La Redoute

UNIVERSITE LILLE 1

LEM – LILLE ÉCONOMIE MANAGEMENT (UMR 9221)

ÉCOLE DOCTORALE SÉSAM

PERSONALIZATION IN E-COMMERCE: A PROCEDURE TO CREATE AND EVALUATE BUSINESS RELEVANT RECOMMENDATION SYSTEMS

THESE EN VUE DE L'OBTENTION DU TITRE DE DOCTEUR EN SCIENCES ÉCONOMIQUES

Stijn GEUENS

25 avril 2017

SOUS LA DIRECTION DE PROF. DR. KRISTOF COUSSEMENT ET PROF. DR. KOEN W. DE BOCK

MEMBRES DU JURY

Directeurs de thèse :

Dr. Kristof COUSSEMENT, Full professor, IESEG School of Management

Dr. Koen W. DE BOCK, Associate professor, Audencia Business School

Rapporteurs :

Dr. Dietmar JANNACH, Full professor, TU Dortmund

Dr. Wouter VERBEKE, Assistant professor, VU Brussels

Examinateur :

Dr. Dominique CRIÉ, Professeur des universités, Université Lille 1, LEM UMR CNRS 9221

Invité :

Arnaud BOUTELIER, Responsable big data analytics, La Redoute

UNIVERSITE LILLE 1

LEM – LILLE ÉCONOMIE MANAGEMENT (UMR 9221)

ÉCOLE DOCTORALE SÉSAM

PERSONALIZATION IN E-COMMERCE: A PROCEDURE TO CREATE AND EVALUATE BUSINESS RELEVANT RECOMMENDATION SYSTEMS

THESE EN VUE DE L'OBTENTION DU TITRE DE DOCTEUR EN SCIENCES ÉCONOMIQUES

STIJN GEUENS

25 avril 2017

SOUS LA DIRECTION DE PROF. DR. KRISTOF COUSSEMENT ET PROF. DR. KOEN W. DE BOCK

MEMBRES DU JURY

Directeurs de thèse :

Dr. Kristof COUSSEMENT, Full professor, IESEG School of Management

Dr. Koen W. DE BOCK, Associate professor, Audencia Business School

Rapporteurs :

Dr. Dietmar JANNACH, Full professor, TU Dortmund

Dr. Wouter VERBEKE, Assistant professor, VU Brussels

Examinateur :

Dr. Dominique CRIÉ, Professeur des universités, Université Lille 1, LEM UMR CNRS 9221

Invité :

Arnaud BOUTELIER, Responsable big data analytics, La Redoute

L'université Lille 1 n'entend donner aucune approbation ni improbation aux opinions émises dans cette thèse. Ces opinions doivent être considérées comme propres à leur auteur.

LABORATOIRE DE RATTACHEMENT :

Lille Économie Management (LEM – UMR CNRS 9221), Laboratoire de recherche rattaché à l'Université de Lille et à la Fédération Universitaire Polytechnique de Lille (FUPL)

Préparation de la thèse sur le site de l'IESEG School of Management, 3 Rue de la Digue, 59000 Lille

ACKNOWLEDGEMENTS

Three years ago, I decided to go back to academia and started my Ph.D. During this journey, a lot of people contributed, directly and indirectly, to my dissertation and I would like to take a moment to thank them.

First, I would like to express my gratitude to Prof. dr. Kristof Coussement and Prof. dr. Koen W. Bock, my two supervisors. They shared a lot of knowledge, give great advice, contributed actively to performed studies and reviews, and were always ready to help. I remember a quote dating back to one of the first days of my Ph.D.: "We are your employees. Feel free to contact us whenever you need to." Besides great advisors, they have both congenial personalities. Working with them and going on conference trips together were very enjoyable experiences. Kristof and Koen, thank you very much for the great experience during the past three year.

Further, I would like to thank the members of the Ph.D. exam committee for all the time and effort they spent in reading and evaluating this dissertation: Prof. dr. Dietmar Jannach, Prof. dr. Wouter Verbeke, Prof. dr. Dominique Crié.

I would like to express my gratitude to La Redoute as a company for giving me the opportunity to work with their data, supporting me financially and operationally in conducting my Ph.D., and facilitating field experiments. I want to thank especially my colleagues of the BI team and in specific the big data analytics team. Finally, I want to acknowledge the contribution of Arnaud. You were a great mentor guiding me at La Redoute and contributing to my Ph.D. by delivering excellent operational insights. Merci à vous tous!

A second institution I want to thank, is IESEG School of Management. They provided me with the opportunity to pursue my academic goals by offering operational and financial support. I would like to thank everyone from general management, over research department, to IT department.

Third, I would like to thank Université Lille 1, Ecole Doctorale SÉSAM, and LEM for allowing me to complete my Ph.D. research. Finally, I would like to thank ANRT for donating a CIFRE grant. This grant served as the perfect glue to allow a smooth collaboration between academic research and industrial operationalization.

Special thanks go to my colleagues of the marketing and negotiations department at IESEG. They are always ready for good advice and a friendly chat. I especially want to thank the Flemish colleagues for the many inspiring lunches and great moments on the second floor of building B.

A group I could/should not forget are my fellow Ph.D. students, as we spent a lot of time together in the same office. I would like to thank Kristine, Libo, Salim, Koi, Zhyang, Annabelle, Marion, Karina, Albane, and of course the marketing and negotiations guys and girls: Jenny, Helen, Adrian, Arno, Christina, and Steven. Even though we spent most of our days together in the same room, lunches and occasional after hours' drinks were always a fun experience. Thanks to you all!

I would also like to thank my family. Especially my two brothers, Sep and Neel, and my parents. Without them I would not have been here. They gave me the opportunity to pursue my studies and even today they are morally supporting me, especially during weekends when working and living in Retie.

An important group to recognize are my friends. Even though they might not have contributed directly to my dissertation, they helped me to relax and to take the edge off in stressful times. Thanks to 'den Angel', 'de Hodonk', and all the others for the occasional visits in Lille and fun weekends.

Stijn Geuens
Lille, January 29th, 2017

RESUME GLOBAL

Les systèmes de recommandation sont un sujet très étudié dans la littérature sur l'apprentissage automatique, ce qui a permis la création de nombreux algorithmes de pointe. Cette thèse doctorale va au-delà de simples propositions de nouveaux algorithmes de recommandation en tirant parti des toutes dernières techniques et en étudiant les interactions de ces techniques avec diverses sources de données de différents types. Nous nous sommes penchés sur la création de canevas capables d'aider les universitaires et les décideurs du marché dans le cadre du développement, de l'évaluation et du test des systèmes de recommandation dans le contexte du commerce en ligne. Dans ce but, cette thèse se penche d'abord en profondeur sur un algorithme spécifique (filtrage collaboratif) dans un environnement hors-ligne portant sur des données historiques, puis élargit son champ d'investigation pour étudier les systèmes de recommandation hybrides et de maximisation du chiffre d'affaires et effectuer une expérience de terrain. Concrètement, cette thèse apporte une nouveauté à la littérature de sept manières différentes. Premièrement, nous décrivons dans le chapitre I un cadre devant servir à évaluer l'influence des caractéristiques d'entrée, d'une matrice d'achat binaire, sur le meilleur algorithme de filtrage collaboratif en termes de précision, de diversité et de temps de calcul. Nous avons validé le cadre proposé et les résultats obtenus grâce à lui en traitant hors-ligne des jeux de données réelles issues d'un grand magasin en ligne européen, La Redoute. Deuxièmement, nous proposons dans le chapitre II un cadre en cinq étapes destiné à développer et à évaluer des systèmes de recommandation hybrides qui analysent différentes sources de données, que nous validons à partir de données historiques réelles tirées du site de La Redoute. Troisièmement, le chapitre II introduit l'importance des caractéristiques dans la littérature sur les systèmes de recommandation. Quatrièmement, les algorithmes offrant les meilleurs résultats dans les tests hors-ligne sont utilisés dans le chapitre III afin de servir de base pour la création de deux systèmes de recommandation pour la maximisation du chiffre d'affaires. Cinquièmement, nous proposons, au chapitre III, un cadre pour étudier trois effets des systèmes de recommandation (pour la maximisation du chiffre d'affaires). Il ressort de ce cadre que les systèmes de recommandation exercent une influence positive sur les indicateurs de conversion tout au long du funnel d'achat. De plus, ce cadre suggère que l'inclusion du facteur « chiffre d'affaires » influence de manière positive la valeur de chaque commande. Par conséquent, la performance des systèmes à maximisation du chiffre d'affaires dépasse celle des systèmes de recommandation traditionnels en termes de chiffre d'affaires, au vu de la synergie entre le taux

de conversion et l'effet « valeur par commande ». Sixièmement, nous validons notre cadre par une expérience de terrain à grande échelle, en collaboration avec La Redoute. Enfin, une étude de cas montre que les machines à factorisation hybrides offrent le plus haut potentiel en termes de taux de conversion et de chiffre d'affaires. Si une machine à factorisation hybride traditionnelle donne un plus grand nombre de commandes, une machine à factorisation à maximisation du chiffre d'affaires offre un potentiel plus élevé en termes de chiffre d'affaires.

Mots-clés : *E-commerce, Systèmes de recommandation, Filtrage collaboratif, Données d'achat binaires, Hybridation ; Machines à factorisation ; Importance des caractéristiques, Maximisation des recettes ; Expérience de terrain*

GENERAL ABSTRACT

Recommendation systems are a heavily investigated subject within machine learning literature, resulting in the creation of many state-if-the-art algorithms. This doctoral dissertation goes beyond merely proposing new recommendation algorithms by leveraging state-of-the-art techniques and investigating the interaction of these techniques with different data sources having distinct characteristics. The focus lies upon the creation of frameworks guiding both marketing decision makers and academics in developing, evaluating, and testing recommendation systems in an e-commerce context. To create these frameworks, this dissertation starts by first investigating a specific algorithm in depth (collaborative filtering) in an offline setting on historical data and afterwards opening the scope to hybrid - and revenue maximization recommendation systems and field experiments. Concretely, this dissertation adds to literature in seven distinct ways. First, a framework evaluating the influence of input characteristics, of a binary purchase matrix, on the best collaborative filtering algorithm in terms of accuracy, diversity, and computation time is designed in Chapter I. The proposed framework and it findings are validated on real-life offline data sets of a large European e-tailer, La Redoute. Second, a five-step framework to develop and evaluate hybrid recommendation systems combing different data sources is proposed and validate on real-life historical data of La Redoute in Chapter II. Third, Chapter II introduces feature importance in the recommendation systems literature. Fourth, the best performing algorithms in the offline tests are leveraged to serve as basis for creating two revenue maximization recommendation systems in Chapter III. Fifth, a framework investigating three effects of (revenue maximization) recommendation systems is proposed in Chapter III. In this framework it is argued that recommendation systems have a positive influence on conversion business metrics throughout the purchase funnel. Additionally, the framework suggest that revenue inclusion positively influences value per order. Consequently, revenue maximization recommenders outperform traditional recommendation systems in terms of revenue, driven by synergy between conversion and value per order effect. Sixth, the framework is validated in a large-scale field experiment executed in collaboration with La Redoute. Finally, a business case shows that hybrid factorization machines have the highest potential in terms of conversion and revenue. A traditional hybrid factorization machine results in the highest number of orders and a revenue maximization factorization machine has the highest potential in terms of revenue.

Keywords: *E-commerce, Recommendation systems, Collaborative filtering, Binary purchase data, Hybridization; Factorization machines; Feature importance, Revenue maximization; Field experiment*

RÉSUMÉ DÉTAILLÉ

1 Description

Cette thèse est divisée en cinq chapitres. Après une introduction générale, les chapitres II à IV constituent le corps de cet ouvrage. Un cinquième et dernier chapitre conclut la thèse en suggérant des pistes pour de nouvelles recherches. Bien que les chapitres II à IV soient basés sur des recherches indépendantes les unes des autres, ils font partie d'un processus uni pour le développement de système de recommandation. Pour Gunawardana et Shani (2009), le processus de développement des systèmes de recommandation comprend trois étapes : le test hors-ligne, les expériences de terrain et l'obtention de résultats fiables. C'est à partir de cette définition du procédé que nous avons orienté notre thèse pour la création de systèmes de recommandation, se fondant essentiellement sur des sources de données implicites, dans le contexte du commerce en ligne B2C. Premièrement, le chapitre II considère le cas d'un algorithme de recommandation, CF, dans son traitement en profondeur d'une source de données implicites, les données d'achat, lors d'un test hors-ligne. Deuxièmement, le chapitre III incorpore les résultats du chapitre II et élargit son champ d'investigation à des systèmes de recommandation hybrides et à des sources de données multiples dans un environnement hors-ligne à partir de données historiques. Enfin, les systèmes de recommandation hybrides donnant les meilleures performances dans les tests hors-ligne sont repris au chapitre IV pour servir de base au développement de systèmes de recommandation avec maximisation du chiffre d'affaires. En outre, nous menons une expérience de terrain à grande échelle pour comparer en conditions réelles les performances des systèmes de recommandation traditionnels et de ceux à maximisation du chiffre d'affaires. Dans chaque chapitre, nous appliquons les tests statistiques qui s'imposent pour l'obtention de résultats fiables.

Nous avons adopté la même structure pour chacun des chapitres. Chaque chapitre reprend tout d'abord plusieurs questions de recherche qui n'ont pas encore été abordées dans la littérature avant de rapprocher contribution académique et pertinence pratique, posant ainsi les fondations de chaque étude. Les objectifs de recherche ayant été ainsi définis, une méthodologie est développée et présentée en tant que cadre pratique. Enfin, les résultats sont analysés et des conclusions sont formulées, ainsi que des implications pratiques.

2 Objectif global

Dans la littérature sur les systèmes de recommandations, beaucoup d'études se concentrent sur de nouvelles méthodes de développement d'algorithmes dans un contexte d'apprentissage automatique. Ce large corpus de recherche nous fournit un très grand nombre d'algorithmes extrêmement avancés. Tirant parti de ces algorithmes, notre thèse va au-delà du simple développement de nouveaux algorithmes ou du compte-rendu des dernières avancées en créant des cadres pratiques pour le développement et l'évaluation de systèmes de recommandation basés sur des sources de données multiples.

Le choix du meilleur algorithme et d'une configuration optimale est une tâche difficile pour les universitaires et plus encore pour les professionnels du marketing. Nous sommes convaincus que « le » bon algorithme qui conviendrait à chaque situation n'existe pas ; il convient en réalité de bien choisir son algorithme en fonction des sources de données disponibles et de leurs caractéristiques. Cette thèse vise à accompagner les universitaires et les professionnels dans leurs efforts de création de systèmes de recommandation pour le commerce en ligne. Concrètement, nous accomplissons cela en :

1. proposant un cadre décisionnel pratique permettant de choisir le meilleur algorithme de filtrage collaboratif en fonction de la caractéristique d'entrée d'une matrice d'entrée binaire et des indicateurs devant être optimisés (chapitre II) ;
2. validant les cadres proposés sur des données historiques (chapitre II) ;
3. proposant un cadre décisionnel pratique pour développer, évaluer et interpréter des systèmes de recommandation hybrides combinant différentes sources de données (chapitre III) ;
4. validant les cadres proposés sur des données historiques (chapitre III) ;
5. ouvrant la boite noire des systèmes de recommandations hybrides en introduisant l'importance des fonctionnalités dans la littérature sur les systèmes de recommandation (chapitre III) ;
6. concevant des systèmes de recommandation à maximisation du chiffre d'affaires (chapitre IV) ;

7. proposant un cadre pour identifier :

   a. l'effet des systèmes de recommandation sur les indicateurs de conversion tout au long du funnel d'achats (chapitre IV) ;
   b. l'effet de l'inclusion du facteur « chiffre d'affaires » sur la valeur par commande (chapitre IV) ;
   c. l'effet de la synergie de la conversion et de l'effet « valeur par commande » des systèmes de recommandations avec maximisation du chiffre d'affaires sur le chiffre d'affaires (chapitre IV) ;
8. menant une expérience de terrain et une étude de cas pour évaluer les résultats et valider le cadre proposé en termes d'indicateurs d'entreprise (chapitre IV).

3 Principaux résultats

Dans le reste de cette section, nous discutons tour à tour des principaux résultats de chaque chapitre.

Le *chapitre II* analyse l'effet des caractéristiques d'une matrice d'entrée binaire (telles que la rareté, le taux objet/utilisateur et la répartition des achats) sur la configuration optimale de l'algorithme CF, qui est caractérisé par une étape de réduction des données, une étape de méthode CF et une étape de mesure de la similarité. Les principaux résultats sont au nombre de trois. Premièrement, les indicateurs d'évaluation sont influencés par la configuration de l'algorithme. Concrètement, la précision et la diversité des recommandations générées sont influencées par la technique de réduction des données, la méthode CF et la mesure de similarité. Le temps de calcul dépend uniquement de la technique de réduction des données. Il a été observé que les méthodes exactes, la décomposition de la valeur singulière et l'analyse de la correspondance sont plus rapides comparées aux procédures itératives, à l'analyse du composant principal logistique et à la factorisation de la matrice non négative, puisqu'elles prennent plus de temps à converger. Deuxièmement, nous avons analysé l'influence des caractéristiques d'entrée sur la performance. L'algorithme qui offre les meilleures performances (un algorithme basé sur l'analyse de la correspondance, sur la CF à base objet, sur la similarité de cosinus ou de corrélation) en termes de précision reste le même quelles que soient les caractéristiques des données d'entrée. Cependant, pour la diversité et le temps de calcul, le modèle qui offre la meilleure performance varie en fonction des caractéristiques d'entrée. Troisièmement, la configuration d'algorithme optimale est influencée par les caractéristiques d'entrée. La précision dépend de la rareté, tandis que la diversité varie avec la

rareté, la distribution d'achats et le taux utilisateur/objet. Le temps de calcul n'est influencé que par la répartition des achats et le taux utilisateur/objet.

Le *Chapitre III* nous donne cinq principaux résultats. Premièrement, les données comportementales brutes et les données produit sont les sources de données les plus prédictives dans les systèmes de recommandation basés sur une seule source de données. Les données client sont la troisième source la plus importante et enfin, les données comportementales agrégées sont la contribution la moins prédictive. Deuxièmement, le fait de combiner différentes sources de données augmente la performance des systèmes de recommandation. Troisièmement, le gain de rendement obtenu par l'ajout de sources de données supplémentaires diminue au fur et à mesure qu'on ajoute des données. Quatrièmement, la combinaison de caractéristiques à partir d'une machine à factorisation est préférée par rapport à la pondération *a posteriori* en tant que technique d'hybridation pour combiner le nombre optimal de sources de données. Enfin, les scores d'importance à cette fin des sources de données et des caractéristiques individuelles suivent une tendance nette : les données comportementales brutes sont la source de données la plus importante (39,38 %), suivies par les données produit (33,52 %), les données client (26,56 %) et enfin, les données comportementales agrégées (8,43 %). En termes d'importance des caractéristiques individuelles, les caractéristiques des données comportementales brutes implicites sont très importantes. Les évaluations explicites sont notablement les moins importantes des caractéristiques des données comportementales brutes, essentiellement de par le fait que ces informations ne sont disponibles qu'en petites quantités. Les données d'achat sont également des informations importantes à collecter, surtout en ce qui concerne les données de division produit et de marque. Bien qu'elles ne soient pas aussi importantes, les caractéristiques des données client individuelles peuvent ajouter de la valeur aux systèmes de recommandation. Enfin, les caractéristiques des données comportementales agrégées ont relativement peu d'importance.

Le *chapitre IV* valide le cadre proposé en identifiant les effets des systèmes de recommandation (avec maximisation du chiffre d'affaires) à différentes étapes du funnel d'achat dans une expérience à grande échelle, dont les résultats sont au nombre de cinq. Tout d'abord, le chapitre IV montre que la personnalisation a un effet positif sur les indicateurs de conversion tout au long du funnel d'achat. Deuxièmement, il est démontré qu'un système de recommandation hybride basé sur un modèle prenant en compte une combinaison de différentes caractéristiques offre de meilleures performances en termes de taux de clics, de taux

de visites, de taux d'ajout au chariot et de taux de conversion qu'un système qui combine *a posteriori* une seule source de données ainsi que les systèmes de recommandation basés sur la mémoire. Troisièmement, l'inclusion du facteur « chiffre d'affaires » accroit la valeur par commande. Quatrièmement, les systèmes de recommandation avec maximisation du chiffre d'affaires offrent de meilleures performances que les systèmes traditionnels en termes de chiffre d'affaires à l'étape de la commande, étant donné la synergie entre la conversion et l'effet « valeur par commande ». Enfin, notre étude de cas a montré que, par rapport aux chiffres habituels de l'entreprise, le meilleur système de recommandation traditionnel engendre une hausse du nombre de commandes de 350 % pour la gamme de produits recommandés et de 9,58 % pour l'ensemble des produits, tandis que le meilleur système de recommandation à maximisation du chiffre d'affaires engendre quant à lui une hausse du chiffre d'affaires de 442 % pour la gamme de produits recommandés et de 14,62 % pour l'offre de produit complète comparé à la stratégie de recommandations actuelle de l'entreprise.

Référence

[1] A. Gunawardana, G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, 2009, J. Mach. Learn. Res., 10 2935-2962.

