve items corresponding to the categories, thanks to the tradeoff table. When necessary, the constraints can be relaxed iteratively until the desired number of recommendations per category are found.

EPC as used in both experiments of this chapter was established with seven main categories calculated. These were:

1. more popular and cheaper
2. more popular but more expensive
3. same brand and cheaper
4. same brand but more expensive
5. just as popular and cheaper
6. same price range and just as popular
7. people who like this also like

Although seven categories of recommendations were available, we chose to only display five at any moment in the RS column of the detail page, and in a random order, such as to reduce users' habituation to screen position. Two boxes were always left out since EPC categories 1 & 2, and 3 & 4 are in opposition: only one of each can be displayed at any time. Each box of recommendations contained up-to six items, which were horizontally scrollable without



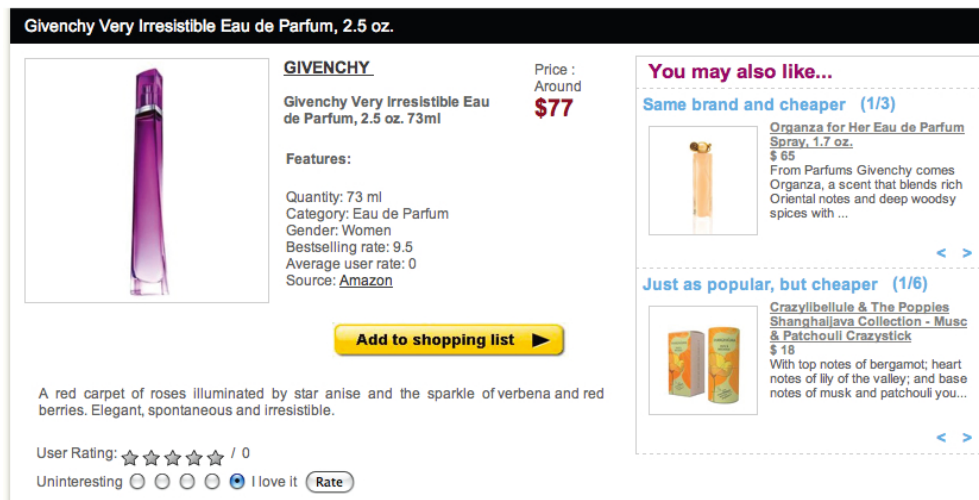


Figure 6.3: Snapshot of the *detail* page of the perfume recommender.

reloading the page. The “People who like this also like” group relied on a standard collaborative filtering algorithm. Other recommendation categories were related to products’ features. Please note that in Figure 6.3 , only two categories are shown for space reasons.

**Why the perfume domain?** The perfume domain was chosen as it is a slightly above-norm field in terms of complexity. Had a very common domain been selected, users would have felt less engaged in their interactions, possibly resulting in some “shortcut” behaviours. Furthermore, as this thesis aims to understand acceptance mechanisms, it was important to maximise opportunities of putting users in situation that they had not already experienced, similar to ones where acceptance has a key role to play. Finally, by having a less conventional domain, users are forced to rely more on the tools proposed, helping us to evaluate the efficiency of the different parts tested.

### 6.3.1 Diversity Metrics

The term *diversity* can have a very broad meaning in the field of recommender systems, as highlighted in the related work. In the two experiments that follow, we define diversity and novelty as follows.

**DEFINITION: DIVERSITY** *We define recommendation diversity as the fact that two (or more) products demonstrate strong dissimilarities in terms of their respective features.*

**DEFINITION: NOVELTY** *We define recommendation novelty as the fact that a product proposed to a user is unknown to him.*

Under these definitions, it appears that novelty and diversity can easily be linked, since

an unknown item would at the same time introduce novelty and diversity among a list of recommendations. We acknowledge this link, and operate under the assumption that novelty is a sub-element of diversity. We do not go into any higher level of detail, as it is in no way the purpose of this thesis to provide a hierarchy or organisation between these two terms. For more details on such a discussion, please refer to works such as [86].

The definition of diversity highlights that we consider diversity as a measure which relies directly on the similarity between products, as suggested in [144], where the more two items are similar, the less there is diversity between them. Drawing our inspiration from [127], we compute the similarity between two products  $p_1$  and  $p_2$  by using a weighted mean of the similarities between  $p_1$  and  $p_2$  for each of the five attributes that characterise a perfume (brand, price, quantity, category, and popularity), as shown in Equation 6.1.

$$Sim(p_1, p_2) = \frac{\sum_{i=1..5} w_i * sim_{attribute=i}(p_1, p_2)}{\sum_{i=1..5} w_i} \quad (6.1)$$

In order to obtain a diversity score among a collection of recommendations, we then computed the average intra-list similarity (ILS) of each recommendation category  $C$ , as defined in [144] (see Equation 6.2).

$$ILS(C) = \frac{\sum_{p_i \in C} \sum_{p_j \in C, p_i \neq p_j} Sim(p_i, p_j)}{2} \quad (6.2)$$

Of course, higher ILS scores denote lower diversity. We can extend these diversity calculations by also computing the diversity RD of a perfume  $p_i$  relatively to a set of  $n$  perfumes  $P$ , thanks to Equation 6.3 [86].

$$RD(p_i, P) = \frac{\sum_{j=1..n} (1 - Sim(p_i, P_j))}{n} \quad (6.3)$$

An ILS score tells us how similar items are within a cluster, keeping in mind that the clusters change depending on the currently viewed item. However, the set of perfumes  $P$  represents products that already caught the attention of the active user on a given page. The relative diversity (RD) consequently allows to measure the added value of considering a new perfume, compared to the sequence of past consultations. At last, we can measure the need for diversity of the active user as time goes on, by adding up the relative diversities of each new considered alternative compared to the user history.

In order to give an overview in terms of diversity, of the different recommendations categories of EPC, we calculated the average intra-list similarities (ILS) for each of the seven recommendation categories. They are summarised in the Table 6.1. We calculated these average values by considering that every category is composed of six recommendations on the detail page. This is often true but it is not always the case, since this number corresponds to the maximal number of recommendations for a category. It is worth pointing out that the fewer recommendations there are in a category, the more diverse this category is. The Table 6.1 also displays the average similarity (Sim) and relative diversity (RD) between a recommendation and its related product. In the case where we only consider the relative diversity between two products, we can say that  $RD = (1 - Sim)$ . We can see from the table that the categories which provide

Table 6.1: Average Intra-List Similarity for six recommendations (ILS), Average Similarity between a recommendation and the perfume of the current detail page (Sim), and Relative Diversity of a recommendation relative to the current perfume (RD).

| Category                          | ILS        | Sim | RD         |
|-----------------------------------|------------|-----|------------|
| More popular and cheaper          | 4.5        | 0.4 | <b>0.6</b> |
| More popular but more expensive   | 4.5        | 0.4 | <b>0.6</b> |
| Same brand and cheaper            | 4.5        | 0.6 | 0.4        |
| Same brand but more expensive     | 4.5        | 0.6 | 0.4        |
| Just as popular and cheaper       | 4.5        | 0.6 | 0.4        |
| Same price range, just as popular | <b>6.0</b> | 0.8 | 0.2        |
| People who like this also like    | <b>3.0</b> | 0.4 | <b>0.6</b> |

the most diversity (or the least similarity) are “More popular and cheaper”, “More popular but more expensive”, and “People who like this also like”. As expected, the category “Same price range and just as popular” supplies on the contrary very poor diversity, but a higher similarity.

## 6.4 Experiment 5: Diversity in a Content vs. Layout Approach

### 6.4.1 Experiment Setup

In order to evaluate how content and interface layout play a role in *how* users perceive diversity, we setup a real-user evaluation where we used and extended the perfume framework detailed in 6.3. The study focused on the second page, the *detail page* which displayed recommendations in the right-hand column. The following considerations on content and layout concern only this column of recommendations.

#### Content Selection


One of the main focuses of this study is on how content is selected in recommenders. As pointed out in Section 6.2, works like [67] show that even in popular recommenders like Amazon, many dimensions, including diversity, are crucially missing in the content proposed. In order to get a better perception of the role of content in the user’s appreciation of system interaction, we chose two algorithms for selecting content.

*Amazon* (AZ) is today the historical reference in terms of commercial recommender systems. Its selection of items labelled “Users Who Bought Also Bought” has become a classic in terms of recommenders and is often used as a comparison in studies, like in Experiment 3. We chose to use it here as one of our algorithms for selecting content. While crawling Amazon’s set of perfumes to make sure our database was up-to-date (as explained later in Section 6.4.2) we also made sure we crawled all of the available recommendations. We started out by collecting items from “Users Who Bought Also Bought” but we ended up using those from “Users Who Viewed Also Viewed” mainly for two reasons. Firstly, several items in the first category were not products from the perfume domain (i.e. other items that Amazon sells). Secondly, there are

### List view


**You may also like...**

Previous




**Philosophy Amazing Grace**  
\$ 40.00 - 60.00  
Amazing Grace fragrance for women is a uniquely feminine blend of soft, floral blossoms a...

---




**TOMMY GIRL 10 For Women By TOMMY HILFIGER eau de toilette**  
\$ 14.99  
TOMMY GIRL 10 was launched by the designer house of Tommy Hilfiger in 2006. This scent po...

---




**Emporio White for Her by Giorgio Armani 3.4 oz Eau de Toilette Spray Limited Edition**  
\$ 45.99  
White for her - a gentle breeze bathed in light, calm yet incredibly freshe, with notes of...

---



**Eau Des Merveilles by Hermes for Women 1.6 oz Eau de Toilette Spray**  
\$ 39.99  
Eau des Merveilles (French for Water of Wonders) is a refreshing, sweet and light floral-b...

---



**SUNFLOWERS by Elizabeth Arden EDT SPRAY 3.3 oz for Women**  
\$ 14.50  
Introduced by the design house of Elizabeth Arden in 1993, SUNFLOWERS by Elizabeth Arden


Next

Displaying (1-5) of 29 perfumes.

### Organized view

**You may also like...**

**More popular and cheaper (1/6)**




**Pink Sugar FOR WOMEN by Aquolina - 3.4 oz EDT Spray (Tester)**  
\$ 24.00  
Pink Sugar is stylish and lively, with a distinctive personality, Pink Sugar takes you on ...

< >

---

**Same brand and cheaper (1/1)**




**HUGO by Hugo Boss EDT SPRAY 1.3 OZ**  
\$ 29.49  
Launched by the design house of Hugo Boss in 1997, HUGO by Hugo Boss for WOMEN possesses a ...

< >

---

**Just as popular, but cheaper (1/6)**




**TOMMY GIRL 10 For Women By TOMMY HILFIGER eau de toilette**  
\$ 14.99  
TOMMY GIRL 10 was launched by the designer house of Tommy Hilfiger in 2006. This scent po...

< >

---

**Same price range and just as popular(1/6)**




**Ck One Summer 2009 3.4 EDT Unisex By Calvin Klein**  
\$ 29.31  
CKO Summer captures the energy of a refreshing splash in the pool. CKO Summer is a limited...

< >

---

**People who like this also like (1/5)**



**Gucci By Gucci by Gucci for Women. Eau De Toilette Spray 2.5-Ounces**  
\$ 50.15  
Launched by the design house of Gucci in 2007, GUCCI BY GUCCI by Gucci for WOMEN possesses ...

< >

Figure 6.4: Comparison of List and Organised view.

nearly twice as many items in the second category as in the first, reducing the item-set where no recommendations were available. The second algorithm that we used is *Editorial Picked Critiques* which is detailed in the previous section, 6.3.

## Interface Design

The other main motivation in this study was the way in which the interface design can influence users' perceptions of recommendations, following our encouraging results in Experiment 4. Studies by Chen and Pu like [30, 31] have shown that preference-based organisation achieves high prediction accuracy while introducing classification and diversity. We wanted to extend this finding in our context, by comparing the default organised view of EPC to a more traditional list view. These two alternatives are described hereafter and shown in Figure 6.4.

### List View (LV)

A list view is a standard way of representing recommendations. Amazon's longstanding way of displaying "Users Who Viewed Also Viewed" items is in a horizontal list, scrollable without reloading the page. We chose to implement the same layout, but vertically in order to keep it positioned as closely as possible to the organised view (see next subsection). The scrolling without reloading was also implemented. Recommended articles from Amazon were simply displayed as on Amazon, and will hereafter be referred to as *AZLV*. The EPF perfumes were randomly selected among the calculated recommendation item-set, and will be referred to as *EPFLV*.

### Organised View (OG)

The organised view we chose relies directly on the five categories used in EPC. The categories are displayed one after the other vertically, each one presenting the title, thumbnail and short description of a perfume, with a link to the detail of the concerned item. For each category, up to six perfumes are scrollable horizontally without reloading the page. We refer to them as *EPCOG*. They are shown in the right column of Figure 6.4.

In the case of the recommendations coming from Amazon, we re-mapped them to the five categories. EPC's "people who like this also like" obviously took the "Users Who Viewed Also Viewed" perfumes. For the categories using a popularity measure, we used Amazon's best-selling score as a replacement. For categories with price information such as "same brand and cheaper", we took into consideration Amazon's data format: most perfumes have a single price, but a small subset is proposed as a single product but with a range of prices (corresponding to different perfume volumes). In such cases, cheaper products were those with a price inferior to the lower bound of the range, and more expensive ones above the upper bound. All these Amazon items are later referred to as *AZOG*.

The four possible setups of content and layout are summarised in Figure 6.5.

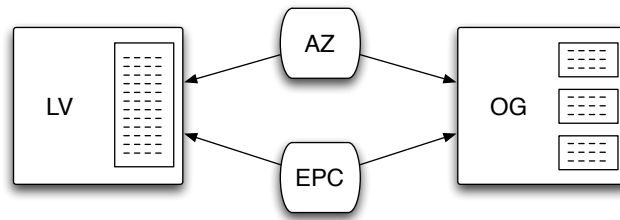


Figure 6.5: Four possible combinations of content and layout of Experiment 5.

Table 6.2: Design of the real-user evaluation.

| Group   | First sessions |           | Second sessions |           |
|---------|----------------|-----------|-----------------|-----------|
|         | Interface      | Algorithm | Interface       | Algorithm |
| 8 users | List           | Amazon    | List            | EPC       |
| 8 users | List           | Amazon    | Organised       | Amazon    |
| 8 users | Organised      | Amazon    | Organised       | EPC       |
| 8 users | Organised      | Amazon    | List            | Amazon    |
| 8 users | List           | EPC       | List            | Amazon    |
| 8 users | List           | EPC       | Organised       | EPC       |
| 8 users | Organised      | EPC       | Organised       | Amazon    |
| 8 users | Organised      | EPC       | List            | EPC       |

## 6.4.2 Experiment Procedure

In this section we first present the performance-evaluation criteria before outlining the procedure of the real-user evaluation and introducing the dataset and participants. The material given to participants is detailed in Appendix E.

### Evaluation Criteria

In order to evaluate the system tested, we chose to rely on both objective criteria from the interaction logs, and subjective criteria from users' opinions. In this real-user evaluation, we mainly concentrate on the following objective criteria: the average session length and the average number of pages viewed in the interface. Participants' subjective opinions include ease of use, satisfaction, usefulness, etc. These are obtained through a post-stage assessment questionnaire detailed in the next section.

### Evaluation Setup

For this user-study we adopted a within-subjects design of the real-user evaluation where each participant is asked to evaluate the two different interfaces in sequence before giving us their

preference. The interface and algorithms were randomly assigned so as to equilibrate any potential bias. In order to reduce the learning effect, users were given a different setup for the second session, and each combination of setups saw an equal number of people take the same combination but in reverse order. With two interfaces, and two algorithmic approaches, we had four (2 x 2) conditions in the experiment, depending on the session order. Only one parameter was changed between the first and second session, either the interface or the algorithm. In order to further reduce any potential bias, a one-week lapse was imposed between the two sessions. The combinations of users and experiments are presented in Table 6.2.

We implemented a wizard-like online website containing all instructions, interfaces and questionnaires so that subjects could remotely participate in the evaluation. The general online procedure consisted of the following steps.

**Step 1** The users are asked to fill in a short questionnaire about their background profile. They are questioned about their age, gender, nationality and profession. A few specific questions aimed at determining their computer literacy and perfume expertise are asked such as “I use Internet very frequently”.

**Step 2** A brief explanation of the experiment-to-come is shown to the users. They are informed that they will be directed to an e-commerce site with more than 5000 popular and common perfumes. The goal will be to search and select three perfumes that they would be willing to purchase for themselves, given the opportunity.

**Step 3** The testers can then start the experiment. Each time they put a perfume in their basket, the user is reminded of how many more they must select.

**Step 4** When finished, each user is asked to fill in a post-stage assessment questionnaire to evaluate the system that was just tested. The questions are listed Table 6.3.

**Step 5** A week later, the users are asked to repeat the experiment, this time having to search for three perfumes for someone of the opposite gender.

At the end of the second session, a final preference question S16 is added to the assessment questionnaire, asking which website they preferred. All questions in this study (except S16) are statements to which a user can indicate a level of agreement on a five-point Likert scale, ranging from  $-2$  (strongly disagree) to  $+2$  (strongly agree);  $0$  is neutral.

### **Dataset and Participants**

The dataset of perfumes used in this experiment was updated just before launching the study. The perfumes were crawled from Amazon, making sure that we had a dataset containing the most recent and popular fragrance products available on the market. 6,529 items were accessible, covering 3,969 items for Women (1,936 Eau de Toilette, 2,033 Eau de Parfum) and 2,560 items for men (1,642 Eau de Toilette, 525 Aftershave & 393 Cologne). Sample perfumes (which were often priced below USD 10) were removed, as they might have fuelled unconventional online purchase behaviours in the experiment.

Table 6.3: Post-stage assessment questionnaire.

| ID   | Statement   |
|------|---|
| S1   | The website gave me good recommendations.                                     |
| S2   | The recommendations were novel.   |
| S3   | The recommendations were diverse.   |
| S4   | The recommendations were better than what I may receive from a friend.        |
| S5   | The recommendations were better than what I may get from Amazon.com.          |
| S6   | I trust the recommendations.  |
| S7   | Given the opportunity, I would buy the recommended perfumes.                  |
| S8   | My overall satisfaction with the website is high.                             |
| S9   | Looking for a perfume using this website required too much effort.            |
| S10  | I found this website easy to use.   |
| S11  | I enjoyed using this website.   |
| S12  | I found this website useful for finding perfumes I might like.                |
| S13  | I felt in control in telling the system what I like.                          |
| S14  | I am confident that the three perfumes I selected are the best choice for me. |
| S15  | If this were a real website, I would use it again.                            |
| S16* | Which recommendation interface do you prefer?                                 |

The user study was carried out over a period of three weeks. As an incentive, four USD 100 gift vouchers were proposed in a draw to users who had completed the study, in order for them to purchase one of the perfumes they had selected. Finally we obtained 64 users in total who completed the whole evaluation process. Their demographics information is shown in Table 6.4. We also asked three questions to determine users' expertise in terms of computer knowledge, internet use and perfumes. These are reported in 6.6. It appears that users are quite regular computer users and that they do not feel that they use the internet too frequently. Concerning perfumes, 16% seems to have very little knowledge, but overall we believe the distribution is acceptable.

### 6.4.3 Analysis of Results

As explained in Section 6.4.2, we relied on both objective and subjective measures to understand users' behaviour in this experiment. Hereafter we first present the objective measures before the subjective ones. We also present users' final preferences and discuss correlations and users' comments.

#### Objective Measures

To be successful, a RS must be able to efficiently guide a user through a product-space and, in general, short recommendation sessions are preferred as low effort is often correlated with



Table 6.4: Demographic characteristics of participants.

| Characteristics |   | Users (total: 64) |
|-----------------|---|-------------------|
| Age             | < 20                                      | 8                 |
|                 | 21-30                                     | 46                |
|                 | 31-50                                     | 8                 |
|                 | >51                                       | 2                 |
| Gender          | Female / Male                             | 25 / 39           |
| Nationality     | Swiss                                     | 27                |
|                 | Spanish                                   | 5                 |
|                 | Italian                                   | 3                 |
|                 | Romanian                                  | 3                 |
|                 | Other                                     | 26                |
|                 | (American, French, Indian, Moroccan, ...) |                   |

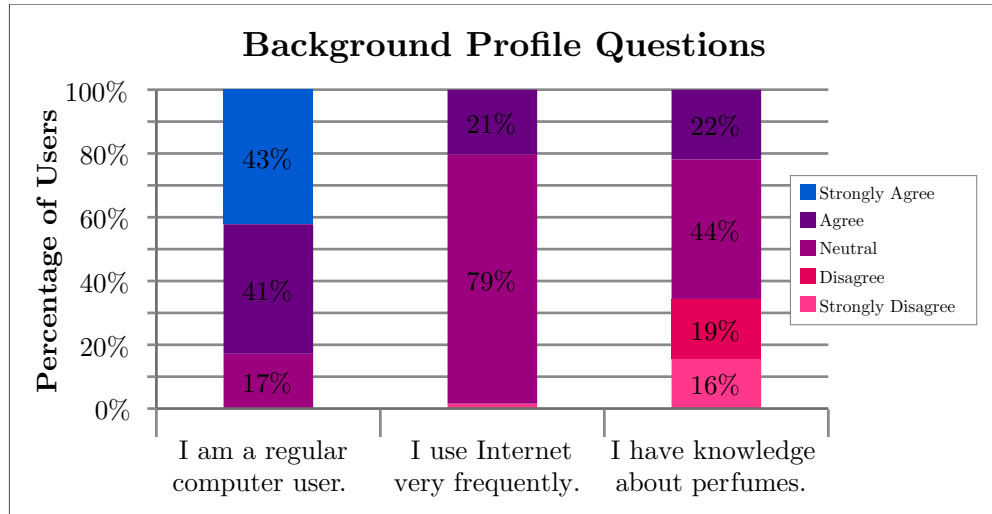


Figure 6.6: Background profile of users.

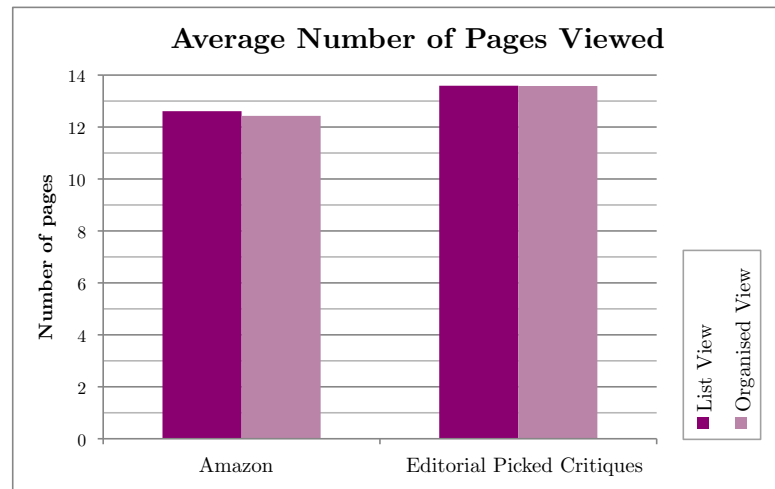


Figure 6.7: Average number of pages viewed.

purchase intentions. For this evaluation we rely on the average session lengths and the number of pages viewed. For each recommendation interface and algorithm combination we averaged the sessions lengths across all users. It is important to remember that any sequencing bias was eliminated by randomising the presentation order, in terms of interface type or content algorithm.

For these objective measures of timing and page-views, we employed a statistics package which was called (through javascript) after any page had been loaded, rather than relying on access logs. Figure 6.8 presents the results of the average session lengths for the different interfaces and algorithms. There are two important observations to be made. Firstly, the organised view appears to yield longer sessions than the list view. The average time for EPCOG is 07:56 against 05:34 in EPCLV. Secondly, both averages for AZ are smaller than their counterparts with EPC.

Figure 6.7 shows the average number of page views for the different groups of users. Clearly the interface dimension does not seem to influence the number of pages viewed. The main difference can be seen between AZ and EPC. On average, EPC users viewed one page more than on AZ.

### User Experience

In addition to the objective evaluation results, we were interested in understanding the quality of the user experience afforded by the two interfaces and algorithms. A post-stage assessment questionnaire was given when each system had been evaluated. The fifteen statements are listed in Table 6.3.