LIST OF PUBLICATIONS

This doctoral dissertation is based on three individual studies:

Geuens S., Coussement K., De Bock K.W., A framework for configuring collaborative filtering-based recommendations derived from purchase data, revise and resubmit at *European Journal of Operational Research*

- The preliminary results of this study were published in the proceedings of the Ph.D. session of *ECML/PKDD* (2014), Nancy, France
- The preliminary results of this study were presented at *ECML/PKDD* (2014), Nancy, France and *BAFI* (2015), Santiago, Chili

Geuens S., Coussement K., De Bock K.W., A Decision Support System to Evaluate Data Source Combinations and Feature Importance in E-Commerce Recommendations Systems, submitted to *Decision Support Systems*

- The preliminary results of this study were published in the proceedings of *RecSys* (2015), Wien, Austria
- The preliminary results of this study were presented at the doctoral symposium of *RecSys* (2015), Wien, Austria, *BAFI* (2015), Santiago, Chili, and *AMS World Marketing Congress* (2016), Paris, France

Geuens S., De Bock K.W., Coussement K., Boutelier A., The Effect of Revenue Maximization Recommendation Systems on the Purchase Funnel Metrics: A Field Experiment, working paper to be submitted to *Marketing Science* (Special issue: Marketing Science and field experiments)

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **1**
RÉSUMÉ GLOBAL **3**
GENERAL ABSTRACT **5**
RESUME DETAILLE **7**
1 DESCRIPTION 7
2 OBJECTIF GLOBAL 8
3 PRINCIPAUX RESULTATS 9
RÉFÉRENCE 11
LIST OF PUBLICATIONS **13**
TABLE OF CONTENTS **15**
CHAPTER I GENERAL INTRODUCTION **19**
1 BACKGROUND 21
2 OUTLINE 23
3 RESEARCH OBJECTIVES & RESEARCH QUESTIONS 24
4 MAIN FINDINGS 27
REFERENCES 28
CHAPTER II A FRAMEWORK FOR CONFIGURING COLLABORATIVE FILTERING-BASED RECOMMENDATIONS DERIVED FROM PURCHASE DATA **31**
ABSTRACT 33
1 INTRODUCTION 33
2 RELATED RESEARCH 36
3 EXPERIMENTAL SETUP 38
3.1 Input Data *39*
3.2 Collaborative Filtering Algorithms *40*
3.2.1 Data Reduction as a Preprocessing Step 40
3.2.2 CF Methods 40
3.2.3 Similarity Measure 41
3.3 Evaluation Metrics *42*
3.3.1 Accuracy 42
3.3.2 Diversity 43
3.3.3 Computation Time 43
4 RESULTS 43
4.1 RQ1. How Does CF Algorithm Configuration Affect Performance? *43*
4.2 RQ2. How Do Input Data Characteristics Influence the Optimal CF Configuration(s)? *45*
4.3 RQ3. How Sensitive Are the Optimal CF configuration(s) To Variations in the Input Data Characteristics? *47*
4.4 Empirical Validation *49*
5 DISCUSSION 51
6 CONCLUSIONS, LIMITATIONS, AND FURTHER WORK 53
REFERENCES 55
APPENDICES 58
Appendix A: Data Generation Process *58*
Step 1: Obtain correlation structure from real-life user-item matrix 58
Step 2: Determine user-item input matrix dimensions to set user-item ratio 59
Step 3: Determine the desired number of non-zero elements in user-item input matrix and item vectors to govern sparsity level and purchase distribution 59
Step 4: generate and discretize user-item matrix 63
Appendix B: Statistical results for RQ1 ..... *65*
Appendix C: Best performing models for each data set (RQ2) *66*
Appendix D: Statistical results for RQ3 ..... *68*
References *71*

CHAPTER III A DECISION SUPPORT SYSTEM TO EVALUATE RECOMMENDATIONS SYSTEMS COMBING MULTIPLE DATA SOURCES AND IDENTIFY FEATURE IMPORTANCE IN E-COMMERCE....73
ABSTRACT....75
1 INTRODUCTION....75
2 LITERATURE REVIEW....78
2.1 Recommendation Calculation....78
2.1.1 Single Data Source Algorithms....78
2.1.2 Hybrid Algorithms Combining Different Data Sources....80
2.2 Evaluation and Interpretation....80
2.2.1 Evaluation....80
2.2.2 Interpretation....81
3 EVALUATION OF DATA SOURCE COMBINATION AND INTERPRETATION OF FEATURE IMPORTANCE: AN EMPIRICAL DECISION FRAMEWORK....81
3.1 The Empirical Decision Framework....81
3.1.1 Step 1: Data Collection....82
3.1.2 Step 2: Recommendation Calculation....85
3.1.3 Step 3: Evaluation & Interpretation....87
3.1.4 Step 4: Deployment....89
3.2 Empirical Results....90
3.2.1 RQ1a. Do recommendation systems based on different single data sources differ in performance?....90
3.2.2 RQ1b. Does combining different data sources enhance predictive performance?....91
3.2.3 RQ1c. What is the optimal order in which to add data sources to a recommendation system?....91
3.2.4 RQ2. Which hybridization technique performs best for recommendation models with the optimal number of data sources?....92
3.2.5 3.RQ3. Which are the most important predictors in the best performing recommendation model?....93
4 CONCLUSION....95
REFERENCES....96
CHAPTER IV THE EFFECT OF REVENUE MAXIMIZATION RECOMMENDATION SYSTEMS ON THE PURCHASE FUNNEL METRICS: A FIELD EXPERIMENT 101
ABSTRACT....103
1 INTRODUCTION....103
2 FRAMEWORK AND RESEARCH QUESTIONS....106
3 RELATED RESEARCH....108
3.1 Algorithms....108
3.1.1 Traditional Hybrid Recommendation Systems....108
3.1.2 Revenue Maximization Recommendation Systems....109
3.2 Purchase funnel....111
3.3 Field Experiments in Recommendation Systems....112
4 FIELD EXPERIMENT....113
4.1 Setting....113
4.2 Data....114
4.3 Algorithms....114
4.3.1 Non-Personalized Recommendation Algorithm....115
4.3.2 Personalized Traditional Recommendation Algorithms....115
4.3.3 Personalized Revenue Maximization Recommendation Systems....116
4.4 Evaluation....118
5 RESULTS....119
5.1 Analysis....119
5.2 Click Through Stage....120
5.3 View Stage....120
5.4 Cart Addition Stage....121
5.5 Order Stage....121
6 DISCUSSION....122
7 BUSINESS CASE....123
8 CONCLUSIONS, LIMITATIONS AND FUTURE WORK....124
REFERENCES....125
APPENDIX....129

CHAPTER V GENERAL CONCLUSION & FURTHER WORK 131
1 GENERAL CONCLUSION 133
2 LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH 135
REFERENCES 139
CONCLUSIONS GÉNÉRALES 141
LIST OF TABLES 145
LIST OF FIGURES 147

CHAPTER I

GENERAL INTRODUCTION

CHAPTER I
GENERAL INTRODUCTION

1 Background

Throughout history, consumer society has outlived some important evolutions. Looking only at the 20th century a shift from customization over mass production to mass customization is observed. Concretely, in the beginning of the 20th century people went to a local craftsman, like for example a tailor or a shoemaker, who would fabricate a custom product. Starting from the second half of the 20th century, the standard shifted towards mass production of standardized products. Both forms of societies have advantages and disadvantage. In a mass production society products are fast and readily available at lower prices, whereas in customized societies products have a longer throughput time and are typically higher priced, but have a better fit with the consumer's needs.

At the end of the 20th century a shift towards a mass customization society, combining advantages of both customization and mass production, is observed [1]. Evolution in technology makes it possible to adapt mass production processes to more customized products in a lower throughput time and at a lower price. Additionally, globalization facilitates worldwide distribution and marketing of products. Globalization makes it possible for companies to be represented around the world with a wide range of products. From a consumer perspective, globalization leads to a large offering of mass production products marketed by a wide range of companies. Assuming a consumer is well-informed about the total offer, most probably a suitable product exists. As this item is standard product, the customer gets the advantages of both mass production and customization.

The level of available information about the global product offering has increased significantly with the surge of the internet and nowadays many companies are present on the World Wide Web with their own e-commerce website. Whilst these evolutions made mass customization possible, they initiated a novel problem, i.e. information overload [2]. Concretely, it is impossible for a consumer to scan the entire internet, or even a whole website to find the product perfectly fitting his needs, therefore a customer should be guided in his efforts to find customized product sets. A typical e-commerce website produces a lot of data about customers, products, and customer-product interactions. Data science allows to transform this collected data into information about the consumer and his needs. One of the

methods used by e-commerce companies to overcome the information overload and to guide customers in finding a suitable product are *recommendation systems* [3].

Recommendation systems are data science tools designed to recommend relevant product sets to customers being beneficial for both customers and companies. Recommendation systems aid the selection process of customers by narrowing down the options for each customer and helping them explore less obvious products [3]. Such recommendation systems increase customer satisfaction [4] and benefit the e-commerce company, because greater satisfaction leads to increased sales, revenue, and loyalty [5, 6].

Starting from the early years of recommendation systems' research, studies have focused on machine learning and the creation of state-of-the-art-algorithms [7]. Together with the rise of e-commerce, the interest in recommendation systems has grown in both the academic and business world. In academia this phenomenon has been shown by the large body of research in machine learning and increasing interest in other domains (operations research, information systems, marketing, etc.). The organization of an annual conference on recommendation systems (RecSys) and the shift towards the development of procedures for testing and evaluating recommendation systems in real-world contexts [8-10] explify the increasing interest in the academic world. In business a number of big players like Netflix.com, Amazon.com, iTunes, Last.fm, and Yahoo! contribute to the development of recommendation systems [11]. These companies are dedicated to research and have helped narrowing the gap between academics and business by publishing research papers [12, 13], organizing competitions, i.e. the Netflix Prize Contest [14] and Kaggle competitions [15], and making data sets publicly available [16].

This dissertation contributes to the reduction of the gap between academic research and business application. The Ph.D. project is a collaboration between Université Lille 1 and IESEG School of Management, two academic institutions and La Redoute, a large European e-commerce company specialized in apparel and home decoration. La Redoute offers their data as research source in exchange for the creation of operational recommendation systems. This collaboration drives innovative academic research that is applicable and deployable at a company level.

2 Outline

This dissertation is divided into five chapters. After a general introduction, Chapters II – IV construct the body of this work. A fifth and final chapter concludes the dissertation and gives some paths for further research. Even though the Chapters II - IV are based on independent studies, they are part of a unified process for recommender development. Gunawardana and Shani (2009) split the process of developing recommendation systems into three stages, i.e. offline testing, field experiments, and drawing reliable results. This dissertation takes the proposed process as guidance to create recommendation systems, mainly based on implicit data sources, in a B2C e-commerce context. In recommendation systems, implicit data sources refer to data sources containing customer actions. These actions, like purchases or view, are used as implicit proxies for preference of a customer.

First, Chapter II investigates one recommendation algorithm, CF, on one implicit data source, purchase data, in depth in an offline test. Second, Chapter III incorporates the results of Chapter II and opens the scope to recommendation systems combining multiple data sources, i.e. hybrid recommendation systems, in an offline setting on historical data. Finally, the best performing hybrid recommendation systems of the offline tests are leveraged in Chapter IV to serve as basis for revenue maximization recommendation systems. Additionally, a large-scaled field experiment is executed to evaluate traditional and revenue maximization recommendation systems in a real-life setting. In each chapter reliable results are drawn by applying the appropriate statistical tests. Table I.1 gives an overview of the input data, algorithms and validation data used in the chapters.

The within structure of each chapter is also comparable. First, each chapter proposes several research questions, not yet addressed in literature and relevant for both academics and practitioners, serving as a foundation for each study. Based on these research goals, a methodology, presented as a practical framework, is developed. Finally, the results are analyzed and conclusions together with practical implications are formulated.

Chapter	Title	Input Data	Algorithms	Validation Data
II	**A Framework for Configuring Collaborative Filtering-Based Recommendations Derived from Purchase Data**	Behavioral data: Purchase data	32 CF configurations	54 synthetic data sets 2 real-life offline data sets
III	**A Decision Support System to Evaluate Data Source Combinations and Feature Importance in E-Commerce Recommendations Systems**	Customer data Product data Behavioral data Aggregated behavioral data	A posteriori weighting of best performing algorithms in Chapter II (HHR) Factorisation machines (HFM)	8 real-life offline data sets
IV	**The Effect of Revenue Maximization Recommendation Systems on the Purchase Funnel Metrics: A Field Experiment**	Customer data Product data Behavioral data Aggregated behavioral data	Best performing HHR and HFM algorithms of Chapter III Revenue maximization variations of HHR and HFM	Email field experiment with four different waves

Table I.1: Input data, algorithms, and validation data used in the different chapters.

3 Research Objectives & Research Questions

In recommendation systems' literature a lot of studies have focused on novel algorithm development in a machine learning context. This extensive body of research has resulted in an enormous variety of state-of-the-art algorithms. This dissertation leverages these algorithms by going beyond the development of merely new algorithms and providing an overview of the state-of-the-art by creating practical frameworks for development and evaluation of recommendation systems.

Deciding on the optimal algorithm and its exact configuration to implement is a difficult task for academics and even more so for marketing practitioners. I believe there is no 'one size fits all algorithm' and a good algorithm choice dependents on the data sources available and their characteristics. This dissertation aims to guide academics and professionals in their efforts to create recommendation systems. In accordance with Gunawardana and Shani (2009), this dissertation exemplifies the staged development process of a recommendation system in which the first two chapters focus on offline testing, while Chapter IV evaluates results of Chapter II and mainly Chapter III in a field experiment.

Regardless this global structure, each chapter in this dissertation contributes in three distinct ways. First, *decision frameworks* guiding and improving the creation and evaluation of recommendation systems applicable in both academy and industry are constructed. Second, the *data* and more specific the available data sources and their characteristics are often taken for

granted in literature. This dissertation does not overlook this aspect and investigates the effect of data characteristics, data sources, and feature importance, as it is critical in practical implementations. Finally, this study evaluates and validates results on *real-life data sets* either in offline tests on historical behavioral data (Chapter II and III) or in a field experiment (Chapter IV) to increase external validity. The deployed validation procedures could be used as guidance for companies and further academic research. In the remainder of this section the research objectives and specific research questions are discussed for each chapter separately.

Chapter II focuses on the creation of a *framework* to guide marking scientists in creating collaborative filtering (CF) recommendation systems. CF is a popular and one of the most successful recommendation techniques [17]. The popularity of the techniques implies that a lot of different configurations have been developed during the last 15 years [7]. Even though CF is heavily studied, no clean overview or framework to identify the best CF algorithm exists. Identify THE best CF algorithm is too opportunistic as the optimal configuration is case dependent. Therefore this study investigates the interaction between *input data layout* and CF algorithm configuration to create a framework guiding academics and practitioners in identifying the best input characteristic – CF algorithm combination. Concretely the interaction between three input characteristics, i.e. sparsity, item-user ratio, and purchase distribution, and 30 CF algorithm configurations are evaluated in terms of three distinct metrics, i.e. accuracy, diversity, and computation time. The proposed framework allows e-commerce companies to decide on the optimal CF configuration as a function of their specific binary purchase data sets and desired (combination of) metrics to optimize. The reader also gains insight into the impact of changes in the input data set on the preferred algorithm configuration. Additionally, the proposed framework is tested on 54 synthetic data sets and *validated on two real-life historical data sets*. To concretize these research objectives, they are presented as topical research questions:

RQ1. How does CF algorithm configuration affect performance?

RQ2. How do input data characteristics influence the optimal CF configuration(s)?

RQ3. How sensitive are the optimal CF configuration(s) to variations in the input data characteristics?

Chapter III goes beyond CF and introduces hybrid recommendation systems in this dissertation. The focus of Chapter III lies on data source combination as this strategy overcomes issues related to single data source algorithms and improves accuracy [18]. In total

three main objectives are identified. First, this chapter designs a *framework* guiding academics and practitioners in developing and evaluating recommendation systems with *multiple data sources* as input. Second, the *feature importance* scoring framework developed by Breiman (2000) is introduced in a recommendation setting to open the recommender system's black box. Third, the impact of hybridization strategies – i.e. a posteriori weighting and feature combination [18] – applied to four distinct data sources i.e. product data, customer data, raw behavioral data, and aggregated behavioral data is empirically *validated using eight real-life historical data sets* offered by a large European e-commerce company. Specific research questions proposed in this chapter are:

RQ1a. Do recommendation systems based on different single data sources differ in performance?

RQ1b. Does combining different data sources enhance predictive performance?

RQ1c. What is the optimal order in which to add data source groups to a recommendation system?

RQ2. Which hybridization technique performs best for recommendation models with the optimal number of data sources?

RQ3. Which are the most important predictors in the best performing recommendation model?

Chapter IV has four main objectives. First, Chapter IV leverages the best performing hybrid recommendation systems proposed in Chapter III by using them as basis to design revenue maximization recommendation systems. Second, a framework identifying three effect of traditional recommendation systems and revenue maximization recommenders on business metrics is proposed. This framework argues that both traditional - and revenue maximization recommendation systems influence conversion metrics in every stage of the purchase funnel, i.e. click through -, view -, cart addition -, conversion rate. Additionally, the framework suggests that revenue maximization recommendation systems have an effect on value per order as revenue inclusion drives a value effect. Consequently, it is argued that revenue maximization recommendation systems outperform traditional recommenders in terms of revenue in the order stage as these systems are driven by both conversion and value effect. Third, a large-scaled field experiment executed in collaboration with La Redoute validates the proposed framework. Finally, a business case demonstrates the added value of recommendation systems in terms of numbers of orders and revenue. The validity of proposed framework is investigated by answering the following research questions:

RQ1a. Is there an effect of personalization on conversion metrics throughout the purchase funnel?

RQ1b. Is there an effect of hybridization method on conversion metrics throughout the purchase funnel?

RQ2. Is there an effect of revenue inclusion on value (value per order)?

RQ3a. Is there an effect of personalization on revenue (value per visit)?

RQ3b. Is there an effect of hybridization method on revenue (value per visit)?

RQ3c. Is there an effect of revenue inclusion on revenue (value per visit)?

4 Main Findings

The main findings of each chapter are separately discussed in the remainder of this section.

Chapter II analyses the effect of characteristics of a binary input matrix, i.e. sparsity, item-user ratio, and purchase distribution on the optimal CF algorithm configuration, which is characterized by a data reduction step, a CF-method step, and a similarity measure step. The main findings are threefold. First, evaluation metrics are influenced by algorithm configuration (RQ1). Concretely, accuracy and diversity of the generated recommendations are influenced by data reduction technique, CF-method, and similarity measure. Computation time depends only on data reduction technique. It is observed that exact methods, singular value decomposition and correspondence analysis, are faster compared to iterative procedures, logistic principal component analysis and non-negative matrix factorization, as they take longer to converge. Second, the influence of input characteristics on performance is analyzed (RQ2). The best-performing algorithm (algorithm based on correspondence analysis, item-based CF, and cosine or correlation similarity) in terms of accuracy remains consistent regardless of the input-data characteristics. However, for diversity and computation time, the best-performing model varies with the input characteristics. Third, the optimal algorithm configuration is influenced by input characteristics (RQ3). Accuracy depends on sparsity, while diversity varies with sparsity, purchase distribution, and user-item ratio. Computation time is only influenced by purchase distribution and user-item ratio.

Chapter III has five main finding. First, raw behavioral data and product data are the most predictive sources in a single data source recommendation systems. Customer data is the third most important source and finally aggregated behavioral data has the least predictive contribution (RQ1a). Second, combining different data sources increases the performance of a recommendation systems (RQ1b). Third, the incremental return of adding additional data

sources diminishes (RQ1c). Fourth, factorization machine based feature combination is preferred over a posteriori weighting as hybridization technique for combining the optimal number of data sources (RQ2). Finally, the importance scores of both data sources and individual features therein are showing a clear pattern. Raw behavioral data is the most important data source (39.38%) followed by product data (33.52%), customer data (26.56%), and finally aggregated behavioral data (8.43%) (RQ3). In terms of the importance of individual features, implicit RBD features are very important. Explicit ratings are notably the least important RBD features, mainly because this information is only available in smaller amounts. Furthermore, PD is important information to gather, especially product division and brand data. Although somewhat less important, individual CD features can add value to recommendation systems. Finally, ABD features have relatively little importance.

Chapter IV validates the proposed framework identifying effects of (revenue maximization) recommendation systems in different stages of the purchase funnel in a large-scaled field experiment and results are fivefold. First, Chapter IV finds that personalization has a positive effect on conversion business metrics throughout the entire purchase funnel (RQ1a). Second, it is demonstrated that a hybrid, feature combination, model-based recommendation system outperforms a recommender a posteriori combining single data sources, memory-based recommendation systems in terms of click through rate, view rate, cart addition rate, and conversion rate (RQ1b). Third, revenue inclusion increases the value per order (RQ2). Fourth, revenue maximization recommendation systems outperform traditional recommenders in terms of revenue in the order stage due to synergy between the conversion and value per order effect (RQ3a – RQ3c). Finally, a business case shows that the best performing traditional recommendation system results in an increase in the number of orders with 350% for the set of recommended items and 9.58% for the all products compared to the company benchmark. Additionally, the business case indicates that the optimal revenue maximization recommender generates an increase in revenue of 442% for the set of recommended products and 14.62% for the complete product offering compared the current company's recommendation strategy.

References

[1] B.J. Pine, S. Davis, Mass Customization: The New Frontier in Business Competition, 1999, Harvard Business School Press

[2] J.B. Schafer, J. Konstan, J. Riedl, Recommender systems in e-commerce, 1999, Proceedings of the 1st ACM conference on Electronic commerce, ACM, Denver, Colorado, USA, pp. 158-166.

[3] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Recommender systems: An introduction, 2010, Cambridge University Press

[4] Y. Jiang, J. Shang, Y. Liu, Maximizing customer satisfaction through an online recommendation system: A novel associative classification model, 2010, Decis. Support Syst., 48 470-479.

[5] Asim Ansari, C.F. Mela, E-Customization, 2003, J. Marketing Res., 40 131-145.

[6] V.Y. Yoon, R.E. Hostler, Z. Guo, T. Guimaraes, Assessing the moderating effect of consumer product knowledge and online shopping experience on using recommendation agents for customer loyalty, 2013, Decis. Support Syst., 55 883-893.

[7] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey, 2013, Knowl.-Based Syst., 46 109-132.

[8] R. Kohavi, R. Longbotham, D. Sommerfield, R.M. Henne, Controlled experiments on the web: survey and practical guide, 2009, Data Mining and Knowledge Discovery, 18 140-181.

[9] A. Gunawardana, G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, 2009, J. Mach. Learn. Res., 10 2935-2962.

[10] U. Panniello, M. Gorgoglione, A. Tuzhilin, Research note—In CARSs we trust: How context-aware recommendations affect customers' Trust and other business performance measures of recommender systems, 2016, Inform. Syst. Res., 27 182-196.

[11] N. Sahoo, R. Krishnan, G. Duncan, J. Callan, On Multi-component Rating and Collaborative Filtering for Recommender Systems: The Case of Yahoo! Movies, 2012, Inform. Syst. Res., 23 231-246.

[12] G. Linden, B. Smith, J. York, Amazon.com Recommendations: Item-to-Item Collaborative Filtering, 2003, IEEE Internet Comput., 7 76-80.

[13] Y. Koren, R. Bell, C. Volinsky, Matrix Factorization Techniques for Recommender Systems, 2009, IEEE Comput., 42 30-37.

[14] Netflix, Netflix Prize, 2009, http://www.netflixprize.com/.

[15] Kaggle, Kaggle: Your Home for Data Science, 2016, https://www.kaggle.com/.

[16] GroupLens, GroupLens Datasets, 2016, http://grouplens.org/datasets/.

[17] E. Turban, D. King, K.J. Lee, T.-P. Liang, D.C. Turban, Electronic Commerce: A Managerial and Social Networks Perspective, 2015, 8 ed., Springer

[18] R. Burke, Hybrid recommender systems: Survey and experiments, 2002, User Modeling and User-Adapted Interaction, 12 331-370.

[19] L. Breiman, Random forests, 2001, Mach. Learn., 45 5-32.

CHAPTER II

A FRAMEWORK FOR CONFIGURING COLLABORATIVE FILTERING-BASED RECOMMENDATIONS DERIVED FROM PURCHASE DATA

CHAPTER II

A FRAMEWORK FOR CONFIGURING COLLABORATIVE FILTERING-BASED RECOMMENDATIONS DERIVED FROM PURCHASE DATA

Abstract

This study proposes a decision support framework to help e-commerce companies select the best collaborative filtering algorithms (CF) for generating recommendations on the basis of online binary purchase data. To create this framework, an experimental design applies several CF configurations, which are characterized by different data-reduction techniques, CF methods, and similarity measures, to binary purchase data sets with distinct input data characteristics, i.e., sparsity level, purchase distribution, and item-user ratio. The evaluations in terms of accuracy, diversity, computation time, and trade-offs among these metrics reveal that the best-performing algorithm in terms of accuracy remains consistent regardless of the input-data characteristics. However, for diversity and computation time, the best-performing model varies with the input characteristics. This framework allows e-commerce companies to decide on the optimal CF configuration as a function of their specific binary purchase data sets. They also gain insight into the impact of changes in the input data set on the preferred algorithm configuration.

Keywords: E-commerce, Recommendation systems, Collaborative filtering, Binary purchase data, Data generation

1 Introduction

In a typical e-commerce setting, a customer receives an abundance of product-related information. Consider, for example, a customer shopping for a pair of pants. A web shop typically contains hundreds of pairs, making it impossible for the customer to scan every product and make an optimal product evaluation and the best purchase decision. To help visitors cope with information overload, websites often feature recommendation systems that create personalized product sets for each customer [1]. This personalized selection facilitates the choice process and leads to greater satisfaction [2]. E-commerce also benefits from these systems because greater satisfaction leads to increased sales, revenue, and loyalty [3, 4].

Neighborhood-based collaborative filtering (CF) is one of the most widely used algorithms in recommendation systems [5, 6] in both literature and industry. A number of recent studies in literature investigate CF [7-9]. Rich Relevance, Certona [10], Predictad, Bommerang, Criteo, and Coversant Media [11] are some providers of commercial recommendation engines heavily relying on CF. Additionally other companies also deploy CF as recommendation tool i.e. Netflix.com, Amazon.com, iTunes, Last.fm, and StoryCode.com [12]. The exclusive use of the user-item matrix as input is an important advantage of this technique in that no data beyond actual customer behavior are needed to produce the recommendations [13]. Accordingly, this study focuses on neighbourhood-based CF, which is referred to as CF for the remainder of this paper.

The performance of recommendation systems is influenced by various characteristics of the input matrix and the algorithm configuration. In terms of input characteristics, sparsity, purchase distribution, and item-user ratio influence the final recommendation and are three important input characteristics to investigate. First, highly sparse data weakens the performance of CF algorithms because less information is available for calculating similarities [6, 14, 15]. Second, the purchase distribution influences the recommendation system's performance because CF algorithms tend to be biased towards recommending more popular products. This is because these items have more historical data [14, 16]. This "long-tail problem" may lead to overspecialization. Finally, in terms of the item-user ratio, many CF algorithms are designed to make recommendations in a setting in which the number of users is high compared to the number of items, which generally results in better performance of item-based methods when compared to user-based techniques [17].

Next to input data characteristics, the algorithm's configuration affects the recommendation results. CF algorithms consist of different steps, each of which affects the final recommendations. Sarwar et al. (2000) define a CF procedure that divides the recommendation framework into three distinct steps: data reduction as preprocessing, CF method, and similarity measures. In the data reduction step, a dimension reduction method is applied to the initial input matrix. This step is often used in the preprocessing phase of the algorithm's building process, and it results in a smaller, denser matrix [19]. The reduced size might positively influence the model's efficiency, while the lower sparsity level can have a positive impact on the model's accuracy. Second, with regard to the CF method, two possible distinctions exist. An algorithm is either user-based [18] or item-based [6, 20]. The former calculates similarity between customers, and products are proposed on the basis of the behavior of a user's nearest

neighbors [13]. In contrast, an item-based system calculates similarity between products and proposes items that are similar to those purchased by the customer. In the third step, identification of nearest neighbors requires calculating similarity. Many such similarity measures appear in the extant literature [21]. Notably, the effect of this step on performance has exhibited less significance in the literature. Concretely, Breese et al. (1998) evaluate mutiple similarity measures, amongst which correlation and cosine similarity, on multiple datasets. They find that the best performing similarity measure is not consistant and depends on the dataset.