From the overall results listed in Figure 6.9, we can see that all four approaches generated similar results across the fifteen questions. S4 and S9, the comparison with recommendations of a friend and the feeling that this website required a lot of effort, both seem to score lower than the middle score of the five-point Likert scale. At the same time, S3 and S10, diversity of

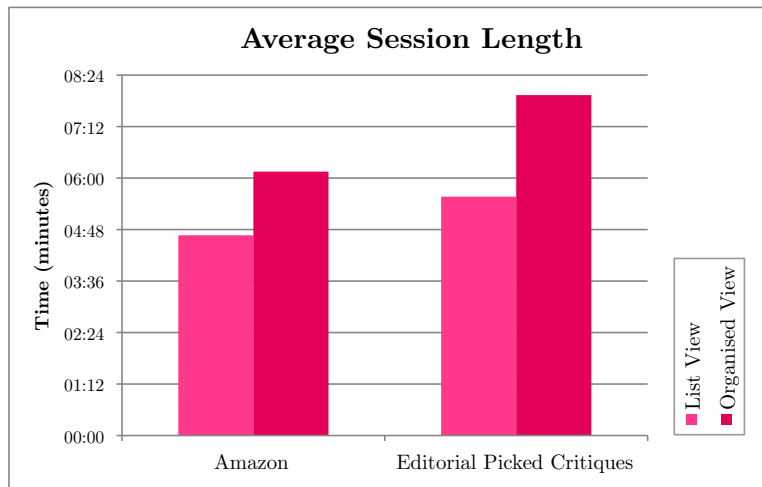


Figure 6.8: Average session length.

Table 6.5: Diversity scores for each configuration. (diversity =  $1 - ILS$ )

|    | AZ    | EPC   |
|----|-------|-------|
| LV | 0.572 | 0.628 |
| OG | 0.531 | 0.553 |

recommendations and ease of use for this site, seem to score higher (between 0.5 and 1.0) than the other questions which often range between 0 and 0.5 for all four setups.

Statistical analysis with Anova reveals that S3 is the only marginally statistically significant question ( $p = 0.10$ ), pointing out that the two organised views are perceived as providing a more diverse set of items. Pair-wise T-tests also show a few marginally significant differences. For S2, AZLV and AZOG differ by 0.37 ( $p = 0.10$ ), and for S14, AZLV and AZOG differ by 0.42 ( $0.05 < p < 0.1$ ). In order to obtain more insight on this observation, we used Ziegler *et al.*'s ILS score to calculate the diversity proposed in each of our four configurations. We report in Table 6.5 the average amount of diversity which users encountered in each four configurations. Please note that ILS is a score of similarity, hence the use of  $1 - ILS$  here to obtain a score of diversity. The table shows us that participants in the evaluation encountered (on average) 3.9% more diversity when the recommendations came from EPC rather than AZ. However, the LV produced a 5.8% more diverse set of item than the OG view. The most diverse configuration was EPCLV, and the least was AZOG.

As part of a more detailed analysis of results, and profiting from the fact that this was a within group study, we reverted to analysing the changes of averages between both sessions. We measured whether a user gave a higher, equal or lower score the second time to each question. Whether Amazon or EPC, users who first tested the list view and then the organised one saw an increase of 0.5 on average for S1 ( $p < 0.05$  for AZ, and  $p = 0.08$  for EPC); the increase for those exposed in the other order was smaller and not significant. The question about novelty,

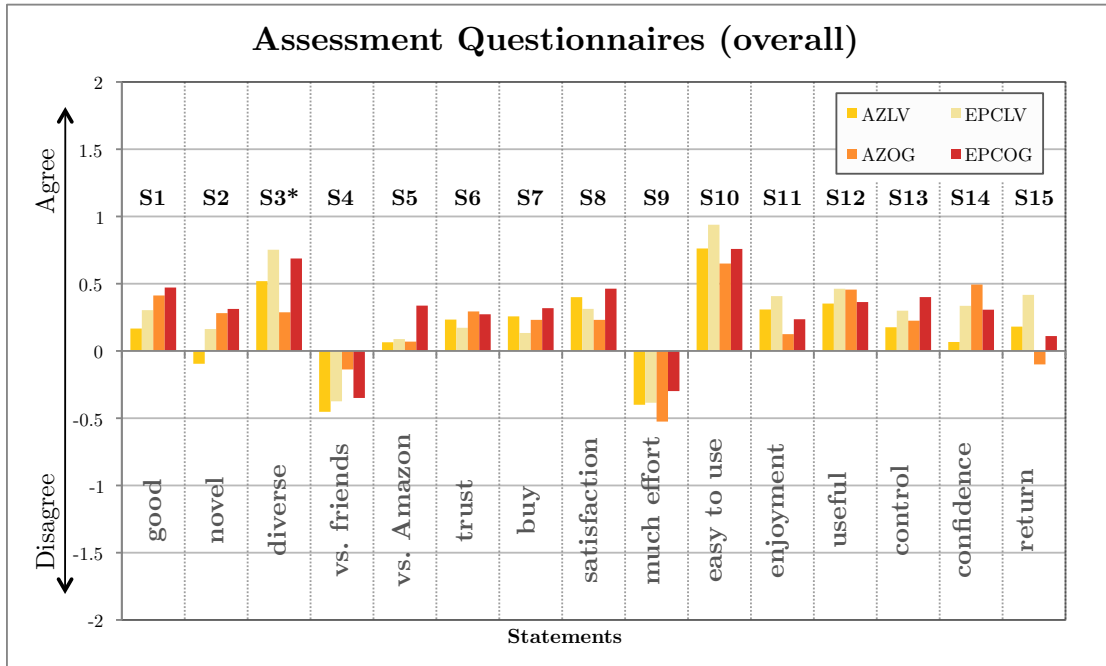


Figure 6.9: Results from the post-stage assessment questionnaires (sessions 1 & 2).

S2, sees many significant scores. The increase from AZOG to EPCOG for novelty is 0.43 ( $p < 0.05$ ) and 0.6 ( $p < 0.05$ ) in reverse order. Just like for S1, users who went from LV to OG in both algorithms saw a significant increase in novelty: 0.83 ( $p < 0.05$ ) for AZ users, and 0.25 ( $p < 0.05$ ) for EPC. The reverse is inconclusive with lower averages and non-significant differences. Novelty, in list view (AZLV to EPCLV) also increases by 0.13 ( $0.05 < p < 0.1$ ). Diversity, S3, also presents two significant results. AZOG to EPCOG shows an increase of 0.14 ( $p < 0.05$ ), and overall users who experience AZ before EPC feel an increase of 0.40 ( $p < 0.05$ ).

We extended our results' analysis by separating users into those who experienced the list view and those who tested the organised layout (independently of the content they experienced). This gave us the graph of Figure 6.10 . Two dimensions (S2, S15) show significant results ( $0.05 < p \leq 0.1$ ). LV seems to provide less novel results than OG (by 0.25), and both scores are relatively neutral. LV users however seem to be more inclined to use it again if this were a real website (by 0.15). We also did the orthogonal analysis, consisting of regrouping data into AZ and EPC testers, independently of the interface experienced (Figure 6.11). This produced two dimensions, which showed significant differences: diversity (S3) and the perception of high required effort (S9) ( $0.05 < p < 0.1$ ). EPC users clearly felt more diversity, although both averages are above the score trend of most other questions. For S9, Amazon users felt it required less effort, although again EPC users also scored below the neutral 0 score.

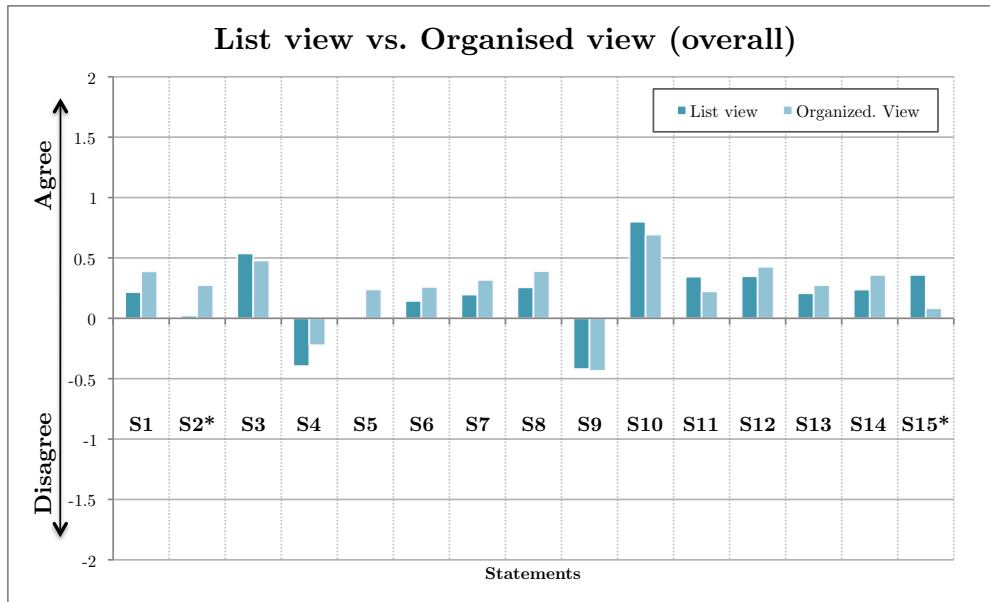


Figure 6.10: Results from the assessment questionnaires divided into List and Organised view.

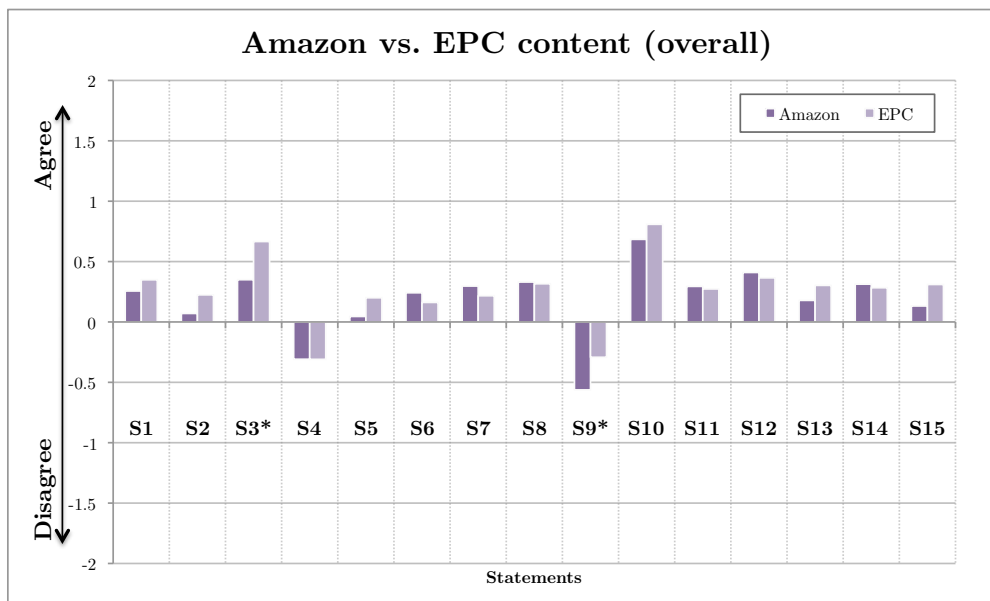


Figure 6.11: Results from the assessment questionnaires divided into Amazon and Editorial Picked Critiques content.

### Final Preference

After both sessions we asked users to tell us which system they preferred. The three possible answers (and scores) were “the first one I tested” (-2), “the second one I tested” (+2) or “both were equal” (0). The results are in Table 6.6. It is interesting to see that users going from LV to OG score an average of 1.14 for this question, against 0.24 for those who tested in the other order (significant,  $p = 0.02$ ). When users test AM before EPF, they score 0.35, whereas the other order scores 0.26 ( $p > 0.1$ ). This seems to indicate that for users’ overall preference, the content make less of a difference than the layout.

### Correlations

The analysis of results also led us to computing the correlations  $r$  across all the data collected. All correlations reported below are significant at the 0.01 level, two-tailed. We have decided to report the five main dimensions that present the strongest set of correlations. We would like to point out that the correlation scores range between 0.4 and 0.7 which can be seen as medium to strong tie-ups. Considering that our experiment did not solely test the recommendations, but the whole website and user-experience of purchasing, we believe that these are acceptably strong.

First of all S1 (the website gave me good recommendations) shows two strong correlations. Both S2 (novelty) and S3 (diversity) correlate at respectively  $r = .50$  and  $r = .53$ , further supporting our earlier results. The two also show a correlation between them of  $r = .35$ .

The second set of correlations is centred on the intention to buy. S7 correlates with the three previously mentioned variables, S1 ( $r = .57$ ), S2 ( $r = .46$ ) and S3 ( $r = .40$ ). The intention to purchase also shows links with trust, S6, which brings us to the next cluster. Trust has always been a key dimensions in recommender systems, and an important result here is that it correlates with S1 ( $r = .60$ ), S7 ( $r = .54$ ), but above all S2: novelty ( $r = .51$ ) and S3: diversity ( $r = .49$ ).

The fourth large set of correlations concerns users’ satisfaction S8. As expected it correlates with the previously mentioned dimension, trust ( $r = .37$ ) and with the intention to buy ( $r = .51$ ). Two other questions that strongly correlate with trust are enjoyment S11 ( $r = 0.65$ ) and usefulness S12 ( $r = .58$ ). More importantly, ease of use S10 shows a link of  $r = .64$  while a high perception of required effort S9 correlates negatively at  $r = -.54$  with trust.

This brings us to the fifth and last group of correlations. This time we look at connections where “users felt that looking for a perfume using this website required too much effort”, S9. All the correlations reported in this paragraph are negative correlations with S9. In a similar way as for satisfaction, enjoyment S12, perceived usefulness S11 and ease of use S10 correlate with high effort at respectively  $r = -.60$ ,  $r = -.45$  and  $r = -.58$ . Furthermore, S9 also is linked with users perceptions of being in control S13 at  $r = -.36$  and finally the intention to use the website again S15 with  $r = -.44$ .

### User Comments

A fifth of the participants left comments at the end of either the first or second experiment. We hereafter report a selection of them. Several comments in the second sessions indicate that users felt more relaxed towards the system they tested. However, it is worth pointing out that this

Table 6.6: Which system did users prefer between both sessions.

| Which system did you prefer? |          |
|------------------------------|----------|
| AZLV                         | 4 users  |
| AZOG                         | 9 users  |
| EPCLV                        | 7 users  |
| EPCOG                        | 7 users  |
| Both were equal              | 35 users |

should not bias the results as there were always two groups testing the same two configurations but in reverse order.

Three users in the first session pointed out that they had not really noticed the recommendation column. We consider this to be acceptable. When we were designing the study, we chose not to focus solely on the recommendations' part, but we preferred to design the experiment on a whole website, in order to observe how recommenders participate in users global interactions with e-commerce websites.

A few comments were made on the layouts. One user of the list view reported that he "hated to go through a long list of recommended perfumes". At the same time, two users found items within each category of the organised view to be too similar. One of them even preferred the list view because "in a sense they are less obvious recommendations". And while one participant expressed the wish to have more explanations on why a specific item was being recommended, two others reported that among all recommendations, the "people who bought also bought" category (EPCOG) was "by far the most interesting".

#### 6.4.4 Discussion

When looking at the objective measures, some important differences are revealed. It appears that the number of pages viewed does not seem to be influenced by layout but by the algorithm, unfavourably for EPC. At the same time, the organised view and EPC seem both accountable for increased session times, with the OG looking to be the most important of the two. At first, this might appear to be problematic since short recommendation sessions are to be preferred as low effort is often correlated with user satisfaction, as in Experiment 3. Here S15 further supports this tendency, with LV users having a higher intention to use the website again. However, when we consider users' final preference, we see that users who have been exposed to the LV first before the OG show a strong preference in favour of the OG, whereas the reverse does not produce any preference, despite the fact that OG requires longer interactions. That being said, it does seem that in all four configurations, users find the system to require a relatively low amount of effort, since the averages for S9 in Figure 6.9 are all lower than the neutral 0 score. We believe that these results indicate that as long as the effort required is low, users are prepared to spend more time when they perceive other benefits such as those of an organised layout.

At this stage, it seems important to understand the benefits of an organised layout and if the users prefer it. The analysis of changes between session one and two is very revealing. Independently of the content tested, users' scores show a strong increase in S1 (good recommendations)

and in S2 (novelty) when users first tested the LV and then the OG. When users do the opposite, the difference is smaller and not significant. Furthermore, the biggest score increase, especially for novelty, is noticed for users of AZ content, which is known to lack diversity and novelty [66]. This is an important result which supports the tendency observable in Figure 6.10 that compares LV vs. OG. On this graph, S2 (novelty) shows that users perceived the organised layout as providing more novelty than the list. These results not only show that users prefer organisation interfaces over traditional list views, when having been exposed to both, but also one of the reasons why: the organisation helps to discover novel items, a real benefit for the user satisfaction.

Novelty is however not the only point worth discussing about the LV vs. OG comparison. This graph shows us that the layout does not seem to influence the measure of diversity, S3. Nevertheless, when we compare users' statements on a AZ versus EPC axis (Figure 6.11) we can observe a strong difference for S3 in favour of EPC. If we put these observations in perspective with the diversity calculations reported in Table 6.5, we can see that EPC was more diverse than AZ, supporting the observed difference. The importance of the content is further highlighted by the analysis of how users' scores evolved between the two sessions, where people first exposed to AZ before EPC scored a 0.4 increase on average, against no increase in reverse order. We believe that this is a fair result, not only because it shows how important diversity is in users' satisfaction, but also because it shows that content plays a strong role in encouraging diversity.

Several correlations are reported in this study. In our opinion, the main contribution here is that for the first time within the same study, novelty and diversity are shown to link directly to many key dimensions mentioned in much of the related work, such as satisfaction and the intention to purchase. More importantly it correlates with trust. This result is interesting because it goes against works by Swearingen and Sinha such as [131] where they had established that the inclusion of previously liked items in the recommendation set increased users' perceptions of trust.

To conclude this discussion, we would like to put these results into perspective. This experiment sheds light on elements which lead to true user satisfaction. Our approach and analysis allowed us to highlight the contributions of both the content selection process and the layout design. However, it is valuable to keep in mind that study "only" tested four specific configurations. We chose to use a classical and popular website design which is as close as possible to a majority of recommender e-commerces in use today, and we crawled Amazon's recommendations which are often seen as a baseline reference. Under this setup we showed that diversity was most influenced by the content algorithm, and that novelty was inspired by the layout. Considering the choices we made in setting up the experiment, we are confident that these results could be applied in other recommenders.



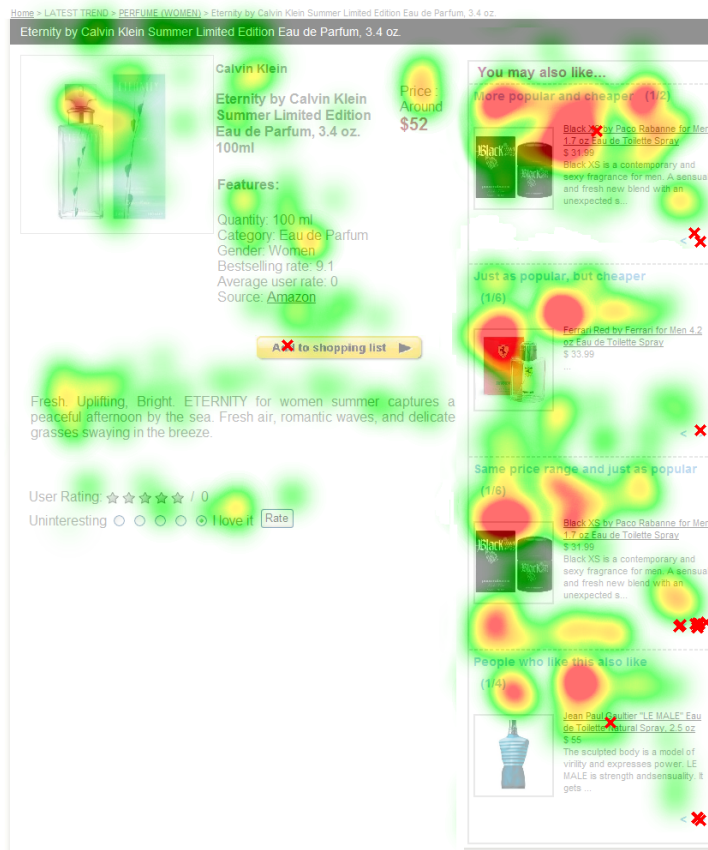


Figure 6.12: Hotspots recorded by the eye tracker on a detail page of a perfume.

## 6.5 Experiment 6: Diversity in Buyers' Decision Process

The second experiment we set up consisted of an in-depth real-user evaluation, with an eye tracker, on a perfume e-commerce framework described earlier in Section 6.3. We observed eighteen people. The eye tracker used in our experiment was a Tobii 1750. This device consists of a computer screen with a camera installed on the top. A software is then capable of capturing the user's points of gaze. Except a short calibration phase, the setup allows users to look at the screen in a natural way without the need for a head mount. The perfume website was setup in its default configuration with the EPC algorithm and the organised layout. The database used was similar to that of Experiment 5, and contained fewer items (3'500 perfumes), crawled again from Amazon's perfume section, and proposed all popular brands and perfumes that are available in regular perfume shops. Using an eye-tracker allowed us to record all of users' gazes. An example of gaze recording is shown in Figure 6.12. The detail of how we analysed the data from the eye tracker is explained in Section 6.5.2.

### 6.5.1 Hypotheses

This section is dedicated to hypotheses that we hope to validate thanks to this experiment.

We expect that the recommender will have an impact on two general aspects of consumer decision making in an online shopping environment: (1) choice strategies, and (2) consideration sets. *Choice strategies* can be thought of as methods (sequences of operations) for searching through the decision problem space [95]. *Consideration sets* are conceptualised as sets of alternatives (one for each user) that consumers consider seriously for purchase [57]. Since our framework does not include a comparative tool of items' features, we will suppose in this chapter that an item is in the consideration set as soon as the user paid attention by looking at it. As explained in Section 6.2, literature shows that these choice strategies can be decomposed into two steps [55]. The goal of users during the first step is to discover adapted search criteria according to their needs and capabilities of the interface. The second step then consists in comparing products which fit to these criteria. After having shown that the recommender impacts these two stages, the goal of this experiment will be to go deeper into detail to understand the exact role of the recommender within the decision process. In particular, we aim at understanding which sub-processes of the product brokering stage are directly involved in the increasing influence of the recommender (e.g. diversity), and what aspects of the recommender help the users to make the purchase decision simpler.

In order to clarify our expectations, we propose a selection of four hypotheses. The first two address the *when* question, and the other two the *why*. To start, we hope to show that the influence of the recommender is independent from the two product brokering steps identified in [55].

**HYPOTHESIS 5** *The influence of the recommender system is independent from the two product brokering steps identified in [55].*

We expect that the influence of the recommender system will escape the subdivision of choice strategies into two steps, and will progressively replace the MCF tool. Furthermore, we believe that this influence should increase without a break during each session until the active user adds a product to the basket, whatever the number of search cycles required to refine criteria for the lexicographic ordering.

**HYPOTHESIS 6** *The influence of the recommender system continuously increases across time at the product brokering stage.*

Escaping the subdivision of choice strategies, we also hypothesise that the recommender should help users to both: (1) refine criteria to use in the lexicographic ordering, and (2) find valuable alternatives to a product in the consideration set. Moreover, this influence should last from one product search to another made consecutively by a user, progressively replacing the MCF tool. We therefore expect the RS's influence to grow continuously. The next two hypotheses are about the *why* question.

**HYPOTHESIS 7** *Users seek opportunities in order to reach a decision, and this need results in a new way of scoping products.*

We believe that classifying recommendations in diverse categories should help within the decision process. Thus, users should pay attention to different categories of recommendations before taking a decision. Browsing the site through the recommender will then become an opportunity of discovering new interesting alternatives. Nevertheless, we expect the users to subconsciously choose one or two categories in accordance with their needs, thus truly influencing their decision. At last, non-experts should use known items as a starting point and count on the recommender to get novelty.

**HYPOTHESIS 8** *Recommendations help users to increase their confidence, when comes the time to make a decision, by fulfilling their need for diversity.*

Beyond believing that the influence of the recommender is maximal when users are close to making a decision, we hypothesise that products added to the basket more often come from the interactions with the recommender than from pure interactions with the multi-criteria search tool. This would confirm that the recommender increases the confidence of users. More importantly, we believe that the influence of the recommender is explained by the users' need for diversity. We consequently expect the users to prefer the recommendation categories providing the greatest diversity.

## 6.5.2 Experiment Setup

### Diversity Metrics

In addition to the diversity metrics presented in Section 6.3.1, we ran diversity calculations on the MCF which is used in this second experiment on diversity. Our predicted values are reported in Table 6.7. We computed the average relative diversity between two products coming from the multi-criteria search tool. This value is dependent on the number of criteria selected by users, but also on the way products are sorted in the ordered list. If for instance, users choose to sort products alphabetically, then the default display which only shows 16 items per list, will have a high probability of resembling a classification by brand, thereby reducing diversity. We therefore distinguish between having the sort criterion linked to selected MCF criteria or not.

Of course, the use of these metrics supposes that we are able to measure the interest of users for the different products proposed through the interface. In the next subsection, we will introduce the indicators of usage that we used.

Table 6.7: Average Relative Diversity between two perfumes coming from MCF tool.

|                                      |     |     |     |     |     |
|--------------------------------------|-----|-----|-----|-----|-----|
| Number of selected criteria          | 0   | 1   | 2   | 3   | 4   |
| Sort criteria $\in$ MCF selection    | -   | 0.8 | 0.6 | 0.4 | 0.2 |
| Sort criteria $\notin$ MCF selection | 0.8 | 0.6 | 0.4 | 0.2 | 0   |

### Indicators of Usage

In order to measure the concrete impact of the recommender on users' behaviour, we chose to combine implicit and explicit criteria.

The first implicit indicator consisted of a collection of access logs stored on the server in a CLF format.<sup>4</sup> These files notably contain information about the pages that users viewed, and the time that users spent on each of the consulted pages. We also used a javascript to instantly track all users' clicks and collect them in the access log files. The clicks highlight deliberate relations between the users and the system.

In addition to log files, we used the data from the eye tracking system. The latter records users' eye movements, allowing us to see both where a person is looking at any given time, and the sequence in which their eyes are shifting from one location to another. Tracking people's eye movements usually helps HCI researchers to understand the factors that may impact upon the usability of system interfaces [51]. In our case, we applied this technology to catch users' usage patterns and better understand decision making sub-processes. Indeed, what a person is looking at is assumed to indicate the "on top of the stack" thought of cognitive processes [68]. The practice of inferring useful information from eye movement recordings involves defining areas of interest over certain parts of a display or interface under evaluation, and to analyse the eye movements which falls within such areas. We used two metrics to reach this last goal: the *fixations* and the *reading heat maps* [26]. Fixations designate moments when a user looks at a particular area for a fair amount of time. Reading heat maps provides an overall view of activities on a page (see Figure 6.12). To create the heat map, data from each person looking at the page is combined to show what percentage of people viewed each part of the page. The colours reference the proportion of participants whose eyes fixated on certain elements of the page. The red areas are where the larger percentages of users looked most.

Apart from the usual gaze plots and heat maps which can be collected with an eye tracker, we decided to rely on a large palette of *Areas Of Interest* (hereafter AOIs). Exporting the time spent on each AOI is important as it is objective data about users' actions on the website. We defined 27 generic types of AOIs. First we created four main AOIs: one for the *multi-criteria* filtering tool (MCF), one for the ordered *list-view* of choices (L), and on the detail page we defined one for the detailed *description* about the desired perfume (D), and finally one for the whole *recommender* part (RS). Secondly we defined many other smaller AOIs which include the recommendation categories, the photo, title, price, quantity & category, rating, the descriptive comment and rating for the detail of any perfume. Likewise we defined AOIs in the multi-criteria search for each category brand, price, quantity, category. A selection of all these AOIs is detailed in Figure 6.13.

### 6.5.3 Experiment Procedure

This second study was designed as an in-depth one hour lab-study. At all times, participants could ask questions to the available assistant conducting the study. The material given to participants is detailed in Appendix F. The general online evaluation procedure consisted of the following steps:

---

<sup>4</sup><http://www.w3c.org/Daemon/User/Config/Logging.html>

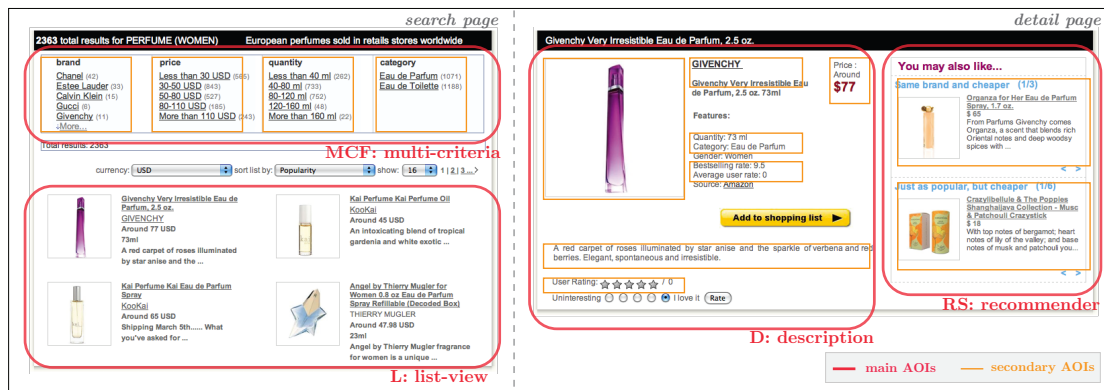


Figure 6.13: Snapshots of the AOIs on the *search page* (a) and the *detail page*.

**Step 1** The participants are welcomed by the assistant. They are briefly introduced to the topic of the experiment. They are informed that the perfume e-commerce website they will test contains over 3,500 most commonly used and sold perfumes in the world. The users are also told about the incentive.

**Step 2** A detailed set of questions is asked to each user, so as to collect their background information (age, sex, etc.). Several questions are included to measure the users' familiarity with the perfume domain. The answers to this step are detailed in the following section.

**Step 3** Before starting the experiment, the eye tracker is calibrated to the user's eyesight. The experiment can now start and the tracking session is launched by the assistant, who encourages the users to explore the system before fully launching into the first task.

**Step 4** The users have then two separate tasks which they are asked to complete.

**Session A** One goal is to select up to three perfumes that they have never heard of or used before, but that they would be prepared to buy for themselves. They are asked to put them in the basket, and are informed that they may select more than three and delete some at the end. In the rest of this chapter, this recording will be called *Session A*.

**Session B** The other goal consists in searching for one perfume that they would like to offer to someone, preferably from the opposite sex, such as to reduce potential bias of product habituation. We call this part *Session B*.

In order to reduce another bias linked to the fatigue, we alternate the order of sessions. Half of the users complete the task of Session A, before Session B. Others start with the Session B and end with Session A. These two sessions are recorded in two different files.

**Step 5** To conclude the study, fourteen preference questions are asked in order to explicitly assess users' overall perceptions of the system after the experiment. This allows us to match explicit and implicit data to confirm our hypotheses.

Table 6.8: Post-stage assessment questionnaire.

| ID   | Statement   |
|------|---|
| P1.  | The items under “You may also like” are attractive.                         |
| P2.  | The items under “You may also like” are educational.                        |
| P3.  | The items under “You may also like” appeared to be a good deal.             |
| P4.  | The items under “You may also like” appeared to be marketing material.      |
| P5.  | The items under “You may also like” influenced my selection.                |
| P6.  | The items under “You may also like” will influence my future selection.     |
| P7.  | The names of the categories are useful and adequate.                        |
| P8.  | I am satisfied with the overall quality of the interface.                   |
| P9.  | I found the interface easy to use.  |
| P10. | I would buy the perfumes recommended to me, given the opportunity.          |
| P11. | If this were a real website, I would use it in the future to find perfumes. |
| P12. | I believe that the recommender algorithm is efficient.                      |
| P13. | The recommended perfumes were diverse.                                      |
| P14. | The recommended perfumes were novel.  |

The preference questions were, as in the other study, statements to which users could indicate their level of agreement on a five-point Likert scale, ranging from  $-2$  to  $+2$ , where  $-2$  means “strongly disagree” and  $+2$  is “strongly agree”,  $0$  is neutral. The post-stage questions are listed in Table 6.8. The questions were asked in a random order, such as to make sure no ordering bias appeared.

### Participants’ Background

The user study was carried out over a period of three weeks. Immediate incentives, chocolate or wine, were offered right after the study. More importantly, users who had completed the study took part in a draw for a CHF 100.- voucher to buy one of the perfumes they had added to the basket. By telling users about this price before the study, we maximised chances that users would behave truthfully. A total of eighteen volunteers were recruited as participants. They were from three different continents, with different professions (student, worker, Ph.D. student) and educational backgrounds (high school, graduate school). Table 6.9 provides some of their demographic characteristics.

All of the participants expressed the opinion that *fragrance* was an important feature, necessary in order to describe a perfume as shown in Figure 6.14. Other important aspects included price, brand, quantity & design. Amid the background questions, two assessed users’ experience levels with regard to the Web (online media, information retrieval, Internet communication, online communities, online entertainment and e-commerce) and online shopping. Both revealed that all users’ had strong web experience, although online shopping experience remained limited to classical items such as books, music, travel and electronic items, much less so for food, drinks, groceries or clothes.

Table 6.9: Demographic characteristics of participants.

|                            |  |  |                    |
|----------------------------|--|--|--------------------|
| Gender                     | Female<br>10 (56%)   | Male<br>8 (44%)  | Total<br>18 (100%) |
| Age                        | 17-25<br>5 (28%)   | 26-40<br>12 (66%)  | 41-55<br>1 (6%)    |
| Education                  | High school, Graduate school                                 |  |                    |
| Profession                 | Student, Ph.D. student, Worker                               |  |                    |
| Nationality                | American, Chinese, French, German, Russian, Serbian, Swiss   |  |                    |
| Online Shopping Experience | travel<br>books<br>music<br>electronics<br>food<br>groceries | 94% a few times a year<br>56% a few times a year<br>22% monthly<br>39% a few times a year<br>88% never<br>28% a few times a year |                    |

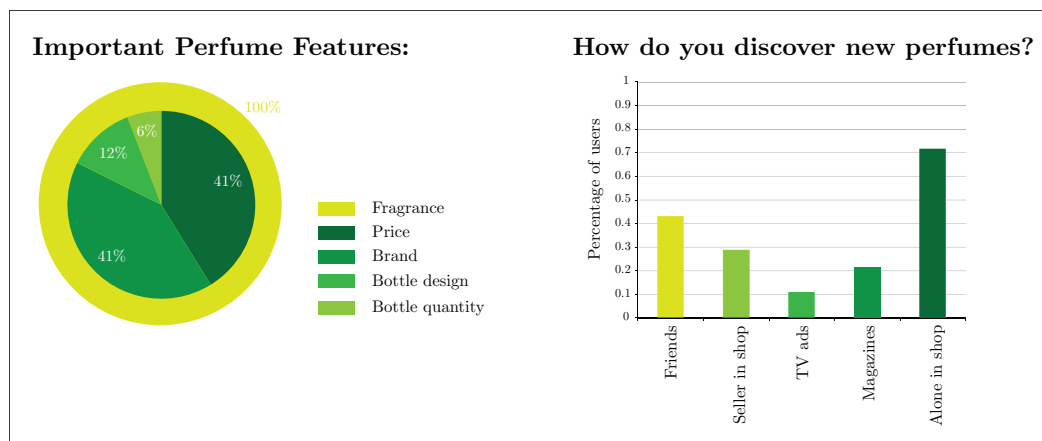


Figure 6.14: Users' background knowledge about perfumes.

The second part of the background questions surveyed users' predisposition towards perfumes. Six of the eighteen participants considered they were knowledgeable about perfumes. In the rest of the chapter, we will call them "perfume experts". Six participants said they bought perfumes about once a year, nine a few times a year and one nearly monthly. When questioned about how they discovered new perfumes, 61% said that they preferred to just test perfumes alone in a shop. Other answers included relying on friends' advice, or the seller in the shop, as shown in the second graph of Figure 6.14. Most users also told us that they were prepared to reveal information such as previously liked & disliked perfumes, and price, in order to get recommendations (from a seller). Interestingly, all users were prepared to reveal information about smells that they liked, but told us it was difficult to describe them and hence relied on other aspects.

#### 6.5.4 Analysis of Results

Thanks to the eye tracking system, we first aimed at measuring how users' interest for the different parts of the website evolves as time goes on. Averaging 1,350 fixations per user and per session, we recorded 48,891 fixation points throughout the study. We defined 7,720 AOIs, sorted into 593 different web pages and corresponding to 27 variables. These pages were of two sorts as explained above: the search pages, and the detail pages. The Table 6.10 synthesises the number of average pages seen by users and the average session times. The statistics from the group of users who experienced session A before session B are very similar to that who experienced the sessions in the other order (around 10% of difference as regards the number of search pages and detail pages viewed in each case). Because the experiment was quite long, we checked this potential difference in order to dismiss the fatigue as a factor influencing the users' behaviours.

Table 6.10: Statistics of sessions for the overall set of users.

|                                | Session A | Session B | Both  |
|--------------------------------|-----------|-----------|-------|
| Average Number of Search Pages | 6.53      | 4.61      | 11.22 |
| Average Number of Detail Pages | 10.67     | 4.72      | 15.39 |
| Average Session Time (Minutes) | 10.43     | 5.17      | 7.80  |

We then computed the total fixation duration for each user  $\in U$  on the different AOIs over time  $t$ . We particularly paid attention to duration of four variables: the multi-criteria box  $M$ , the lexicographic ordered list  $L$  (list-view), the description of perfumes  $D$  and the recommender system  $R$ . Two representations were chosen to show the usage of these elements. Firstly, we used a bubble view where the circles' diameters represent the duration spent on the selected element; this is shown in Figure 6.15. A shift from the MCF to the RS can quite convincingly be seen on the graph, where the MCF and list-view seemed to be most used at the start, but where the recommender seems to be more and more used as time passes.

Secondly, we focused on the specific usage of the MCF tool versus the RS agent over time for the overall set of users  $U$ . We considered the time spent looking at  $M$  and  $R$  boxes respectively. We summed the cumulative fixation duration of these two AOIs and made them explicit in Figure 6.16. Consequently, the curves tend to flatten when the users stop looking at the cor-



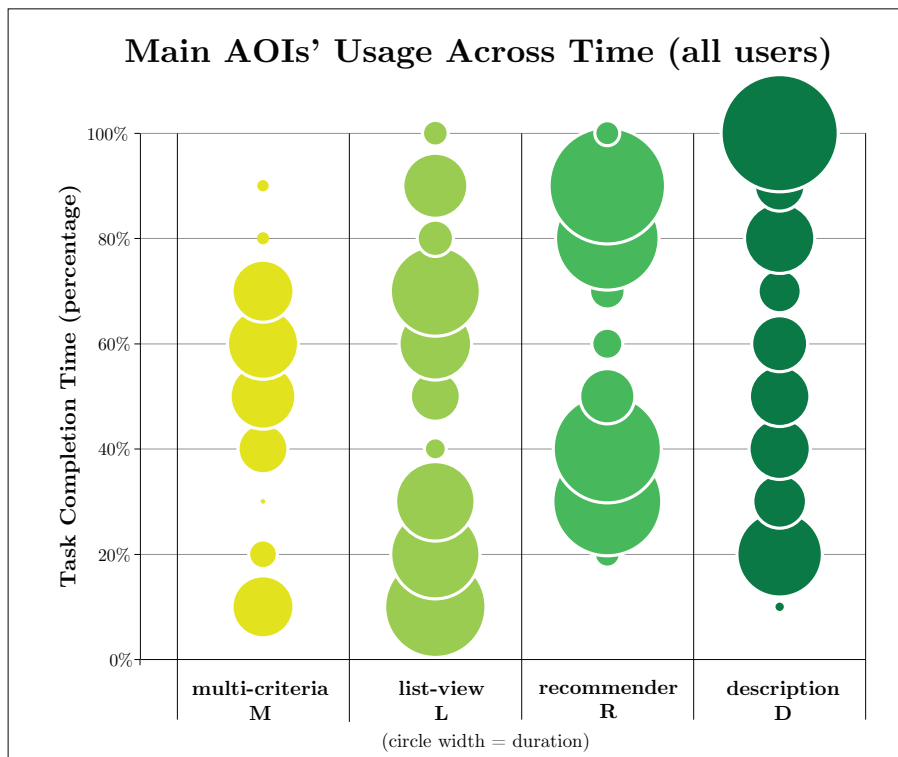


Figure 6.15: Bubble view of time spent using the four main tools.

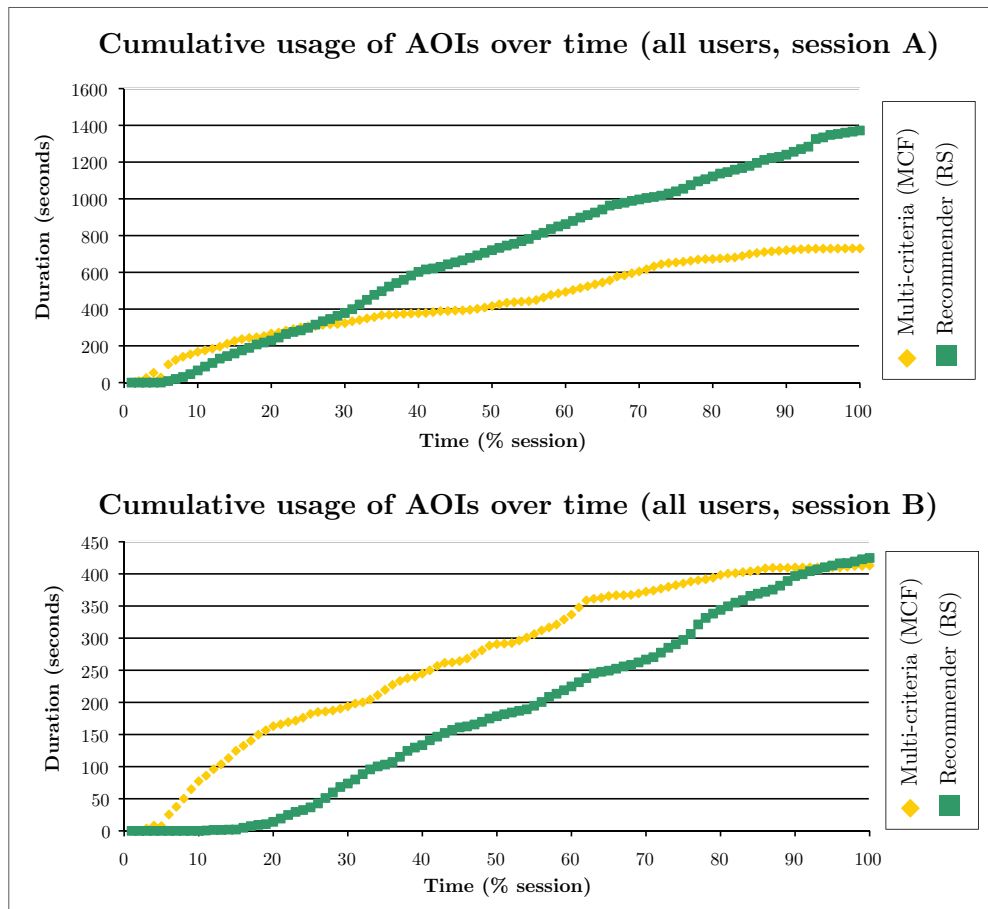


Figure 6.16: Usage in terms of cumulative time of Recommendations compared to Multi-Criteria.

responding AOIs. We noticed that the use of RS increases much faster than the use of MCF as time goes on. The RS has a linear progression as soon as users start going on detail pages (the first visit on a detail page occurred on average around 18% of the total time in session A, and around 37% in session B), while MCF followed a curve closer to a logarithmic one.

As the previous result could have been induced by different sizes of AOIs, we normalised these duration measures. We will use this normalisation in the rest of this chapter. In order to do so, we defined the variable  $R: U \times \mathbb{R} \rightarrow \mathbb{R}$  as:

$$R_u(t) = \frac{\text{fixation duration of user } u \text{ on } R \text{ in } (t-1; t]}{\text{total fixation duration of } u \text{ on } R} \quad (6.4)$$

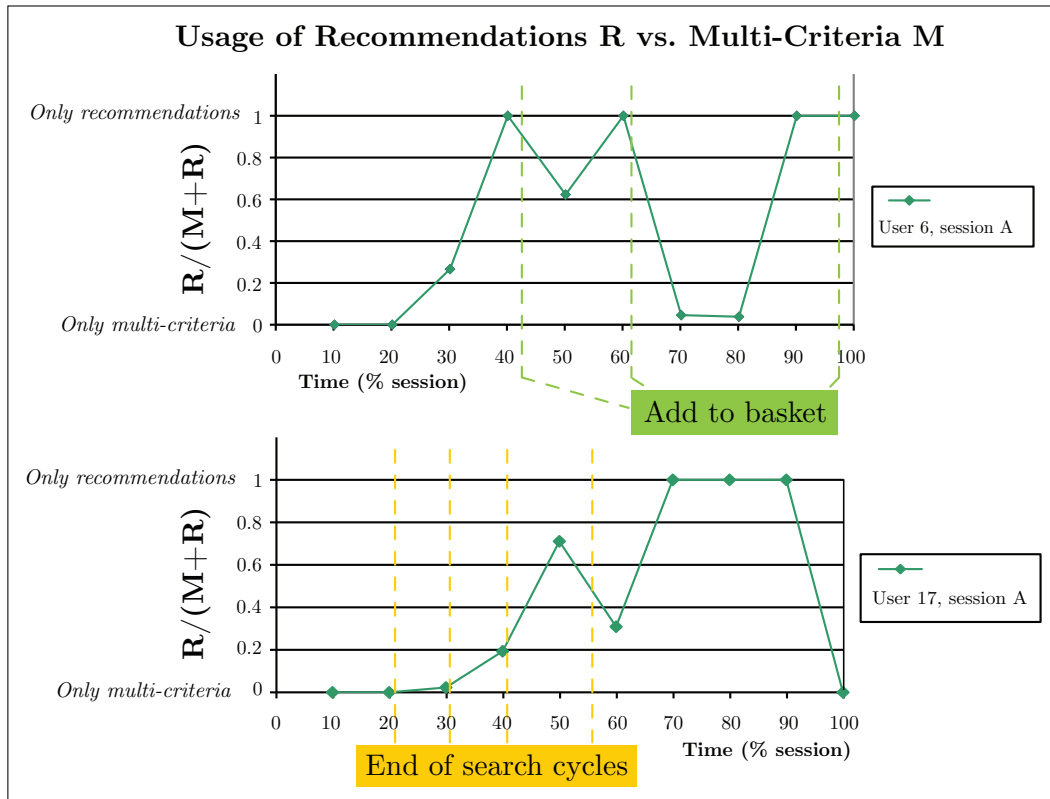


Figure 6.17: Two examples of MCF and RS over time, with purchase decision peaks.

We then measured the usage of recommendations in comparison with the multi-criteria box for each user  $u$  and each session  $s$ , by computing the function  $f$ :

$$f_{u,s}(t) = \frac{R_u(t)}{M_u(t) + R_u(t)} \quad (6.5)$$

In order to increase our understanding, we also analysed the actions of users from the implicit access logs to determine the time at which users added products to their basket. This allowed us to cross data from the eye tracking system and the access logs. Through this approach, we managed to show that, for most of users (i.e. for 86% of products added to the basket for both sessions), the function  $f_{u,s}$  was increasing constantly until a product was added to the basket. Each peak of the function consequently corresponds to a purchase decision. For illustration purposes, we chose two representative examples of the function  $f$  and displayed them in Figure 6.17 (Users no. 6 and 17, Session A). The light green vertical lines in the first graph of Figure 6.17 correspond to times at which the user added a product to the basket (yellow lines explained hereafter).

Table 6.11: Average numbers of cycles (all users).

|                       | Both sessions concatenated in the order<br>in which users did the experiment |           |           |           |
|-----------------------|--|-----------|-----------|-----------|
|                       | Product 1  | Product 2 | Product 3 | Product 4 |
| Search cycles         | 2.1  | 0.9       | 1.2       | 2.2       |
| Recommendation cycles | 0.7  | 1.0       | 1.7       | 1.8       |

Thanks to action logs, we then computed the number of *search cycles* and *recommendation cycles*, which we define as follows:

**DEFINITION: SEARCH CYCLE** *We define a Search Cycle as a period of time within the session, for which the selected MCF criteria are fixed. Changing one or several criteria implies to begin a new search cycle.*

**DEFINITION: RECOMMENDATION CYCLE** *We define a Recommendation Cycle as a period of time between two clicks on a recommendation.*

For purposes of illustration, we displayed the search cycles (in yellow) of the user no. 17 during the session A in the second graph of Figure 6.17. This example can be generalised in the sense that for each user, none of the search cycles end at a time where there is a peak (or a hollow) of the function  $f_{u,s}$ . Compared to the 86% of cases where an addition to the basket is neighbouring a recommendation peak, it seems clear that the RS is critical in users' decision process. We calculated the average number of search cycles and recommendation cycles that users accomplished during session A and session B. Results are shown in Table 6.11, where we decided to order results based on the order in which users took the experiments. This way we can look at whether the RS's influence continues to increase across both experiments. Results support this, since no pattern can be seen in the search cycles results, whereas the average number of recommendation cycles increases from one session to another.