Most studies evaluate recommendation systems based on the accuracy of their recommendations, but other important metrics exist as well [17]. E-commerce companies need recommendation systems that offer not only good accuracy but also a certain level of diversity. As indicated by [22], accuracy and diversity are often a trade-off because systems optimizing accuracy tend to propose popular products while disregarding diversity, thereby leading to overspecialization. In addition to diversity, the algorithm's computation time is important [17]. In an e-commerce setting, the volume and velocity of data are very high. Therefore, it is important to be able to calculate recommendations in a reasonable time span. Whether the computation time is reasonable depends on the application. When, for example, a system is used to send personalized e-mails once a day, its efficiency is less important than when it is used to provide real-time recommendations on an e-commerce site[1].

Previous studies have not addressed the combined effect of input characteristics and algorithm configurations on the evaluation of recommendation systems based on binary purchase data. Therefore, this study complements the extant literature by simultaneously combining different configuration levels of input characteristics—sparsity, purchase distribution, and item-user ratio—with different CF algorithm configurations, identified by data reduction technique, the CF method, and similarity in terms of accuracy, diversity, and computation time. Our goal is to guide e-commerce companies to select optimal CF algorithm configurations based on the characteristics of their binary purchase data sets and (the combination of) the most important evaluation metrics. To create this framework, an experimental design is proposed, aimed at analyzing the impacts of three input characteristics

[1] Note that profitability is also an important evaluation metric. Due to the offline character of the tests executed in Chapter II, this metric is not tested in this chapter. Profitability is evaluated in detail in Chapter IV.

of a purchase data set on the optimal algorithm configuration. 54 experimental and two real-life validation data sets are included as basis for the analysis. The three central research questions are:

RQ1. How does CF algorithm configuration affect performance?

RQ2. How do input data characteristics influence the optimal CF configuration(s)?

RQ3. How sensitive are the optimal CF configuration(s) to variations in the input data characteristics?

Results show that the optimal CF algorithm (RQ1) is stable across input characteristics (RQ2) in terms of accuracy, but the absolute level of the metric depends on the data characteristics (RQ3). The overall most diverse model is affected by the purchase distribution. Changes in input characteristics will alter the level of diversity. Finally, the overall most time-efficient models are influenced by the item-user ratio.

The next section details related research and the value added by this study. After presenting the experimental setup, as well as the characteristics of the input data sets and algorithms, Section 3 formulates the evaluation metrics for analysis. Section 4 contains the results of the study. Finally, this article concludes with a discussion, some limitations, and suggestions for future research.

2 Related Research

CF recommendation systems transform historical user data into a relevant, personalized product offering, often based on *explicit customer ratings* [23], as exemplified by Amazon.com [24], Netflix.com [25], or Movielens [17]. These systems base their product propositions on product ratings explicitly provided by customers on a specified rating scale. Although they clearly represent customers' preferences, the systems demand user effort, time, and cost [19]. Moreover, in online retail settings that are characterized by broad, deep, and fast-changing product offerings, such feedback is often hard to collect in sufficient amounts.

Additionally, [26] identify two critical flaws in explicit ratings. First, explicit data tends to be biased because customers have difficulties expressing their preference, which could lead to arbitrary or incorrect ratings. Second, customers tend to rate only a small fraction of the products they purchase, resulting in a partial view of their preferences and, therefore, sparse data sets. Awarding rating incentives to users is a possible strategy to collect more data [27]. While this strategy results in an increased volume of data gathering, the quality is not

guaranteed as users tend to merely rate items to receive the extrinsic incentive without an intrinsic interest in the rating task [28]. Additionally, this strategy is relatively expensive as a (monetary) incentive needs to be given.

An alternative strategy to overcome the problems related to explicit rating is the use of *implicit ratings* [26]. In contrast to explicit ratings, implicit information does not require direct user feedback, but derive input from user behavior. Specifically, user actions like purchases, views, and additions to cart are considered as implicit expressions of preference. The collection of implicit feedback is objective and non-intrusive, and this form of data is readily available in customer databases [26]. Therefore implicit feedback is more likely to be collected in sufficient amount at low cost to construct a reliable recommendation systems. Despite these important advantages, implicit data only assume a customer preference, whereas explicit ratings are a more direct expression of preference.

A specific form of implicit information is binary purchase data [15]. In this study, information on customers' past purchase behavior, which is collected in large e-commerce logs, is used for product predictions that are likely to fit customers' profiles.

Most research into binary purchase e-commerce settings compares CF configurations with other (newly developed) algorithms. The techniques compared to CF include the popular method [29], association rules [18, 20], Bayesian models [15, 30], graph theory [31], and model-based CF methods, such as matrix factorization [20, 25] or support vector machines [32]. CF methods applied to binary purchase data have taken various forms in the past, as shown in Table II.1.[2]

In brief, Table II.1 offers five notable observations. First, the effects of *controlling for binary purchase characteristics* as experimental factors with different levels remain under-investigated. The only exception is sparsity, which Sarwar et al. (2000) included as a experimental factor in their study. Second, the same conclusion applies to *reduction techniques as preprocessing*. Only [18] includes this step in their research, for which they use singular value decomposition (SVD). Third, item- and user-based *CF methods* have been studied thoroughly, but comparisons of these two CF methods in a binary purchase setting are limited.

[2] Section 3 offers a more in-depth discussion of the different binary purchase characteristics, algorithm configurations, and evaluation metrics.

	Binary Purchase Characteristics Controlled	Reduction as Preprocessing	CF Method	Similarity Measure	Evaluation Metric
Breese et al. (1998)	None	None	User-based	Cosine Correlation	Accuracy
Li et al. (2010)	None	None	User-based	Cosine	Accuracy
Deshpande and Karypis (2004)	None	None	Item-based	Cosine	Accuracy
Linden et al. (2003)	None	None	Item-based	Cosine	Accuracy
Pradel et al. (2011)	None	None	Item-based	Cosine	Accuracy
Sarwar et al. (2000a)	Sparsity	None SVD	User-based	Cosine	Accuracy
Current study	Sparsity **Purchase distribution** **Item/user ratio**	None SVD **CA** **NMF** **LPCA**	User-based Item-based	Cosine Correlation **Jaccard**	Accuracy **Diversity** **Computation Time**

Note: Bolded items indicate novel considerations by the present study.

Table II.1: Previous studies incorporating CF algorithms in a binary purchase setting

Fourth, cosines and Pearson correlations are standard *similarity measures* in a binary purchase context, while a Jaccard similarity measure is less common. Fifth, accuracy is the most popular *evaluation metric*.

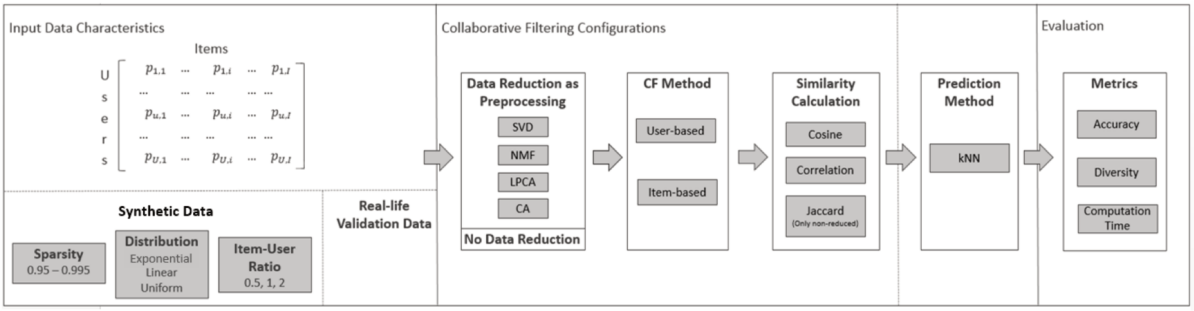
This study seeks to extend existing literature by simultaneously controlling for three input characteristics (i.e., sparsity level, purchase distribution, and item-user ratio) and implementing different algorithm configurations. More specifically, the current study implements four reduction techniques as preprocessing step—SVD, correspondence analysis (CA), nonnegative matrix factorization (NMF), and logistic principal component analysis (LPCA)—and offers comparisons of item- and user-based CF and cosine, Pearson correlation, and Jaccard similarity. Finally, diversity and computation time are incorporated as evaluation metrics, in addition to accuracy.

3 Experimental Setup

The experimental design focuses on answering the three research questions posed in the introduction. It investigates the impact of two aspects of the recommendation process: the input-data structure of the binary purchase matrix and the recommendation systems' CF algorithms. To analyze the link between the input characteristics and algorithm configurations, this study relies on a $5 \times 2 \times 3$ between-subjects experimental design in which the experimental conditions are the algorithm configurations. All algorithm configurations are tested on 54 synthetic data sets with different input characteristics and validated using two real-life data

sets. The results and a comparison of the algorithms allow us to identify the best algorithm combinations for given input characteristics. Figure II.1 depicts the experimental design.

Figure II.1: Overview of the experimental design



3.1 Input Data

In general, recommendation systems suffer from problems related to the input characteristics of the binary purchase matrix [33]. This study investigates the possible influence of controlling for the three input characteristics—sparsity, purchase distribution, and item–user ratio—in terms of the effect on the proposed evaluation metrics. Carefully designed synthetic data sets, with explicitly simulated data input characteristics, are a good basis to analyze the effect of input characteristics, as we are able to control the desired input configurations to guarantee internal validity. This study analyses the effect of input configuration on synthetic data sets in a fist stage, and validates the results using two real-life data sets in a second stage to increase external validity [34].

Typically, an input matrix is *sparse* because each customer buys a limited number of products and each product is purchased by a limited number of customers. For this study, the sparsity levels of the synthetic data sets are artificially set to mimic real-life situations based on figures of publically available data sets, like Netflix, EachMovie, LastFM, and MovieLens. As a typical recommendation setting consists of very sparse input matrices [6, 35], higher sparsity levels are more realistic. By implementing levels of 95%, 96%, 97%, 98%, 99%, and 99.5%, realistic settings are created.

The second input characteristic is the *distribution of purchases* [14]. Although some items may be very popular, most products are bought only a few times. CF tends to be less accurate towards the long tail and to promote only the most popular products, which leads to overspecialization [16]. To vary input, this study uses three purchase distributions: an exponential distribution that mimics the long tail, a linear distribution with a moderate tail, and

a uniform distribution. Whereas the first distribution is the most realistic, the two latter distributions relax the long-tail assumption, thereby allowing us to investigate the impact on the algorithm's performance.

The third factor influencing CF performance is the *item-user ratio*. Most applications utilize a very low item-user ratio. This study varies the item-user ratio by adjusting the number of rows in the binary input matrix [36]. The synthetic data sets consist of 1,000 items. Setting the number of customers equal to 500, 1,000, and 2,000 produces variations in the item-user ratio equal to 2, 1, and 1/2, respectively.

The combinations of these three input characteristics create 54 synthetic data sets, which are generated based on a four step process. First, the correlation structure is determined. Second, the optimal user-item ratio is designed by controlling for the number of users in the input matrix. Third, the purchase distribution and overall sparsity of the input matrix is simultaneously set. Finally, the resulting synthetic user-item matrix is generated and discretized. The detailed data generation process is discussed in Appendix A.

3.2 Collaborative Filtering Algorithms

As described by [18], the CF algorithms are characterized by variations in three steps: data reduction technique as preprocessing, CF method, and similarity measure.

3.2.1 Data Reduction as a Preprocessing Step

Four data reduction techniques have been identified for the preprocessing phase of the algorithm building process. Three popular techniques are singular value decomposition (SVD) [19, 25], nonnegative matrix factorization (NMF) [37], and logistic principal component analysis (LPCA) [38]. A fourth technique, correspondence analysis (CA), is also used. The latter method is conceptually similar to principal component analysis but can be only applied to binary data [39]. To the best of our knowledge, CA has never been benchmarked in recommendation settings.

3.2.2 CF Methods

This study implements and compares user- and item-based CF. In a user-based system, similarities between users are calculated to predict $\hat{p}_{ki}$ or the purchase probability for item *i* for user *k*, while item-based CF algorithms consider the similarity between items to calculate $\hat{p}_{ki}$. Mathematically, for a user-based CF approach [18],

$$\hat{p}_{ki} = \frac{1}{|L|} \sum_{l \epsilon L} sim(k,l)\, x_{li}\,, \tag{1}$$

where L is the set of nearest neighbors or users most similar to user k who purchased item i; *sim(k,l)* is the similarity between users k and l, of whom the latter is in the neighborhood of user k; x_{li} is a binary variable that indicates whether user l purchased item i (i.e. x_{li}= 1) or not $(i.e. x_{li} = 0)$.

For an item-based approach [6],

$$\hat{p}_{ki} = \frac{1}{|J|} \sum_{j \epsilon J} sim(i,j)\, x_{kj}\,, \tag{2}$$

where J is the set of nearest neighbors or items similar to item i; *sim(i,j)* is the similarity between items i and j of which the latter is an item in the neighborhood of item i; x_{kj} is a binary variable that indicates whether user k purchased item j (i.e. x_{kj}= 1) or not $\left(i.e. x_{kj} = 0\right)$.

3.2.3 Similarity Measure

This study uses the cosine, Pearson correlation, and the Jaccard similarity to calculate *sim(k, l)* and sim(i, j) (see Equation (1) and (2)). Mathematically, for a user-based approach, the cosine similarity is defined as

$$sim(k,l) = \frac{\sum_{z \in Z_{kl}} x_{kz}\ x_{lz}}{\sqrt{\sum_{z \in Z_k} x_{kz}^2}\ \sqrt{\sum_{z \in Z_l} x_{lz}^2}}, \tag{3}$$

the Pearson correlation similarity [15] equals

$$sim(k,l) = \frac{\sum_{z \in Z_{kl}} (x_{kz} - \overline{x_k})\ (x_{lz} - \overline{x_l})}{\sqrt{\sum_{z \in Z_{kl}} (x_{kz} - \overline{x_k})^2\ \sum_{z \in Z_{kl}} (x_{lz} - \overline{x_l})^2}}, \tag{4}$$

and the Jaccard similarity [40] is defined as

$$sim(k,l) = \frac{z_{kl}}{z_k \cup z_l}, \quad (5)$$

where the input matrix is represented as a set of user vectors having a length equal to the number of items in the data set; x_{kz} and x_{lz} are the binary purchase indicators of item z by users k and l; $\overline{x_k}$ and $\overline{x_l}$ are the mean purchase rates of these users; Z_k is the set of items bought by user k; Z_l is the set of items bought by user l; Z_{kl} is the set of items bought simultaneously by users k and l. Equivalent expressions for cosine, Pearson correlation, and the Jaccard similarity between items i and j, *in an item-based setting,* can be easily derived.

Note that the Jaccard similarity measure is only applicable on binary purchase vectors. Given that during data preprocessing, the data reduction step transforms the input matrix to a non-binary matrix, the Jaccard similarity is only applied to the non-reduced input matrix.

3.3 Evaluation Metrics

To assess the results of different CF configurations given distinct binary input purchase data sets, this study considers several performance metrics: accuracy, diversity, and computation time. The calculation of the metrics involves a randomly drawn test sample consisting of 20% of the input data set [18, 20, 32].

3.3.1 Accuracy

In this study accuracy is evaluated by means of ranked classification accuracy. Specifically, for every user a top-N of recommendations, consisting of 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, and 200 items is constructed by sorting the items descending based on recommendation score. The items in the top-N are considered as true prediction, while all other products are considered negative predictions. The results of the predictions for each recommendation size (*RS*) are based on the confusion matrix and expressed in terms of the *F*1 measure [17]. *F*1 is chosen as evaluation because it is described as the harmonic mean of recall, and precision [41], consequently giving an indication for both completeness and exactness. Additionally, the true negative element (tn) of the confusion matrix is not a part of the formula to calculate *F*1. This is important because the binary input matrix is very sparse and consequently tn would be very high, distorting the evaluation metric. Equation 6 shows the formula to calculate *F*1:

$$F1_{RS} = 2\frac{precision\ recall}{precision+recall} = 2\frac{\left({}^{tp}/_{tp+fp}\right)\left({}^{tp}/_{tp+fn}\right)}{\left({}^{tp}/_{tp+fp}\right)+\left({}^{tp}/_{tp+fn}\right)}, \quad (6)$$

where *tp* is the number of recommended items that have been purchased by the user; *fp* is the number of recommended items that have not been purchased by the user; *fn* is the number of items not recommended that have been purchased by the user.

3.3.2 Diversity

A key characteristic of a recommendation system is the degree of diversity offered to customers [22]. Depending on the situation, the set of predicted items should either be very similar or more diverse. Intra-list similarity (ILS) indicates the similarity among a set of offered products, resulting in a (negative) measure for the diversity of a recommendation system [42]. This metric, calculated for each customer or averaged over all clients, indicates the diversity for a fixed number of recommended items:

$$ILS_{RS} = 0.5 \sum_{i \in Rec} \sum_{j \in Rec, j \neq i} sim(i,j), \quad (7)$$

where *RS* indicates the recommendation size, which varies between 5 and 200; *Rec* indicates the set of recommended items; and *sim(i, j)* indicates the similarity between two items *i* and *j* in the recommended list. The cosine similarity measure is used to calculate ILS.

3.3.3 Computation Time

Algorithms have three important computation time components: reduction, similarity calculation, and prediction time. This study analyzes the computation time of each algorithm configuration in each data set.

4 Results

4.1 RQ1. How Does CF Algorithm Configuration Affect Performance?

The answer to RQ1 involves estimating the following analysis of covariance (ANCOVA) model:

$$Metric_i = \beta_{0i} + \beta_{1i}RM + \beta_{2i}CFM + \beta_{3i}SM + \beta_{4i}RS + \varepsilon_i, \quad (8)$$
$$where\ i \in \{F1,\ ILS,\ Computation\ Time\},$$

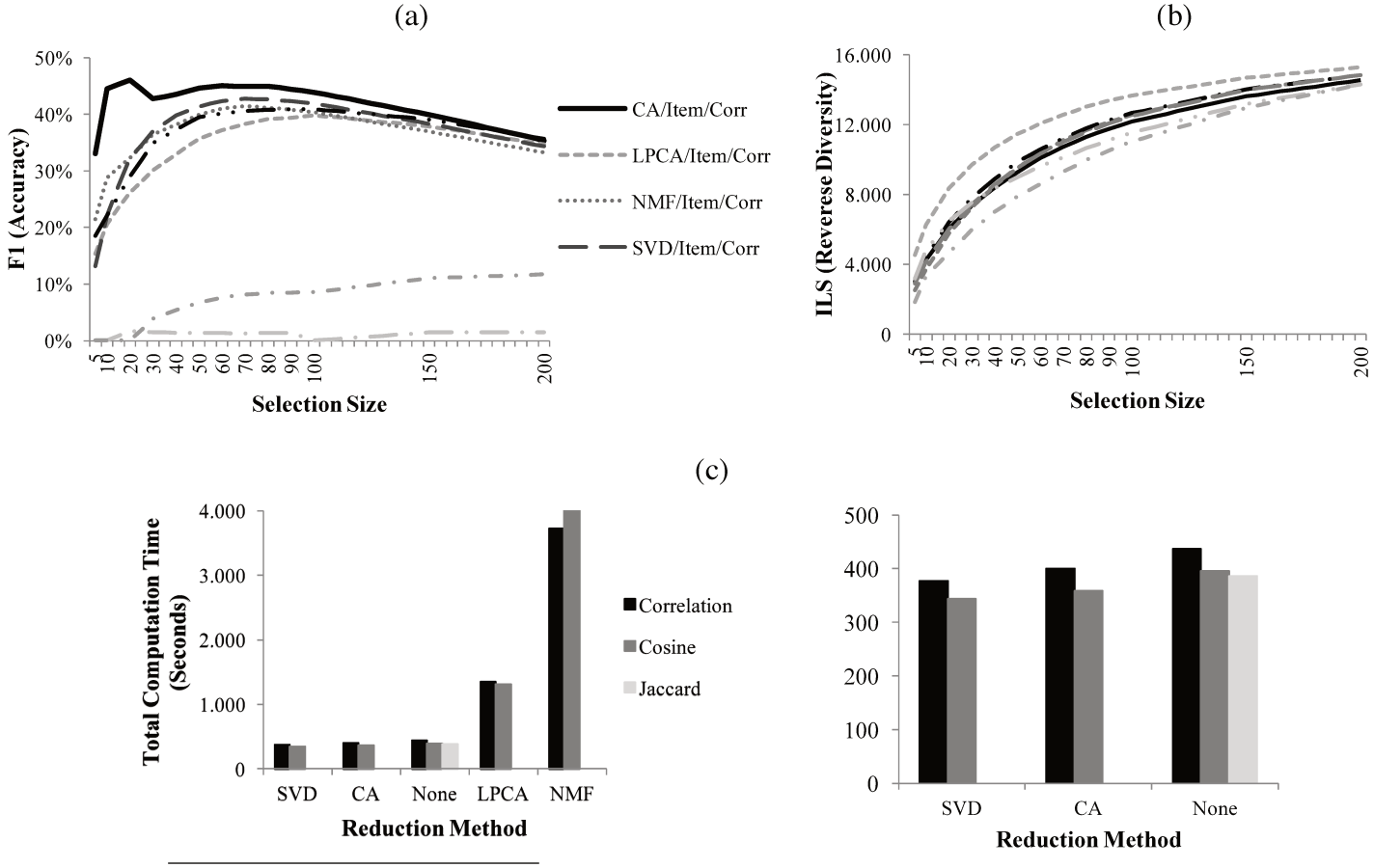
CF method and cosine similarity as similarity measure. The reduction method, CF method, and to determine the main effects of the reduction technique (*RM*), CF method (*CFM*), and similarity measure (*SM*). Recommendation size (*RS*) is a covariate that controls for different Top-N selections.[3] The statistical results of the AN(C)OVA analysis are presented in Table II.B.1 of Appendix B. Figure II.2 and Table II.2 briefly summarize the findings. he pairwise t-

Evaluation Metric	Data Reduction	CF Method	Similarity Measure
Accuracy	CA > NMF, SVD > LPCA > None[1]	Item > User	Cos, Corr > Jaccard
Diversity	CA, None, NMF, SVD > LPCA	User > Item	Jaccard > Cos, Corr
Time	SVD, CA, None < LPCA < NMF	/	Jaccard < Cos, Corr

Notes: A > B indicates a significantly higher value on the given metric for method A compared to method B; A < B indicates a significantly lower value on the given metric for method A compared to method B; A, B indicates method A obtains a better value on the given metric compared to method B, but the differences are not significant; / indicates no significant influence of the method on the given metric.
[1] None refers to algorithm configurations without data reduction as preprocessing.

Table II.2: Overview of effects of CF algorithm variations as function of the evaluation metric

Figure II.2: F1 Measure (a), ILS (b), and Computation Time (c) as function of selection size



[3] As the algorithms produce continuously ordered, personalized product sets, they do not need to be run for each selection size. Each algorithm only needs to be run once, and computation time is represented by a single statistic. For this evaluation metric, the analysis is an analysis of variance (ANOVA) without recommendation size as a covariate.

test results, which indicate differences across the levels of the parameters, are presented in TablesII.B.2-II.B.4 of Appendix B. In Figure II.2a, and in the remainder of this study, algorithms are represented by the three CF configuration settings, i.e. "Data reduction technique/CF method/Similarity Measure". For instance, the "CA/Item/Cos" algorithm representation refers to the algorithm with CA as data reduction technique, item-based CF as similarity measure have significant impacts on accuracy and diversity. However, computation time is only significantly affected by the reduction technique and similarity measure, while the CF method does not have a significant effect.

Specifically, a comparison of the different *reduction methods* as a preprocessing step shows that CA performs significantly best in terms of accuracy. The algorithms based on no reduction (None), CA- (most accurate reduction technique), NMF-, and SVD-reduced input matrices do not differ significantly in terms of diversity, although their diversity (lower ILS) is greater than that achieved by algorithms based on the LPCA reduction technique. Algorithms based on SVD, CA, and unreduced matrices are the fastest and not significantly different from each other. Moreover, the models based on these three reduction methods outperform LPCA and NMF. With regard to the *CF method*, the item-based version produces significantly better accuracy, but a lower diversity than user-based CF. Finally, the *similarity measure* shows significantly better accuracy of the cosine and Pearson correlation similarity measures when compared with Jaccard similarity. However, Jaccard similarity results in the highest diversity and fastest computation time.[4] Cosine and correlation similarity are equally diverse and computationally intensive, but they are less diverse and slower than the Jaccard measure.

4.2 RQ2. How Do Input Data Characteristics Influence the Optimal CF Configuration(s)?

An ANCOVA similar to the one used to assess RQ1 provides the results for RQ2.[5] To evaluate the performance for each data set, the analysis is repeated 54 times:

$$Metric_{ij} = \beta_{0ij} + \beta_{1ij}RM + \beta_{2ij}CFM + \beta_{3ij}S + \beta_{4ij}RS + \varepsilon_{ij}, \quad \text{where } i \in \{F1, ILS, Computation\ Time\} \text{ and } j \in \{1, \ldots, 54\}. \tag{9}$$

[4] The Jaccard similarity results in the fastest computation time because of the interaction effect with reduction technique. Jaccard similarity is only calculated for algorithms based on the non-reduced matrix, which is a fast reduction technique. Therefore, Jaccard is also perceived as fast. Cosine and correlation similarity are also calculated for algorithms preprocessed by NMF and LPCA reduction techniques, which are slow. Thus, cosine and correlation measures appear slow. However, a comparison of all three similarity measures that includes only the three fastest algorithms (CA, SVD, and none) reveals no significant differences.

[5] Similar to RQ1, the analysis of calculation time is limited to an ANOVA, without recommendation size as a covariate.

Appendix C contains tables with the optimal model(s) for each data set, as well as an indication of the models that do not differ significantly from the best algorithm. Table II.3 provides a general summary of the results.

Data Characteristics	*Evaluation Metric*	*CF Characteristics*		
		CF Method	**Similarity Measure**	**Reduction Technique**
Sparsity		Sparsity has no impact on the optimal model configuration in terms of accuracy, diversity, or computation time		
Purchase distribution	**Accuracy Time**	Purchase distribution has no impact on the optimal configuration in terms of accuracy or computation time		
	Diversity	Exponential: item	Exponential: cosine, correlation	Exponential: no clear pattern
		Linear: user	Linear: Jaccard	Linear: no reduction
		Uniform: low sparsity: item Uniform: high sparsity: user	Uniform: cosine, correlation	Uniform: no clear pattern
Item–user ratio	**Accuracy Diversity**	Item–user ratio has no impact on the optimal configuration in terms of accuracy or diversity		
	Time	0.5 (2000 users): item	Item–user ratio has no impact on the similarity measure in terms of computation time	0.5 (2000 users): exponential: CA 0.5 (2000 users): linear: SVD 0.5 (2000 users):uniform: SVD
		1 (1000 users): user		1 (1000 users): no reduction
		2 (500 users): user		2 (500 users): no reduction

Table II.3: Overview of effects of CF algorithm variations and input data characteristics as a function of the evaluation metric

The best *CF method* in terms of accuracy remains steady regardless of the input characteristics. Item-based CF always achieves better accuracy than user-based CF. For diversity, an interaction emerges between CF method and purchase distribution. For exponential distributions, item-based algorithms are the most diverse, while user-based CF in combination with no reduction and Jaccard similarity gives the most diverse results for linear distributions. For the uniform distribution, a third-order interaction with sparsity arises. For lower-sparsity configurations, item-based CF is dominant. However, at a higher level of sparsity, user-based CF takes over. Finally, in terms of computation time, user-based CF is preferable in data sets with item- user ratios of 2 and 1, which are the smallest data sets. For those with more users and an item-user ratio of 1/2, item-based algorithms are faster.

The effect of the input characteristics on the best *similarity measure* depends on the metric. For accuracy, cosine and correlation similarity perform better than Jaccard similarity. Input characteristics do not influence this preference. In terms of diversity, cosine and correlation measures result in the most diverse results for exponential and uniform distributions. However, for data sets with a linear purchase distribution, Jaccard is the most diverse similarity measure. The fastest similarity measure is not significantly influenced by input characteristics, although

cosine configurations are often not significantly faster, especially when given higher item-user ratios.

Finally, the effects of input characteristics on the best *data reduction technique* as a preprocessing step reflect the evaluation metric. In terms of accuracy, CA is the best-performing data reduction technique. For diversity, purchase distribution exerts an effect, such that for linear distributions, algorithms based on the non-reduced matrix (in combination with Jaccard similarity) lead to the highest diversity. For exponential and uniform distributions, less structure arises. The fastest reduction technique also depends on the item–user ratio. For data sets with high ratios (2 and 1; fewer users), algorithms based on the non-reduced matrix are fastest. For larger data sets (1/2), algorithms based on reduced input matrices are faster. With an exponential distribution, CA is the fastest reduction technique, while SVD is the fastest for linear and uniform distributions.

4.3 RQ3. How Sensitive Are the Optimal CF configuration(s) To Variations in the Input Data Characteristics?

The previous sections describe the models that perform best in static circumstances. However, what happens when the input characteristics of the binary purchase data change? To answer RQ3, this study relies on an analysis of the β_0 coefficients of the ANCOVA models from RQ2 (Equation 9), with the optimal model as a reference category:

$$\beta_{0i} = \gamma_{0i} + \gamma_{1i}\, iur + \gamma_{2i}\, spar + \gamma_{3i}\, dist + \varepsilon_i, \text{ where } i \in \{F1, ILS, Computation\ Time\}. \quad (10)$$

Equation 10 represents the ANOVA model used to evaluate the sensitivity of the β_0 coefficient. It indicates the performance of the optimal model in terms of the item-user ratio *(iur*), the sparsity level (*spar*), and purchase distribution (*dist*).

The results show that the accuracy of the optimal models—CA/Item/Cos and CA/Item/Corr[6]—is only significantly influenced by the sparsity level. The diversity of the most diverse models[7] depends on all three input characteristics. Purchase distribution and item-user

[6] As CA/Item/Cos and CA/Item/Corr reveal identical impacts on changes in input characteristics, this study only provides the statistical results and graphs for CA/Item/Corr.

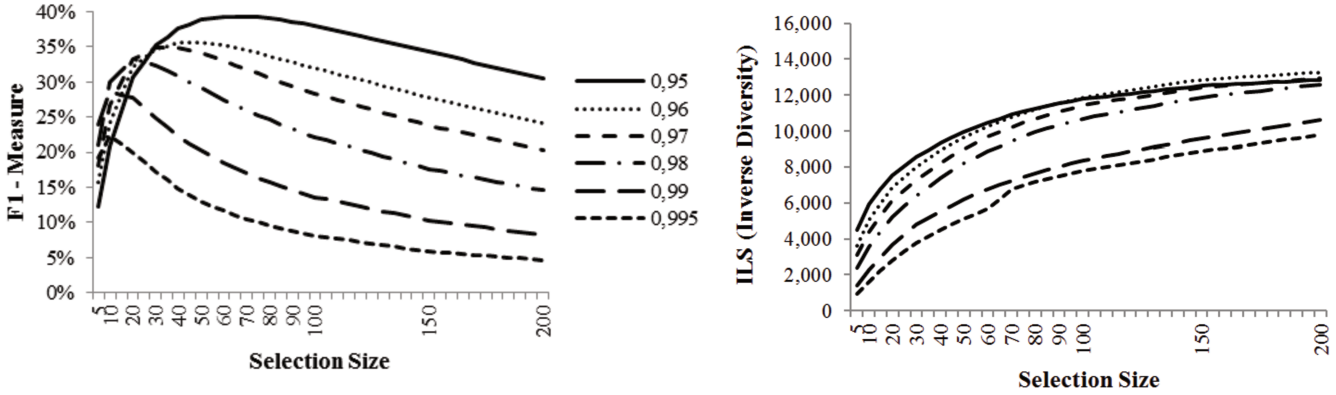
[7] The most diverse models are None/User/Jaccard, CA/Item/Cos, Corr, SVD/Item/Cos, Corr, and NMF/Item/Cos, Corr. For the three last model types, the discussion focuses on the correlation configuration because the impacts of input characteristics on the cosine configurations of the algorithms are similar.

ratios also significantly affect the computation time of the fastest models.[8] As the findings in Table II.4, Figure II.3, and Appendix D show, *sparsity level* has a negative impact on both accuracy and diversity. In other words, the accuracy of the optimal algorithm decreases with sparsity, whereas diversity rises as sparsity increases.

		CA/Item/Corr	*SVD/Item/Corr*	*NMF/Item/Corr*	*None/User/Jaccard*
If Sparsity	**Accuracy**	-0.0165	/	/	/
+ 1%	**ILS**	-1 128.30	-1 044.46	-1 147.31	-669.23

Table II.4: Influence of sparsity on best performing models: accuracy and diversity

Figure II. 3: Accuracy (a) and diversity (b) for different sparsity levels as a function of selection size for the CA/Item/Corr algorithm



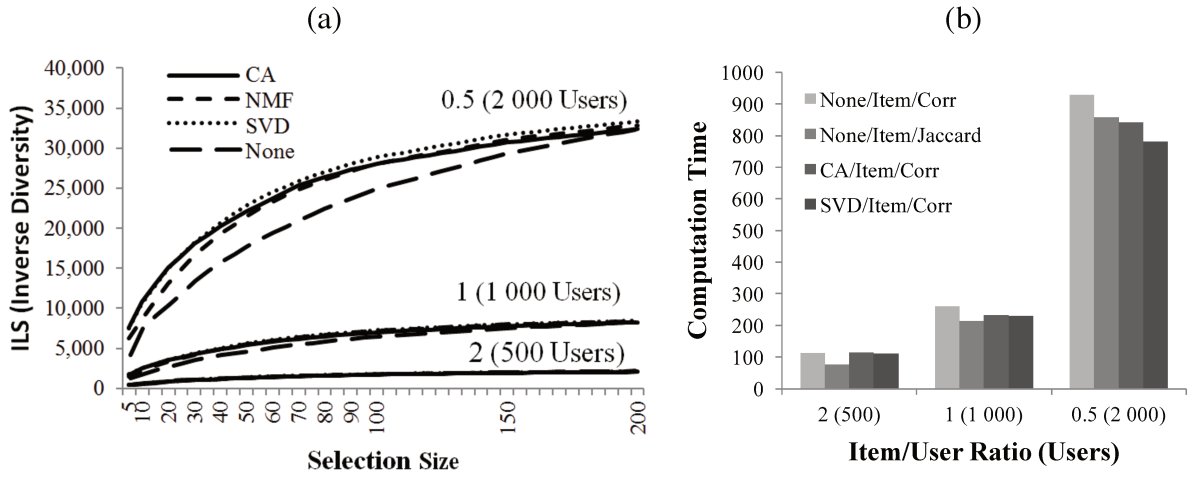
The effects of the *purchase distribution* on the performance of the optimal models (Appendix D) reveal that the highest diversity (for the None/User/Jaccard model) occurs when the purchase distribution is linear. It significantly differs from diversity in data sets with exponential and uniform distributions. The latter two distributions do not differ significantly in terms of diversity. For the CA, NMF, and SVD/Item/Corr algorithms, the ANOVA indicates a significant effect, although none of the pairwise t-test results indicate differences between two levels individually. The exponential distribution results in the lowest diversity but does not significantly differ from the linear or uniform distributions, nor do the two latter distributions differ significantly. The effects of purchase distribution on computation time appear

[8] The fastest algorithms consist of four types: (1) CA/User, Item/Cos, Corr; (2) SVD/User, Item/Cos, Corr; (3) None/User, Item/Cos, Corr; and (4) None/User, Item/Jaccard. The discussion of these models focuses on the Item/Corr configurations of the algorithms for the first three types and on the Item/Jaccard configuration for the fourth type. User-based CF has the same time sensitivity toward input characteristics, and cosine tendencies are comparable to correlation sensitivity.

significant, but an analysis of the differences across individual levels highlights no significant impact (see Appendix D).

The *item–user ratio* shows a significant effect on diversity and computation time, as Figure II.4 shows. A higher item–user ratio (fewer users) leads to greater diversity (lower ILS) for all four types of optimal models. For computation time, a higher ratio (fewer users) leads to faster runs. For item-user ratios of 2 and 1, all four optimal algorithms exhibit comparable calculation times. For an item-user ratio of 1/2, the algorithms based on the reduced matrix CA and SVD are (not significantly) faster than the non-reduced models.

Figure II.4: Diversity (a) and computation time (b) as a function of item–user ratios for the None/User/Jaccard and CA, NMF, SVD/Item/Corr algorithms



4.4 Empirical Validation

To validate the results derived from the synthetic data, we calculated the ANCOVA model for two real-life data sets covering a European e-commerce firm. In this regard, we analyze data for two product categories—women's clothing and furniture. The observed characteristics of these data sets are provided in Table II.5.

Data set	Women's Clothing	Furniture
Purchase distribution	Exponential	Exponential
# Items	33,599	22,016
# Users	595,737	187,305
Item-user ratio	6%	12%
# Purchases	22,017,784	2,886,595
Sparsity	99.89%	99.93%

Table II.5: Real-Life Data sets and their Characteristics

Evaluation Metric	*Data set*	*Reduction Technique*	*CF method*	*Similarity Measure*
Accuracy	**Women's Clothing**	**$F_{4, 228} = 12.15$ ($p < 0.001$)**	**$F_{1, 228} = 117.70$ ($p < 0.001$)**	**$F_{2, 228} = 7.72$ ($p = 0.02$)**
	Furniture	**$F_{4, 228} = 12.89$ ($p < 0.001$)**	**$F_{1, 228} = 33.99$ ($p < 0.001$)**	**$F_{2, 228} = 6.84$ ($p = 0.04$)**
		CA > NMF, SVD, LPCA > None	Item > User	Cos, Corr > Jaccard
Diversity	**Women's Clothing**	**$F_{4, 228} = 51.91$ ($p < 0.001$)**	**$F_{1, 228} = 28.12$ ($p < 0.001$)**	$F_{2, 228} = 0.04$ ($p = 0.84$)
	Furniture	**$F_{4, 228} = 83.20$ ($p < 0.001$)**	**$F_{1, 228} = 324.06$ ($p < 0.001$)**	$F_{2, 228} = 1.74$ ($p = 0.19$)
	Women's Clothing Furniture	None, CA, NMF, SVD, CA, None, SVD, NMF <	User < Item	Jaccard, Cos, Corr
Time	**Women's Clothing**	**$F_{4, 228} = 1{,}621.35$ ($p < 0.001$)**	**$F_{1, 228} = 7.60$ ($p = 0.02$)**	**$F_{2, 228} = 6.14$ ($p = 0.03$)**
	Furniture	**$F_{4, 228} = 16{,}839.10$ ($p < 0.001$)**	**$F_{1, 228} = 32.38$ ($p < 0.001$)**	**$F_{2, 228} = 37.33$ ($p < 0.001$)**
		SVD, CA, None < LPCA < NMF	Item < User	Jaccard < Cos, Corr

Notes: A > B indicates a significantly higher value on the given metric for method A compared to method B; A < B indicates a significantly lower value on the given metric for method A compared to method B; A, B indicates method A obtains a better value on the given metric compared to method B, but the differences are not significant

Table II.6: Overview of Effects of CF Algorithm Configurations

Table II.6 shows the results of the ANCOVA analyses, (comparable to the ANCOVAs calculated in section 4.1), which evaluate the effects of the reduction technique, the CF method, and the similarity measure on accuracy, diversity, and computation time (seconds) for the two real-life data sets. Recommendation size is included as a covariate.

The *reduction technique* has a significant impact on accuracy, diversity, and computation time. Consistent with the results from the synthetic data sets, CA reduction is more accurate than the other techniques. In terms of diversity, LPCA reduction leads to the most diverse recommendations. This difference is significant for both the synthetic and the furniture data sets. Although the women's clothing data set exhibits a similar pattern, the t-tests are not significant. The findings on the fastest reduction technique from synthetic data sets are validated using the real-life data sets. In addition, models based on SVD, CA, and no reduction have the lowest computation time.

The *CF method* significantly affects accuracy, diversity, and computation time. For accuracy and diversity, the results found for the synthetic data sets are confirmed. Item-based CF are more accurate than user-based CF, but the latter CF method leads to higher diversity. In the analysis of synthetic data sets, the CF method does not have a significant effect on computation time. However, the real-life data sets show that item-based CF is significantly faster than user-based CF.

This is mainly because the real-life data sets have a low item-user ratio. The distinction among the CF methods becomes more important because of the bigger size of the data sets. This finding confirms earlier research indicating that item-based CF is faster than user-based CF in settings with a small number of items compared to the number of users.

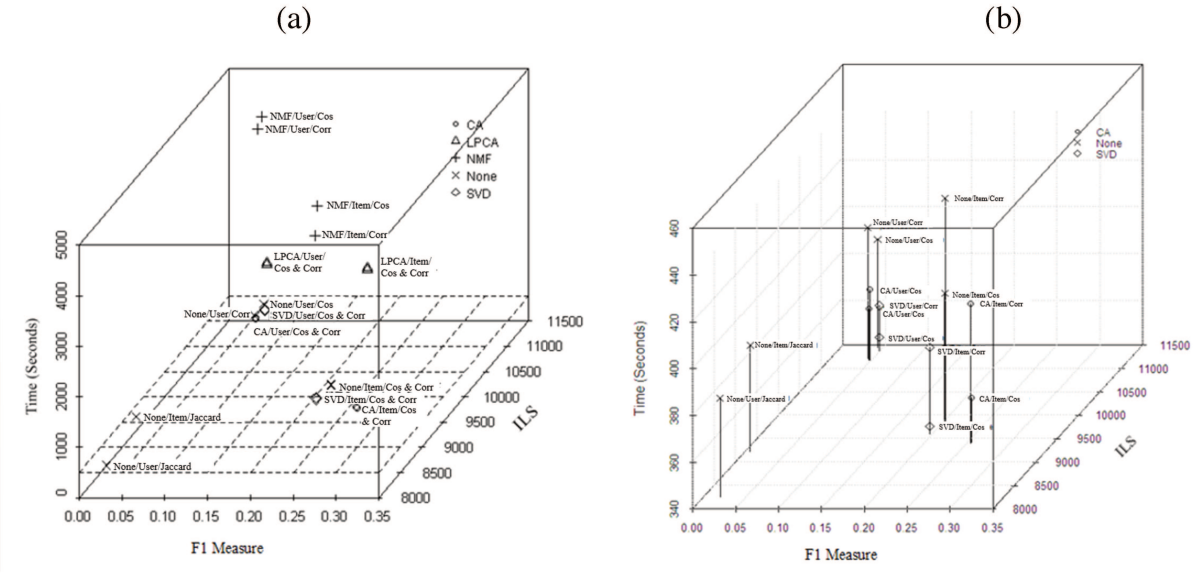
The *similarity measure* influences accuracy and computation time, but has no impact on diversity for the two real-life data sets. In agreement with the results for the synthetic data sets, cosine and correlation similarity are more accurate compared to Jaccard similarity. For diversity, a pattern indicating that Jaccard similarity is more diverse than the cosine and correlation measures is found in both synthetic and real-life data sets. The results are significant for the real-life data sets. Finally, Jaccard similarity is significantly faster compared to cosine and correlation measures. This is mainly due to the second-order effect between reduction technique and similarity measure, as explained in Section 4.1.

5 Discussion

E-commerce companies constantly try to maximize returns on their websites by, for example, using recommendation systems based on CF. Therefore, finding the right balance among accuracy, diversity, and calculation time is essential for selecting a suitable system. This study investigates the performance of several CF models and produces a decision support framework that can guide e-commerce firms in their efforts to select the best algorithm.

The optimization of recommendation systems' accuracy is very important for companies. In addition, given the impressive size of client and product databases, computation time is an important metric to keep in mind and to balance in relation to system accuracy. Diversity, a third metric, should be assessed together with the other two metrics to reach the desired level. The preferred level of diversity most likely depends on the system's purpose in terms of whether it focuses on upselling or cross-selling [43]. If the e-tailer wishes to upsell, the

Figure II.5: Algorithm variations as a function of accuracy, diversity, and computation time



recommended item set should be less diverse. In contrast, in systems aimed at cross-selling, a more diverse set is favorable. In addition, the company must avoid overspecialization. The trade-off graph across these three evaluation metrics is shown in Figure II.5a and indicates that LPCA and NMF are far slower than the other three algorithms. To emphasis this, an extra plane is drawn at the time of 500 sec. The remainder of this discussion, therefore, focuses on the three fast models, which are represented in Figure II.5b.

The fastest algorithms—SVD, CA, and no reduction based CF—reveal four performance groups. First, algorithms that use no reduction technique and Jaccard similarity lead to low accuracy, high diversity, and a fast computation time. Second, algorithms that employ user-based CF and cosine or correlation similarity show low accuracy and low diversity with a fast computation time. Third, the None, SVD/Item/Corr, Cos algorithms achieve high diversity and medium to high accuracy, along with fast computations. Fourth, the CA/Item/Corr and Cos models provide the best results—high diversity, the highest accuracy, and a low calculation time. Table II.7 provides an overview of the average rankings of the different algorithm configurations over the 54 synthetic and 2 real-life data sets. A lower ranking indicates a better average value for the focal metric. The lower the average rank, the better the algorithms scores on the respective dimension.

The optimization of the algorithm configurations also depends on the characteristics of the binary input data set available to an analyst. Given a data set with specific input characteristics, the proposed framework offers an indication of which algorithm configurations are most

Algorithm	Average Ranking		
	Accuracy	Diversity	Time
CA/Item/Cos, Corr	1.28	3.28	2.83
NMF/Item/Cos, Corr	2.53	3.26	10.55
SVD/Item/Cos, Corr	3.74	3.45	1.57
None/Item/Cos, Corr	3.81	4.68	5.02
LPCA/Item/Cos, Corr	6.25	7.66	8.42
LPCA/User/Cos, Corr	6.62	7.38	8.58
NMF/User/Cos, Corr	7.89	6.57	10.45
SVD/User/Cos, Corr	8.04	6.94	2.45
CA/User/Cos, Corr	8.79	6.60	3.85
None/User/Cos, Corr	8.81	7.08	4.60
None/User/Jaccard	9.55	3.83	4.08
None/Item/Jaccard	10.58	5.28	3.60

Note: Lower average rank corresponds to better performance (higher accuracy, higher diversity, lower running time).

Table II.7: Comparison of Accuracy, Diversity, and Computation Times for Different Algorithm Configurations

appropriate in a situation with a certain goal, based on a data set with specific input characteristics.

Changes in the binary data set probably alter the performance of the algorithm configurations. The proposed framework can estimate these alterations. For example, if an e-commerce shop becomes more popular, leading to an expanded client base and more products sold per customer, then the sparsity, item-user ratio, and purchase distribution are likely to change. Table II.8 details the effects of the input characteristics on the evaluation metrics.

	Sparsity	Purchase Distribution	Item–User
Accuracy	Negative	/	/
Diversity	Positive	Uniform, Linear ≥ Exponential[a]	2 > 1 > 0.5
Time	/	/	2 < 1 < 0.5

Notes: A > B indicates a significantly higher value on the given metric for method A compared to method B; A < B indicates a significantly lower value on the given metric for method A compared to method B; A, B indicates method A obtains a better value on the given metric compared to method B, but the differences are not significant; / indicates no significant influence of the method on the given metric.

[a] The uniform purchase distribution achieves significantly higher diversity than the exponential. The linear form does not differ significantly from uniform or exponential distributions. This relationship is not valid for the None/Jaccard algorithms, in which case linear purchase distributions result in more diverse solutions, but no difference arises between uniform and exponential distributions.

Table II.8: Effects of input characteristics on three evaluation metrics

Finally, scalability is a big issue in CF. In this study, we leverage big data technology which allows to run CF algorithms even on the real-life data set with 22,017,784 purchases by distributing computations over the memory of many machines. Distributed computing tecnhiques allow the execution of the CF algorithm in memory within an reasonable time.

6 Conclusions, Limitations, and Further Work

The ability to provide personalized recommendations is important for e-commerce firms. Along these lines, a recommendation system helps marketing departments establish an item set for each customer. While many companies cannot collect sufficient explicit feedback or product ratings from customers, they possess large transaction logs. The data in the latter can serve as low-cost input for a recommendation system [44]. From this view, binary purchase data provides an interesting basis for recommendations, and they relate directly to firm performance. A recommendation system, as an automated decision support system, can help marketing departments establish an item set for each specific customer.

This study provides a framework that can guide marketers in building better recommendation systems and avoiding trial-and-error processes in their attempts to find a suitable recommendation algorithm. The framework not only identifies the most accurate model but also gives an indication of the diversity and calculation times of different models. To do so, this study analyzes the performance of different CF algorithm configurations. In this regard, we use synthetic data sets with different binary purchase input characteristics as well as two real-life validation sets.

We find that the accuracy and diversity of the generated recommendations depend on the data reduction technique, the CF method, and the similarity measure. Computation time is influenced only by the data reduction technique, mainly in the sense that LPCA and NMF are based on multiplicative updating algorithms, whereas SVD and CA are calculated in one step. Second, different input characteristics can lead to other optimal algorithms. For accuracy, the optimal model is stable (CA/Item/Cos, Corr), but various characteristics lead to different optimal configurations for diversity and computation time. In addition, the optimal model configuration for each data set is influenced by input characteristics. For example, accuracy depends on sparsity, while diversity is influenced by sparsity, the purchase distribution, and the item-user ratio. Computation time is only influenced by the purchase distribution and the item-user ratio.