To complete this picture, we considered the number of MCF criteria which were active each time a product was added to the basket. It appears that the average number of MCF criteria used to reach a sought-after product is 2.0 during session A, and 1.6 during session B. We analysed user's overall clicks in the interface. Table 6.12 synthesises the average number of clicks on MCF criteria and recommendation categories, all users taken together. A selection of values is highlighted for the discussion later in this chapter.

After computing the global influence of RS as regards to explicit clicks, we aimed at figuring out if eye movements revealed the same dominating and influential recommendation categories within the recommender system. We defined the following:

Table 6.12: Average numbers of clicks (all users).

|     |                                   | Session A  | Session B  | Both sessions |
|-----|-----------------------------------|------------|------------|---------------|
| MCF | Brand                             | <b>2.1</b> | <b>1.3</b> | <b>1.7</b>    |
|     | Price                             | 0.7        | 0.4        | 0.6           |
|     | Quantity                          | 0.4        | 0.1        | 0.3           |
|     | Category                          | 1.2        | 0.6        | 0.9           |
| RS  | More popular and cheaper          | <b>0.9</b> | <b>0.4</b> | <b>0.7</b>    |
|     | More popular, more expensive      | 0.2        | 0.0        | 0.1           |
|     | Same brand and cheaper            | 0.7        | 0.1        | 0.4           |
|     | Same brand, more expensive        | 0.3        | 0.2        | 0.3           |
|     | Just as popular and cheaper       | 0.3        | 0.1        | <b>0.2</b>    |
|     | Same price range, just as popular | 0.5        | 0.1        | <b>0.3</b>    |
|     | People who like this also like    | <b>0.9</b> | <b>0.3</b> | <b>0.6</b>    |

**DEFINITION: DOMINATING CATEGORY** We define a recommendation category as dominating if:

$$\% \text{ of usage of a category} > \frac{100\%}{\text{No. of categories}} \quad (6.6)$$

**DEFINITION: INFLUENTIAL CATEGORY** We define a recommendation category as influential if a product added to the basket directly comes from a recommendation of the considered category.

Heat maps allow us to measure the impact of each of the seven recommendation categories on the overall set of users. We computed the number of times each user looked at a recommendation category. We normalised by dividing these values with the number of AOIs linked to each recommendation category. Figure 6.18 shows that each category has a significant importance when we merge all users' data (weighted means according to session duration) and there were four dominating categories. However, eye movements confirmed data from Table 6.12 according to which customers seem more attracted by more popular items and products from "people who like this also like".

The arithmetic means of recommendation categories' usage for all users are displayed in Table 6.13. We compared the number of dominating recommendations with the number of recommendation categories looked at, the number of recommendation categories used (when users click on items of these categories), and the number of influential recommendation categories.

At last, we evaluated the proportion of products added to the basket thanks to RS, in comparison with those which came from MCF (see Table 6.14). Among the 69 products globally added to the basket, 31 perfumes only came from the MCF tool without any interaction with RS. 28 products had been added just after a click on a recommendation (influential category). 10 perfumes were added following one or several comparisons with some recommendations (by going on detail pages of recommendations and then going back to the previous page). Of course, most of the perfumes coming from the MCF tool have been added after taking a look at the dif-

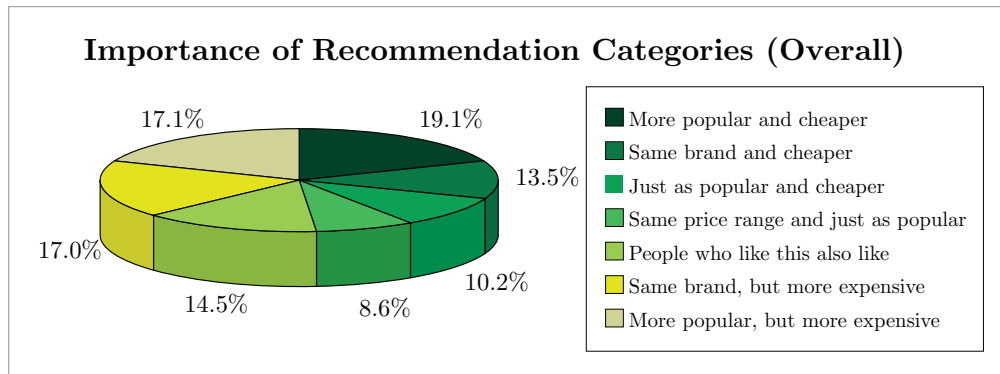


Figure 6.18: Proportion of time spent looking at each recommendation category (all users).

Table 6.13: Average effects of recommendation categories.

|           | Number of dominating categories | No. of categories looked at | No. of categories used | No. of influential categories |
|-----------|---------------------------------|-----------------------------|------------------------|-------------------------------|
| Session A | 2.74                            | 5.89                        | 2.74                   | 1.37                          |
| Session B | 1.95                            | 4.42                        | 0.84                   | 0.26                          |
| Both      | 2.34                            | 5.16                        | 1.79                   | 0.82                          |

ferent recommendations on the detail page, but without clicking on them. Users rely more on the recommender RS than on MCF (38 vs. 31 in Table 6.14). The testers completed two tasks through sessions A and B, in a random order. When we consider solely the first task, independently of which session is concerned, we note that 23 selected products came from RS and only 11 from MCF (significant,  $p = 0.049$ ). The difference is less strong for the second task (19 RS vs. 17 MCF,  $p = 0.7$ ), possibly due to habituation or fatigue. Overall, if we take the ordering into account, RS scores 42 (60%) vs. 28 (40%) for MCF.

Table 6.14: Number of products added to the basket that came from RS vs. MCF.

|   | Session A | Session B | Both      |
|---|-----------|-----------|-----------|
| Added after MCF selection without RS                  | <b>19</b> | <b>12</b> | <b>31</b> |
| Added after having been influenced by RS              | 25        | 3         | 28        |
| Added after interacting with RS                       | 7         | 3         | 10        |
| Total number coming from RS (influence + interaction) | <b>32</b> | <b>6</b>  | <b>38</b> |

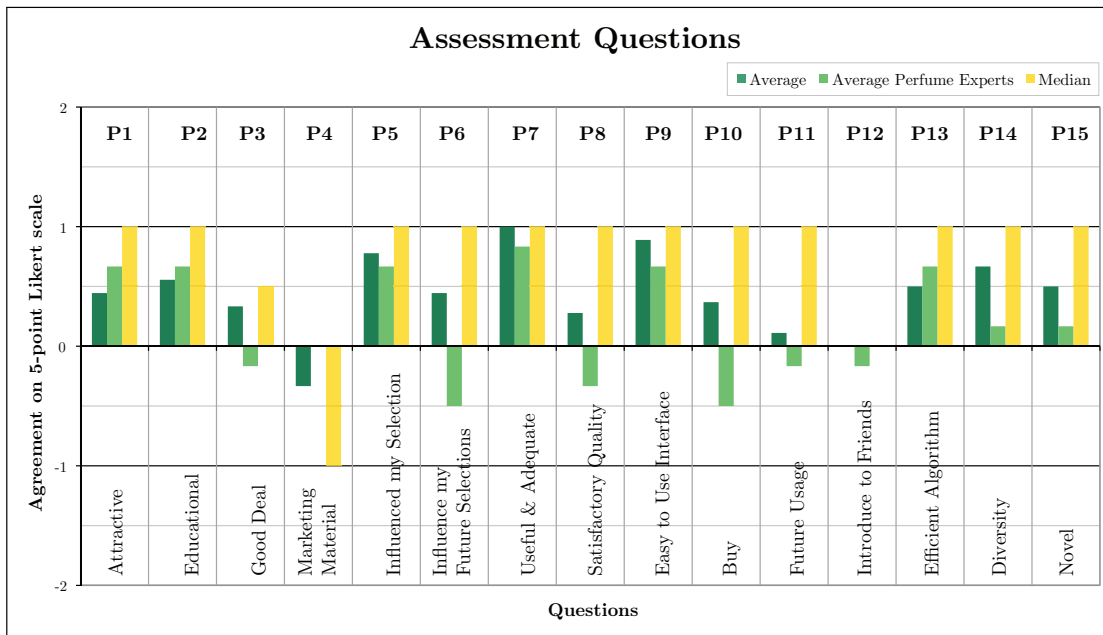


Figure 6.19: Answers of the assessment questionnaire.

### User Preferences

After the experiment, we asked users to fill the assessment questionnaire displayed in Table 6.8. Answers of this post-study survey are summarised in Figure 6.19 where we report the average scores of users. We also display the average answer of perfume experts (determined according to the initial background questionnaire) and the overall median.

In order to measure and ensure the veracity of decisions to add products in the basket, we asked users if they would buy the chosen perfumes given the opportunity or at least go in a perfume shop to smell them and learn more about them. 56% of users agreed that they would be prepared to buy given the opportunity; 17% were not sure to buy them, but agreed to smell them before making a final decision. The answers of the assessment questionnaire (see Figure 6.19) showed that both standard and expert users found the recommender attractive (P1), educational (P2), useful (P7), easy to use (P9) and efficient (P13). However, regular users seemed more influenced by the RS than experts (P5 and P6). Encouragingly, average users did not seem to perceive recommendations as marketing material (P4). Standard users strongly expressed their satisfaction as regards diversity (P14), whilst experts much less. The correlation analysis revealed that the recommendation diversity (P14) is strongly correlated to intention to buy (P10), and to influence of selection (P5 and P6). These correlations are strong and statistically significant with respectively:  $r = .547$  ( $p = 0.015$ ) and  $r = .579$  ( $p = 0.009$ ). At last, diversity is correlated with satisfaction ( $r = .466$ ,  $p = 0.044$ ) and ease of use ( $r = .487$ ,  $p = 0.034$ ). In complement, we ran a factor analysis to uncover the latent structure of the assessment ques-

tionnaire's variables. The cluster composed of the factors "diversity" ( $r = .777$ ), "intention to buy" ( $r = .872$ ) and "future influence" ( $r = .704$ ) explains for the most part the variability of answers. The reliability of this result is assessed by the Cronbach's alpha ( $\alpha = .756$ ), with a significance of the Bartlett' Test of Sphericity equal to 0.001. The perceptions of users are consequently in accordance with the hypotheses proposed.

The next section is dedicated to a discussion about these results, which strongly support our hypotheses. The key results of the correlation analysis which was computed, are made explicit and also discussed in the next section.

### 6.5.5 Discussion

Looking at the users' final preferences after the study, we were somewhat surprised. We were expecting to have some more contrasted results, especially regarding diversity and novelty. Unfortunately, P14 and P15 do not stick out from other results. The only observation regarding these two measures, is the fact that expert users score lower on both questions. We suspect that this is mainly due to the fact that more expert users know a larger number of perfumes, and thus feel that there is less diversity among the selection of recommended perfumes than someone less knowledgeable on the topic. We believe that the more noteworthy results are not among the subjective assessment questions, but more among the objective measures recorded by the eye-tracker.

These results show that, for every single user and for the overall set of users, the use of the recommender system progressively replaces the use of multi-criteria search over time according to fixations (cf. Figure 6.16), and average number of search cycles compared to recommendation cycles (see Table 6.11). As the number of search cycles remains relatively stable, the number of recommendation cycles constantly increases. This tendency is clearly visible in Figure 6.15. All these elements support our HYPOTHESIS 6, according to which the influence of the recommender system continuously increases across time at the product brokering stage.

We also successfully reproduced the hypotheses of [55] about the purchase decision making process' division into two steps: the first one to identify a subset of products, the second to explore alternatives. We indeed noticed that each participant started the experiment by identifying adequate search criteria. Thanks to users' access logs, we know that this step required an average number of two search cycles. During these search cycles, the user does not simply stay on the first page. We see from our data that, having chosen some search criteria, participants are then reading examples of corresponding perfumes' detail pages, before going back to the search page to refine the search criteria. We were not able to establish any explicit link showing that RS influenced user's selection of criteria. All interviewees then used the selected search criteria in order to consider alternatives, moving to the second step of [55], before making a decision. It therefore appears that the two-step division of decision process does not show any dependance on the RS, as shown in Figure 6.17. At the same time however, as we pointed out at the beginning of this discussion, our study results suggest that the recommender has a strong influence, and that this influence increases throughout the whole time. These elements tend to support our HYPOTHESIS 5, which states that the impact of the recommender system is independent from the two product brokering steps identified in [55].



Our experiment also highlights that this effect is greater at the end of the decision process (see the typical schemes in Figure 6.16). In nearly all product searches we can see that the function  $f_{u,s}$  is continuously increasing, from the beginning of the search to the moment the user added the product to the basket. As a continuation of these observations, we examined the impact of the different categories in our recommender systems. We computed both the clicks and time spent looking at each of these categories individually and for the overall set of users. Results in Table 6.9 highlight a high visual interest of users for almost every category, even if only two of them influenced their choices on average. Figure 6.18 and Table 6.12 outlined the preferences of users for categories “More popular and cheaper” and “People who like this also like”, despite the fact that the most selected MCF criterion is the brand. As shown earlier in Table 6.1, these two categories are those which propose the highest diversity. We believe this shows that users are open-minded to diversity. Our results also pointed out the disinterest for categories “Same price range and just as popular” and “Just as popular and cheaper”. By consequent, the increasing influence of RS is characterised by an attraction for categories providing the biggest levels of diversity (see Tables 6.1 and 6.7). The most popular categories of RS offer the same level of diversity than the list-view (L) with one selected MCF criterion, but with a higher accuracy. This explains why the users’ need for diversity led them to use the recommender, rather than the MCF tool. Moreover, Table 6.14 shows that the majority of products added to the basket came from interactions with the recommender. This strongly supports our HYPOTHESIS 8, according to which RS increases the confidence of users by providing diversity.

Finally, it appears that users were often browsing the site through the recommender, which thus becomes an opportunity for discovering new interesting alternatives. However, users do not envisage to be influenced by the recommender system (P6). Moreover, both the number of recommendation categories used (in terms of clicks) and the number of influential recommendations correlate strongly with the intention to use in the future (P11). These correlations are respectively  $r = .966$  ( $p < 0.001$ ) and  $r = .810$  ( $p = 0.027$ ). In consequence, it seems that users reached a satisfying decision by taking a look at different categories and unconsciously prioritise them according to their needs, in order to find the appropriate product. These results come to support our HYPOTHESIS 7 as it means that users are scoping products in a new way, seeking opportunities to reach a decision.

## 6.6 Conclusions

In this chapter we have investigated three key questions about diversity in recommendations: *how* do users perceive diversity, *why* do they need it and *when* is diversity useful in users’ buying decision process. We set-up two large user studies which allow us to bring some answers to these questions.

To start, we explored how both layout and content might influence users’ perceptions of recommendations in terms of diversity. Thanks to an in-depth within-group user study, we opposed the performance of Amazon’s recommendations against Editorial Picked Critiques, thus giving us two very different algorithmic approaches. In the same study, we also compared a traditional list-view representation to an organised view. Organisation interfaces have been shown to be effective in improving users’ perceptions of recommendation quality, and increasing

their system-acceptance. Under the setup of this experiment, we argue that diversity and novelty are two important dimensions leading to user satisfaction. Our data shows that diversity is most influenced by content, whereas novelty is enhanced through layout. Furthermore, when exposed to both a list and organised layout, users' prefer the organised view despite longer interactions. Finally, our work supports that diversity increases trusting intentions, leading to purchase intentions.

Specifically, our results show that when exposed to both Amazon's recommendations and Editorial Picked Critiques, users do not have a preference for either system. Nonetheless, when being exposed to two layouts, a list and an organised view, users show a stronger preference for the organised view. The detailed analysis of data also points out that participants found EPC as being more diverse than Amazon's recommendations, although the latter was seen as requiring less effort. The orthogonal inspection of the list versus organised view found that organised was considered more novel, while still being diverse. However, despite these observed differences, the overall result in terms of users' perceived qualities, is that the four setups are perceived very similarly when they are just integrated into a shopping website. Users do notice diversity but not in an obvious way.

In our second study, we examine the impact of a recommender system on customers' decision process. We went by the two-steps model of Häubl *et al.* [55] with the aim of discovering new usage patterns and decision sub-processes, using an eye-tracking system. Ultimately, our goal was to identify the role of diversity in this process. Relying on a similar perfume simulation website as in the first study, we examined the impact of RS all over the product brokering stage and showed how this technology changes the way customers come to an online purchase decision. With the aim of discovering these new usage patterns, we looked very precisely at behaviours of eighteen users confronted to a perfume e-commerce website during a one-hour lab-study. During this experiment, we were able to track users' actions, and to collect almost 49,000 fixation points with the eye tracking system. This data allowed us to approximate (with a strong degree of confidence) users' cognitive paths, that led them to a purchase decision through interactions with the system. We then paid attention to how the influence of recommender systems integrates into the purchase decision making model. The analysis of our results, cross-checked with the users' assessment questions, leads to three major conclusions. First, the influence of the recommender does not appear to be constrained by the two steps of the decision process [55]. It just increases in an interrupted process until the purchase decision is close to being taken. Second, users rely on the recommender to enhance their confidence in the purchase decision. Third, we outlined the need of users for diversity when making a decision. This not only comes to support the theory according to which it is necessary to find a good compromise between accuracy and diversity in order to increase quality of recommendations, but it also provides an excellent guideline about when this diversity is needed. Said otherwise, we highlight a new exploration model where customers efficiently use categories of recommendations to obtain diversity. This model consists of two elements: looking at different categories to envisage new alternatives, and then to unconsciously prioritise these categories until one of them sufficiently fits the user's needs so that it provides confidence and induces the decision.

These studies constitute a major step towards the understanding and formalisation of users' online behaviours. We integrate these findings into our discussions of Section 7.5 and propose a

theoretical model for maximising users' satisfaction by balancing users' needs for accuracy and diversity throughout the decision process.



## Chapter 7

# Design Guidelines and Diversity Model

From a certain perspective, the work of this thesis can be seen as being somewhat unconventional. The ultimate goal is less the exploration of a low-level phenomena, but more the general understanding of factors leading users to accepting recommendations and adopting the whole system. In this chapter we aim to put forward the lessons learned from the different topics studied, thanks to a series of design guidelines. The specificity of these guidelines, and to what extent they can be generalised, is discussed in Section 7.4.1.

The challenge while designing a recommender system, especially for entertainment products, is often to make delicate trade-offs between opposing requirements. We have attempted to address these issues through several approaches. We studied user's behaviours and perceptions leading to the *acceptance* of recommendations, and possibly their long-term *adoption*. We also considered how users' control, through a comparison of *explicitly* and *implicitly* revealed preferences, changed their overall satisfaction. Another approach we studied was how influential the interface design could be compared to content-driven solutions. Finally we explored users' need for *diversity* and how it leads to *confidence*. Motivated by these experiments' results, certain repeated observations and user comments, we derived a set of design guidelines which we elaborated around four primary axis: user effort, purchase intentions, complex systems and diversity. These axis were chosen as they were the main aspects we studied in all our experiments. We believe they will be helpful to the research community for the future development of recommender systems.

### 7.1 User Effort

In the design of our experiments destined to explore and characterise the concepts of acceptance and ultimately adoption (Chapter 3), we ended up considering a large number of dimensions. These were selected for their potential to influence users and also because they belonged to (or extended) the TAM. The dimensions included quality measures such as enjoyability, novelty or accuracy, but also measures of effort such as registration time, time to recommendations and ease of use in various cases. The user effort measures were chosen as they represent one of the key elements of the TAM: perceived ease of use.

In both user studies (Experiment 1 and 2) user effort demonstrated some significant results

from users. In the first experiment, the two music recommenders *Pandora* and *Last.fm* were compared in a within-subject comparative user study involving 64 participants. We investigated users' initial acceptance of recommender technology and users' subjective perception of the respective systems. Without making a direct link with the TAM, our results point out that the initial effort when using a system for the first time is a key factor in users' overall satisfaction. We show that *Pandora* gets users started in three times less time on average, that the time needed to get the first personalised recommendations is much shorter. Besides, user's comments point out that installation difficulties and interface complexity were the two biggest reproaches that could be made to *Last.fm*. The correlation analysis further supports these observed effects. One of the conclusions of this first experiment was to minimise user effort in terms of the registration time, the download time and the time to get recommendations. In order to expand on these results and to explore dimensions of the TAM more deeply, we setup a new user study with *Pandora* and *Last.fm* where we tested promising factors in a higher level of detail. We separated user effort (perceived ease of use) into several dimensions thanks to six questions. We included system complexity and the amount of involvement needed to obtain desired songs. Like in the first study, our results showed that one of the biggest differences between both systems came from the effort questions. Five of six EOU questions showed significant results, including the time needed to reach the first few enjoyable songs. The correlation analysis further endorsed these observations with Q6, a question on users' satisfaction presenting the strongest correlation with Q23, the assessment of initial time needed to reach interesting music.

This collection of observations linked to user effort in the first stages of using a system, led us to propose the following first guideline:

**Guideline 1.** *Making sure the initial time to recommendation is short helps to increase user satisfaction.*

Another observation central to both studies was made about the number of features proposed and the general ease of use. In Experiment 1, users commented on difficulties met with *Last.fm*'s interface. They described it as being not intuitive, not clear or not comfortable. In Experiment 2, we took the opportunity to directly assess this issue by asking (Q26) if the website offered too many features irrelevant to music recommendations. Results indicate a high number of features are a problem for *Last.fm*. The correlation analysis further supports this claim by linking Q26 with satisfaction (and mood). In addition, and at a more general level, the assessment of ease of use for the two tasks of the system (listening to and discovering music) shows how strongly users feel about this dimension. In the third experiment, we compared Amazon's traditional user-controlled interface with more recent personalised systems using recommendations, computed on an implicitly collected profile. Again, the results pointed out how important the user effort factor is in recommendations. The behavioural recommender was perceived by users as requiring significantly less effort, as postulated in HYPOTHESIS 3, although the overall satisfaction was similar to that of the search & browse. The correlation analysis further supports this claim, as the users' of the behavioural recommender with a big profile answered that they were satisfied when they found the system to require low amounts of effort.

Seen individually, these results might simply lead to the general observation that systems should not be too complicated. Because of their repeated occurrence, and because of the direct

assessments and correlations of Q26, we are convinced that they form a key principal in building user-centric recommender systems. We propose the following Guideline 2:

**Guideline 2.** *Focusing the system on the main tasks by minimising the number of features helps to increase user satisfaction.*

One more reason we see for minimising the number of features, was spotted in Experiment 2 where the number of features correlates inversely with having songs suited to a user's mood. The importance of emotional aspects such as mood, in recommendations is only starting to be studied. A few works like [132] exist, but the topic is still very open. There are several psychology studies linked to music, which have come up with different music classification schemes related to emotions. They propose dimensions such as *arousal* and *valence* [77, 71] which allow to express musical preferences in a new way. There is undoubtedly a strong interest in being able to model how elements like emotions, a user's mood or even personality, might be determinant factors influencing recommendation accuracy. Due to the fact that items may be experienced differently depending on our mood, it is reasonable to believe that moods affect the way people accept recommendations. The importance of the context in which a recommendation is accepted, is currently a hot topic in the research community [98] and a user's mood clearly contributes to this context. In Experiment 2, our results show that the number of features (Q26) and navigable links (Q27) correlate inversely with having songs suited to a user's mood. Furthermore, the questions which presents the biggest difference between both system is that about recommendations suiting a user's mood (Q2). This is highly in favour of *Pandora* which has been shown to be above all perceived as much easier to use. We hence conclude Guideline 3:

**Guideline 3.** *Maximising the ease of use is likely to make the system more suited to a user's mood, thereby increasing the perceptions of accuracy.*

## 7.2 Purchase Intentions

One of the other key dimensions which is found in the TAM is intention to use, which can be assimilated to the acceptance of recommendations, as described in our model of Section 3.2. When we consider e-commerce websites, where the recommender is integrated into a whole set of other features, the closest analogy is in our opinion the act of purchasing an item. For this reason, we systematically asked users if they were interested in buying the considered item(s). As we only ran studies where users simulated the final purchase, we relied on questioning purchase intentions. We have come up with a series of guidelines for this dimension.

Purchase intentions came up as an important result in several of our studies. Evaluating purchase intentions is delicate, especially in simulated environments. Several studies such as [139, 100] showed that users perceived significant differences in the experiments, but that in the end it did not correlate with, or influence their intention to buy. In our first experiment, one of the subjective questions asked users if the system was good compared to recommendations they might receive from a friend. This measure gave some promising results. First of all it showed that the number of songs users were prepared to purchase was correlated with recommendation quality. Secondly, correlation analysis among the measured variables showed that enjoyability

of songs, interface satisfaction, and the number of songs loved were the most important factors in predicting the relative quality of recommendations as being better than what the user may get from their friends. In Experiment 2, we took time to refine what users perceived as being the quality of recommendations, which we mapped to the perceived usefulness dimension of the TAM. One of the strongest results was that users perceived *Last.fm*'s recommendation technology as being less accurate than that of its counterpart. Although only moderately significant statistically, this was supported by post-study interviews where users often reflected negatively on the accuracy. The correlation further endorsed this view by showing that the perceived accuracy of the underlying algorithm was the only value which correlated significantly with the intention to purchase the recommended song.

The combination of these results leads us to emphasise the importance of proposing good quality suggestions. Quality being a vague concept, we show that accuracy is one of its key components. However, we see from our results that there are some domain and goal specific elements to take into account. Keeping in mind our goal of being user-centric, we therefore suggest using a common metric to define accuracy, such as measuring recommendation accuracy compared to suggestions made by users' friends. We propose the following guideline:

**Guideline 4.** *Maximising the quality of recommendations, such as accuracy, relative to a commonly shared measure increases purchase intentions.*

At the same time, and without being contradictory, there are other dimensions which are important in users' acceptance of a system. Quality is one part of the equation, and we would here like to highlight another: enjoyability. Enjoyability is a dimension which can easily be misinterpreted, especially when considering entertainment recommender systems. The goal of providing an enjoyable experience seems obvious, but also somewhat secondary to a user's fundamental task. Our results tell a different story.

In both of the first experiments, enjoyability score are highly in favour of *Pandora*. In Experiment 1, the factor that shows the highest correlation with recommendation quality is the enjoyability of recommendations. In Experiment 2, *Pandora* testers thought that the recommended songs were more enjoyable and more satisfying, and again we see a correlation between enjoyability and the direct assessment of perceived usefulness (Q18 & Q19). More generally, many of the questions testing the quality dimensions (including enjoyment) correlate with the direct assessment of usefulness. Data from Experiment 5, where we use the content vs. layout approach on our perfume recommender, further supports these conclusions. Indeed, a strong correlation can be seen between trusting intentions and both enjoyment and perceived usefulness.

To us, such results show that enjoyment is not just a fanciful secondary component of recommender systems, but an important dimension which stimulates users' perceptions of usefulness. This is valuable since perceived usefulness has been shown to increase purchase intentions earlier in our results, but also in many studies including large scale research [78]. We propose the following guideline:

**Guideline 5.** *Providing an enjoyable system is likely to increase the perceived usefulness, known to increase purchase intentions.*

The last guideline of this section concerns how important it is to understand what users want to get out of your system. In the case of our first two experiments, we were putting users in



front of music recommendations websites. The main goal of people when using those systems is naturally to *find* music<sup>1</sup>, where by find we either mean to hear known songs, or also to discover new songs. In the case of our experiments, our instructions were focused on the discovery task. In Experiment 2, we directly assessed whether users found that the site was easy to use as a recommender system for discovering music. It turns out that users' answers correlated with users' will to own the discovered songs. In parallel with this, throughout the user studies of the thesis, and as shown in Chapter 6, diversity and novelty are two highly important elements in users' final satisfaction.

We believe that these observations support the need to provide interfaces where it is easy to discover new items. This need was derived from observations in the field of entertainment recommender systems, but is possibly applicable in other fields, especially as domains get complex. Users become overwhelmed with information. For this reason, we acknowledge that the following guideline is specific to the domain of music recommender systems but are confident that it has a larger scale of application.

**Guideline 6.** *Providing users with an interface where it is easy to discover new items will increase their will to own the item, and thus their intention to purchase or return to the system.*

### 7.3 Complex Systems

The following guidelines concern recommendations across multiple or complex systems.

In Experiment 2, where we compare the two music recommender systems *Pandora* and *Last.fm* at a higher level of detail, another result about novelty caught our attention. We indeed took the opportunity in this study to insert into the acceptance questions, a direct assessment of users' intentions to use such technology in the future. We asked participants: if a similar technology existed for recommending other things to them (books, movies), would they use it? The results were not significantly different between both music systems, and the two averaged around the middle score of the Likert scale. However the only question that showed any correlation with this question was novelty.

Such a result, which at the time of the study was quite surprising, is much less so now for the following reasons. First we need to consider previous work on trust by Chen. Chen studied trust building in recommender systems. We can see that one of her key results is that she managed to show that trust intentions have a significant causal relationship with users' intentions to return [29]. Second we must consider our more recent results from Experiment 6. In Experiment 6 we run a between-group user-study on our perfume recommender platform, where the recommendations either come from Amazon or Editorial Picked Critiques, and the interface is either a list or organised view. In the detailed analysis of data, initial suppositions on correlation analysis later confirmed by a factor analysis, highlighted that novelty, diversity and trust had strong correlations and formed a key cluster. The combination of Chen's work and our observations supports that novelty helps to build trust, which itself increases people's intentions to return to

---

<sup>1</sup>Arguably, Last.fm is by default a more social website where users can also interact with each other. However, this should not alter the validity of our findings, as we focused all of our experiments on the recommender tool, excluding all other features destined at other user goals.

the system. In addition, because the relevant question of Experiment 2 questioned multiple domains, we suggest that novelty may be applied to multiple product domains. We suggest the following guideline:

**Guideline 7.** *In systems that cover multiple product domains, providing novel suggestions can encourage users to return, and to use the recommender on the other provided domains.*

In Experiment 4, we tried a different approach to that of our previous studies. We setup an experiment where we opposed layout versus content, a method that we kept for Experiment 6. As explained in Chapter 5, we decided to rely on a critiquing-based recommender system, and designed an experiment where traditional compound critiques, represented textually, were compared to a visual interface which represented various critiques by a set of meaningful icons. The two product domains (in the dataset) were computer laptops and digital cameras. The main difference between the two datasets was that the laptop assortment is more complex. It contained more products and each had more features than the cameras. Results from the study showed that for all preference questions, the majority of users preferred the visual interface. Furthermore, although the two systems had exactly the same algorithm to generate compound critiques, the visual interface enhanced users' perceptions on the recommendation quality (see Q5). These results pointed out that the visual interface had gained a much stronger support from end-users during the online shopping process.

At the same time, we noticed that while the visual interface performed better than the textual interface with both laptop and camera datasets, the visual interface had achieved higher performance improvements with the laptop dataset than with the camera dataset. When the product domain is rich, the textual interface will generate very long strings of text to describe the compound critiques, which are not easy for users to read. By comparison, the visual interface could provide an intuitive and effective way for users to make decisions (for example by simply counting the number of positive and negative icons). We believe that as the system or product space gets more complex, long textual descriptions become a burden, and the synthetic nature of a visual solution can become a real tool with real advantages.

A similar kind of observation can be made upon considering results from Experiment 6. The detailed analysis resorted to considering ordering effects between users groups in the within-subject study. The two layouts proposed clearly demonstrated different levels of complexity: the list view is a default and traditional way of representing items (not only in the virtual world of websites and computers, but also in the real, physical world) and the organised view introduces a new categorisation and a different (horizontal) scrolling mechanisms. As a consequence it is more complex than a list view. The ordering analysis showed that users did not demonstrate any clear preference when first exposed to the organised view before the list view. On the contrary, those having first tested the list before discovering the organised counterpart significantly preferred the second.

These observations have led us to propose two guidelines which try to take into account the advantages of the observed solution when confronted with more complex product domains.

**Guideline 8.** *Favouring visual to textual representations when a key element of a system demonstrates complexity can reduce users' effort and encourage them to use the element.*

**Guideline 9.** *When considering new features, more complex than current counterparts, introducing them progressively helps users to understand them and ultimately prefer them.*

## 7.4 Diversity

In this last section, we propose a series of guidelines about diversity. They are mainly based on the conclusions of the last two user studies, when those were supported by results from our earlier studies. In Experiment 5 we ran a within-subject study where we opposed content and layout. Amazon’s recommendations and the EPC algorithm selected the content, and a list or organised view were proposed. These first results did not show very big differences among users’ perceptions, but they did highlight that diversity was the dimension where users perceived the most differences. In Experiment 6 we setup an in-depth evaluation with an eye tracker on the same perfume e-commerce framework as for the previous experiment. We used its default configuration with the EPC algorithm and the organised layout. Through the usage of an eye-tracker we were able to analyse users’ behaviours with a much finer lever of detail. Thereby we managed able to explain *how*, *when* and *why* users needed diversity in recommendation systems, and we hereafter propose one guideline for each of these elements.

Let us first consider the *when* aspect. In Experiment 2, analysis of the data had led us to questioning whether recommendation accuracy alone was sufficient to satisfy users’ needs, or whether elements generating novelty or diversity might be just as valuable. Our results were that for both music systems, the score for the perceived accuracy of the recommendation technology was below that of the average quality question. Yet, questions such as Q1 or Q5 and the post-study interviews revealed that overall users were satisfied with either system, and dimensions such as novelty showed good correlations with acceptance questions. Based on these results, it was reasonable to imagine that in order to please users, the system only needed to have a “minimal” recommendation quality (in terms of accuracy), which could then be complemented by dimensions such as novelty. A similar idea had already been proposed, but with diversity instead of novelty in [144, 86]. These observations became much clearer when we obtained results from Experiment 6. In this last experiment we outlined how users only progressively used the recommender system, and how as they got closer to their desired item, they needed to explore alternatives thereby requiring to see diverse recommendations. Said otherwise, the temporal dimension influences users’ needs. At first they need accurate recommendations in order to get a product as close as possible to their preferences or to the ideal they had in mind. Then progressively, users work towards establishing that this is a confident choice. We thus propose the following guideline:

**Guideline 10.** *Favour an approach where accuracy is provided at the beginning, and where diversity is introduced as a second step or progressively.*

We also considered what lessons could be learned from the *how* element of users’ need for diversity. In order to do so, we looked back at results from Experiment 3 which can be seen as somewhat disappointing. Indeed, the theoretical benefits of the recommendations made from profile information gathered implicitly were perceived by users. The data showed us that users felt the behavioural recommender required less effort for finding items, and was trust-worthy.

Despite these positive aspects, users felt both recommendation approaches were equally satisfactory. In addition and somewhat more importantly for our upcoming guideline, users found the normal search & browse to provide more diversity and a better control. We found these results disappointing. There is much to bet that when given the choice between both mechanisms, most users will resort to the search & browse, the implicit recommender needing time to build a reasonable profile size for any user before providing good recommendations. Beyond this negative observation, there was an encouraging correlation that was found: correlation analysis suggested a link between diversity and confidence, both among the search & browse users and the implicit big-profiled users. The factor analysis we ran afterwards pointed out more clearly that confidence and diversity (and trust and control) belonged together as a cluster, explaining the main trends of the overall data. This is interesting because it was further supported by results of Experiment 6 and more specifically our HYPOTHESIS 8 which we validated. We showed that recommender systems increase users' confidence by providing diversity. Users need to see a certain diversity in order to make sure they've chosen the best product available, logically helping them to develop confidence in their choice. We propose the following guideline:

**Guideline 11.** *Including diversity can help users have confidence in the items they have selected.*

Finally, we finish our selection of guidelines with a more personal take on the results seen throughout this thesis. It evolves around the *why* questioning about diversity, and in general ways to improve recommender systems. As mentioned on several occasion like in Section 6.2, much of the work on recommender systems has aimed at improving algorithmic accuracy. The now famous Netflix competition where the challenge was to reach a 10% increase of accuracy, lasted almost three years and more than 43,000 entries from over 5,100 teams were submitted<sup>2</sup>. One is entitled to ask whether this is really useful. Do users perceive differences in algorithms which supposedly are a few percent better than others? Throughout this thesis, we have taken a broader stance, seeking to understand the essence that users really extracted from websites that include a recommender system. In at least two of our studies, despite being confronted to technically very different systems (expected to yield distinct levels of accuracy), our users did not express significant differences of satisfaction. This has inspired us to propose an open guideline to conclude. We believe people should keep in mind the fundamental goal of their system, and make sure they take into account the specifics of their item-domain.

**Guideline 12.** *Upon reaching a reasonable level of recommendation accuracy, consider what might most satisfy users' needs in order to generate the overall biggest impact.*

### 7.4.1 Generalisation of Guidelines

It is legitimate to question whether the results obtained throughout the thesis might not be too specific to be generalised into guidelines: the experiments covered multiple product domains and systems. In order to help put into perspectives these guidelines, we hereafter take the opportunity to discuss briefly two considerations which we believe are useful.

---

<sup>2</sup><http://www.netflixprize.com/>

### **How Obvious is the Obvious?**

It is possible that to some extent, parts of these guidelines might appear to be obvious or even trivial. Whilst we believe that this is not true, we would like to point out that many of the problems raised by the numerous users in our six studies, are themselves trivial. User problems often look easy to solve, when seen from outside. The challenge is making an environment where sometimes opposing goals can be optimally handled by all level of users, from beginners to experts. In a recent interview<sup>3</sup>, J. Spool the CEO of User Interface Engineering explained the same problem very nicely, when speaking about usability testing:

The primary benefit of any usability testing project is not the report at the end or the list of recommended changes. Our research shows it's the exposure the team has with observing real users work with their designs. The more exposure, the better the products that come out.

The issue is the same with our twelve proposed guidelines. Even if it appears that a system already takes care of one of these guidelines, designers should maximise time spent with or observing people as they use the system. This will allow them to make sure that a feature implemented because of a guideline, is being perceived by users. It is only at this condition that any guideline will make sense and become useful. If not, the measure is of no use.

### **Validity of Guidelines**

The proposed set of guidelines was the outcome of our six user-studies where we collected a comprehensive amount of data about how users perceive recommendation systems. All guidelines have at least one precise observation or result at its root. However, we do not always have enough evidence to support that these would remain valid in other domains and conditions. Further work is required for a stronger generalisation.

We used several correlations at the basis of many guidelines. One difficulty with correlations is causality. Common statistical interpretation suggests that correlations do not imply causality. Yet, rational reasoning can put forward a “most likely” direction in an observed correlation between two variables. If, for example, a relation is found between “hot temperatures” and “dehydration”, it is not fanciful to envisage that the first is the cause of the second, and probably not the opposite. Upon proposing our guidelines, we often reasoned by proposing a direction in order to facilitate the potential understanding of the detailed relationships. Readers should remain aware that the causality is not proven.

In order to have more robust guidelines, repeated studies would be necessary, and in each potential domain where the guidelines might be applied. This was not possible in the frame of this thesis. There are however three aspects which we would like to put forward to justify why we believe that our guidelines are solid and will scale well. As pointed out in Chapter 1, despite the fact that experiments were carried out on three product domains, they all share three common characteristics which support our choice in using them. Firstly, they are all everyday consumer products, also called public taste products, which users are accustomed to. Secondly, they are

---

<sup>3</sup>Interview in UI Trends, January 7<sup>th</sup>, 2010 by Russell Wilson

low risk or low involvement products. Through their low price range, they can very easily be bought by users, without spending a lot of time thinking about the purchase, contrary to the acquisition of a car or an apartment [129]. Thirdly, these product domains are complex in features. They all three can be classified according to a high number of individual features. Beyond the considerations on the product domains, it is important to keep in mind that the systems compared were always identical in terms of features. In all our studies where we confronted users with two systems, we made sure they only encountered the parts of the system which proposed the same features. Finally, and possibly most importantly, many of the observations in terms of the qualities perceived by users, are similar from one study to another. Despite being presented with different interfaces, systems or product domains, users felt strongly about similar qualities across studies. Even in Experiment 5 where the content and layout change, users perceive few changes. We are therefore confident that at least the core message behind each guideline can be applied to other fields.

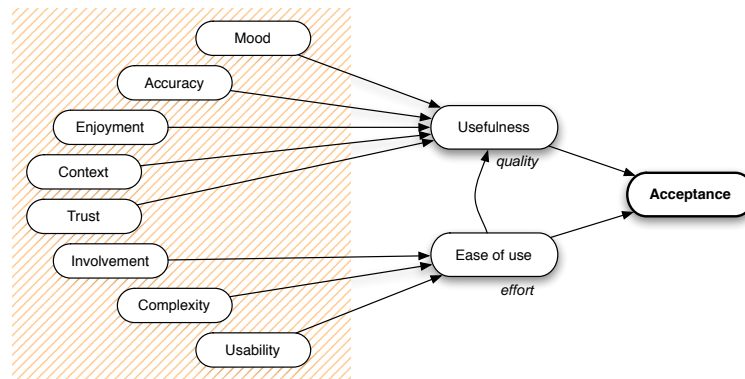


Figure 7.1: Dimensions of the original research model.

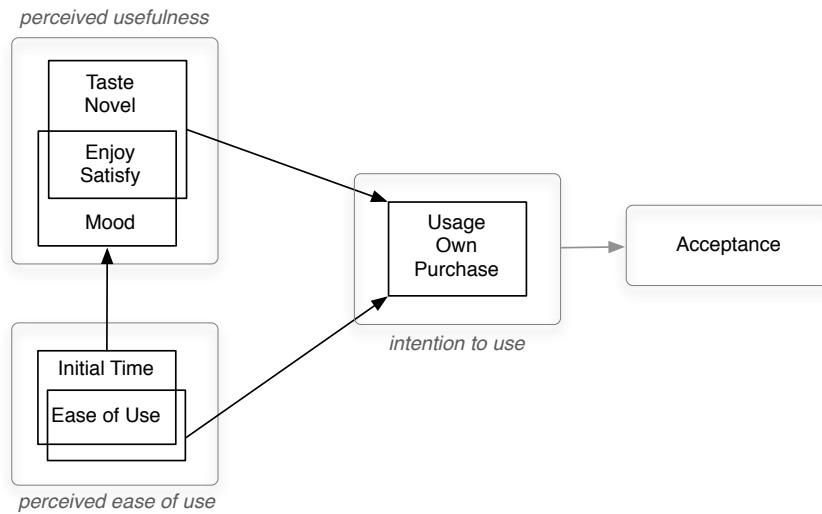


Figure 7.2: Key correlations of Experiment 2 modelled according to the TAM.

## 7.5 Revised Research Model

Early on in the thesis, we introduced Davis *et al.*'s TAM and we discussed in Section 3.2 how it formed a starting point for our upcoming research. We hereafter discuss how our overall results can be put into perspective with such a model, what new representations could be proposed and how we ultimately propose to model the role of diversity in recommender systems.

In Section 3.2.1 we introduced Figure 3.3 (which we here repeat in Figure 7.1). This schema presented a first research model, inspired by the TAM (Figure 3.1). Reasoning around the meaning of Perceived Usefulness and Perceived Ease of Use, and taking inspiration from the approaches chosen in other studies (detailed in the same section), we tried to extend and detail what lies behind the dimensions of the TAM in the case of entertainment recommender systems. The idea was to express which “individual” dimensions were influencing users’ perceptions of usefulness and ease of use. In this first proposed model, we balanced both evident dimensions such as enjoyment and satisfaction, with less obvious ones like having songs tailored to one’s mood.

Unfortunately the picture that results give in our second study, is only partly in line with our first model. Despite having strong differences in averages for the preference questions between both systems, there is little common evidence. We decided to rely on the correlation analysis in order to evaluate what possible tendencies were appearing, hoping to get close to our proposed model. We draw in Figure 7.2 our best effort. The reported dimensions were selected based on the strength and statistical significance of correlations. They were then grouped on a similar plan as that of the TAM, and the causality between groups is implied following the TAM.

Among the possible dimensions supporting the PEOU category were the two direct assessments of Ease of Use (Q24 easy to use as an internet radio, and Q25 easy to use as recommender),

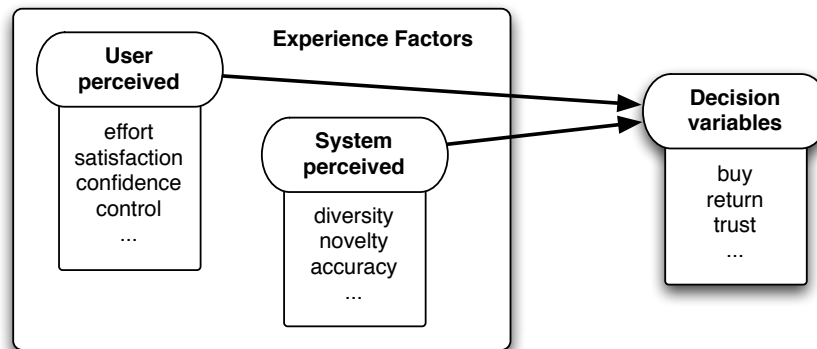


Figure 7.3: Proposed research model focused on experience and decision variables.

quite naturally. The only other dimension which showed a small link with a dimension supposedly belonging to PU was the initial time to recommendation. When compared to Figure 7.1, Initial Time might be seen as a sub-component of Involvement. If we consider which PU dimensions of our model are supported in the reported correlations, Enjoyment and Satisfaction show good correlations. Mood, Taste and Novelty show smaller correlations and only the first was proposed in our model. Finally the elements belonging to the Intention to Use of the TAM show many strong correlations, which is expected but not very insightful. Overall we believe that this study is not supportive of our model since only a few dimensions were confirmed, often just partially. However several dimensions did show a tendency, that could lead them to being included in such a model under certain conditions. We further analyse such possible elements through results from the following experiment.

Seeking to get a better understanding of the limitations of our previous model, we imagined a new model whilst designing our Experiment 3, focused on experience factors and decision variables. We chose this fresh perspective as it evolves around user's perceptions more directly. Our reasoning was the following. When potential buyers visit a website like Amazon, they encounter a certain *experience* with the system before making a certain number of *decisions*. During this process, they *perceive* some key features (directly or indirectly through their experience) which influence their decisions. This holds whether users reveal their preferences explicitly or implicitly. We relied on these three dimensions to propose an alternative model which is depicted in Figure 7.3.

Among the Experience Factors, we distinguish those which are perceived by the user, and those which are perceived as belonging to the system. For example, users' experience, what they directly encounter while using a system, is influenced by components such as their level of confidence, or the users' perception of the amount of effort required. At the same time, users discern certain dimensions such as accuracy or diversity which symbolise what they feel about the system. We considered that these two categories belong together under the "experience" tag. Indeed both system- and user-perceived variables are immediately sensed, often without any



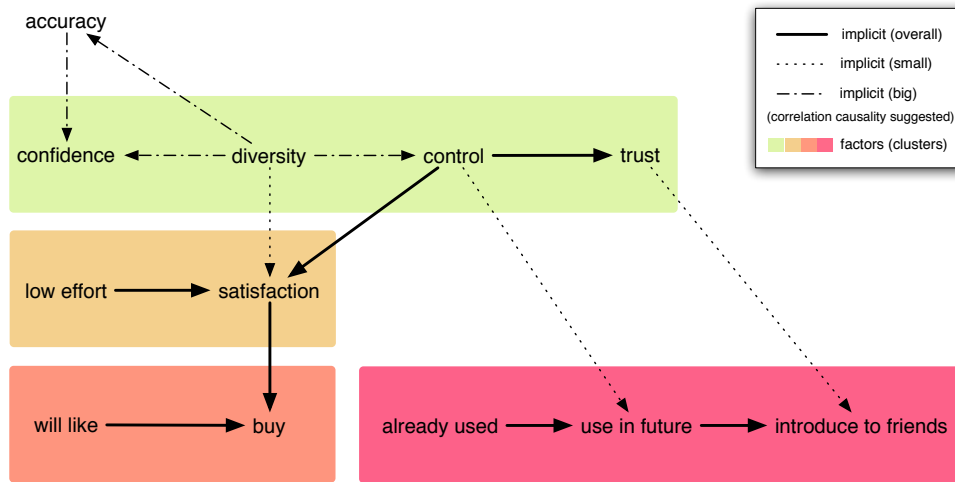


Figure 7.4: Factors and correlations from implicit profiles.

conscious reflection. The third cluster of the figure characterises the decision process through variables such as the intention to return to the system, and the intention to purchase. We also include in this category the dimension of trust. Albeit trust can be understood as an element pertinent to the user’s experience, we chose to classify it under “decision” because it is a variable that can only be assessed after a certain time spent using the system. It is an opinion built across time, whereas confidence is something that we perceive on the spot, on the very moment of finding a product or receiving a recommendation.

Unfortunately, when we conducted Experiment 3, where we compared the performance of *Amazon’s* recommendation (based on implicitly gathered profile information for small and large profiles) with its normal search & browse interface, we obtained results which strongly dismissed our second model. Both the factor and correlation analysis pointed to a different clustering of dimensions. We represent these ties between elements in Figure 7.4. It should be noted that the causality of correlations in the figure are suggested through our interpretation.