As shown in the extant literature, binary purchase data serve as a good basis for recommendation systems because purchase information not only reflects clear customer actions that indicate preferences but is also directly linked to firm performance. In addition to purchase data, e-commerce firms usually log and gather information about other customer actions, such as clicks, views, and additions to the cart or wish lists. Such information can also serve as a basis for recommendations. Therefore, investigations of combinations of multiple data sources as inputs could highlight ways to boost performance relative to a system based solely on purchase information.

CF algorithms are very popular and successful (Bobadilla et al., 2013) because they offer a strong basis for recommendation systems. However, model-based CF, as well as content- and demographic-based systems, are becoming more popular and could be applied as benchmarks against the CF algorithms used in this study. Along these lines, an interesting extension might be to combine several techniques to create a hybrid decision-support system.

In this study, reduction techniques are used in the preprocessing step of the CF procedure in order to increase efficiency and memory. Although model-based CF methods have deployed direct-imputation methods based on decomposed matrices, this study did not replicate this technique, mainly due to the structure of the input matrix. In other words, direct imputations seek to estimate blanks in the original matrix. As the input matrix for this study only contained 0s and 1s (no purchase/purchase), and had no missing values, direct imputation is less useful. Nevertheless, direct-imputation techniques should be tested in future research.

References

[1] S. Olafsson, X. Li, S. Wu, Operations Research and Data Mining, 2008, Eur. J. Oper. Res., 187 1429-1448.

[2] J. Alba, J. Lynch, B.A. Weitz, C. Janiszewski, R. Lutz, A. Sawyer, S. Wood, Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Marketplaces, 1997, J. Marketing, 61 38-53.

[3] S.M. Weiss, N. Indurkhya, Lightweight collaborative filtering method for binary-encoded data, 2001, in: L. De Raedt, A. Siebes (Eds.) Principles of data mining and knowledge discovery, Springer Berlin Heidelbergpp. 484-491.

[4] K.-W. Cheung, J.T. Kwok, M.H. Law, K.-C. Tsui, Mining customer product ratings for personalized marketing, 2003, Decis. Support Syst., 35 231-243.

[5] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, 2005, IEEE Trans. Knowl. Data Eng., 17 734-749.

[6] M. Deshpande, G. Karypis, Item-based top-N recommendation algorithms, 2004, ACM Trans. Inf. Syst., 22 143-177.

[7] B. Loni, A. Said, M. Larson, A. Hanjalic, 'Free lunch' enhancement for collaborative filtering with factorization machines, 2014, 8th ACM Conference on Recommender Systems, ACM, Foster City, CA, pp. 281-284.

[8] K. Verstrepen, B. Goethals, Unifying nearest neighbors collaborative filtering, 2014, Proceedings of the 8th ACM Conference on Recommender systems, ACM, Foster City, Silicon Valley, California, USA, pp. 177-184.

[9] S. Larrain, C. Trattner, D. Parra, E. Graells-Garrido, K. Nørvåg, Good Times Bad Times: A Study on Recency Effects in Collaborative Filtering for Social Tagging, 2015, Proceedings of the 9th ACM Conference on Recommender Systems, ACM, Vienna, Austria, pp. 269-272.

[10] S.E. Aldrich, Recommender Systems in Commercial Use, 2011, Artificial Intelligence Magazine, 32 28-34.

[11] E. Turban, D. King, K.J. Lee, T.-P. Liang, D.C. Turban, Electronic Commerce: A Managerial and Social Networks Perspective, 2015, 8 ed., Springer

[12] N. Sahoo, R. Krishnan, G. Duncan, J. Callan, On Multi-component Rating and Collaborative Filtering for Recommender Systems: The Case of Yahoo! Movies, 2012, Inform. Syst. Res., 23 231-246.

[13] M. Papagelis, D. Plexousakis, Qualitative Analysis of User-Based and Item-Based Prediction Algorithms for Recommendation Agents, 2005, Eng. Appl. Artif. Intell., 18 781-789.

[14] G. Adomavicius, J. Zhang, Impact of data characteristics on recommender systems performance, 2012, ACM Trans. Manage. Inf. Syst., 3 1-17.

[15] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, 1998, 14th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Madison, WI, pp. 43-52.

[16] H. Steck, Item popularity and recommendation accuracy, 2011, 5th ACM Conference on Recommender Systems, ACM, Chicago, Illinois, pp. 125-132.

[17] J.L. Herlocker, J.A. Konstan, K. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, 2004, ACM Trans. Inf. Syst., 22 5-53.

[18] B.M. Sarwar, G. Karypis, J.A. Konstan, J.T. Riedl, Analysis of recommendation algorithms for e-commerce, 2000, 2nd ACM Conference on Electronic Commerce, ACM, Minneapolis, Minnesota, pp. 158-167.

[19] M. Kellar, C. Watters, J. Duffy, M. Shepard, Effect of task on time spent reading as an implicit measure of interest, 2004, 67th Asis&T Annual Meeting, Medford: Information Today Inc, Providence, RI, pp. 168-175.

[20] B. Pradel, S. Sean, J. Delporte, S. Guerif, C. Rouveirol, N. Usunier, F. Fogelman-Soulie, F. Dufau-Joel, A case Study in a Recommender System Based on Purchase Data, 2011, 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, ACM, San Diego, California, USA, pp. 377-385.

[21] D. Iacobucci, P. Arabie, A. Bodapati, Recommendation Agents on the Internet, 2000, J. Interact. Marketing, 14 2-11.

[22] G. Adomavicius, Y. Kwon, Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques, 2012, IEEE Trans. on Knowl. and Data Eng., 24 896-911.

[23] X. Su, T.M. Khoshgoftaar, A survey of collaborative filtering techniques, 2009, Advances in Artificial Intelligence, 2009 2-2.

[24] G. Linden, B. Smith, J. York, Amazon.com Recommendations: Item-to-Item Collaborative Filtering, 2003, IEEE Internet Comput., 7 76-80.

[25] Y. Koren, R. Bell, C. Volinsky, Matrix Factorization Techniques for Recommender Systems, 2009, IEEE Comput., 42 30-37.

[26] K. Palanivel, R. Sivakumar, A Study On Implicit Feedback In Multicriteria E-Commerce Recommender System, 2010, J. Electron. Commer. Res., 11 140-156.

[27] P. Resnick, H.R. Varian, Recommender systems, 1997, Commun. ACM, 40 56-58.

[28] B. Shapira, P.B. Kantor, B. Melamed, The effect of extrinsic motivation on user behavior in a collaborative information finding system, 2001, Journal of the American Society for Information Science and Technology, 52 879.

[29] Y. Li, J. Hu, C. Zhai, Y. Chen, Improving One-Class Collaborative Filtering by incorporating Rich user Information, 2010, 19th ACM International Conference on Information and Knowledge Management, ACM, Toronto, ON, Canada, pp. 959-968.

[30] A.V. Bodapati, Recommendation systems with purchase data, 2008, J. Mark. Res., 45 77-93.

[31] K. Dutta, D. VanderMeer, A. Datta, P. Keskinocak, K. Ramamritham, A fast method for discovering critical edge sequences in e-commerce catalogs, 2007, Eur. J. Oper. Res., 181 855-871.

[32] P. Rong, Z. Yunhong, C. Bin, N.N. Liu, R. Lukose, M. Scholz, Y. Qiang, One-Class Collaborative Filtering, 2008, Eighth IEEE International Conference on Data Mining (ICDM), pp. 502-511.

[33] B.M. Sarwar, G. Karypis, J.A. Konstan, J.T. Riedl, Application of dimensionality reduction in recommender system -- A case study, 2000,

[34] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, 2004, ACM Trans. Inf. Syst., 22 5-53.

[35] A. Popescul, L.H. Ungar, D.M. Pennock, S. Lawrence, Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments, 2001, 17th Conference in Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., pp. 437-444.

[36] C.C. Aggarwal, J.L. Wolf, K.-L. Wu, P.S. Yu, Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering, 1999, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Diego, California, USA, pp. 201-212.

[37] X. Su, T.M. Khoshgoftaar, A Survey of Collaborative Filtering Techniques, 2009, Advances in Artificial Intelligence, DOI 10.1155/2009/421425 1-19.

[38] S. Lee, J.Z. Huang, J. Hu, Sparse Logistic Principal Components Analysis For Binary Data, 2010, The Annals of Applied Statistics, 4 1579-1601.

[39] M. Greenacre, Correspondence Analysis in Practice, 2007, Chapman and Hall/CRC

[40] L. Candillier, F. Meyer, F. Fessant, Designing Specific Weighted Similarity Measures to Improve Collaborative Filtering Systems, 2008, in: P. Perner (Ed.) Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects, Springer Berlin Heidelbergpp. 242-255.

[41] Z.C. Lipton, C. Elkan, B. Naryanaswamy, Optimal thresholding of classifiers to maximize F1 measure, 2014, in: T. Calders, F. Esposito, E. Hüllermeier, R. Meo (Eds.) Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelbergpp. 225-239.

[42] C.-N. Ziegler, S.M. McNee, J.A. Konstan, G. Lausen, Improving Recommendation Lists Through Topic Diversification, 2005, 14th International Conference on World Wide Web, ACM, Chiba, Japan, pp. 22-32.

[43] I. Bose, X. Chen, Quantitative models for direct marketing: A review from systems perspective, 2009, Eur. J. Oper. Res., 195 1-16.

[44] G.I. Doukidis, K. Pramatari, G. Lekakos, OR and the management of electronic services, 2008, Eur. J. Oper. Res., 187 1296-1309.

Appendices

Appendix A: Data Generation Process

This appendix discusses the data generation process to create the synthetic data sets and is divided into four steps, i.e. (i) determining correlation structure, (ii) determining the user-input matrix dimensions to set the user-item ratio, (iii) designing sparsity and purchase distribution, and finally, (iv) generating and discretizing the user-item matrix. The first step is executed once, while steps 2 until 4 are repeated for each of the 54 synthetic data sets that are created for the experiments in this study, exhaustively combining all the variations of three input characteristic configurations.

Consider the following notation. The input of a recommender system is provided in the format of a user-item matrix, denoted X. X consists of entries x_{ki}, binary values that indicate whether user (customer) k purchased item (product) i (i.e., $x_{ki} = 1$) or not ($x_{ki} = 0$). This logical matrix provides purchase information for n users and m items and hence, $X \in \mathbb{Z}_2^{n \times m}$. Equivalently, one can say that X consists of m item vectors $x_{*i}; i \in \{1, \dots, m\}$ that each represents purchase information of n users for an item i.

Step 1: Obtain correlation structure from real-life user-item matrix

In a realistic setting the purchase of some items correlates with the purchase of others. For results to generalize well to real-life applications, a realistic purchase correlation structure is vital to account for variety in terms of cross-elasticity of demand and accommodate the existence of complementary, substitute and independent goods. Hence, this study opts to transfer the correlation structure by calculating a correlation for a real-life data set obtained from a European e-tailer and replicate this correlation structure in the synthetically generated data sets.

The first step involves the calculation of a correlation matrix for the real-life user-item matrix. The Pearson correlation coefficient can be reduced to the phi coefficient (φ_{ij}) for two binary item vectors of items *i* and *j* [1, 2]. The phi coefficient is based on the confusion matrix of two binary variables. In the case of two item vectors for items *i* and *j*, this confusion matrix is given by Table II.A1.

		Item j		
		$x_{kj} = 0$	$x_{kj} = 1$	$x_{kj} \in \{0,1\}$
Item i	$x_{ki} = 0$	a	b	$n - \|x_{*i}\|_1$
	$x_{ki} = 1$	c	d	$\|x_{*i}\|_1$
	$x_{ki} \in \{0,1\}$	$n - \|x_{*j}\|_1$	$\|x_{*j}\|_1$	n

Table II.A.1: Confusion matrix of item vectors for items *i* and *j*

Based upon the quantities defined in Table II.A1, the phi coefficient can then be obtained as

$$\hat{\varphi}_{ij} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \quad \text{(A1)}$$

The correlation coefficients are calculated for every pair of items *i* and *j* in the real life data set to construct a full correlation matrix *R* with values $\hat{\varphi}_{ij}\ \forall i, j \in \{1, \dots, m\}$. *R* is used to govern the data generation and discretization in Step 4.

Step 2: Determine user-item input matrix dimensions to set user-item ratio

To create a complete user-item input matrix, the number of item vectors is equal to the number of items, each having a length equal to the number of users. The number of items is constant for all data sets; m=1000; and the item-user ratio $\frac{m}{n}$ is adjusted by the varying the number of users and *n* is set to 2,000, 1,000, and 500 to obtain item-user ratios of respectively 1/2, 1, and 2.

Step 3: Determine the desired number of non-zero elements in user-item input matrix and item vectors to govern sparsity level and purchase distribution

In step 3 the overall sparsity level and purchase distributions are set simultaneously through determining how dense each item vector should be, i.e. how many non-zero elements each item vector should contain. The sparsity level is a negative function of the number of non-zero elements which, for a logical matrix or vector, is given by its entrywise 1-norm:

$$\|X\|_1 = \sum_{k=1}^{n}\sum_{i=1}^{m} x_{ki}\,. \tag{A2}$$

Sparsity level (denoted $spar$) is then given by

$$spar = 1 - \frac{\|X\|_1}{mn}. \tag{A3}$$

Note that the overall desired sparsity level for a data set is set up-front and thus the required number of non-zero elements in user-item matrix, $\|X\|_1$, is easily derived from equation A3:

$$\|X\|_1 = mn(1 - spar). \tag{A4}$$

The purchase distribution dictates whether all items are bought equally or not, and to which extent variation exists. Put differently, the purchase distribution prescribes how and to which extent the sparsity level varies across item vectors. Constant item success requires a constant sparsity level for all item vectors, while differences in sparsity level across item vectors reflect variations in relative item success. To translate the overall desired sparsity of the data set to the sparsity levels of the item vectors, three alternative purchase distributions are imposed in the synthetic data sets: a uniform, a linear, and an exponential distribution.

Under a uniform purchase distribution, all the individual items are assumed equally successful, i.e. they have all been purchased by an equal number of customers. In this case, item vectors should have the same level of sparsity and hence, $\|x_{*i}\|_1$, the number of non-zero elements in the item vector of item i, is equal for all items:

$$\|x_{*i}\|_1 = \frac{\|X\|_1}{m}\ ;\ \forall i \in \{1, \dots, m\} \tag{A5}$$

and so, from equation A3 we derive:

$$\|x_{*i}\|_1 = n(1 - spar);\ \forall i \in \{1, \dots, m\} \tag{A6}$$

Under a linear purchase distribution, some items are considered more successful than others and success, i.e., the number of purchasing users, increases constantly over items when ordered from least to most often purchased. Consequently, the sparsity level for the corresponding item

vectors decreases linearly over items. The total number of non-zero items in the user-item matrix is now calculated as follows:

$$\|X\|_1 = \sum_{i=1}^{m} \delta i \tag{A7}$$

where δ is the slope of the linear function that represents the constant increase in item success, i.e. the increase in number of non-zero elements in the item vectors when comparing an item to the next more successful one.

The value of parameter δ depends on the values of m, n, and $spar$ and can be deterministically calculated. By combining A3 and A7, an expression for the value of δ can be derived:

$$\sum_{i=1}^{m} \delta i = mn(1 - spar)$$

$$\Leftrightarrow \tag{A8}$$

$$\delta \frac{m(m+1)}{2} = mn(1 - spar)$$

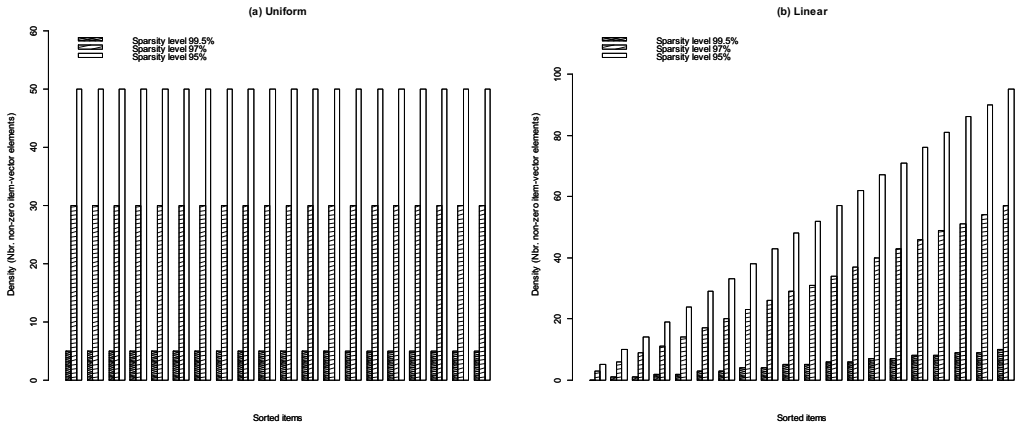
$$\Leftrightarrow$$

$$\delta = \frac{2n(1 - spar)}{m+1}$$

And thus, $\|x_{*i}\|_1 = \frac{2n(1-spar)}{m+1} i;\ \forall i \in \{1, \dots, m\}$

Figure II.A1 shows through a visualization (a) how a uniform purchase distribution compares to (b) a linear one for three alternative global sparsity levels.

Figure II.A.1: Example of uniform (a) and linear (b) purchase distributions for three alternative sparsity levels (m=20; n=1,000)



Finally, to obtain an exponential pattern, in which most items are only bought a few times and a small number of products are very popular, a natural exponential function governs the number of non-zero elements in each item vector:

$$\|X\|_1 = \sum_{i=1}^{m} e^{\mu i} \tag{A9}$$

where the parameter μ is introduced to control the range of the natural exponential function that is projected onto the set of item vectors. Larger values of μ increase this range, allowing for denser (less sparse) item vectors.

By combining equations A4 and A9, an expression for the value of μ (which depends upon the values of m, n, and $spar$) can be derived:

$$\sum_{i=1}^{m} e^{\mu i} = mn(1 - spar). \tag{A10}$$

via the formula for the sum of the first n terms of a geometric series:

$$\sum_{k=0}^{n-1} vr^k = v\frac{1-r^n}{1-r} \tag{A11}$$

and Equation A10 one can derive:

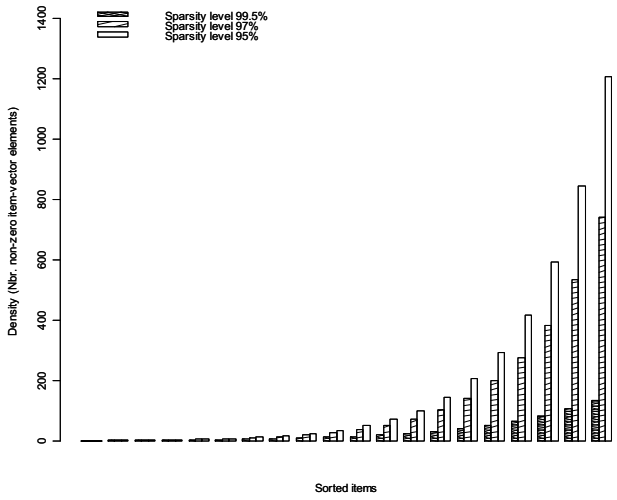
$$\frac{1-e^{(m+1)\mu}}{1-e^{\mu}} - 1 = mn(1-spar). \quad \text{(A12)}$$

For each value of $mn(1-spar)$ a value for μ can be calculated based upon Equation A12. Table II.A2 gives an overview of the μ-values obtained for each combination of *n* and *spar*, keeping *m* constant at 1,000 users.

		spar					
		95%	**96%**	**97%**	**98%**	**99%**	**99.5%**
	500	0.00481	0.00452	0.00413	0.00362	0.00262	0.00159
n	**1,000**	0.00565	0.00535	0.00504	0.00452	0.00362	0.00262
	2,000	0.00648	0.0062	0.00583	0.00535	0.00452	0.00362

Table II.A.2: $\boldsymbol{\mu}$ –values as function of *n* and *spar* for *m* = 1,000

Figure II.A.2: Example of an exponential purchase distribution for three alternative sparsity levels (m=20; n=1,000)



Note that regardless of the requested purchase distribution, values of $\|x_{*i}\|_1$ are rounded to obtain discrete quantities and a minimum value of 1 is imposed.

Step 4: generate and discretize user-item matrix

In this final step, the synthetic user-item matrix is created. First, a real-valued, intermediary user-item matrix is generated to respect the correlation structure (*R*) identified in Step 1 and the desired dimensions (i.e., *n* and *m)* identified in Step 2. Second, discretization is applied to enforce the desired sparsity level and purchase distribution (i.e., item vector sparsity levels obtained in Step 3) in order to obtain a logical user-item matrix.

An intermediary $n \times m$ – matrix S_1 is first populated with random, normally distributed values s_{ki} ; $s_{ki} \sim N(0,1)$ for every $i \in \{1, \dots, m\}$ and $k \in \{1, \dots, n\}$. Then, through applying a Cholesky factorization to the correlation matrix *R* (a positive-definite matrix),

$$R = R_{CF} {R_{CF}}^T \tag{A13}$$

where R_{CF} is the Cholesky factor of *R* and ${R_{CF}}^T$ the transpose of R_{CF} [3], the desired correlation structure can be enforced onto S_1 . Specifically, a second intermediary matrix S_2 is obtained by enforcing the correlation structure *R* onto S_1 through the following matrix multiplication [4]:

$$S_2 = {R_{CF}}^T S_1. \tag{A14}$$

The item vectors in S_2 now correlate in correspondence to R and the values of S_2, denoted ${s'}_{ki}$, are still real-valued. Since we want to obtain a logical matrix, a final step transforms S_2 into a logical user-item matrix *X* [5]. This discretization is applied by, for every item *i,* defining a threshold u_i that is used to dichotomize values ${s'}_{ki}$ $\forall k \in \{1, \dots, n\}$ of matrix S_2:

$$x_{ki} = I({s'}_{ki} > u_i);\ \forall i \in \{1, \dots, m\};\ \forall k \in \{1, \dots, n\}. \tag{A15}$$

where u_i is a threshold obtained by solving

$$u_i = \underset{u \in \mathbb{R}}{arg\ min} [(\sum_{k=1}^{n} I({s'}_{ki} > u) - \|x_{*i}\|_1)^2];\ \forall i \in \{1, \dots, m\}. \tag{A16}$$

Hence, the threshold u_i used to dichotomize the item vector for item *i* is chosen so that the number of non-zero values equals $\|x_{*i}\|_1$ the quantity calculated in Step 3.

The result is a synthetic logical user-item matrix that satisfies the imposed user-item ratio, sparsity level and purchase distribution and exhibits a purchase correlation structure from real-life data.

Appendix B: Statistical results for RQ1

The first research question investigates which algorithm variations perform best, without accounting for any input characteristics. The analysis of the effect of different algorithm variations relies on AN(C)OVA, in which the algorithm variations are the independent variables, and accuracy, diversity, and computation time are the dependent variables. Table II.B.1 shows the AN(C)OVA results of the effect of algorithm configuration steps (columns) on the evaluation metrics (rows). The *F*-statistic and *p*-value for reduction technique are show in the first column. If the effect of reduction technique on the evaluation metric is significant ($p < 0.05$), the cell content is set in bold. Equivalently the second and third column indicate the effect of CF method and similarity measure on the evaluation metrics. The results indicate which parameters—data reduction techniques, CF methods, and similarity measures—influence performance. The differences within levels of parameters are analyzed using a pairwise t-test, with the results shown in Tables II.B.2–II.B.4.