There are several parts which we believe are interesting in this new representation. First of all, we note that accuracy only shows correlation in the implicit large group, and interestingly connects with both diversity and confidence. It however does not appear in any of the significant clusters. Secondly, we note that intention to purchase does not seem to present any link with the cluster of future usage (already used, use in future, introduce to friends). This connection had been highlighted in Figure 7.2 but somehow not here. This being said, we wonder if the most notable observation is not that the links of implicit small and big groups are very different.

Upon considering the fact that our revised model was strongly dismissed, and that within this last analysis there seemed to be two different “graphs” depending on either the implicit small or big group, we felt obliged to question whether user’s needs were not changing. Obviously, the detailed setup of our experiments is not the same from one case to another and the systems used present algorithmic and layout variations. That being said, we always positioned our analysis

on the same user-level, focusing on users' perceptions. Said otherwise, we studied very similar questions in all studies, if not the same dimensions, taking as explained in the first chapters of the thesis, a slightly higher level approach but resulting in a coherent transversal research.

### 7.5.1 Time dependent Accuracy vs. Diversity model

Experiment 5, which highlights some important results concerning diversity, provides further support for part of our investigation. Behind what arrears at first sight as similar appreciation by the users of all of the four systems tested, the factor analysis we ran highlighted clusters which are only in parts coherent with those we highlighted in Figure 7.4. The strongest group concerns low effort, ease of use and enjoyment, whereas in our previous graph low effort is united with satisfaction. Another group contains novelty, diversity and trust, which can also be found in our earlier schema but next to confidence and control. Control itself is clustered with usefulness in our latest diversity experiment, not in Figure 7.4.

There are several possible reasons that might stimulate this observed outcome. One probable explanation is that some of these dimensions which we see appear and disappear from one analysis-graph to another, are linked to the observed domain or interface. In which case, despite our higher level approach to these studies, the dataset or website's look & feel would be influencing some of the variables we are measuring. However, this is not the justification that we believe is the most probable. Another interesting explanation is the possibility that users' needs are changing throughout the course of their interaction with the system. It has already been shown that one important issue regarding users' preferences is the fact that they might not yet know their preferences as they start to use a system, and often change them in different contexts [70, 101]. It is this questioning that led us to consider the dimension and role of time with regard to diversity in Experiment 6.

The results of this last experiment, led us to propose a time-dependent diversity-model for maximising user's overall satisfaction, shown in Figure 7.5. The graph expresses in which way *accuracy* and *diversity* should ideally be interleaved during the decision process of a user's interaction with a system. The graph should be understood as follows.

At first, up until  $t_1$ , the system knows little or nothing about the user's quest. We propose that the optimal strategy here is to maximise diversity. Our experiment confirmed the first step (of two) of Häubl's choice strategies [55], whereby users' goal during the first step is to discover adapted search criteria according to their needs and capabilities of the interface. Diversity is more likely than accuracy to satisfy this goal, since accuracy would above all provide similar items possibly wrongly (since little is known about a user's search goal at first). At the same time, diversity can help users to develop a certain confidence in the system, as they will note certain items with which they are familiar. After  $t_1$ , we suggest that the importance of accuracy strongly increases, leading to a decrease in diversity. The study results supported our HYPOTHESIS 6, according to which the influence of the recommender system continuously increases across time at the product brokering stage. Since at this point we are still in the first step of Häubl's choice strategies, where users seek to identify a subset of products, it seems clear that users need recommendations which are as accurate as possible. We symbolise this increase in accuracy by a curve of convex nature, in order to highlight that the more a user interacts with a system, the more a system should know about a user's goal. As time gets closer to  $t_2$ , users move into the

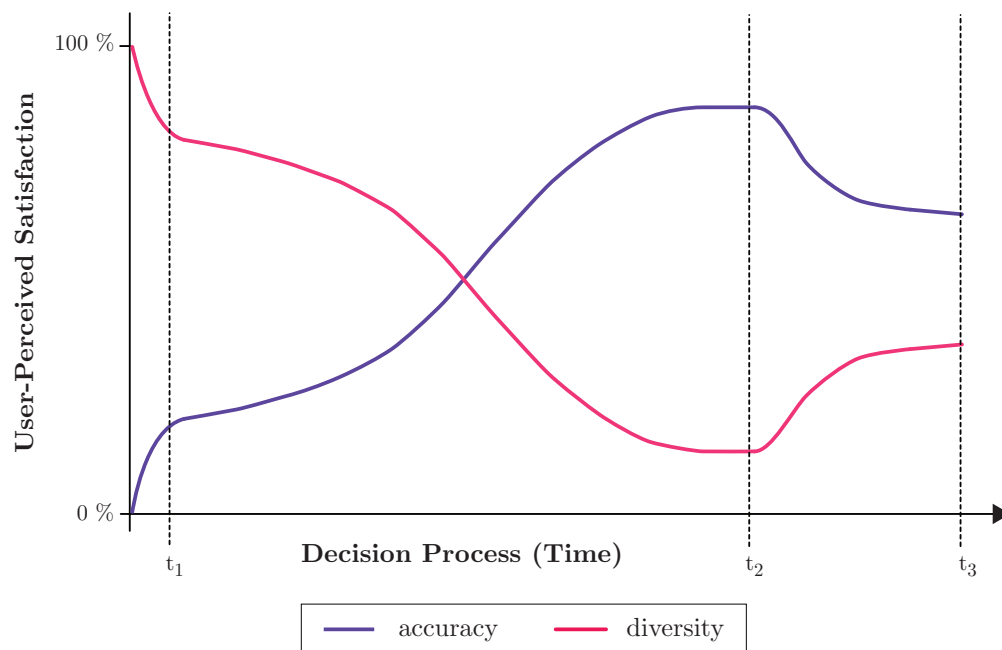


Figure 7.5: Time dependent Accuracy vs. Diversity model.

second step, where their aim is to explore alternatives. Our model suggests a plateau around  $t_2$  before presenting a clear decrease (respectively increase) for accuracy and diversity up to the moment where a user decides to purchase the item,  $t_3$ . This was proposed because our results validated HYPOTHESIS 7, which states that users are scoping products in a new way, seeking opportunities to reach a decision. The fact that the accuracy curve does not return down to zero is logical, but also supported by HYPOTHESIS 5. We believe that the more important message here, is that having reached a certain accuracy among the recommended items, users seek alternatives and thereby need diversity, especially as this reinforces their confidence. This was supported by our data which validated our HYPOTHESIS 8, according to which recommender systems increase the confidence of users by providing diversity.

We believe that this model constitutes a big step towards the understanding and formalisation of users' online patterns of behaviour. It derives directly from the results of the last study, but we argue that it has the potential to be compatible with several of our other studies. As demonstrated previously, the different revisions of the TAM that we proposed all shared a certain common ground, but specific results from each pushed us to change them each time. Our model provides one likely explanation: our studies were directed at different parts of the decision process, leading to observations made in conditions where accuracy and diversity were different from one study to another.

The results in this thesis show that diversity is an important dimension in the recommendation process. The last two experiments we conducted, specifically designed to observe the importance of diversity, allowed us to model a first part of the sub-processes involving recom-

mender systems within the framework of purchasing decisions, in a time-dependant way. We discuss possible future work around our proposed model in the Chapter Conclusions of this thesis, including the proposal to use such a model to try and guess, dynamically, how to improve the recommendations proposed to users. Such an adaptive idea is highly promising. However, we would like to highlight that before being able to adapt to a users' interactions, we believe that a system should make sure it offers some diversity at all time: the opportunity of diversity should always be present. We showed that users seek opportunities, and that diversity can help them to build confidence. In that sense, it is important for users to always have the possibility to view, or select, some more diverse items.

## Chapter 8

# Conclusion

The goal of this thesis was to study how users perceived the different qualities of recommender systems, leading them to accept the proposed recommendations and ultimately driving them to the adoption of the system. Throughout our investigations, we managed to single out one essential quality: the diversity of recommendations. When an e-commerce presents a set of recommendations to a user, the later is readily able to perceive the degree of diversity offered. We then studied this effect in greater detail, allowing us to understand the role of diversity in the process leading users to accept the recommendations. We designed six experiments where we evaluated participants' reactions to different systems. Overall, we observed 306 users in real-life scenarios, collecting a large amount of data covering users' perceptions of recommendations. We used the Technology Acceptance Model as a baseline model, and refined it throughout our transversal investigations. We revealed how accuracy and diversity of the recommendations were interleaved, introducing *time* as a new element in a user-satisfaction model. Moreover, based on the experimental results, we derived a set of design guidelines.

After considering some large-scale differences between two music-recommender systems, we looked at how perceptions of control, through explicitly and implicitly revealed preferences, could influence users' satisfaction. Differences between layout and content were also explored and tested, before focusing on the need for diversity in the different stages of the purchase process. These investigations have allowed us to gain a detailed understanding of how users perceive and act upon recommendations, when they are confronted with them for the first time in online shopping.

## 8.1 Contributions

### 8.1.1 Attracting New Users to Recommender Systems

In order to launch our investigations into which qualities lead users to accept recommendations and ultimately persuade them to adopt the regular use of the system, we chose to use two music recommenders. We compared them in the first two experiments of this thesis, in both a within-subject and between-subject study. These studies were, to the best of our knowledge, the first to investigate acceptance issues in recommendation-seeking systems, for entertainment products.

Our data revealed that Davis *et al.*'s initial Technology Acceptance Model [41] can be successfully adapted and applied to such recommender systems. Despite the model being very simple, we proved that it suffices to capture the most important interaction mechanisms which lead users to accept recommendations. Specifically, our results showed that the perceived usefulness of a recommender system in terms of quality, and the perceived ease of use in terms of effort, are directly correlated with the user's acceptance of the recommendations. We also took this opportunity to evaluate some more detailed domain-specific dimensions, proposed in our research model. Many of them were corroborated by results. Measures of quality such as the perceived accuracy, the enjoyment of using the system, the satisfaction procured and the impression of having music tailored to a user's taste, were directly correlated with acceptance. We also observed that measures of effort like the initial time needed to get interesting recommendations and the ease of use for discovering music were linked with acceptance.

As it is naturally important to focus on making recommendations as accurate as possible, it is necessary to realise that from a user's perspective, *more accurate* does not always equate with *better liked* or *more useful*. Our data reinforced the idea that it is crucial to understand users' real and current preferences. For both systems, the participants strongly appreciated the fact that the system elicited preferences for them. However, we showed that the most reactive system was perceived as much better than its counterpart at tailoring songs to users' tastes. Users even believed the system was able to customise songs in accordance with their mood. Our studies show the necessity for low-involvement recommenders to be highly reactive, helping to take the users' context into account.

### **8.1.2 Evaluating Implicitly Expressed User Preferences**

To understand more accurately how important it is to model users' preferences accurately, we decided to compare two opposing approaches. We carried out a between-subject study where one group of users tested a *search & browse* interface, and two other groups for which recommendations were computed on a profile collected implicitly through purchase histories. As much of the work until now on implicit recommendations had been incremental and focused on accuracy metrics, we wanted to evaluate how well this compared to traditional solutions.

Our experiment revealed that users do not perceive much contrast between the two mechanisms with regard to their overall satisfaction. The only true difference they noticed was the amount of effort required to operate in both setups, and clearly in favour of the behavioural recommender. The explicit search & browse is seen by users as proposing a more diverse set of items, and gives them more confidence about their choices in the search. The positive result for the implicit recommender was that users are trusting these implicitly generated recommendations, as soon as their profile reaches a certain size, which is encouraging. Within the framework of our experiment, our results showed that although users may obtain limited benefits from recommenders, they perceive them as being trustworthy and are willing to accept them.

### **8.1.3 The Impact of Visual Renderings in Critiquing-Based Recommenders**

Another approach we tested was to use more visual renderings, in order to understand whether such user interface design changes had more impact on users' perceived qualities than algorithm-

mic improvements. We chose to work on critiquing-based recommender systems, where user interface design is an important issue because of the high amount of information presented to users. Traditionally the compound critiques are presented as sentences of plain text. In this thesis we proposed a new *visual* interface which represented compound critiques by a set of meaningful icons. We extended an online web application to evaluate this new interface using a comparative real-user study.

We showed that the visual interface was more effective than the textual interface. Several dimensions participating in user's overall satisfaction showed a strong preference for the iconised representation. It is capable of significantly reducing users' interaction efforts and attracts them to select and apply the compound critiques more frequently in complex product domains. Users' subjective feedback also showed that the visual interface was highly promising in enhancing users' shopping experience. With this study, we clearly showed that a strong impact could be made in users' perceived qualities through the use of an enhanced visual rendering.

#### **8.1.4 The Role of Diversity in Recommendations**

To conclude the series of experiments of this thesis, we ran two studies addressing the role of diversity in recommendations, understanding how users' perceived benefits from having more or less diverse recommended items. We oriented our investigations in order to address the questions *how*, *when* and *why* to suggest diversity in recommendations. We prepared two experiments with an online perfume e-commerce prototype, using the form of critiquing called Editorial Picked Critiques.

The first of our diversity-experiments relied on a similar approach as the previous study in this thesis: we arranged a Layout against Content within-subject study. We tested two algorithmic approaches for making recommendations and two visual designs for critiques. We used Amazon's well established "Users who bought also bought" recommendations compared to those generated through Editorial Picked Critiques. We then displayed them either as a simple list of items, or as a layout organised by categories. Our results showed that users' overall satisfaction did not differ much between either of the configurations. However, the dimension which users' perceived with the biggest contrast was the diversity of recommendations. Results show that layout did seem to have an important effect: when users have been exposed to interfaces, users develop a preference for organisation interfaces over traditional list views. The analysis of the data provided at least one reason for this: the organised layout is perceived by users as helping them to discover novel items, a real benefit for the user satisfaction, and this independently of the recommendation algorithm. We also noticed that this preference occurred, even though this interface required more effort in terms of session lengths. We believe that our results indicate that as long as the effort required is low, users are prepared to spend more time when they perceive other benefits such as those of an organised layout. Finally, our study also reported a large number of correlations. The main contribution here was that for the first time within the same study, novelty and diversity were shown to link directly to many key dimensions such as satisfaction, trust and for the first time to the intention to purchase.

In the second diversity-experiment, we used the same perfume framework with the Editorial Picked Critiques and an organised layout. We used an eye-tracker to carry out an in-depth lab-study of users' purchase decision processes. We especially paid attention to the way users

perceived and expressed a need for diversity. This experiment allowed us to make four key contributions. First, from the 45,000 fixation points that we recorded throughout the study, we successfully reproduced Häubl *et al.*'s two-step purchase decision making process [55]: first users identify a subset of products before exploring alternatives. However, we highlighted the fact that the influence of the recommender does not seem constrained to these two steps. Second, we showed how it increases, in an interrupted process, until the purchase decision is close to being taken. Third, we observed that users rely on the recommender to enhance their confidence in the purchase decision. Finally, we outlined users' need for diversity when making a decision. Our findings not only supported a hypothesis raised earlier in the thesis, according to which it is necessary to find a good compromise between accuracy and diversity in order to increase quality of recommendations perceived by users, but they also provided an excellent guideline as to when this diversity is needed by users.

As result of this last experiment, we showed a new exploration mode where customers efficiently use categories of recommendations to obtain diversity. This scheme consists of two elements: looking at different categories to envisage new alternatives, and then unconsciously prioritising these categories until one of them sufficiently fits the user's needs so that it provides confidence and induces the decision.

### **8.1.5 Guidelines and Diversity Model**

Based on the six experiments in this thesis, and with the insight gained through the users' qualitative comments on each study, we have derived a set of twelve guidelines that could be helpful in re-thinking recommender systems from the perspective of the users. The suggested item-lists would be tailored to the users' needs and in particular to their current information seeking task. In this way one would greatly increase the chances that they accept the recommendations. These twelve guidelines were elaborated around four primary axes which reflected the main aspects that were studied in the experiments. These axes were: reducing user effort, increasing the intentions of purchasing, acceptance of recommendations in complex systems and the role of diversity in users' purchase decisions. In the first axis we pointed out, above all, how and why it is necessary to keep user effort as low as possible. In the second we recommend working to improve the enjoyability of a system, as it has several benefits that lead to intentions to purchase and to user satisfaction. In the third group of guidelines, we summarise the observations from studies which were made either on complex datasets or with recommender systems across multiple domains. Finally we report three rules for providing users with suitable diversity.

One of the more important general contributions of this thesis is the deeper understanding of acceptance factors, and especially diversity. There are two points which we would like to discuss here. Through our investigations, we first managed to show that the Technology Acceptance Model fits the field of recommender systems very successfully. With its simplistic nature, it really captured the dimensions which users felt most strongly about. This observed strength is also its Achilles' heel, since a greater level of detail would be welcome for optimising recommenders in multiple environments. In our research model, as with McNee's Human Recommender Interaction model, we sought to propose and validate domain specific dimensions, hoping to provide this more detailed view. This was only partly successful. However, and this brings us to the second point, we were able to explain one good reason why this was the case. Through our last



experiment on diversity, we singled out that diversity had a time-dependent dimension. We believe that users do not always need diversity in their purchase process, and that this certainly also applies to other dimensions, helping to explain why they are so difficult to capture and verify with a static acceptance model.

Finally, through these observations, we were able to synthesise the different research models considered throughout our experiments. We proposed a diversity-model for maximising user's overall satisfaction, which expresses how *accuracy* and *diversity* should ideally be interleaved during a user's interaction with a system, in order to maximise their user experience and perceived system qualities.

## 8.2 Future Research Directions

In this section, we outline several promising research directions which would be helpful for extending the results and observations made so far in this thesis.

### 8.2.1 Adoption after Acceptance

In our initial work on acceptance, in particular through the comparison of two online music recommender systems, the experiments carried out showed some potentially promising directions. Unfortunately, both studies that we ran were conducted over a rather short period of time. A few weeks after both studies, we tried to contact several of the participants to gather some further comments about their experiences with either system, for instance seeking to see if they had continued to use them. This turned out to be somewhat unsuccessful for a number of reasons, including that many of the participants were students. Given their commitment to academic work, they had little time to listen more attentively to these online music radios. We are convinced that a long term study of such systems would be highly promising. Our investigations always considered users' acceptance of recommendations, and their intention to use or purchase in the future. However, we never directly assessed the *adoption* of the system. Therefore a study running across a longer timeline would be highly instructive.

Another issue observed in these first two studies, was the apparent lack of impact of social features. We are tempted to ask: why are the social features of websites like *Last.fm* so difficult for users to accept and adopt? Since carrying out the experiments, the general adoption of social-features across the Web has grown tremendously, and it is unclear if we would obtain similar observations on the music systems today. At the time, the post-study interviews revealed that users showed little interest in the websites' social features, and were even sometimes annoyed by them. Such findings further support how important it is to understand the preferences of users. The current Web 2.0 model highly encourages these so-called *social* tools where users interact and give opinions in all sorts of ways, but it remains unclear what motivates users to contribute to social recommenders, and how users perceive the possible benefits of such social contributions in recommenders. Such are the questions which were left open after our experiments on acceptance of recommendations.

## 8.2.2 Explicit and Implicit User Preferences

Through our work of Chapter 4, where we compared a *search & browse* interface, with recommendations computed on implicitly gathered profile data, we can envisage extensions which have the potential to yield fruitful observations. On a model close to that used by Swearingen *et al.* in [131], we believe it would be interesting to run a large-scale version of our study. Our experiment only tested two preference mechanisms among many, and on one dataset. Recommender systems using explicitly expressed preferences, such as rating information, have been around for many years now. We believe that including Amazon’s “Users who bought also bought” in our study could for instance have contributed to the overall evaluation of users’ preferences. At the time of designing our experiment, we had not selected this solution for two reasons. First of all we needed a baseline measure, since no other work on the topic had been established. Secondly, our will to evaluate the more recent approaches being developed (behavioural recommenders) drove us to choose two implicit groups (small and big profile). Other types of recommender systems could also be added, such as demographic or ontology based recommenders.

## 8.2.3 Refining Diversity Explorations

In this thesis, we have often observed the importance of diversity in users’ perceptions. The last two experiments we conducted, specifically designed to observe the importance of diversity, allowed us to model a first part of the sub-processes involving recommender systems within the framework of purchasing decisions.

The perspectives collected in the experiments consisted of elements which pointed to the impact of recommendation categories and the way in which these offered opportunities to discover new and interesting alternatives by providing greater diversity. However, we believe that one should now envisage some longer term studies. Future experiments should try to exploit our studies in order to attempt to anticipate users’ needs. It is clearly possible to envisage a dynamic algorithm which would improve the satisfaction of the users through a real-time usage analysis. This would almost certainly require the prior development of a mathematical formulation as explained below.

One can envisage many possible experiments linked to the time-dependent nature of accuracy and diversity as suggested in our model. One of the more interesting of these which we would consider for investigation in future work is a formalisation of the effect of diversity on users’ satisfaction. Our research model was established on the basis of our observations with the eye-tracker, but only represents the tendencies of the accuracy-diversity effects over time. In a future study, we would like to find a method for measuring accuracy and above all diversity. With the identification of the time-dependent sections, it would now be possible to run multiple studies where users are, in each time-region, presented with multiple combinations of accuracy and diversity. Through repeated and methodological measures, a mathematical formalisation of the two key dimensions should be possible. One of the challenges here would certainly be to find an effective way of measuring users’ satisfaction repeatedly, without disrupting the user from a reasonable purchase pattern. Such an investigation would also give the opportunity to explore if other dimensions have a role at any time in the recommendation process.

### 8.3 Take Home Message

In this dissertation, we have conducted several user studies. Despite what appears as very different settings and systems, we observed many behavioural patterns which were common across all users, especially regarding the acceptance of recommendations. Our investigations demonstrated the importance of reducing users' effort, but also how an enjoyable system can become useful to the user. Equally, we showed that understanding users' preferences, in a dynamic and contextual way, is a key to success, and that the quest for the highest accuracy in recommendations is not always the most valuable asset for users: diversity can be just as useful and indispensable in terms of users' perceptions.

To conclude this thesis, we choose to write just one take home message: we need to define new user-centric metrics for improving user-perceived qualities of recommenders. One strong message from this thesis, is that different systems rely on different recommender algorithms. Despite their strengths and weaknesses, each one can provide high quality accurate recommendations. Our online user studies reveal that users' trust them and are willing to use them. However, the precision of the algorithms is not what makes the real difference for users at the end of the day, hence our take home message. Users perceive qualities in a system through an ensemble of factors. In such a scenario, the enjoyability, the necessary effort or the diversity among suggestions, are elements which are just as important. Whilst continuing to strive to improve algorithms, we encourage researchers and technologists to understand users' preferences more dynamically. In order to be able to assist users at all times and in rich ways, we first need to determine what kinds of recommendations are valuable to users. It is only by putting the users first, at the centre of our research, that we will re-define adequate metrics leading to users' satisfaction. Let us not forget that, as H. D. Thoreau said: it's not what you look at that matters, it's what you see...



# Appendix A

## Appendix: Experiment 1

### A.1 Description of the User Study

This is the descriptive text of Experiment 1, which was given to the participants. It presented them with the instructions to complete the experiment.

#### Introduction

##### What is this user study about?

The goal of this study is to measure the success of music recommender systems.

Music Recommender Systems are programs which attempt to predict music that a user may be interested in, given some information about the user's profile. We need you to help us evaluate such a system. Don't try and be a "good" user, just interact as you would if you had discovered the product yourself. The study is separated in four steps.

**Step1: Get started** Tell us about your background, create a profile, (download,).

**Step2: Be a user** Listen to music, rate songs and artists, fill in our template.

**Step3: Tell us about it** Answer our 15min questionnaire, and tell us your thoughts.

**Step4: Summary** Finalise your statement by summarising your opinion.

The system that you will discover plays music (like a radio) through internet. The system is free and lets you give feedback on whether you like or not a song/artist, and gives you the chance to skip a song and go straight to the next one. Rating a song & artist is important as it will give the system the possibility to recommend more music to you. Detailed information is given on each system in the following sections.

### Technical requirements

The music recommender you will discover is compatible PC & Mac. (Windows 2000 and XP with Internet Explorer 6 and Firefox. It also works on the MacOS X 10.3+ with Safari and Firefox)

For an optimal experience, we recommend you conduct this study at EPFL or with a fast internet connection as it streams music constantly at a high quality. Flash 7 or 8 is required (8 recommended).

## A.2 Quick Presentation of Last.fm

This is the descriptive text which summarises how to get started with Last.fm, in the framework of this experiment. An equivalent for Pandora is detailed in the next section.

### Getting started with Last.fm

#### Last.fm

Last.fm is a music engine based on a massive collection of Music Profiles. Each music profile belongs to one person, and describes their taste in music. Last.fm uses these music profiles to make personalised recommendations, match you up with people who like similar music, and generate custom radio stations for each person. You can also tell Last.fm a song or artist you like to get started, but more options are available.

#### To start...

In order to get started, go to `http://www.last.fm` where you will first need to create a user by signing up (step 1), then download their application (step 2) (as shown in Figure A.1) and finally enter your username and password in the application (step 3).



Figure A.1: Snapshot from Last.fm

Downloading the program will create a folder with the application. Just launch it (no installation is required), enter the username and password you just created in step 1. Last.fm will then ask you to give the name of a artist or song, and that's it, you should be listening to music.

### Giving feedback with Last.fm

Giving feedback in Last.fm is really easy. The application features a “heart” button, and a “forbidden” button, which allow you to respectively indicate whether you like or hate a song/artist. The program records what you listen to, to your music profile.

## A.3 Quick Presentation of Pandora

### Getting started with Pandora.com

#### Pandora.com

Pandora is a music discovery service designed to help you find and enjoy music that you’ll love. The idea is that you tell Pandora one of your favorite songs or artists and they’ll launch a streaming radio station, built just for you, to explore that part of the music universe.

#### To start...

In order to start, go to <http://www.pandora.com>. The first time, you won’t need to create a user, nor to download and install an application, in order to listen to music. Their flash application will directly ask you to give the the name of an artist or song that you like, and will immediately create a radio station with similar musical characteristics.

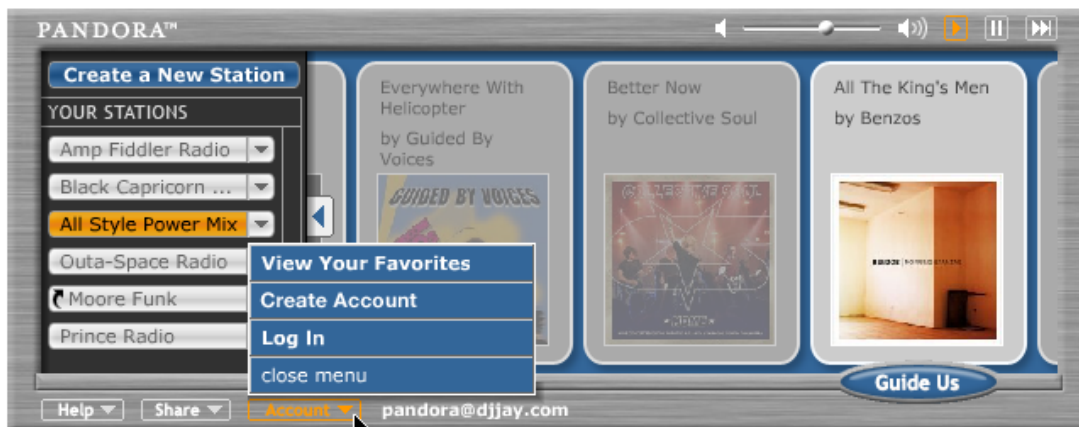


Figure A.2: Snapshot from Pandora

However, we ask you to create an account in order to store your musical preferences. In order to do this, you can click on the the “account” button (bottom left) (as shown in Figure A.2) and select “Create Account” or wait 15 minutes as Pandora.com will automatically propose it to you. Registration requires you give a United States zip code (for licensing reasons). Just invent any 5 digit number, it should work (example: 41179). A cookie will be created to remember your e-mail & password, so that you won’t need to enter this on further visits.

### **Giving feedback with Pandora.com**

Giving feedback in Pandora.com is really easy. You have two ways of doing it. You can either press the ovalshaped “Guide Us” button, or you can rick-click the album picture of a song. In both cases you will have the possibility to indicate your preferences.

## **A.4 Questions**

Users were asked to answer a small set of preference questions after testing a system.

### **Satisfaction and enjoyability**

How satisfied are you of the interaction with the system you tested? If not, what was the biggest interaction problem?

How enjoyable where the recommended songs? If not, what was the biggest problem?

### **General information**

In which conditions did the experiment take place (at work, at home, alone, with friends)

Were you interrupted whilst listening to music? If so, what did you then do?

Had you ever seen a music recommender system such as the one you just evaluated? If so, what is its or their name(s)?

In 30min, how many songs were you able to listen to?

### **Quality of recommendations**

How many songs did you really love ?

How many songs were you ready to purchase online?

What about offline, in a traditional shop?

Did you just appreciate or really discover music?

Do you feel you could have discovered more? If so, then why?

- feedback options were too limited
- heard certain songs twice
- not enough time to listen

### **Your opinion**

How was your first impression after the first 3 songs?



Would you have discovered this system on your own?

Did you try other features proposed by the system? If so, which?

Was the system good, compared to recommendations you may receive from a friend?

Did you find it easy to provide feedback? If not, why?

The system gave more personalized recommendations, based on my feedback.

Did you always listen until the end of songs, or often skip forward?

Do you find listening to music enjoyable whilst performing other tasks on the computer?

## **A.5 Template**

During the evaluation of both systems, users were given a printed template to fill out, allowing to record log-like information about the songs they had heard. The template is shown in Figure A.3.



## Appendix B

# Appendix: Experiment 2

### B.1 Description of the User Study

This is the descriptive text of Experiment 2, which was given to the participants. It presented them with the instructions to complete the experiment. It was declined into two versions, one for Last.fm detailed hereafter, and its equivalent for Pandora.

#### User study on music recommendation systems

Welcome to this user study and thank you for your interest. We need your help and feedback in order to test a music recommendation system.

By taking part and completing our user study you will automatically be entered into a draw for a 100 CHF gift (or voucher).

#### User study outline

This experiment is all about listening to music. So don't worry, it is not very strenuous and should be quite enjoyable. You may even perform other tasks whilst listening to music.

The experiment will happen in two short phases: the first for getting accustomed to the system, and the second to thoroughly test it (30min). At the end you will be asked to answer a short online questionnaire before having a small post-study interview.

#### Background information

Before you start, we would just like to know a little bit more about you. Please answer the following 6 background questions.

- Are you male or female?
- Your age group is?
- How big is your personal collection of music?

- The following adjective(s) best describe your current emotional status: energetic, happy, relaxed, calm, sad, tired, tense.
- Have you ever used Pandora or Last.fm?

### **Part 1 of the experiment**

#### **Last.fm**

Today, you will test *Last.fm*. The scenario is the following. You would like to get some recommendations for discovering new music. Instead of getting some recommendations from your friends, you will use the internet radio Last.fm.

Your goals will be to first get yourself set up with the system by creating an account (username & password). Then spend as much time as you need interacting with the site until the songs recommended to you are enjoyable.

#### **Observations**

- You don't need to download Last.fm's music player, you can use the online one.
- Last.fm has a "radio license" to broadcast music.
- Please be aware of the fact that we are not evaluating you, we are evaluating the system so relax. Don't try and perform "as you should" but just "be your normal self".
- This is not a race. You don't have to skip songs in order to hear as many as possible.

#### **Ready to start**

You are now ready to start listening. This is how you should proceed:

1. Go to <http://www.last.fm> and start the experiment
2. When you find that the music played is enjoyable, then you have finished the first part of the experiment

Thank you - you have finished for today. We will contact you again in ~10 days to fix another session to complete part 2 of the experiment. In the meantime, feel free to use the system at home.

### **Part 2 of the experiment**

Now that you are familiar with the system, we would like you to really use it in a concrete real-life situation.

The scenario here is the same as in the first part of the experiment. You would like to get some recommendations for discovering new music. Instead of getting some recommendations from your friends, you will use the internet radio Last.fm. Your goals this time will be to use the system thoroughly for half an hour and get recommendations for good music, as stated in the scenario. Try and see if you can discover new songs and new unknown artists that you like.

### **Observations**

Remember, we are not evaluating you, we are evaluating the system so relax. Don't try and perform "as you should" but just "be your normal self".

### **Ready to start**

You are now ready to start listening. This is how you should now proceed:

1. Go to <http://www.last.fm> , login and start the experiment (scenario)
2. In half an hour, stop listening
3. Then, go to <http://hci.ep.ch/study> and complete the main questionnaire
4. A short interview will conclude the study

Thank you very much for your time and consideration.

## **B.2 Post Study Interview**

After each study, we took the time to speak openly to the participants. The guideline-questions we used to assess their overall opinion are detailed hereafter.

- First Impressions
- Transparency; Control (feedback); Confidence?
- Did it work? Was it accurate?
- Did you feel that your feedback was taken into account?
- What did you expect from such a system?
- Last.fm: Was it useful to see friends and their profile?
- What disturbed you? What was wrong in the interaction?
- Did you use the social features?
- Will you continue using? Why?



## Appendix C

# Appendix: Experiment 3

### C.1 Description of the User Study

This is the descriptive text of the experiment, which was given to the participants. It presented them with the instructions to complete the study. It was declined into two versions, one for using Amazon without recommendations, and one for using the implicit recommender. The text bellow is for the recommender.

Welcome to this user study and thank you for your interest. We need your help and feedback in order to test a recommendation mechanism. By taking part and completing our user study you will automatically be entered into a draw for one of 10 x 20 CHF vouchers.

#### User study outline

This experiment explores users' perception of recommendations in different scenarios. It is a relatively leisurely experiment where you will be asked to search for five desired books on Amazon's website, and then rate them. Should be easy, right?

The experiment will happen as such: you will first search for five books, before rating them and finally answering a short global questionnaire. Feel free to ask any questions during the experiment.

#### Background information

Before you start, we would just like to know a little bit more about you. Please answer the following background questions.

- Are you male or female?
- Your age group is?
- Do you have an Amazon shopping profile? If so, how many books have you bought so far?

- How often do you surf or shop on Amazon?
- Do you read a lot of books?

### **The experiment**

Based on your past purchases, Amazon can recommend a whole selection of products. We would like to get your opinion on the quality of these recommendations.

The scenario we propose for this step of the experiment is the following. You would like to get some recommendations for discovering good *books*, so you go to Amazon's recommendation page, tailored to your profile. You browse through this list of recommendations and select five books, maybe glancing at the book details and short description. Based on this selection, you rapidly create your opinion on how good Amazon's recommendations are.

### **Observations**

- If Amazon indicates that “they don't have any recommendations for you today”, please contact the assistant.
- Please be aware of the fact that we are not evaluating you, we are evaluating the system so relax. Don't try and perform “as you should” but just “be your normal self”.

### **Experiment Procedure**

This is how you should now proceed:

1. please write them the time it is now
2. go to your usual Amazon website, and log in to your account (amazon.com or amazon.ch or amazon.fr or amazon.de)
3. head to your recommended product page on amazon.com/yourstore (or .de .fr .ch)
  - (a) please navigate through the *book* categories from the left menu. If you wish, you may refine this selection by navigating one level deeper into the subcategories, but be aware that if your profile is too small, Amazon will not be able to make any recommendations in the subcategories.
4. once you have found an item which you like, please write it down. For this we provide a template (see next section) which you can complete.
  - (a) for each item, we need your evaluation. For products that you don't already know, please build your opinion by reading the detailed information of each item before answering the rating questions on the template.
  - (b) please repeat step 3. such as to select five books in total.
5. to conclude, please answer the set of preference questions (detailed in Chapter 4) to indicate your overall appreciation of this kind of recommendations



## C.2 Template

During the experiments, users were given a printed template to fill out, allowing to record log-like information about the books they had found. The template is shown in Figure C.1.

**TEMPLATE**

**1** Book name : .....

By : .....

I have never heard of this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I think I will like this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I am willing to buy this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I selected this book because: .....

.....

**2** Book name : .....

By : .....

I have never heard of this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I think I will like this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I am willing to buy this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I selected this book because: .....

.....

**3** Book name : .....

By : .....

I have never heard of this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I think I will like this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I am willing to buy this book:   strongly disagree    -2    -1    0    +1    +2   strongly agree

I selected this book because: .....

.....

Figure C.1: Template provided.



## Appendix D

# Appendix: Experiment 4

Experiment 4 was carried out as an online study, where users had instructions which accompanied them throughout their tasks. Hereafter we detail the introductory text which they were given. It was mainly designed to present the study, its interface, and to make sure that users understood the iconic representation of critiques that they would be testing. The steps encountered are detailed in Chapter 5.

### D.1 Introduction to the User Study

Thank you for participating in this evaluation. The purpose of this trial is to evaluate various types of recommender feedback. For this trial we have developed a recommender system that uses a special type of feedback called a *critique*. Figure D.1 is a screenshot of a typical recommendation. There are two types of feedback options, *unit critiques* and *compound critiques*.

The panel on the left-hand side of the screen allows you to provide feedback on individual features of a recommendation - *unit critiques*. So for example, you can specify that you want a cheaper laptop by clicking on the button on the left-hand side of the price feature. Or you can tell the system you would prefer an Apple laptop by selecting “Apple” from the drop-down box for the Brand Feature.

On the bottom right-hand side of the screen are alternative feedback options called *compound critiques*. They illustrate other types of products available. If one of them matches your preferences you can click on the “I Like This” button.

This study proposes two types of compound critiques: *textual* and *visual* compound critiques. These are highlighted in Figure D.2. As the names suggest, the textual compound critiques use text to describe the properties of the recommendations, such as “Faster CPU, Larger Screen and Smaller Memory”. The visual compound critiques use icons to represent the same dimensions.

The icons each describe one feature (for instance: weight, speed, memory, etc.). The *colours* are used to describe positive (i.e. green) and negative (i.e. red) improvements. *Up and down arrows* are used to indicate if the value increases or diminishes. Gray and the equal sign indicate

The screenshot displays a product recommendation system interface. On the left is a 'Product Features' sidebar with dropdown menus for Brand (Sony), ProcessorType (Core Duo), ProcessorSpeed (2 GHz), ScreenSize (13.3 inches), HardDriveCapacity (120 GB), Weight (3.7 lbs), BatteryLife (6 hours), and Price (\$2999.99). Red arrows point from these features to the 'Unit Critiques' box. The main area is titled 'Our Recommendation' and features a Sony VAIO SZ280P/C laptop with a price of 2999.99 USD. Below this are 'Main Features' and a 'Product Description'. The 'Other Options' section lists five alternatives with their respective trade-offs. A 'Compound Critiques' box highlights the trade-offs for options 3, 4, and 5. A yellow box labeled 'Unit Critiques' is overlaid on the left, and another yellow box labeled 'Compound Critiques' is overlaid on the right.

Figure D.1: Screenshot of the system with Unit and Compound Critiques.

This close-up shows two critique boxes. The first is a 'Textual Compound Critique' for '1. More Optical Zoom.' which lists trade-offs: 'But Different Brand, More Expensive, Less Flash Memory, Smaller Screen and Thicker'. The second is a 'Visual Compound Critique' for '2. Kodak' which features a row of icons representing different camera models. Red arrows point from the text in the first critique to the icons in the second. Yellow boxes with red text label each critique type.

Figure D.2: Textual vs. Visual Compound Critiques.



Figure D.3: Iconic representation of weight variations.

an identical value. For example, if the weight of a digital camera increases, it will have an up arrow and red color, since this is rather negative. An example is shown in Figure D.3.



# Appendix E

## Appendix: Experiment 5

### E.1 Description of the User Study

This is the descriptive text for Experiment 5, which was given to the participants. It presented them with the instructions to complete the study.

#### Consumer Study of an E-commerce Site

We ask you to evaluate a perfume e-commerce site on how it helps you search and find products, not how it looks. We are also interested in how the site recommends products to you: when you are looking at a product's detail page, there will be some recommendations on the right hand side. The evaluation will take 20 minutes to complete and can be done in front of your own computer.

#### Reward for your participation!

By participating, you get a chance to win a USD 100\$ voucher to buy one of the perfumes that you will have selected in this experiment. Only 40 people are expected to take the study, so you have a strong chance of winning. Moreover, your participation helps us understand consumer behavior and improve interaction functions of the online stores such as Amazon.com. We thank you in advance for your participation and good will.

#### 3 steps involved in the study

1. We ask you some background questions. Such information is strictly confidential and will not be revealed to a third party.
2. You will simulate the searching and selection of three perfumes, which you are prepared to buy given the opportunity.
3. To finish, you will answer a short assessment questionnaire to tell us what you think of the e-commerce site.

Please be aware of the fact that we are not evaluating you, we are evaluating the system so relax. Don't try to perform "as you should" but just "be your normal self". (This study is optimised for Firefox, Safari, and the latest Internet Explorer 8. If you have an earlier version of Internet Explorer, please upgrade or use Firefox. Thank you.)

### **Step1: Let's get started!**

I'm ready to start the study. A link to the background questionnaire was provided. The questions are detailed in Chapter 6.

### **Step 2: Search and Select Three Perfumes**

Please *carefully read* the following instructions before you proceed.

- You will be directed to a fashion e-commerce website with more than 5000 popular and common perfumes, for both men and women.
- Your goal is to choose *three* new perfumes (for yourself) and put them into the shopping basket. This means that you are prepared to buy them given the opportunity.

Important: please select perfumes that you do not yet own or know of. Please also make sure you press the yellow *add to shopping list* button each time you select a perfume: don't worry, the purchase is a simulation but we do record the "add to shopping list" action for the purpose of the study.

*I have read the instructions, I want to start the experiment. A link to the interface was provided.*

### **Step 3: Nearly finished...**

To wrap up the experiment, would you please be so kind as to answer a small set of final preference questions? Ok, take me to the final preference questions. a link to the questionnaire was provided. Questions are listed in Chapter 6.



# Appendix F

## Appendix: Experiment 6

### F.1 Description of the User Study

This is the descriptive text for Experiment 6, which was given to the participants. As this was an in-depth study, a larger amount of background questions were asked before actually performing the evaluation in front of the eye-tracker. An assistant was present at all times during the study. The final preference questions are detailed in Chapter 6 directly.

#### Introduction

The Human computer interaction group of EPFL is conducting a user experiment in which we try to understand and evaluate how an e-commerce website effectively helps users find perfumes that they may want to buy for themselves, friends, and/or family members. This perfume catalogue contains more than 3500 most commonly used and sold perfumes in the world. Without the product search and other functionalities available on the website, it may be very difficult to go through the entire catalogue unless you already have an idea of what you want to buy.

Please select up to three perfumes that you have never heard or used before, but you are prepared to buy either for yourself or as gifts. Please put them in the basket. You may select more than three and delete some at the end. All participants will take part in a draw for a 100 CHF voucher, which the winner will be able to use to purchase one of these perfumes he will have added to his basket. Your answers will be handled in the strictest confidence. Thank you for your time and assistance.

#### Question Group no.1: Demographics questions

- What is your gender?
- What is your age group?
- Which continent are you originally from?

- What is your education level?
- What is your kind of profession?

**Question Group no.2: Background questions (related to their experience of perfume)**

Starting from here, we will start to ask you some questions about your online experience and your interest towards perfume.

1. Are you an internet user? If so, which of the following online applications do you usually use while getting on internet?
  - Online media. (For example, browsing news online using the website of BBC, CNN)
  - Information retrieval (For example, using Google to search for a hotel, a job)
  - Internet communication (For example, email, instant messenger, like MSN, Skype)
  - Online community (For example, using blog, forum, facebook)
  - Online entertainment (For example, playing online game, browsing youtube)
  - E-commerce (For example, purchasing air tickets, and using eBay, Amazon)
  - E-learning (For example, learning language, cooking skills)
2. For interviewee who has online purchase experience: As you have online purchase experience, we want to ask what products you have bought and how often. In the following, we will present you several choices. Please tell us how often you buy them. If you have never bought them, you could choose the option “rarely or never”.
  - For travel (Air tickets and accommodations)
  - Books
  - Music (including purchase of songs and CD), DVDs
  - Electronic items.
  - Food and Drinks
  - Other groceries like make-up and clothes
3. For interviewee who doesn't have any online purchase experience: As you haven't purchased online before, here we would like to present you some categories of things. And please choose the ones that you are planning to buy online in the next 3-6 months.
  - For travel (Air tickets and accommodations)
  - Books
  - Music (including purchase of songs and CD), DVDs

- Electronic items.
- Food and Drinks
- Other groceries like make-up and clothes

Ok, from now on, we would like to ask you some questions related to your interest and past experience about perfume.

1. How often do you buy perfume?
2. Which features of perfume significantly matter to you?
3. What are the perfumes you like/bought?
4. Have you used any online perfume website before?

**Question Group no.3: How do you get offline perfume recommendation?**

1. How do you discover new perfumes?
  - Friends
  - Advises from the seller in a perfume shop
  - TV ads
  - Magazines
  - Others (please specify)
2. Which kinds of information will you always reveal to the seller, in order to get his recommendations?
  - The price range that you could accept.
  - You will ask the seller to pick a perfume out of some specific brands, smell, personality.
  - You will tell him the perfume that you liked before.
  - You will tell him the perfume that you really disliked.
  - Others (please specify)
3. Were you always satisfied with their recommendation or not?
4. Have you met any problems when you want to get a perfume recommendation? What were they?

Now, you can start the experiment. Take the time to discover the functionalities and the interface. Imagine that you are looking for a new perfume. If you are satisfied at the end of the process or want to learn more about one or several products, you can add them in the basket.

The final preference questions are detailed in Chapter 6 directly.



# Bibliography

- [1] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on* 17, 6 (2005), 734–749.
- [2] AGGARWAL, C. C., WOLF, J. L., WU, K.-L., AND YU, P. S. Horting hatches an egg: a new graph-theoretic approach to collaborative filtering. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 1999), ACM, pp. 201–212.
- [3] AGRAWAL, R., IMIELIŃSKI, T., AND SWAMI, A. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (New York, NY, USA, 1993), ACM, pp. 207–216.
- [4] AGRAWAL, R., AND SRIKANT, R. Fast algorithms for mining association rules in large databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases* (San Francisco, CA, USA, 1994), Morgan Kaufmann Publishers Inc., pp. 487–499.
- [5] AHLBERG, C., AND SHNEIDERMAN, B. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1994), ACM Press, pp. 313–317.
- [6] AHN, T., RYU, S., AND HAN, I. The impact of web quality and playfulness on user acceptance of online retailing. *Inf. Manage.* 44, 3 (2007), 263–275.
- [7] AJZEN, I., AND FISHBEIN, M. *Understanding Attitudes and Predicting Social Behavior*, facsimile ed. Prentice Hall, March 1980.
- [8] ANAGNOSTOPOULOS, A., BRODER, A. Z., AND CARMEL, D. Sampling search-engine results. In *WWW '05: Proceedings of the 14th international conference on World Wide Web* (New York, NY, USA, 2005), ACM, pp. 245–256.
- [9] ANDERSON, C. *The Long Tail: Why the Future of Business is Selling Less of More*. No. ISBN 1-4013-0237-8. Hyperion, 2006.

## BIBLIOGRAPHY

---

- [10] ARMENGOL, E., PALAUDÀRIES, A., AND PLAZA, E. Individual prognosis of diabetes long-term risks: a cbr approach. *Methods of Information in Medicine* 40, 1 (March 2001), 46–51.
- [11] BAGOZZI, R. P., DAVIS, F. D., AND WARSHAW, P. R. Development and test of a theory of technological learning and usage. *Human Relations* 45, 7 (1992), 659–686.
- [12] BALABANOVIC, M., AND SHOHAM, Y. Combining content-based and collaborative recommendation. *Communications of the ACM* 40, 3 (March 1997).
- [13] BELKIN, N. J., AND CROFT, W. B. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM* 35, 12 (1992), 29–38.
- [14] BERNERS-LEE, T. Information management: A proposal. Tech. rep., CERN, 1990.
- [15] BERNERS-LEE, T., AND CAILLIAU, R. WorldWideWeb: Proposal for a hypertext project. Tech. rep., CERN, 1990.
- [16] BILLSUS, D., AND PAZZANI, M. J. User modeling for adaptive news access. *User Modeling and User-Adapted Interaction* 10, 2-3 (2000), 147–180.
- [17] BONHARD, P., HARRIES, C., MCCARTHY, J., AND SASSE, A. M. Accounting for taste: using profile similarity to improve recommender systems. In *Proc. CHI '06* (2006), ACM Press, pp. 1057–1066.
- [18] BONNIN, G., BRUN, A., AND BOYER, A. Using skipping for sequence-based collaborative filtering. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 775–779.
- [19] BONNIN, G., BRUN, A., AND BOYER, A. A low-order markov model integrating long-distance histories for collaborative recommender systems. In *IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces* (New York, NY, USA, 2009), ACM, pp. 57–66.
- [20] BRADLEY, K., AND SMYTH, B. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science* (2001).
- [21] BREESE, J. S., HECKERMAN, D., AND KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. pp. 43–52.
- [22] BURKE, R. D., HAMMOND, K. J., AND YOUNG, B. C. Knowledge-based navigation of complex information spaces. In *AAAI/IAAI, Vol. 1* (1996), pp. 462–468.
- [23] BURKE, R. D., HAMMOND, K. J., AND YOUNG, B. C. The findme approach to assisted browsing. *IEEE Expert* 12 (1997), 32–40.