		Reduction Technique	***CF Method***	***Similarity Measure***
Evaluation Metric	**Accuracy**	**$F_{4, 15,437} = 709.97$ ($p < 0.001$)**	**$F_{2, 15,437} = 2{,}932.11$ ($p < 0.001$)**	**$F_{3, 15,437} = 2\,196.67$ ($p < 0.001$)**
	Diversity	**$F_{4, 15,437} = 5.84$ ($p < 0.001$)**	**$F_{2, 15,437} = 41.80$ ($p < 0.001$)**	**$F_{3, 15,437} = 12.76$ ($p < 0.001$)**
	Time	**$F_{4, 1,181} = 237.27$ ($p < 0.001$)**	$F_{2, 1,181} = 0.10$ ($p = 0.66$)	**$F_{3, 1,181} = 20.60$ ($p < 0.001$)**

Table II.B.1: AN(C)OVA results for RQ1

				Reduction Method 1			
				CA	**NMF**	**SVD**	**LPCA**
Accuracy			**Mean / SD**	0.1498 / 0.1297	0.1262 / 0.1029	0.1224 / 0.1117	0.1149 / 0.0989
	Reduction Method 2	**None**	0.0860 / 0.1021	**$t_{4,940} = 21.73$ ($p < 0.001$)**	**$t_{5,874} = 15.96$ ($p < 0.001$)**	**$t_{5,537} = 13.71$ ($p < 0.001$)**	**$t_{5,703} = 11.73$ ($p < 0.001$)**
		CA	0.1498 / 0.1297		**$t_{5,238} = -7.47$ ($p < 0.001$)**	**$t_{5,390} = -8.41$ ($p < 0.001$)**	**$t_{5,149} = -11.24$ ($p < 0.001$)**
		NMF	0.1262 / 0.1029			$t_{5,473} = -1.34$ ($p = 0.18$)	**$t_{5,502} = -4.18$ ($p < 0.001$)**
		SVD	0.1224 / 0.1117				**$t_{5,431} = -2.64$ ($p = 0.008$)**
Diversity			**Mean / SD**	9,130.62 / 10,866.02	9,065.45 / 10,854.41	9,287.87 / 10,921.52	10,115.76 / 11,648.35
	Reduction Method 2	**None**	8,440.138 /	$T_{5,015} = 2.43$ ($p = 0.23$)	$t_{5,890} = 1.39$ ($p = 0.17$)	$t_{5901} = 1.86$ ($p = 0.06$)	**$t_{5703} = 4.47$ ($p < 0.001$)**
		CA	9,130.62 / 10,866.02		$T_{4,782} = -0.21$ ($p = 0.84$)	$t_{5510} = 0.61$ ($p = 0.54$)	**$t_{4\,758} = 3.03$ ($p = 0.003$)**
		NMF	9,065.45 / 10,854.41			$t_{5510} = 0.42$ ($p = 0.67$)	**$t_{5499} = 2.86$ ($p = 0.005$)**
		SVD	9,287.87 / 10,921.52				**$t_{5497} = 2.45$ ($p = 0.014$)**
Time			**Mean / SD (sec)**	372.66 / 340.03	4,097.01 / 3,589.70	356.38 / 320.95	1,324.09 / 1,160.99
	Reduction Method 2	**None**	397.04 / 361.00	$t_{470} = -0.79$ ($p = 0.43$)	**$t_{214} = 15.11$ ($p < 0.001$)**	$t_{486} = -1.36$ ($p = 0.17$)	**$t_{238} = 11.27$ ($p < 0.001$)**
		CA	372.66 / 340.03		**$t_{215} = 15.04$ ($p < 0.001$)**	$t_{421} = -0.50$ ($p = 0.61$)	**$t_{247} = 11.45$ ($p < 0.001$)**
		NMF	4,097.01 / 3,589.70			**$t_{214} = -15.11$ ($p < 0.001$)**	**$t_{255} = -10.70$ ($p < 0.001$)**
		SVD	356.38 / 320.95				**$t_{243} = 11.70$ ($p < 0.001$)**

Table II.B.2: Accuracy, diversity, and computation time results of the pairwise t-test of data reduction techniques, indicating t_{df} and *p*-values

			User-Based
Accuracy		**Mean / SD**	0.0457 / 0.0503
	Item-based	0.1879 / 0.1098	**$t_{10\,623}$ = -102.46 (p < 0.001)**
Diversity		**Mean / SD**	9,467.60 / 11,542.19
	Item-based	8,508.73 / 10,276.97	**$t_{12\,981}$= 6.61 (p < 0.001)**
Time		**Mean / SD**	1,242.10 / 2,204.31
	Item-based	1,210.87 / 2,091.26	t_{1161} = 0.25 (p = 0.80)

Table II.B.3: Accuracy, diversity, and computation time results of the pairwise t-test of CF methods, indicating t_{df} and *p*-values

			Correlation	*Cosine*
Accuracy		**Mean/SD**	0.1274 / 0.111	0.1271 / 0.110
	Cosine	0.1271 / 0.110	$t_{13.778}$ = 0.14, p = 0.89	
	Jaccard	0.0127 / 0.0151	**$t_{7,876}$ = 82.14, p < 0.001**	**$t_{7,881}$ = 82.25, p < 0.001**
Diversity		**Mean/SD**	9,377.36 / 11,049.90	9,399.1 / 11,078.27
	Cosine	9,399.14 / 11,078.27	$t_{11\,958}$ = -0.108, p = 0.91	
	Jaccard	6,637.31 / 9,328.662	**$t_{1\,929}$ = 8.98, p <0.001**	**$t_{1\,933}$ = 9.04, p < 0.001**
Time		**Mean/SD**	1,288.88 / 2,165.28	1,332.61 / 2,299.35
	Cosine	1,332.61 / 2,299.35	$T_{1\,054}$ = -0.32, p = 0.75	
	Jaccard	383.92 / 354.59	**t_{624} = 9.03, p < 0.001**	**t_{618} = 8.98, p < 0.001**

Table II.B.4: Accuracy, diversity, and computation time results of the pairwise t-test of similarity measures, indicating t_{df} and *p*-values

Appendix C: Best performing models for each data set (RQ2)

This appendix details the best performing models for each data set with its specific input characteristics, item–user ratio, sparsity, and purchase distribution. For each data set (cell), the best performing model is shown together with superscripts. These superscripts refer to the models listed beneath each table. If a superscript is present in a cell, the corresponding model does not significantly differ from the best performing model in that data set. Tables II.C.1, II.C.2, and II.C.3 represent the best performing models for accuracy, diversity, and computation time, respectively.

Item–User Ratio	Sparsity	*Distribution* Exponential	Linear	Uniform
2 (= 500 Users)	**0.95**	CA/Item/Corr [1]	CA/Item/Cos [1,3,4,5]	CA/Item/Cos [2,3,4,5,6]
	0.96	CA/Item/Corr [1]	CA/Item/Cos [1,3,4,5]	CA/Item/Cos [2]
	0.97	CA/Item/Corr [1]	CA/Item/Corr [1]	CA/Item/Corr [1,3]
	0.98	CA/Item/Corr [1]	CA/Item/Corr [1,3]	CA/Item/Corr [1]
	0.99	CA/Item/Corr [1,3]	CA/Item/Cos [2,3]	CA/Item/Corr [1,3]
	0.995	CA/Item/Corr [1,3]	CA/Item/Corr [1,3,4]	CA/Item/Cos [2,3]
1 (= 1,000 Users)	**0.95**	CA/Item/Corr [1,5]	CA/Item/Cos [2,4,5]	CA/Item/Cos [2,3,4,5,6]
	0.96	CA/Item/Cos [2]	CA/Item/Cos [2,4,5]	CA/Item/Cos [2]
	0.97	CA/Item/Cos [2]	CA/Item/Corr [1,5]	CA/Item/Corr [2]
	0.98	CA/Item/Cos [2]	CA/Item/Cos [2]	CA/Item/Cos [2]
	0.99	CA/Item/Corr [1]	CA/Item/Cos [2]	CA/Item/Corr [2,3]
	0.995	CA/Item/Corr [1]	CA/Item/Cos [2,3]	CA/Item/Cos [2,3,4]
0.5 (= 2,000 Users)	**0.95**	SVD/Item/Cos [1,2,5]	SVD/Item/Corr [1,2,4,5]	CA/Item/Cos [2,3,4,5,6]
	0.96	CA/Item/Cos [2,5]	SVD/Item/Cos [1,2,4,5]	CA/Item/Cos [2,3,4,5,6]
	0.97	CA/Item/Corr [1,5]	CA/Item/Cos [2,5]	CA/Item/Cos [2,5]
	0.98	CA/Item/Cos [2]	CA/Item/Cos [2,5]	CA/Item/Corr [1,3]
	0.99	CA/Item/Cos [2]	CA/Item/Cos [2]	CA/Item/Corr [1]
	0.995	CA/Item/Corr [1]	CA/Item/Cos [2]	CA/Item/Corr [1,3]

[1] CA/Item/Cos [3] NMF/Item/Cos, Corr [5] SVD/Item/Cos, Corr
[2] CA/Item/Corr [4] None/Item/Cos, Corr [6] LPCA/Item/Cos, Corr

Table II.C.1: Best performing models for accuracy as a function of item–user ratio, sparsity, and purchase distribution

Item–User Ratio	Sparsity	*Distribution* Exponential	Linear	Uniform
2 (= 500 Users)	**0.95**	None/Item/Cos [13]	None/User/Jaccard [3]	CA/Item/Corr [1,2,5,6,7,10,11]
	0.96	None/Item/Cos [1,2,4,6,7]	None/User/Jaccard [3]	CA/Item/Cos [1,2,5-12]
	0.97	NMF/Item/Cos [1,2,5,6,7]	None/User/Jaccard [3]	CA/Item/Corr [1,2,5 -10]
	0.98	CA/Item/Cos [1,2,4,6,7]	None/User/Jaccard [3]	None /User/Corr [11]
	0.99	CA/Item/Corr [1,2,6,7]	None/Item/Jaccard [4]	LPCA/User /Corr [1,6,8 - 12]
	0.995	NMF/Item/Corr [1,2,6,7]	None/Item/Jaccard [3]	LPCA/User /Cos [8,10,11,12]
1 (= 1,000 Users)	**0.95**	SVD/Item/Cos [1,2,5,6,7]	None/User/Jaccard [3]	SVD/Item/Corr [1,2,5,7,10]
	0.96	None/Item/Cos [1,2,4,6,7]	None/User/Jaccard [3]	SVD/Item/Corr [1,2,5,7,8,9]
	0.97	CA/Item/Cos [1,2,4,5,6,7]	None/User/Jaccard [3]	None/Item/Cos [1,2,5,7,9]
	0.98	CA/Item/Corr [1,2,4,5,6,7]	None/User/Jaccard [3]	None /User/Corr [8,9,10,11,12]
	0.99	NMF/Item/Corr [1,2,4,6,7]	None/User/Jaccard [3]	None /User/Corr [8,10]
	0.995	NMF/Item/Corr [1,2,4,6,7]	None/User/Jaccard [3]	None /User/Corr [9,10]
0.5 (= 2,000 Users)	**0.95**	None/Item/Cos [13]	None/User/Jaccard [3]	SVD/Item/Corr [1,7]
	0.96	None/Item/Cos [1,2,4,6,7]	None/User/Jaccard [3]	SVD/Item/Cos [1,2,5,7]
	0.97	None/Item/Cos [1,2,4,6,7]	None/User/Jaccard [3]	CA/Item/Corr [1,2,5,7]
	0.98	None/Item/Cos [1,2,4,6,7]	None/User/Jaccard [3]	CA/Item/Cos [1,2,5,7]
	0.99	NMF/Item/Corr [1,2,6,7]	None/User/Jaccard [3]	None/User/Corr [8,11,12]
	0.995	SVD/Item/Corr [1,2,6,7]	None/User/Jaccard [3]	CA/User/Corr [8,11,12]

CA/Item/Cos, Corr [4] None/User/Jaccard [7] SVD/Item/Cos, Corr [10] LPCA/User/Cos, Corr [13] All
[2] None/Item/Cos, Corr [5] LPCA/Item/Cos, Corr [8] CA/User/Cos, Corr [11] NMF/User/Cos, Corr
[3] None/Item/Jaccard [6] NMF/Item/Cos, Corr [9] None/User/Cos, Corr [12] SVD/User/Cos, Corr

Table II.C.2: Best performing models for of diversity as a function of item–user ratio, sparsity, and purchase distribution

Item–User	Sparsity	*Distribution* Exponential	Linear	Uniform
	0.95	SVD/User/Cos (50.5) [1]	SVD/User/Cos (69.6) [1]	None/User/Cos (56.4) [2]
2	**0.96**	None/User/Cos (46.4) [2]	None/User/Cos (71.2) [1]	None/User/Jaccard
	0.97	None/User/Cos (48.2) [2]	CA/User/Cos (76.7) [1]	None/User/Jaccard
(= 500 Users)	**0.98**	None/User/Jaccard	CA/User/Cos (74.5) [1]	None/User/Cos (52.8) [1]
	0.99	None/User/Jaccard	CA/User/Cos (76.9) [1]	SVD/User/Cos (51.4) [1]
	0.995	None/User/Cos (47.6) [1]	SVD/User/Cos (77.5) [1]	SVD/User/Corr (51.8) [1]
	0.95	CA/User/Cos (135) [2]	SVD/Item/Cos (181.7) [2]	None/User/Jaccard
1	**0.96**	CA/User/Cos (128.9) [2]	SVD/User/Cos (178.3) [2]	None/User/Jaccard
	0.97	None/User/Cos (151.6) [2]	SVD/User/Cos (163.3) [2]	SVD/User/Cos (183.8) [2]
(= 1,000	**0.98**	None/User/Cos (174.4) [2]	SVD/Item/Cos (187.5) [2]	CA/User/Cos (176.3) [2]
Users)	**0.99**	None/User/Cos (176.3) [2]	SVD/User/Cos (182.5) [2]	SVD/User/Cos (180.4) [2]
	0.995	None/User/Cos (184.3) [2]	SVD/User/Cos (183.9) [2]	SVD/User/Cos (176.8) [2]
	0.95	CA/Item/Cos (586.4) [2]	SVD/Item/Cos (805.7) [1]	SVD/Item/Cos (772.5) [2]
0.5	**0.96**	CA/Item/Cos (583.3) [2]	SVD/Item/Cos (579.1) [2]	SVD/Item/Cos (780.9) [2]
	0.97	CA/Item/Cos (566.6) [2]	SVD/Item/Cos (860.7) [1]	SVD/Item/Cos (687.9) [2]
(= 2,000	**0.98**	CA/Item/Cos (572.4) [2]	SVD/Item/Cos (838.7) [1]	SVD/Item/Cos (791.4) [2]
Users)	**0.99**	SVD/Item/Cos (572.2) [2]	SVD/Item/Cos (851.1) [1]	SVD/Item/Cos (777.1) [2]
	0.995	CA/Item/Cos (580.3) [2]	SVD/Item/Cos (869.4) [1]	SVD/Item/Cos (789.7) [2]

[1] CA, SVD, None/Item, User/Cos, Corr (, Jaccard) [2] CA, SVD, None, LPCA/Item, User/Cos, Corr (, Jaccard)

Table II.C.3: Best performing models for computation time as a function of item–user ratio, sparsity, and purchase distribution

Appendix D: Statistical results for RQ3

This appendix includes the statistical justification for RQ3 for diversity and computation time. More specific, the sensitivity of diversity and computation time to a change in the input characteristics of the best performing models is tested.

Model	*Sparsity*	*Purchase Distribution*	*Item/User Ratio*
CA/Item/Corr	$\mathbf{F_{1,48} = 23.40\ (p < 0.001)}$	$\mathbf{F_{2,48} = 2.58\ (p < 0.087)}$	$\mathbf{F_{2,48} = 2.43\ (p = 0.0099)}$

Table II.D.1: ANOVA results for accuracy sensitivity of the CA/Item/Corr model to sparsity, purchase distribution, and item/user ratio

Model	*Sparsity*	*Purchase Distribution*	*Item/User Ratio*
None/User/Jaccard	$\mathbf{F_{1,48} = 4.81\ (p = 0.03)}$	$\mathbf{F_{2,48} = 22.71\ (p < 0.001)}$	$\mathbf{F_{2,48} = 15.79\ (p < 0.001)}$
CA/Item/Corr	$\mathbf{F_{1,48} = 27.04\ (p < 0.001)}$	$\mathbf{F_{2,48} = 9.59\ (p < 0.001)}$	$\mathbf{F_{2,48} = 62.36\ (p = 0.0056)}$
NMF/Item/Corr	$\mathbf{F_{1,48} = 36.12\ (p < 0.001)}$	$\mathbf{F_{2,48} = 6.30\ (p = 0.004)}$	$\mathbf{F_{2,48} = 81.27\ (p < 0.001)}$
SVD/Item/Corr	$\mathbf{F_{1,48} = 29.51\ (p < 0.001)}$	$\mathbf{F_{2,48} = 9.25\ (p < 0.001)}$	$\mathbf{F_{2,48} = 84.21\ (p < 0.001)}$

Table II.D.2: ANOVA results for diversity sensitivity of the best performing models to sparsity, purchase distribution, and item/user ratio

Model	*Sparsity*	*Purchase Distribution*	*Item/User Ratio*
CA/Item/Corr	$F_{1, 48} = 0.051$ (p = 0.82)	**$F_{2, 48} = 19.72$ (p < 0.001)**	**$F_{2, 48} = 438.35$ (p =**
None/Item/Corr	$F_{1, 48} = 1.01$ (p = 0.32)	**$F_{2, 48} = 3.44$ (p = 0.04)**	**$F_{2, 48} = 438.35$ (p <**
SVD/Item/Corr	$F_{1, 48} = 0.58$ (p = 0.452)	$F_{2, 48} = 2.09$ (p = 0.135)	**$F_{2, 48} = 926.95$ (p <**
None/Item/Jaccard	$F_{1, 48} = 1.84$ (p = 0.365)	$F_{2, 48} = 1.71$ (p = 0.191)	**$F_{2, 48} = 1063.35$ (p <**

Table II.D.3: ANOVA results for computation time sensitivity of the best performing models to sparsity, purchase distribution, and item/user ratio

Tables II.D.4 and II.D.5 indicate the differences, and their significance, across purchase distribution and item–user ratio levels.[9]

		$\gamma_{Exponential}$	γ_{Linear}	$\gamma_{Uniform}$
None/User/Jaccard	**t_{df}, p**	**$t_{23} = 5.13$, p < 0.001**		**$t_{23} = -4.95$, p < 0.001**
	M	5,891.47	1,654.42	5,480.56
	t_{df}, p	**$t_{34} = 0.21$, p = 0.83**		
CA/Item/Corr	**t_{df}, p**	$t_{30} = 1.53$, p = 0.14		$t_{30} = 0.39$, p = 0.70
	M	6,609.48	3,521.83	2,968.38
	t_{df}, p	$t_{25} = 1.98$, p = 0.06		
NMF/Item/Corr	**t_{df}, p**	$t_{30} = 1.30$, p = 0.21		$t_{34} = 0.02$, p = 0.98
	M	6,026.11	3,590.31	3,619.22
	t_{df}, p	$t_{29} = 1.32$, p = 0.20		
SVD/Item/Corr	**t_{df}, p**	$t_{31} = 1.10$, p = 0.28		$t_{30} = 0.83$, p = 0.41
	M	6,416.78	4,225.59	3,054.13
	t_{df}, p	$t_{25} = 1.88$, p = 0.07		

Table II.D.4: Means, t-values (df), and *p*-values related to the sensitivity of diversity to purchase distribution

[9] Although sparsity appears significant, it is not discussed in this appendix. The sparsity results are covered in Section 4.3.

		$\gamma_{iur} = 0.5$	$\gamma_{iur} = 1$	$\gamma_{iur} = 2$
None/User/Jaccard	t_{df}, p	$t_{19} = 2.50, p = 0.022$		$t_{20} = 2.58, p = 0.018$
	M	7,250.59	1,903.21	563.82
	t_{df}, p	$t_{17} = 3.21, p = 0.005$		
CA/Item/Corr	t_{df}, p	$t_{19} = 5.11, p < 0.001$		$t_{20} = 5.18, p < 0.001$
	M	10,045.16	2,440.11	614.46
	t_{df}, p	$t_{17} = 6.50, p < 0.001$		
NMF/Item/Corr	t_{df}, p	$t_{19} = 5.75, p < 0.001$		$t_{20} = 5.88, p < 0.001$
	M	10,112	2,490.16	633.48
	t_{df}, p	$t_{17} = 7.33, p < 0.001$		
SVD/Item/Corr	t_{df}, p	$t_{19} = 5.83, p < 0.001$		$t_{20} = 5.92, p < 0.001$
	M	10,408.55	2,603.39	684.56
	t_{df}, p	$t_{17} = 7.50, p < 0.001$		

Table II.D.5: Means, t-values (df), and *p*-values related to the sensitivity of diversity to item–user ratio

The statistical analyses of the sensitivity of computation time to the input characteristics are in Table II.D.6, which indicates the significant impact of item–user ratio on the performance of all four best performing models. The purchase distribution significantly influences only two best performing models. The pairwise t-tests, which serve to analyze the differences between levels, are inconclusive, so these results are not included in the appendix.

		$\gamma_{iur = 2}$	$\gamma_{iur = 1}$	$\gamma_{iur = 0.5}$
CA/Item/Corr	t_{df}, p	$t_{21} = -8.34, p < 0.001$		$t_{20} = -14.64, p < 0.001$
	M	115.26 sec	232.22 sec	841.52 sec
	t_{df}, p	$t_{17} = -17.84, p < 0.001$		
None/Item/Corr	t_{df}, p	$t_{24} = -9.53, p < 0.001$		$t_{25} = -27.93, p < 0.001$
	M	111.84 sec	260.14 sec	928.34 sec
	t_{df}, p	$t_{19} = -35.59, p < 0.001$		
SVD/Item/Corr	t_{df}, p	$t_{24} = -9.53, p < 0.001$		$t_{25} = -27.93, p < 0.001$
	M	110.49 sec	230.14 sec	781.61 sec
	t_{df}, p	$t_{19} = -35.59, p < 0.001$		
None/Item/Jaccard	t_{df}, p	$t_{21} = -12.23, p < 0.001$		$t_{25} = -29.30, p < 0.001$
	M	76.56 sec	213.65 sec	857.19 sec
	t_{df}, p	$t_{19} = -38.68, p < 0.001$		

Table II.D.6: Means, t-values (df), and *p*-values related to the sensitivity of computation time to item–user ratio

References

[1] H. Cramer, Mathematical Methods of Statistics, 1946, Princeton University Press, Princeton.

[2] J.P. Guilford, Psychometric methods, 1936, McGraw-Hill, New York.

[3] G.H. Golub, C.F.V. Loan, Matrix computations, 1996, 3 ed., Johns Hopkins University Press

[4] V. Madar, Direct formulation to Cholesky decomposition of a general nonsingular correlation matrix, 2015, Statistics & Probability Letters, 103 142-147.

[5] F. Leisch, A. Weingessel, K. Hornik, On the generation of correlated artificial binary data, 1998, SFB ``Adaptive Information Systems and Modeling in Economics and Management Science".

CHAPTER III

A DECISION SUPPORT SYSTEM TO EVALUATE RECOMMENDATIONS SYSTEMS COMBING MULTIPLE DATA SOURCES AND IDENTIFY FEATURE IMPORTANCE IN E-COMMERCE

CHAPTER III

A DECISION SUPPORT SYSTEM TO EVALUATE RECOMMENDATION SYSTEMS COMBINING MULTIPLE DATA SOURCES AND IDENTIFY FEATURE IMPORTANCE IN E-COMMERCE

Abstract

Recommendation systems help marketing decision makers construct personalization strategies for online consumers, who often are overwhelmed by the abundance of product choices available. To extend existing information systems literature on recommenders, this article proposes a decision support system to evaluate recommendation performance in settings with multiple data sources as input. Additionally, this study aims to open the recommendation system's black box by introducing a feature importance scoring framework into the recommendation system research. Finally, this study validates the impact of hybridization strategies (a posteriori weighting and feature combination) across four distinct data sources (product, customer, raw behavioral, and aggregated behavioral data) empirically, using eight real-life data sets obtained from a large, European e-commerce company. The empirical findings of the validation are fivefold. First, it is shown that raw behavioral and product data are the most predictive sources. Second, combining different data sources increases performance. Third, we show that the incremental return of adding additional data sources diminishes. Fourth, a factorization machine–based feature combination is preferred over a posteriori weighting as a hybridization technique for models with the optimal number of data sources. Finally, we deliver managerial implications based on the importance scores for both the various data sources and individual features that we reveal.

Keywords: E-commerce, Recommendation systems, Hybridization, Factorization machines, Feature importance

1 Introduction

In e-commerce settings, consumers confront an abundance of products. Imagine, for example, shopping for a sweater. An apparel web shop typically markets thousands of sweaters, so it is hard for an individual customer to scan every product to make a well-informed product evaluation. To help visitors cope with this information overload, websites often use recommendation systems that create personalized product sets, narrowing down the options for each customer and helping them explore less obvious products [1, 2]. Such recommendation

systems increase customer satisfaction [3] and also benefit the e-commerce company, because greater satisfaction leads to increased sales, revenue, and loyalty [4-6].

In prior literature [7], a widely used classification divides recommendation systems into three main types, according to the data sources they use [7-11]. First, demographic recommendation systems (DRS) rely on customer data (CD) and calculate the similarity between customers to propose products based on the actions of socio-demographically similar people [e.g. 12, 13]. Second, product data (PD) inform content-based recommendation systems (CBRS) [e.g. 12], which calculate the similarity between the products in which a customer has shown interest, such as by rating, viewing, or purchasing them, and other products that have similar characteristics. Products are proposed based on this product similarity. Third, in collaborative filtering recommendation systems (CFRS), product suggestions are based on behavioral data (BD), such that similarity levels are calculated on the basis of customer actions, such as explicit ratings, purchases, or views. Products are proposed to customers based on the products that were interesting to people who exhibit similar behavior [e.g. 14, 15].

All three recommendation systems offer specific advantages and disadvantages, as listed in Table III.1. For example, CFRS are known for their accurate predictions, adaptivity, and novel and serendipitous recommendations, but they are only scalable to a limited extent and often suffer from long-tail and cold-start problems for new users and new items. Although

	Collaborative Filtering (CF)	**Content-based (CB)**	**Demographic (D)**
Advantages	Accurate [10]		
	No metadata engineering needed [10]	No metadata engineering needed [10]	No metadata engineering needed [10]
	Adaptive [10]	Adaptive [10]	Adaptive [10]
	Serendipity and novelty in results [14]	Comparison between items possible [10]	Serendipity and novelty in results [10]
Shortcomings	Long tail problem [16]	Overspecialization [9]	Long tail problem[16]
	Cold Start for new users and items [9, 17]	Cold start for new users [10]	Cold start for new items [12]
	Scalability and Sparsity [15, 18]	Collection of product information [10]	Collection of customer information [10]

Table III.1: Advantages and Shortcomings of the three recommendation systems.

CBRS are adaptive and allow for comparisons across recommendations, they tend to be less accurate, suffer from overspecialization, and also have a new user problem. Finally, demographic systems are adaptive and capable of offering novel and serendipitous results, but they are less accurate, suffer from the long-tail problem, and offer cold starts for new items [9, 10].

Beyond the shortcomings in Table III.1, these recommenders share two major problems: They use single data sources as input and ignore other potentially interesting data sources, and they do not reveal which features drive their recommendations. To overcome the drawbacks of a single source and increase predictive performance, recommendation algorithms are hybridized [10], such that they combine different data sources. For example, combining a CFRS with a CBRS in a hybrid recommendation system could mitigate the cold-start problem for new items and the long-tail issue of CFRS, and the overspecialization problem of CBRS. This research study proposes a decision support system (DSS) that designs a framework to estimate, evaluate, and interpret recommendation systems based on different data source combination methods. Although the proposed DSS is able to incorporate every possible combination strategy, this study offers deeper insights into two specific hybridization methods: an a posteriori combination of different single data source recommendation algorithms and a feature combination approach that combines multiple data sources into a unified model, i.e. a factorization machine (FM) [19].

Hybrid recommenders have the clear advantage of combining data sources, but they still cannot resolve the interpretability issue of most single-source recommendation systems, as they remain black boxes [20]. In contrast to other analytical tools, it is not straightforward to assess the relative importance of features in the recommender. In a business setting, where recommender adoption often depends on management buy-in and alignment with the business logic, the ability of a particular technique to deliver such insights is important. Therefore, we leverage an approach to calculate feature importance scores that appears to have proven its added value in other analytical fields [e.g. 2, 21, 22].

By designing a DSS that overcomes the previously mentioned shortcomings, this study offers three key contributions. First, the proposed DSS suggests how to combine multiple data sources in hybrid recommendation systems. Second, the DSS introduces feature importance calculation into the recommendation literature, by including it in its evaluation and

interpretation step. Third, using an empirical validation of the proposed DSS, this study addresses five pertinent research questions related to the hybridization of recommendation systems in an e-commerce setting:

RQ1a. Do recommendation systems based on different single data sources differ in performance?

RQ1b. Does combining different data sources enhance predictive performance?

RQ1c. What is the optimal order in which to add data source groups to a recommendation system?

RQ2. Which hybridization technique performs best for recommendation models with the optimal number of data sources?

RQ3. Which are the most important predictors in the best performing recommendation model?

The remainder of this study is structured as follows: The next section reviews the literature on recommendation calculation, together with a brief overview of relevant evaluation and interpretation metrics. Section 3 describes and validates the DSS with an empirical case study using on eight real-life data sets from a European e-tailer. Finally, this study concludes with some implications and ideas for further research.

2 Literature Review

2.1 Recommendation Calculation

Recommendation algorithms lie at the heart of any recommendation system estimation and have been extensively investigated, prompting a variety of classifications [8]. The current study focuses on data source hybridization, so a distinction related to the recommendation calculation centers on the difference between single data source algorithms and hybrid algorithms that combine different data sources.

2.1.1 Single Data Source Algorithms

Single data source algorithms typically are categorized according to the type of data they use, as discussed briefly in the introduction. Additionally, prior literature also groups algorithms according to their filtering techniques [8]. An algorithm can be memory or model based. The former constructs a matrix of users' interests in items, then provides a

recommendation prediction based on a similarity calculation that may be user- or item-based. Equation 1 provides the formula for calculating user-based similarity [15], where $\hat{p}_{c,i}$ is the prediction of an item *i* for customer *c*; *K* stands for the set of nearest neighbors, or the set of customers most similar to customer *c* who rated item *i*; and *sim(c,k)* indicates the similarity between customer c and customer *k*, who is a customer in the neighborhood of *c*:

$$\hat{p}_{c,i} = \frac{1}{|K|} \sum_{k \epsilon K} sim(c,k)\, p_{k,i}\,. \tag{1}$$

Equation 2 is similar except that the prediction relies on items instead of users [23], such that *L* indicates the number of nearest neighbors, or items similar to item *i*, and *sim(i,l)* specifies the similarity between product *i* and *l*, which is a product in the neighborhood of *i*:

$$\hat{p}_{c,i} = \frac{1}{|L|} \sum_{l \epsilon L} sim(i,l)\, p_{c,l}\,. \tag{2}$$

Cosine and Pearson correlation [24] are commonly used similarity measures and can be plugged into Equations 1 and 2, as shown in Equations 3 and 4:

$$Cosine\ Similarity(x,y) = \frac{\sum_{i \in I_{xy}} r_{x,i}\, r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2}\sqrt{\sum_{i \in I_y} r_{y,i}^2}}, \text{ and} \tag{3}$$

$$Pearson\ Correlation\ Similarity(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \overline{r_x})\,(r_{y,i} - \overline{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \overline{r_x})^2 \sum_{i \in I_{xy}} (r_{y,i} - \overline{r_y})^2}}. \tag{4}$$

The interpretation of Equations 3 and 4 thus depends on the CF method used. In a user-based setting, $r_{x,i}$ and $r_{y,i}$ represent the ratings[10] of item *i* by customers *x* and *y*, respectively, and $\overline{r_x}$ and $\overline{r_y}$ represent the respective mean purchase rates of these customers. In an item-based setting, $r_{x,i}$ and $r_{y,i}$ represent the ratings of products *x* and *y*, respectively, by customer *i*, and $\overline{r_x}$ and $\overline{r_y}$ refer to the mean ratings of products *x* and *y*, respectively. Furthermore, I_x, I_y, and I_{xy} denote the set of products rated by customer *x*, customer *y*, and both customers *x* and *y*, respectively.

[10] Rating is used as general term and can be explicit or implicit (e.g., purchases, views, orders).

Model-based filtering techniques are engineered machine learning algorithms that can be applied to a single data source to calculate which items to recommend to a specific user [8]. Bayesian classifiers [25], neural networks [26], graph-based techniques [27, 28], and matrix factorization [29] are among the most widely used algorithms.

2.1.2 Hybrid Algorithms Combining Different Data Sources

An existing body of literature reviews hybrid algorithms that combine different data sources. Previous studies have suggested a combination of raw behavioral data (RBD) features, mainly explicit ratings, with PD or CD algorithms to mitigate the drawbacks of CFRS [e.g. 9, 12, 30]. Other studies investigate the combination of RBD with aggregated behavioral data (ABD) [e.g. 31, 32] or PD with CD [e.g. 12]. Some recent investigations also combine some specific RBD, whether explicit or implicit, with customer and product information [33]. However, the unique combination of all four identified data sources has not been investigated previously in a single study.

In addition to the hybridized data sources, the combination technique varies [9], reflecting three main categories:

1. Weighting: Implement different methods separately and combine their predictions [e.g. 12, 34].
2. Cascading: Incorporate some characteristics of recommendation model A into recommendation model B [e.g. 12, 35].
3. Feature Combination: Construct a general, unifying recommendation model that incorporates features of different models [e.g. 10, 19].

2.2 Evaluation and Interpretation

The evaluation and interpretation of recommendation systems are two important aspect of recommendation assessment. Nevertheless, a bias exists in the amount of research dedicated to them: Evaluation is widely investigated [e.g. 14], whereas interpretation remains poorly addressed in the recommendation literature.

2.2.1 Evaluation

The evaluation and comparison of the quality of different recommendation systems can be done for many different dimensions, each with specific evaluation metrics. Deciding which

metrics to use depends on the ultimate purpose and deployment of the recommendation system [14, 36]. Accuracy is an obvious evaluation factor and is extensively researched in literature. Accuracy comprises predictive metrics (e.g., mean absolute error [MAE], root squared mean error [RSME]), classification metrics (e.g., precision, recall, F1, AUC), and ranking metrics (e.g., Kendall's tau, Spearman's rho, NDPM) [14]. Beyond accuracy, other evaluation factors also might be important, depending on the deployment. Related to the case computation time, coverage, diversity, confidence, novelty, or serendipity might be relevant evaluation criteria [14].

2.2.2 Interpretation

Recommendation systems are often said to be black boxes, making it hard for researchers and professionals to gain insight into these models. In recommendation systems, techniques for gaining such insights are under-investigated, though a number of studies in general machine learning literature address techniques for model interpretation [21, 37].

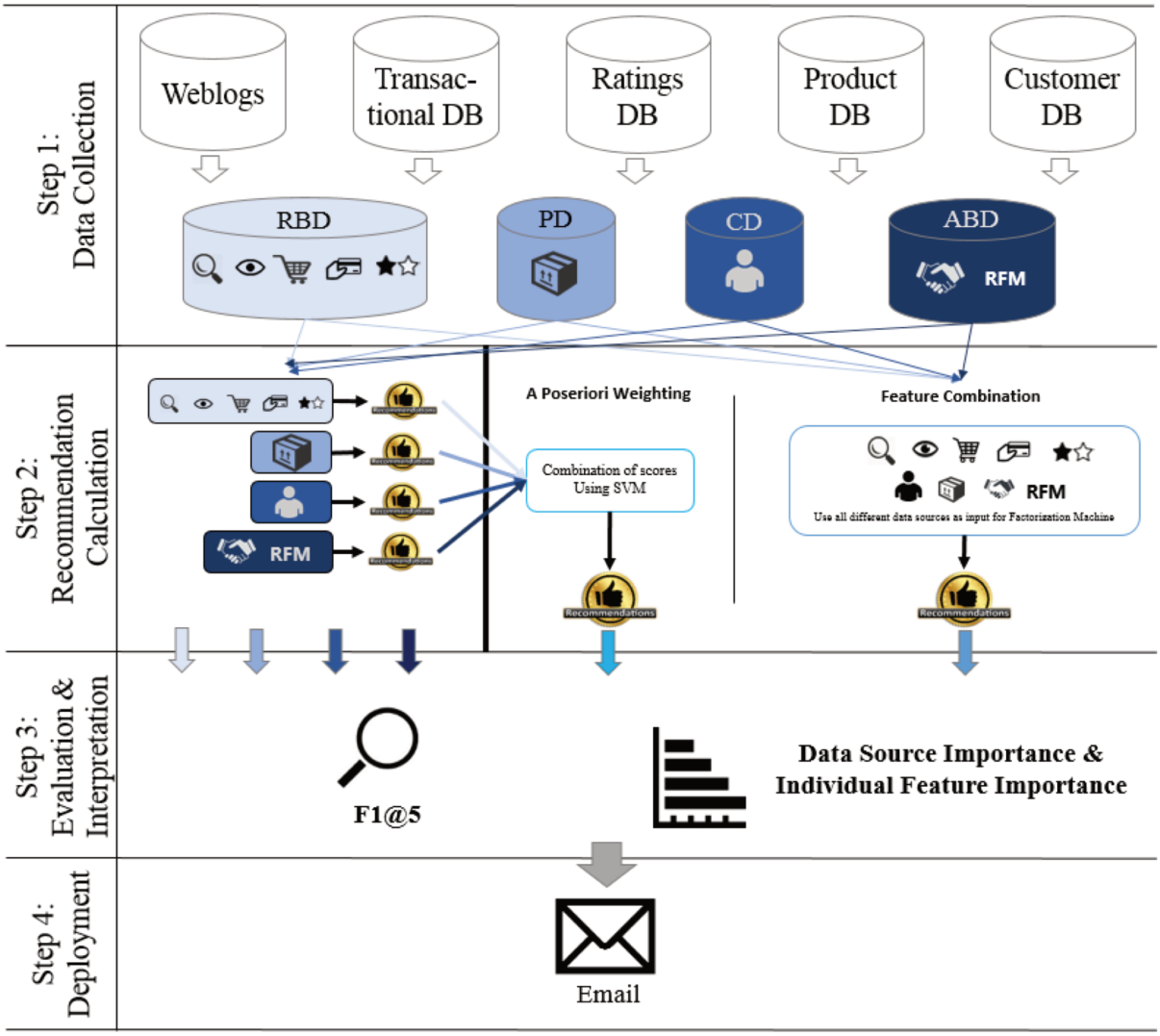
3 Evaluation of Data Source Combination and Interpretation of Feature Importance: An Empirical Decision Framework

Recommendation systems are deployed in many different domains to personalize offers for customers, and e-commerce is particularly notable in this regard [1]. This section proposes a DSS for evaluating data source combination and interpreting feature importance in an e-commerce context.

3.1 The Empirical Decision Framework

The proposed DSS in Figure III.1 consists of multiple steps. First, it collects data from weblogs and company databases that integrate information that serves as input for the recommendation systems. Second, the collected data sources provide input for the recommendation calculation process. In this step, single data source and hybrid algorithms, that exhaustively combine the different data sources, are applied to estimate the different recommendation systems. Third, the DSS evaluates the estimated recommendation systems in terms of an appropriate metric and interprets the models according to feature importance. Fourth, the DSS produces an overview of different ways to deploy the evaluated recommendation systems.

Figure III.1: Empirical DSS.



3.1.1 *Step 1: Data Collection*

The recommendation system first needs to collect data. E-commerce companies store an enormous amount of data in a data warehouse, usually organized into different databases [38]. For example, weblogs typically contain the logs of customer actions, like clicking and viewing behavior, in a web shop [39]. Transactional databases contain the history of customer transactions, such as purchases. Ratings databases provide the explicit rating of customers have offered over time. Product and customer databases consist of up-to-date information about the products, together with their characteristics, and socio-demographic information about customers. These databases are input sources for recommendation systems.

This study leverages data from a large European e-commerce company. The company is active in different markets and sells products in different product categories, so eight distinct

data sets are available, reflecting eight product categories: shoes, children's clothing, decoration, lingerie, furniture, women's clothing, men's clothing, and household linens. An overview of the number of users and items is in Table III.2.

Product Category	Users	Items
Shoes	31,536	11,712
Children's Clothing	16,752	3,956
Decoration	12,747	5,054
Lingerie	11,672	3,514
Furniture	20,507	6,481
Men's Clothing	8,412	4,737
Women's Clothing	50,336	12,979
Household Linens	12,376	2,934

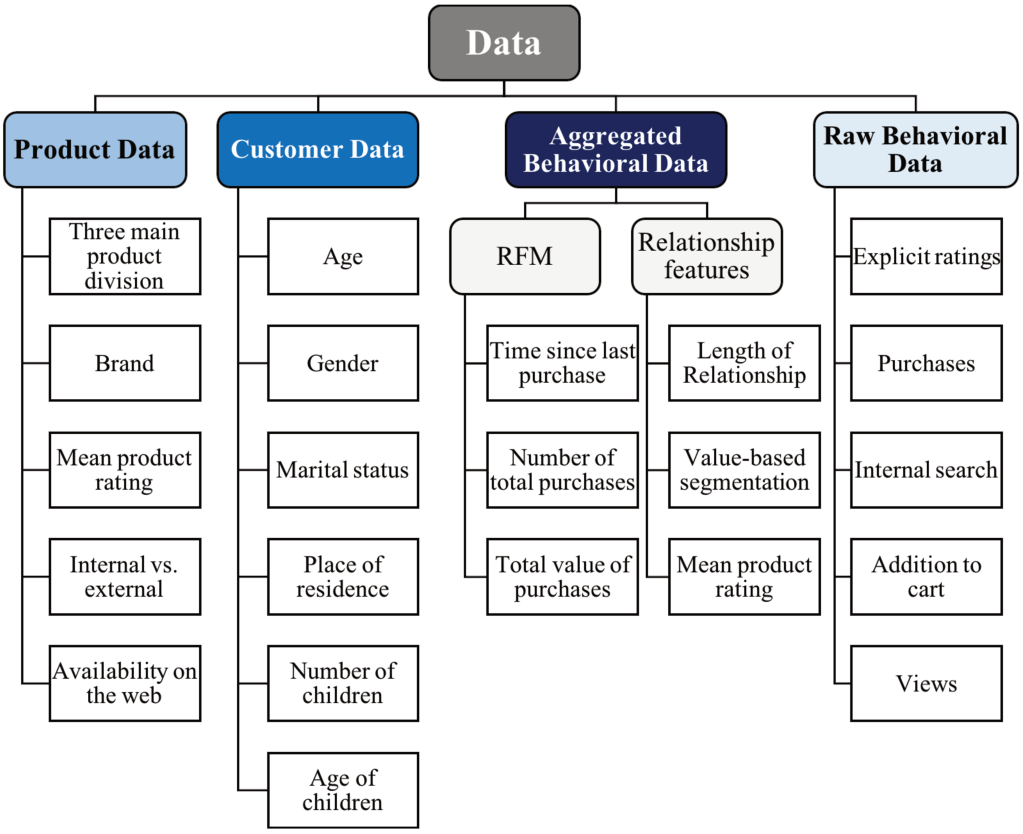
Table III.2: Number of visitors and products in different product categories.

Each data set is hierarchical and contains 23 individual features, divided into four data sources: PD, CD, ABD, and RBD. Additionally, ABD consists of two sub-data sources, namely, recency, frequency, and monetary value (RFM) variables and relationship data. Figure III.2 shows the data structure. The different data sources are discussed below.

Product Data (PD). The first features within PD are the three main product divisions, reflecting catalog information about the products. Depending on the product categories, as discussed in the experimental setup, the interpretation of the product division could differ. For example a wooden garden chair in the furniture category consists of (1) chair, (2) garden, and (3) wood. A women's volleyball shoe could have the divisions (1) sport, (2) women, and (3) volleyball. An indication of the brand also is given, together with a mean product rating. This latter feature is a proxy for popularity. The higher the mean explicit product rating of a product, the better overall ratings it gets. Finally, internal versus external and availability on the web are two features that indicate the product's origin (e.g., house brand vs. external brand) and availability in the web shop.

Customer Data (CD). The CD group six traditional socio-demographic features: age, gender, marital status, place of residence, number of children, and age of children.

Figure III.2:Data structure.



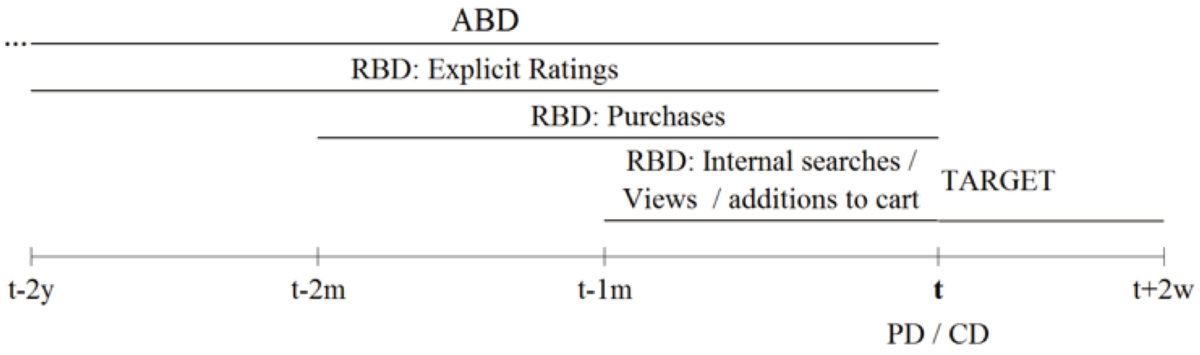
Raw Behavioral Data (RBD). The RBD source is a collection of data sources, frequently used in recommendation systems and especially in collaborative filtering. Prior literature divides it into two main classes: explicit or implicit ratings [57], each of which has advantages and disadvantages. Explicit ratings (e.g., 5 stars) are provided directly by customers, so they offer a clear signal of customer interest [41]. However, explicit systems demand user effort, time, and cost [41]. Moreover, in online retail settings, which are characterized by broad, deep, and fast-changing product offerings, such feedback is often hard to collect in sufficient amounts. In contrast, implicit information does not require direct user feedback but instead derives input from user behavior. The collection of implicit feedback is objective and non-intrusive, and this form of data is readily available in customer databases [57]. This study therefore incorporates both explicit information and four types of implicit information: purchases, internal searches, additions to the cart, and views.

Aggregated Behavioral Data (ABD). This data source falls between RBD and CD. Features in this group are based on behavioral data but aggregated to the customer level.

Whereas purchases in the RBD group indicate a vector of purchases for each customer, ABD gives just one aggregated value. The RFM sub-group contains aggregated features related to purchases, namely, the time since last purchase, number of total purchases, and total value of purchases. The relationship variables also can be aggregated, to offer indications of the length of relationship, value-based segmentation (using an internal analysis in the company), and mean product rating. The latter feature could be a proxy for customer attitude, because it indicates the mean rating a customer gives to products. A higher mean rating might indicate a more positive attitude toward the products.

Figure III.3 shows the data collection timeline for the different data sources. In particular, PD and CD are up to date and collected at time *t*. For ABD and RBD, the collection history varies depending on the feature. That is, for ABD it depends on the customer and is equal to the length of the customer's relationship with the company, whereas for RBD, explicit ratings are collected for two years, purchase information for two months, and internal searches, views, and additions to the cart for a period of one month. Purchases over a period of two weeks (t – t + 2w) offers a target for the creation of model-based systems and model evaluations.

Figure III.3: Data collection timeline.



3.1.2 Step 2: Recommendation Calculation

The second step in the proposed DSS is the calculation of recommendations based on the collected data. The framework allows for the implementation of both single data source and hybrid models. First, for the single data source models, the PD, CD, and ABD sources use content-based, demographic, and aggregated demographic models, respectively, estimated on the basis of memory-based filtering. The models estimate user-based similarity based on the cosine measure and kNN. Recommendation scores for a specific customer are calculated according to the products viewed by similar customers for CD and ABD and products similar

to the ones viewed by the specific customer for PD. For RBD – explicit rating, purchases, internal search, addition to the cart, and views – the memory-based filtering procedure is adapted because of the larger scale of the input matrix. Data reduction, in the form of correspondence analysis [40], initially is applied to make the input matrix denser and smaller, leading to improved efficiency and accuracy [41]. Then item-based similarity is applied, instead of user-based similarity, because it achieves better performance in settings with high user-to-item ratios [14].

Second, for the empirical validation of the proposed DSS, four data sources are considered and exhaustively combined using two alternative, well-researched hybridization techniques: (1) a posteriori combination of predictions, also known as weighting [10], and (2) constructing a unified recommendation model, known as feature combination [10]. The a posteriori weighting approach leverages recommendation systems based on single data sources and combines their scores a posteriori by means of a support vector machine (SVM). The scores of the four single data source recommendation models are estimated and used as input for an SVM for classification, with purchases as the target and is defined as follows:

$$Weighting = SVM(p^D, p^{CB}, p^{CF}, p^{ACF}), \tag{5}$$

where p^D, p^{CB}, p^{CF}, p^{ACF} represents respectively the CF recommendation scores of the demographic-, content-based -, CF -, and aggregated CF memory-based recommendation model. For generalizability and scalability, a linear SVM [42, 43] based on SGD optimization [44] is trained. The regularization parameter is set to 0.01, and L2 regularization is applied because of the differentiability [43]. In contrast, the feature combination approach involves the use of an algorithm that accommodates four different data sources at once and thus realizes hybridization during the estimation of the recommender system. The feature combination technique applied in this study is factorization machines (FM) [19], a model-based technique that works well in situations with mixed data sources, especially for combinations of RBD with other information. Good examples of the use of mixed data sources in FM include context-aware recommendation systems [45-48] and predictions of tweet interactions [49, 50].