## BIBLIOGRAPHY

---

- [24] CARENINI, G., AND MOORE, J. D. Multimedia explanations in IDEA decision support system. In *Working Notes of the AAAI Spring Symposium on Interactive and Mixed-Initiative Decision Theoretic Systems* (Menlo Park, CA, 1998), P. Haddawy and S. Hanks, Eds., pp. 16–22.
- [25] CASTAGNOS, S., AND BOYER, A. Privacy concerns when modeling users in collaborative filtering recommender systems. *Book chapter in “Social and Human Elements of Information Security: Emerging Trends and Countermeasures”* (2008).
- [26] CASTAGNOS, S., AND JONES, N. Recommenders’ Influence on Buyers’ Decision Process. In *In Proceedings of Third ACM Conference on Recommender Systems* (New-York, 2009), ACM.
- [27] CELMA, O., AND CANO, P. From hits to niches? or how popular artists can bias music recommendation and discovery. In *2008 ACM Conference on Recommender Systems* (Las Vegas, USA, 24/08/2008 2008).
- [28] CELMA, O., AND HERRERA, P. A new approach to evaluating novel recommendations. In *2008 ACM Conference on Recommender Systems* (Lausanne, Switzerland, 23/10/2008 2008).
- [29] CHEN, L. *User decision improvement and trust building in product recommender systems*. PhD thesis, EPFL, Lausanne, 2008.
- [30] CHEN, L., AND PU, P. Evaluating critiquing-based recommender agents. In *AAAI’06: Proceedings of the 21st national conference on Artificial intelligence* (2006), AAAI Press, pp. 157–162.
- [31] CHEN, L., AND PU, P. Preference-based organization interfaces: Aiding user critiques in recommender systems. In *UM ’07: Proceedings of the 11th international conference on User Modeling* (Berlin, Heidelberg, 2007), Springer-Verlag, pp. 77–86.
- [32] CHEN, L., AND PU, P. A cross-cultural user evaluation of product recommender interfaces. In *RecSys ’08: Proceedings of the 2008 ACM conference on Recommender systems* (New York, NY, USA, 2008), ACM, pp. 75–82.
- [33] CHOICESTREAM, I. 2006 choicestream personalization survey. Tech. rep., February 2006.
- [34] CHOICESTREAM, I. 2007 annual choicestream personalization survey. Tech. rep., 2007.
- [35] CHOICESTREAM, I. Personalization: An e-commerce business imperative according to choicestream panel audience. Tech. rep., June 2008.
- [36] CLAYPOOL, M., BROWN, D., LE, P., AND WASEDA, M. Inferring user interest. *IEEE Internet Computing* 5 (2001), 32–39.
- [37] COYLE, M., AND SMYTH, B. Enhancing web search result lists using interaction histories. *Advances in Information Retrieval* (2005), 543–545.

## BIBLIOGRAPHY

---

- [38] CRANOR, L. F. 'I didn't buy it for myself': Privacy and ecommerce personalization. In *ACM Workshop on Privacy in the Electronic Society* (2003).
- [39] CUTTING, D. R., KARGER, D. R., PEDERSEN, J. O., AND TUKEY, J. W. Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1992), ACM, pp. 318–329.
- [40] DAVENPORT, T. H., AND BECK, J. C. *The Attention Economy: Understanding the New Currency of Business*. Harvard Business School Press, 2001.
- [41] DAVIS, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (September 1989), 319–340.
- [42] DAVIS, F. D., BAGOZZI, R. P., AND WARSHAW, P. R. User acceptance of computer technology: a comparison of two theoretical models. *Manage. Sci.* 35, 8 (August 1989), 982–1003.
- [43] DEPARTMENT, D. B., AND BRIDGE, D. Product recommendation systems: A new direction. In *Workshop on CBR in Electronic Commerce at The International Conference on Case-Based Reasoning (ICCBR-01)* (2001), pp. 79–86.
- [44] DOODY, J., COSTELLO, E., MCGINTY, L., AND SMYTH, B. Combining visualization and feedback for eyewear recommendation. In *FLAIRS Conference* (2006), pp. 135–140.
- [45] DORSAY, M., MANNING, H., AND CARNEY, C. L. Death by a thousand cuts kills web experience. Tech. rep., Forrester Research, Inc., August 2006.
- [46] FISHBEIN, M., AND AJZEN, I. *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley, Reading MA., 1975.
- [47] FLEDER, D., AND HOSANAGAR, K. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Manage. Sci.* 55, 5 (2009), 697–712.
- [48] FREIERT, M. February top social networks – make way for the new guys. Compete Inc., Web Study, March 2008.
- [49] GÖKER, M. H., AND THOMPSON, C. A. The adaptive place advisor: A conversational recommendation system. In *In Proceedings of the 8th German Workshop on Case Based Reasoning* (2000), DaimlerChrysler, pp. 187–197.
- [50] GOLDBERG, D., NICHOLS, D., OKI, B. M., AND TERRY, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM* 35 (1992), 61–70.
- [51] GOLDBERG, J. H., AND KOTVAL, X. P. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics* 24 (1999), 631–645.



## BIBLIOGRAPHY

---

- [52] GONG, S., YE, H., AND TAN, H. Combining memory-based and model-based collaborative filtering in recommender system. In *PACCS '09: Proceedings of the 2009 Pacific-Asia Conference on Circuits, Communications and Systems* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 690–693.
- [53] GOOD, N., SCHAFER, B. J., KONSTAN, J. A., BORCHERS, A., SARWAR, B., HERLOCKER, J., AND RIEDL, J. Combining collaborative filtering with personal agents for better recommendations. In *AAAI '99/IAAI '99* (1999), pp. 439–446.
- [54] HAIR, JR., J. F., ANDERSON, R. E., AND TATHAM, R. L. *Multivariate data analysis with readings (2nd ed.)*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1986.
- [55] HAUBL, G., AND MURRAY, K. Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents. *Journal of Consumer Psychology* 13(1) (2003), 75–91.
- [56] HAUBL, G., AND TRIFTS, V. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science* 19(1) (2000), 4–21.
- [57] HAUSER, J., AND WERNERFELT, B. An evaluation cost model of consideration sets. *Journal of Consumer Research* 16 (March 1990), 393–408.
- [58] HEAD, M. M., AND HASSANEIN, K. Building online trust through socially rich web interfaces. In *PST* (2004), pp. 15–22.
- [59] HERLOCKER, J. L., KONSTAN, J. A., AND RIEDL, J. Explaining collaborative filtering recommendations. *CSCW'00* (2000), 241–250.
- [60] HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., AND RIEDL, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22, 1 (2004), 5–53.
- [61] HIJIKATA, Y., SHIMIZU, T., AND NISHIDA, S. Discovery-oriented collaborative filtering for improving user satisfaction. In *IUI '09: Proceedings of the 13th international conference on Intelligent user interfaces* (New York, NY, USA, 2009), ACM, pp. 67–76.
- [62] HILL, W., STEAD, L., ROSENSTEIN, M., AND FURNAS, G. Recommending and evaluating choices in a virtual community of use. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems* (1995), ACM Press/Addison-Wesley Publishing Co., pp. 194–201.
- [63] HO, S. Web personalization and its effects on users' information processing and decision making. PhD Thesis of the Hong Kong University of Science and Technology, 2004.
- [64] HO, S., BODOFF, D., AND TAM, K. Timing of adaptive web personalization and its effects on online consumer behavior. *Information Systems Research* (2009).
- [65] HO, S., AND TAM, K. An empirical examination of the effects of web personalization at different consumer decision-making stages. *International Journal of Human-Computer Interaction* 19(1) (2005), 95–112.

## BIBLIOGRAPHY

---

- [66] JONES, N., AND PU, P. User Technology Adoption Issues in Recommender Systems. In *Proceedings of the 2007 Networking and Electronic Commerce Research Conference* (Riva del Garda, Italy, 2007), pp. 379–394.
- [67] JONES, N., PU, P., AND CHEN, L. How Users Perceive and Appraise Personalized Recommendations. In *Proceedings of User Modelling, Adaptation and Personalization* (2009), Springer.
- [68] JUST, M. A., AND CARPENTER, P. A. Eye fixations and cognitive processes. *Cognitive Psychology* 8 (1976), 441–480.
- [69] KAMIS, A. A., AND STOHR, E. A. Parametric search engines: what makes them effective when shopping online for differentiated products? *Inf. Manage.* 43, 7 (2006), 904–918.
- [70] KEENEY, R. L., AND RAIFFA, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc, New York, 1976.
- [71] KIM, S., AND ANDRÉ, E. Composing affective music with a generate and sense approach. In *Proceedings of Flairs 2004 - Special Track on AI and Music* (2004), AAAI Press.
- [72] KONSTAN, J. A., MILLER, B. N., MALTZ, D., HERLOCKER, J. L., GORDON, L. R., AND RIEDL, J. Grouplens: applying collaborative filtering to usenet news. *Commun. ACM* 40, 3 (1997), 77–87.
- [73] KOUFARIS, M., AND HAMPTON-SOSA, W. Customer trust online: examining the role of the experience with the web-site. CIS Working Paper Series CIS-2002-05, Zicklin School of Business, Baruch College, New York, may 2002.
- [74] KRAUSE, S. Pandora and last.fm: Nature vs. nurture in music recommenders. Blog Article, January 2006.
- [75] KRULWICH, B. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine* 18, 2 (1997), 37–45.
- [76] LANG, K. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Machine Learning Conference (ML95)* (1995).
- [77] LANG, P. J. The emotion probe: Studies of motivation and attention. *American Psychologist* 50, 371-385 (1995).
- [78] LEELAYOUTHAYOTIN, L. Factors influencing online purchase intention: the case of health food consumers in thailand. 2004.

- [79] LIN, X., SOERGEL, D., AND MARCHIONINI, G. A self-organizing semantic map for information retrieval. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 1991), ACM, pp. 262–269.
- [80] LINDEN, G., HANKS, S., AND LESH, N. Interactive assessment of user preference models: The automated travel assistant. In *In Proceedings of the Sixth International Conference on User Modeling* (1997), Springer, pp. 67–78.
- [81] LINDEN, G. D., JACOBI, J. A., AND BENSON, E. A. Collaborative recommendations using item-to-item similarity mappings. *United States Patent 6266649* (2001).
- [82] MAES, P., GUTTMAN, R., MOUKAS, A., AND MOUKAS, R. Agents that buy and sell: Transforming commerce as we know it. *Communications of the ACM* 42 (1999), 81–91.
- [83] MARCUS, A. Icon and symbol design issues for graphical user interfaces. 257–270.
- [84] MCCARTHY, K., REILLY, J., MCGINTY, L., AND SMYTH, B. On the dynamic generation of compound critiques in conversational recommender systems. *Adaptive Hypermedia and Adaptive Web-Based Systems* (2004), 176–184.
- [85] MCCARTHY, K., REILLY, J., MCGINTY, L., AND SMYTH, B. Experiments in dynamic critiquing. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces* (New York, NY, USA, 2005), ACM, pp. 175–182.
- [86] MCGINTY, L., AND SMYTH, B. On the role of diversity in conversational recommender systems. In *ICCBR* (2003), pp. 276–290.
- [87] MCNEE, S. M. *Meeting user information needs in recommender systems*. PhD thesis, Minneapolis, MN, USA, 2006. Adviser-Konstan, Joseph A.
- [88] MCNEE, S. M., RIEDL, J., AND KONSTAN, J. A. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems* (New York, NY, USA, 2006), ACM, pp. 1097–1101.
- [89] MCNEE, S. M., RIEDL, J., AND KONSTAN, J. A. Making recommendations better: an analytic model for human-recommender interaction. In *CHI '06 extended abstracts on Human factors in computing systems* (2006), ACM Press, pp. 1103–1108.
- [90] MCSHERRY, D. Diversity-conscious retrieval. In *ECCBR '02: Proceedings of the 6th European Conference on Advances in Case-Based Reasoning* (London, UK, 2002), Springer-Verlag, pp. 219–233.
- [91] MCSHERRY, D. Similarity and compromise. In *Case-Based Reasoning: Research and Development. Proc. ICCBR '03* (2003), K. D. Ashley and D. G. Bridge, Eds., vol. 2689 of *Lect. Notes Artif. Intell.*, pp. 291–305.

## BIBLIOGRAPHY

---

- [92] MONTANER, M., LÓPEZ, B., AND DE LA ROSA, J. L. A taxonomy of recommender agents on the internet. *Artif. Intell. Rev.* 19, 4 (2003), 285–330.
- [93] NEUMANN, A. W. Algorithms for behavior-based recommender systems. *Recommender Systems for Information Providers* (2009), 1–25.
- [94] NICHOLS, D. M. Implicit rating and filtering. In *In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering* (1997), pp. 31–36.
- [95] PAYNE, J., BETTMAN, J., AND JOHNSON, E. *The Adaptive Decision Maker*. Cambridge University Press, 1993.
- [96] PAYNE, J. W., BETTMAN, J. R., AND SCHKADE, D. A. Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty* 19, 1-3 (December 1999), 243–70.
- [97] PENNOCK, D., HORVITZ, E., LAWRENCE, S., AND GILES, C. L. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proc. UAI 2000*, pp. 473–480.
- [98] POMERANTZ, D., AND DUDEK, G. Context dependent movie recommendations using a hierarchical bayesian model. In *Canadian AI '09: Proceedings of the 22nd Canadian Conference on Artificial Intelligence* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 98–109.
- [99] PU, P., AND CHEN, L. Integrating tradeoff support in product search tools for e-commerce sites. In *EC '05: Proceedings of the 6th ACM conference on Electronic commerce* (New York, NY, USA, 2005), ACM, pp. 269–278.
- [100] PU, P., AND CHEN, L. Trust building with explanation interfaces. In *IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces* (New York, NY, USA, 2006), ACM Press, pp. 93–100.
- [101] PU, P., AND CHEN, L. User-involved preference elicitation for product search and recommender systems. *AI Magazine* 29(4) (2008), pp. 93–103.
- [102] PU, P., AND FALTINGS, B. Enriching buyers' experiences: the smartclient approach. In *Proc. CHI '00* (2000), ACM Press, pp. 289–296.
- [103] PU, P., HUAN, Z., AND KUMAR, P. Evaluating example-based search tools. In *EC '04: Proceedings of the 5th ACM conference on Electronic commerce* (New York, NY, USA, 2004), ACM, pp. 208–217.
- [104] PU, P., AND JANECEK, P. Visual interfaces for opportunistic information seeking. *J. Appl. Math* 44 (1987), 665–674.
- [105] PU, P., VIAPPIANI, P., AND FALTINGS, B. Increasing user decision accuracy using suggestions. In *Proc. CHI '06* (2006), ACM Press, pp. 121–130.

## BIBLIOGRAPHY

---

- [106] PU, P., ZHOU, M., AND CASTAGNOS, S. Critiquing Recommenders for Public Taste Products. In *Third ACM Conference on Recommender Systems* (New-York, 2009), ACM.
- [107] RADLINSKI, F., AND DUMAIS, S. Improving personalized web search using result diversification. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2006), ACM, pp. 691–692.
- [108] RASHID, A., ALBERT, I., COSLEY, D., LAM, S., MCNEE, S., KONSTAN, J., AND RIEDL, J. Getting to know you: Learning new user preferences in recommender systems. In *Proc. IUI 2002*, pp. 127–134.
- [109] REILLY, J., MCCARTHY, K., MCGINTY, L., AND SMYTH, B. Dynamic critiquing. In *In Proceedings fo the Seventh European Conference on Case-Based Reasoninb (ECCBR-04)* (2004), Springer, Ed., vol. 3155, pp. 763–777.
- [110] REILLY, J., MCCARTHY, K., MCGINTY, L., AND SMYTH, B. Explaining compound critiques. *Artif. Intell. Rev.* 24, 2 (2005), 199–220.
- [111] REILLY, J., MCCARTHY, K., MCGINTY, L., AND SMYTH, B. Incremental critiquing. *Research and Development in Intelligent Systems XXI* (2005), 101–114.
- [112] REILLY, J., ZHANG, J., MCGINTY, L., PU, P., AND SMYTH, B. A comparison of two compound critiquing systems. In *International Conference on Intelligent User Interfaces, Proceedings IUI (2007)*, Association for Computing Machinery, New York, NY 10036-5701, United States, pp. 317–320. Adaptive Information Cluster, School of Computer Science and Informatics, UCD Dublin, Ireland.
- [113] REILLY, J., ZHANG, J., MCGINTY, L., PU, P., AND SMYTH, B. Evaluating compound critiquing recommenders: A real-user study. In *EC'07 - Proceedings of the Eighth Annual Conference on Electronic Commerce* (2007), Association for Computing Machinery, New York, NY 10036-5701, United States, pp. 114–123. Adaptive Information Cluster, School of Computer Science and Informatics, UCD Dublin, Ireland.
- [114] RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTORM, P., AND RIEDL, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proc. of ACM CSCW'04*, ACM, pp. 175–186.
- [115] RESNICK, P., AND VARIAN, H. R. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.
- [116] RICH, E. User modeling via stereotypes. 329–342.
- [117] SARWAR, B., KARYPIS, G., KONSTAN, J., AND REIDL, J. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web* (New York, NY, USA, 2001), ACM, pp. 285–295.

## BIBLIOGRAPHY

---

- [118] SCHAFER, J. B., KONSTAN, J. A., AND RIEDL, J. Recommender systems in e-commerce. In *ACM Conference on Electronic Commerce* (1999), pp. 158–166.
- [119] SHAPIRA, B., TAIEB-MAIMON, M., AND MOSKOWITZ, A. Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. In *SAC '06* (New York, NY, USA, 2006), ACM.
- [120] SHARDANAND, U., AND MAES, P. Social information filtering: Algorithms for automating “word of mouth”. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems* (1995), vol. 1, pp. 210–217.
- [121] SHEN, E., LIEBERMAN, H., AND LAM, F. What am i gonna wear?: scenario-oriented recommendation. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces* (New York, NY, USA, 2007), ACM, pp. 365–368.
- [122] SHIMAZU, H. Expertclerk: Navigating shoppers buying process with the combination of asking and proposing. In *IJCAI* (2001), pp. 1443–1450.
- [123] SHNEIDERMAN, B., AND MAES, P. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997).
- [124] SIMON, H. A. *The Sciences of the Artificial* (3rd ed.). The MIT Press, 1996.
- [125] SMITH, D., MENON, S., AND SIVAKUMAR, K. Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing* 19, 3 (2005), 15–37.
- [126] SMYTH, B., AND COTTER, P. A personalised tv listings service for the digital tv age. *Knowl.-Based Syst.* 13, 2-3 (2000), 53–59.
- [127] SMYTH, B., AND MCCLAVE, P. Similarity vs. diversity. In *ICCB* (2001), pp. 347–361.
- [128] SMYTH, B., MCGINTY, L., REILLY, J., AND MCCARTHY, K. Compound critiques for conversational recommender systems. In *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 145–151.
- [129] SPIEKERMANN, S., AND PARASCHIV, C. Motivating human–agent interaction: Transferring insights from behavioral marketing to interface design. *Electronic Commerce Research* 2, 3 (2002), 255–285.
- [130] STOLZE, M. Soft navigation in electronic product catalogs. *International Journal on Digital Libraries* 3, 1 (07 2000), 60–66.
- [131] SWEARINGEN, K., AND SINHA, R. Interaction design for recommender systems. *Designing Interactive Systems* (2002).

- [132] TANG, T. Y., AND WINOTO, P. Mood and recommendations: On non-cognitive mood inducers for high quality recommendation. In *APCHI '08: Proceedings of the 8th Asia-Pacific conference on Computer-Human Interaction* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 94–103.
- [133] TATEMURA, J., SANTINI, S., AND JAIN, R. Social and content-based information filtering for a web graphics recommender system. In *ICIAP '99: Proceedings of the 10th International Conference on Image Analysis and Processing* (Washington, DC, USA, 1999), IEEE Computer Society, p. 842.
- [134] TEEVAN, J., DUMAIS, S. T., AND HORVITZ, E. Personalizing search via automated analysis of interests and activities. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2005), ACM, pp. 449–456.
- [135] TVERSKY, A., AND SIMONSON, I. Context-dependent preferences. *Manage. Sci.* 39, 10 (1993), 1179–1189.
- [136] VAN DER HEIJDEN, H. Using the technology acceptance model to predict usage: Extensions and empirical test, 1986.
- [137] WILLIAMS, M. D., AND TOU, F. N. Rabbit: An interface for database access. In *ACM 82: Proceedings of the ACM '82 conference* (New York, NY, USA, 1982), ACM, pp. 83–87.
- [138] YU, C., LAKSHMANAN, L. V. S., AND AMER-YAHIA, S. Recommendation diversification using explanations. In *ICDE '09: Proceedings of the 2009 IEEE International Conference on Data Engineering* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 1299–1302.
- [139] ZHANG, J., JONES, N., AND PU, P. A visual interface for critiquing-based recommender systems. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce* (New York, NY, USA, 2008), ACM, pp. 230–239.
- [140] ZHANG, J., AND PU, P. Survey of solving multi-attribute decision problems. Tech. rep., 2004.
- [141] ZHANG, J., AND PU, P. A comparative study of compound critique generation in conversational recommender systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2006), vol. 4018 NCS, Springer Verlag, Heidelberg, D-69121, Germany, pp. 234–243. Human Computer Interaction Group, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne CH-1015, Switzerland.
- [142] ZHANG, J., AND PU, P. Refining preference-based search results through bayesian filtering. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces* (New York, NY, USA, 2007), ACM, pp. 294–297.

## BIBLIOGRAPHY

---

- [143] ZHANG, M., AND HURLEY, N. Avoiding monotony: improving the diversity of recommendation lists. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems* (New York, NY, USA, 2008), ACM, pp. 123–130.
- [144] ZIEGLER, C.-N., MCNEE, S. M., KONSTAN, J. A., AND LAUSEN, G. Improving recommendation lists through topic diversification. In *Proc. WWW '05*, ACM Press, pp. 22–32.
- [145] ZUKERMAN, I., AND ALBRECHT, D. W. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction 11*, 1-2 (2001), 5–18.



# Nicolas Jones

Human Computer Interaction Group (HCIG)  
School of Computer and Communications Sciences  
École Polytechnique Fédérale de Lausanne (EPFL)

EPFL IC GR-SCI-IC GR-PU  
BC 144 (Bâtiment BC)  
Station 14  
CH-1015 Lausanne

nicolas.jones@epfl.ch

## Research Directions

Recommender systems, interaction design, usability evaluation, acceptance issues, critiquing, user preferences, diversity, entertainment e-commerce, intelligent interfaces, human computer interaction.

## Professional Qualifications

### EPFL - Since January 2006

Ph.D researcher at EPFL in the Human Computer Interactions Group, under the supervision of Dr. P. Pu. Research topic: low-involvement recommender systems. User-studies on user centric issues leading to acceptance and adoption of recommender systems. Exploration of the impact of visualisations for compound critiques. Creation of a music recommendation and visualisation framework coupled with Last.fm.

### IBM - February 2005

Diploma project at IBM Research (Zürich-Rüschlikon, Switzerland). Implementation and performance measures of load balancing algorithms in a detection & pre-treatment architecture for RFID signals.

### Net Oxygen Sàrl - 2000-2007

Co-founder, member of the board and project manager at Net Oxygen Sàrl. The company provides a wide range of services in the IT sector. Core business includes datacenter services, website hosting, information systems and website realisations.

## Education

2006-present Ph.D. in Computer, Communication and Information Sciences  
Swiss Federal Institute of Technology (EPFL), Lausanne

2000-2005 Master in Computer, Communication and Information Sciences  
Swiss Federal Institute of Technology (EPFL), Lausanne

1998-2000 Swiss Federal Certificate of Maturity, Mathematics & Science,  
Cantonal College, Nyon

## Languages

French / English: bilingual

German: fluent

Italian: notions

## Professional Activities

Reviewer at the 2009 and 2010 International Conference on Intelligent User Interfaces.

## Interests

Music: co-founder and leader of a music band of 10 musicians (since 2002).

Harvard Model United Nations: UN simulation, 2000 delegates. Negotiation & public speeches.

Passion for visual design and photography.