The FM technique [19] is based on SVMs and factorization algorithms and combines the advantages of both. Like SVMs, FMs are general predictors, so they work with any real valued feature vector. In contrast though, FMs can estimate interactions, even for problems with huge sparsity like recommender systems, where SVMs fail. The main reason for this interesting property is that feature interaction is calculated on the basis of factorized parameters, so it is possible to calculate the recommendation model in linear time. In contrast with other factorization recommendation algorithms like SVD++ [29], FM works with any real-valued feature vector, creating an opportunity to include different data types. This study adds to FM literature by empirically investigating whether a factorization machine including different RBD, CD, PD, and ABD data sources also outperforms a posteriori weighting of single data source recommendation algorithms. The general model equation of an FM of degree 2 is formulated as follows:

where x_i represent the predictors that are features of the PD, CD, RBD, and/or ABD data sources; w_0 is the global bias; w_i are the parameters related to the first-degree effect of the n

$$Feature\ Combination\ FM = w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} \sum_{j=i=1}^{n} \langle v_i, v_j \rangle x_i x_j, \qquad (6)$$

predictors; and $\langle v_i, v_j \rangle$ represents the dot product of v_i and v_j, such that these two terms are low dimensionality vector representations of rank k of the predictors x_i and x_j. The FM's parameters are set similar to those of the SVM. Concretely, a FM for classification with L2 regularization, a regularization parameter of 0.01 based on SGD optimization results, and 30 latent factors are retained in the recommendation model.

3.1.3 Step 3: Evaluation & Interpretation

After having calculated the recommendations, step 3 evaluates and interprets the results. The general DSS leaves room for different evaluation and interpretation metrics and procedures. This section makes both evaluation and interpretation actionable, in light of the empirical validation. In this study accuracy is evaluated by means of ranked classification accuracy. The items in the top m, based on a descending ordering of the recommendation scores, are considered as true prediction, while all other products are considered negative predictions. The results of the predictions for each recommendation size (RS) are based on the confusion matrix and expressed in terms of the $F1$ measure [17]. $F1$ is chosen as evaluation

because it is described as the harmonic mean of recall, and precision [41], consequently giving an indication of both completeness and exactness. Additionally, the true negative element (tn) of the confusion matrix is not a part of the formula to calculate and *F*1. This is important because the binary input matrix is very sparse and consequently tn would be very high, distorting the evaluation metric. Equation 7 shows the formula to calculate *F*1:

$$F1@m = 2 * \frac{precision@m * recall@m}{precision@m + recoll@m} = 2 * \frac{\left({}^{tp@m}/_{tp@m+fp@m}\right) * \left({}^{tp@m}/_{tp@m+fn@m}\right)}{\left({}^{tp@m}/_{tp@m+fp@m}\right) + \left({}^{tp@m}/_{tp@m+fn@m}\right)}, \quad (7)$$

where *m* is the number of recommended products considered, and

tp@m = Number products in the top *m* of recommended products and purchased by the customer in the evaluation period.

fp@m = Number products in the top *m* of recommended products but not purchased by the customer in the evaluation period.

fn@m = Number of products not in the top *m* of recommended products but purchased by the customer in the evaluation period.

We follow the advice of the e-commerce company and evaluate the recommendation systems based on classification accuracy for five recommendations. Therefore, *m* is set to 5, and the different recommendation systems are evaluated in terms of F1@5. For each customer, an F1@5 value, obtained as described in Equation 6, is calculated. This customer-specific metric then can be averaged over all customers to obtain an average F1@5.

Next, the DSS and models are evaluated using multiple data sets, which increases the validity of the results but requires an adapted approach for statistical testing. Accordingly, the results are analyzed by means of Friedman's aligned rank test [52], allowing for significance and post hoc tests for ranking in a multi–data set setting. Li's procedure with a 95% confidence interval is used for the post hoc testing [52].

To open the black box model generated by different recommendation techniques, the DSS introduces feature importance scores into a recommendation system setting. In black box models, it is not straightforward to identify feature importance, but procedures are developed

for artificial neural networks [21, 22], SVMs [37, 53], and random forests [2, 54]. A well-established procedure to identify feature importance is to calculate the decrease in accuracy of an algorithm when a feature is randomly permutated [2, 54]. The feature importance of the permuted feature then is equal to the decrease in accuracy.

Different algorithms rely on different evaluation metrics to assess decreases in predictive performance. For random forests, it is appropriate to calculate, for example, the mean decrease in impurity or mean decrease in the Gini coefficient [54]. For other classification methods, like SVMs, the decrease in AUC or false positive classification might be calculated instead [55].

For this study, the main evaluation metric for the recommendation systems is F1@5. Therefore, the same procedure is adapted and deployed to measure the feature importance of data sources and individual features of different recommendation systems in terms of F1@5. Equation 8 shows how the feature importance scores are calculated for each feature or data source:

$$FeatImp^{i} = \frac{F1@5_{Full} - F1@5^{i}_{Random\ permutation}}{F1@5_{Full}}, \quad (8)$$

where *i* indicates the feature or data source that is randomly set, $F1@5_{Full}$ is the F1@5 measure for the model including all the features, and $F1@5^{i}_{Random\ permutation}$ indicates the F1@5 measure for the model for which feature *i* is randomly set. Using these feature importance scores, it is possible to calculate an aggregated feature importance score over all data sets by averaging the values, as in Equation 9:

$$FeatImp^{i}_{aggr} = \frac{\sum_{1}^{d} FeatImp^{i}}{d}, \quad (9)$$

where d indicates the number of data sets or product categories.

3.1.4 Step 4: Deployment

Depending on the domain and case, different applications might be suitable [56]. Deployments can vary from applications that require a large amount of products to be ordered and personalized to those for which only a few products need to be recommended. The organization of product listings in a webshop is a good example of an application in which many products need to be ordered. Personalized email campaigns and recommendation zones on a website are on the other hand excellent examples of cases where only a limited number of

products need to be recommended. The final use of the recommendation system in this study is to propose five interesting products, based on PD, CD, ABD, and RBD, to a customer.

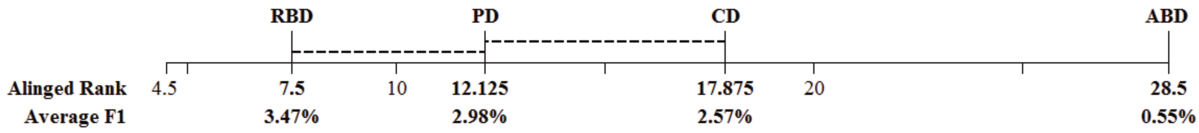
3.2 Empirical Results

To analyze the recommendation evaluation and interpretation process, this section deals specifically with the research questions for the focal case study. It thereby provides insights into the added value of combining different data sources, the optimal hybridization technique, and the important data sources and individual features in the best performing recommendation model.

3.2.1 RQ1a. Do recommendation systems based on different single data sources differ in performance?

The first research question investigates the difference in performance across single data source recommendation models. To answer this research question, for every data source (PD, CD, ABD, and RBD), the F1 scores were averaged by data source over both hybridization approaches (weighting and FM) on the eight data sets. The four average F1 scores constitute the input for the Friedman aligned rank test. The results indicate a significant difference between models with different data sources ($T_3 = 17.69, p < 0.001$). The post hoc results in Figure 3 reveal that the aligned ranks range between 4.5 and 28.5. The minimal value of 4.5 indicates that the data source is optimal in all data sets; the maximum value of 28.5 signals the worst single data source in all data sets. The average F1@5 scores are also plotted below the aligned ranks. Here, RBD is the most predictive data source, followed by PD and CD, and ABD is least predictive in every data set. Figure III.4 also indicates that the difference between RBD and PD and between PD and CD is not significant, as depicted by the dashed lines. All other differences are significant.

Figure III.4: Post hoc test results for different single data source models.

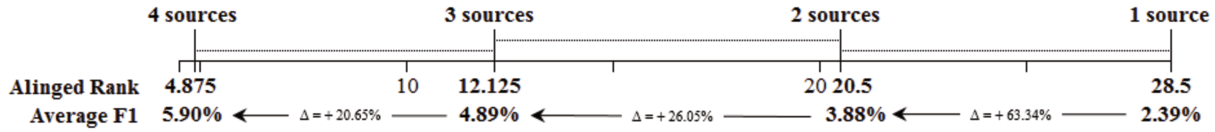


Notes: A lower aligned ranking and a higher average F1@5 indicate better recommendation performance. Dashed lines indicate a non-significant difference, and all other difference are significant (α = 5%).

3.2.2 *RQ1b. Does combining different data sources enhance predictive performance?*

After identifying the most predictive single data source, the next consideration is determining the value of adding further sources. To create the treatments to address this research question, the current study averages all single data source, all two data source, all three data source, and all four data source recommendation models. The results indicate a significant difference in performance among the models with different numbers of data sources ($T_3 = 22.23, p < 0.001$). The results of Li's procedure for post hoc testing in Figure III.5 indicate that more data sources lead to higher predictive performance in terms of aligned ranks and average F1@5. The pattern is clear, yet the impact of adding one extra data source is only marginally significant, indicated by the dotted lines. Figure III.5 also features arrows, indicating the percentage increase in F1@5 associated with adding an extra data source.

Figure III.5: Post hoc test results for algorithms with different numbers of data sources.

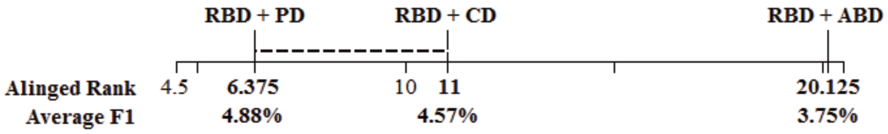


Notes: A lower aligned ranking and a higher average F1@5 indicate better recommendation performance. Dotted lines indicate a marginally significant difference (α = 10%), and all other differences are significant (α = 5%). The arrows indicate the percentage increase in average F1@5.

3.2.3 *RQ1c. What is the optimal order in which to add data sources to a recommendation system?*

Combining data sources adds value. But what is the optimal order for the hybridization procedure? If users have only limited time to develop a recommendation system, which data sources should they include first, second, third, and last? The results for RQ1a demonstrate that RBD is the best single data source, so it should be the first to focus on. Then, to determine the most interesting data source to combine with the RBD, this study ran Friedman aligned rank tests for between RBD + PD, RBD + CD, and RBD + ABD across the eight data sets. Figure III.6 and the related statistics ($T_2 = 11.35, p = 0.0034$) show that it is worthwhile to consider adding a second data source. Although it does not significantly outperform RBD + CD, the best option is to combine RBD + PD as the next step.

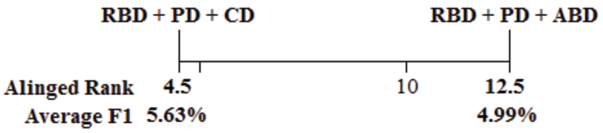
Figure III. 6: Post hoc test results to identify the best second data source to combine with RBD.



Notes: A lower aligned rank and higher average F1@5 indicate better recommendation performance. Dashed lines indicate a non-significant difference, and all other differences are significant (α = 5%).

Figure III.7 and the related statistics ($T_1 = 6.02, p = 0.0141$) show that it is then best to add CD as the third data source; ABD is left as the last data source to integrate.

Figure III.7: Friedman test results to identify the best third data source to combine with RBD and PD.

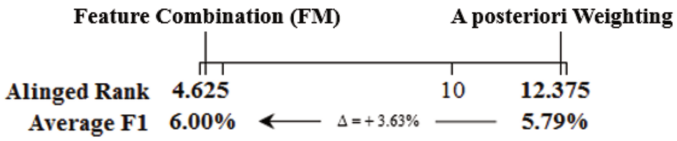


Notes: A lower aligned ranking and higher average F1@5 indicate better recommendation performance.

3.2.4 *RQ2. Which hybridization technique performs best for recommendation models with the optimal number of data sources?*

Combining all four data sources results in the most predictive recommendation system, but the resolution of RQ1b does not distinguish among hybridization techniques. To evaluate the optimal hybridization technique (a posteriori combining versus feature combination using FM, both based on all four data sources), a Friedman aligned rank test was executed. Figure III.8 and the statistics ($T_1 = 5.65, p = 0.0174$) indicate that feature combination significantly outperforms an a posteriori weighting of single data source recommendation models when combining the four proposed data sources. The F1 increases on average by 3.63% when deploying feature combination rather than weighting.

Figure III.8: Friedman test results comparing feature combination (FM) and a posteriori weighting for recommendation models with four data sources.

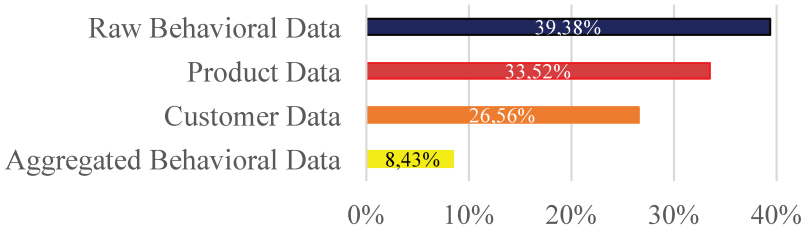


Notes: The arrow indicates the difference in average F1@5.

3.2.5 3.RQ3. *Which are the most important predictors in the best performing recommendation model?*

The previous results indicate that a feature combination based on a FM of all four data sources results in the most predictive recommendation system. To open the black box and look inside the best recommendation model, and thereby find out which predictors are most important, this study leverages the hierarchical data structure to analyze predictor importance on different levels: data sources (PD, CD, ABD, and RBD) and the individual feature level. The aggregated importance score ($FeatImp^{i}_{aggr}$) of each data source can calculated, with the results in Figure III.9.

Figure III. 9: Aggregated data source importance scores.



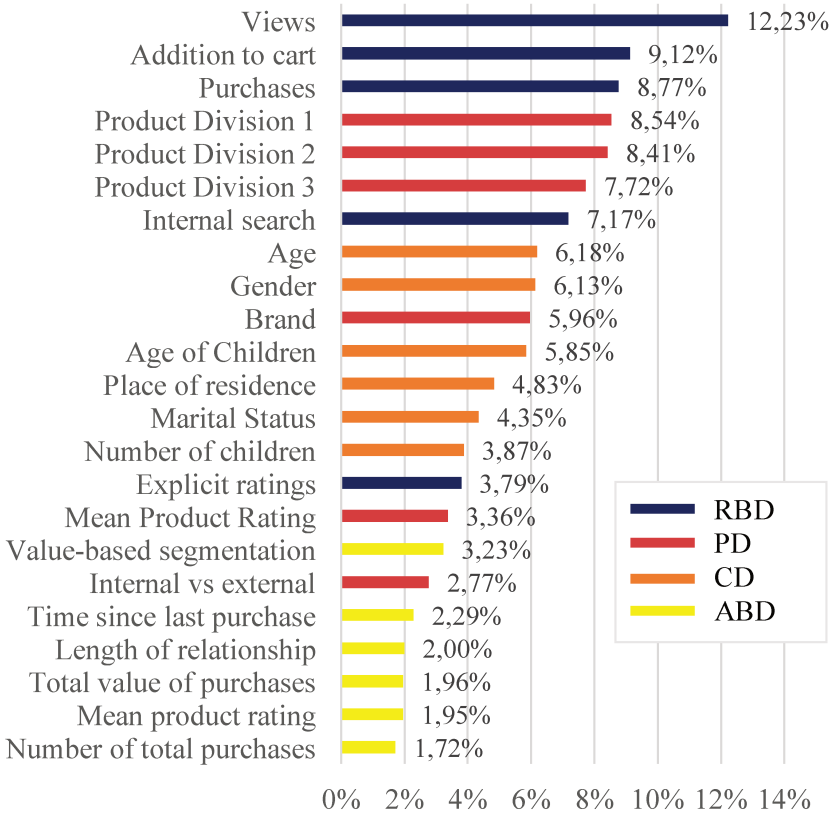
That is, RBD is the most important data source, followed by PD, CD, and finally ABD. A difference also arises across product categories. As Table III.3 shows, for six product categories (furniture, women's clothing, men's clothing, shoes, household linens, and lingerie), the aggregated order of data source importance scores persists. However, for children's clothing, CD exhibits the greatest importance, followed by RBD, PD, and ABD. In the decoration category, PD is the most important data source, followed by RBD, CD, and ABD. Business results validate the logic of these findings.

	$FeatImp^{i}$			
	RBD	**PD**	**CD**	**ABD**
Furniture	46.14%	33.92%	23.92%	9.47%
Children's Clothing	33.93%	28.05%	36.68%	8.44%
Women's Clothing	44.08%	37.21%	17.98%	14.41%
Men's Clothing	44.93%	27.68%	24.16%	9.72%
Shoes	33.89%	32.70%	31.31%	9.33%
Decoration	34.72%	35.65%	30.38%	5.62%
Household Linens	39.36%	37.51%	20.84%	6.03%
Lingerie	38.00%	35.46%	27.24%	4.45%

Table III.3: Data source importance scores per product category.

After identifying the data source importance scores, it is possible to open the black box further by looking at the importance scores of individual features, as listed in Figure III.10.

Figure III.10: Feature importance scores.



Not surprisingly, the three most important features are RBD features. Behavioral data about views, additions to shopping carts, and purchases are critical. In contrast, explicit ratings do not appear to have a top importance, mainly because the data sets contain only a limited number of explicit ratings. The final RBD feature is internal searches, which is the seventh most important feature.

Next, three PD variables, related to product divisions, are important as well. This indicates that recommending products in the same product division results in good predictive results; e.g. suggesting a dress to customers who already have viewed similar outfits will produce relatively good F1@5 scores. Brand is identified as the tenth most important variable. In contrast, mean product rating and internal vs. external product only appear in the second half of the figure. Furthermore, the six CD features exhibit average importance scores and can be found between places eight and fourteen in Figure 9. Finally, the individual ABD features, as

part of the least important data sources, all can be found in the lower end of the figure. Value-based segmentation occupies the seventeenth place and is the most important ABD feature. All other ABD feature have even lower importance scores.

4 Conclusion

This study investigates the personalization strategy provided by a hybrid recommendation systems that e-commerce companies can use. In addition to constructing a DSS for optimizing the hybridization process of data sources, this study and the proposed DSS introduce feature importance into recommendation literature. The validation of this DSS with eight data sets from a European e-commerce company produces five distinct findings. First, RBD and PD are the most predictive sources; companies should focus their efforts to create recommendation systems primarily on these two data sources. Second, combining data sources adds value, and more data sources lead to higher predictive performance. Third, despite the higher predictive performance of recommendation models with four data sources, if a company lacks the ability to invest in all four sources, it can concentrate its efforts. It should do so in the following order: RBD, PD, CD, and then ABD. Fourth, this study suggests using feature combination based on FM for the optimal combination of all four data sources. This technique outperforms an a posteriori weighting of different single data source recommendation models. Fifth and finally, the accuracy of a recommendation system is very important, but beyond having a highly predictive recommendation model, it also insightful to open the black box to determine which data sources and features contribute to recommendation success. According to the current study findings, RBD contributes most to the model, with a data source importance score of 39.38%, followed by PD and CD with importance scores of 33.52% and 26.56%, respectively. Finally ABD is the least important data source, with a feature importance score of only 8.43%.

In terms of the importance of individual features, implicit RBD features are very important, so keeping log data from the e-commerce site is vital. Explicit ratings are less important, mainly because this information is only available in smaller amounts. If a business model does not thrive on ratings (like e.g., Netflix, LastFM), explicit ratings are less important to consider. Furthermore, PD is important information to gather, especially product division and brand data. Although somewhat less important, individual CD features can add value to recommendation systems. Finally, ABD features have relatively little importance and can be less emphasized, if the time and resources available to create recommendation systems are limited.

This study advocates an FM-based combination of data sources, due to the higher accuracy and the ability to mitigate issues related to single data source recommendation models. Although this technique reveals the positive influence of hybridization on accuracy (F1), other aspects that becoming increasingly important in recommendation systems are not investigated herein. Further research could extend the proposed DSS by investigating the added value of data source combination and feature importance in terms of other metrics. The effects on computation time, diversity, novelty, serendipity, and trust also might be analyzed. Moving beyond statistical metrics, a field test might be executed to test the DSS according to business metrics such as the conversion rate and incremental revenue.

References

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

[2] L. Breiman, Random forests, 2001, Mach. Learn., 45 5-32.

[3] Y. Jiang, J. Shang, Y. Liu, Maximizing customer satisfaction through an online recommendation system: A novel associative classification model, 2010, Decis. Support Syst., 48 470-479.

[4] K.-W. Cheung, J.T. Kwok, M.H. Law, K.-C. Tsui, Mining customer product ratings for personalized marketing, 2003, Decis. Support Syst., 35 231-243.

[5] S.M. Weiss, N. Indurkhya, Lightweight collaborative filtering method for binary-encoded data, in: L. De Raedt, A. Siebes (Eds.) Principles of data mining and knowledge discovery, 2001, Springer Berlin Heidelberg, pp. 484-491.

[6] V.Y. Yoon, R.E. Hostler, Z. Guo, T. Guimaraes, Assessing the moderating effect of consumer product knowledge and online shopping experience on using recommendation agents for customer loyalty, 2013, Decis. Support Syst., 55 883-893.

[7] T.-P. Liang, Recommendation systems for decision support: An editorial introduction, 2008, Decis. Support Syst., 45 385-386.

[8] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey, 2013, Knowl.-Based Syst., 46 109-132.

[9] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, 2005, IEEE Trans. Knowl. Data Eng., 17 734-749.

[10] R. Burke, Hybrid recommender systems: Survey and experiments, 2002, User Modeling and User-Adapted Interaction, 12 331-370.

[11] Y.-M. Li, C.-T. Wu, C.-Y. Lai, A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship, 2013, Decis. Support Syst., 55 740-752.

[12] M.J. Pazzani, A framework for collaborative, content-based and demographic filtering, 1999, Artif. Intell. Rev., 13, 393-408.

[13] C. Porcel, A. Tejeda-Lorente, M.A. Martinez, E. Herrera-Viedma, A hybrid recommender system for the selective dissemination of research resources in a technology transfer office, 2012,Inform. Sciences, 184, 1-19.

[14] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst., 22 (2004) 5-53.

[15] B.M. Sarwar, G. Karypis, J.A. Konstan, J.T. Riedl, Analysis of recommendation algorithms for e-commerce, 2000, 2nd ACM Conference on Electronic Commerce, ACM, Minneapolis, Minnesota, pp. 158-167.

[16] T.C.-K. Huang, Y.-L. Chen, M.-C. Chen, A novel recommendation model with Google similarity, 2016, Decis. Support Syst., 89 17-27.

[17] H.-N. Kim, A. El-Saddik, G.-S. Jo, Collaborative error-reflected models for cold-start recommender systems, 2011, Decis. Support Syst., 51, 519-531.

[18] C. Jiang, R. Duan, H.K. Jain, S. Liu, K. Liang, Hybrid collaborative filtering for high-involvement products: A solution to opinion sparsity and dynamics, 2015, Decis. Support Syst., 79, 195-208.

[19] S. Rendle, Factorization Machines, 2010, IEEE International Conference on Data Mining, Sydney, Australia.

[20] S. Dooms, Dynamic generation of personalized hybrid recommender systems, 2013, 7th ACM Conference on Recommender Systems, ACM, Hong Kong, China, pp. 443-446.

[21] M. Gevrey, I. Dimopoulos, S. Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, 2003, Ecol. Modell., 160 249-264.

[22] J.D. Olden, M.K. Joy, R.G. Death, An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, 2004, Ecol. Modell., 178, 389-397.

[23] M. Deshpande, G. Karypis, Item-based top-N recommendation algorithms, 2004 ACM Trans. Inf. Syst., 22, 143-177.

[24] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, 1998, 14th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Madison, WI, pp. 43-52.

[25] M.-H. Park, J.-H. Hong, S.-B. Cho, Location-based recommendation system using Bayesian user's preference model in mobile devices, 4th international conference on Ubiquitous Intelligence and Computing, Springer-Verlag, Hong Kong, China, 2007, pp. 1130-1139.

[26] T.H. Roh, K.J. Oh, I. Han, The collaborative filtering recommendation based on SOM cluster-indexing CBR, 2003, Expert Syst. Appl., 25, 413-423.

[27] K. Dutta, D. VanderMeer, A. Datta, P. Keskinocak, K. Ramamritham, A fast method for discovering critical edge sequences in e-commerce catalogs, 2007, Eur. J. Oper. Res., 181, 855-871.

[28] Y. Wang, W. Dai, Y. Yuan, Website browsing aid: A navigation graph-based recommendation system, 2008, Decis. Support Syst., 45, 387-400.

[29] Y. Koren, Factorization meets the neighborhood: A multifaceted collaborative filtering model, 2008, 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Las Vegas, NV, pp. 426-434.

[30] M.G. Vozalis, K.G. Margaritis, Using SVD and demographic data for the enhancement of generalized collaborative filtering, 2007, Inform. Sciences, 177, 3017-3037.

[31] Y.-Y. Shih, D.-R. Liu, Product recommendation approaches: Collaborative filtering via customer lifetime value and customer demands, 2008, Expert Syst. Appl., 35, 350-360.

[32] A. Albadvi, M. Shahbazi, Integrating rating-based collaborative filtering with customer lifetime value: New product recommendation technique, 2010, Intell. Data Anal., 14, 143-155.

[33] A. Said, S. Dooms, B. Loni, D. Tikk, Recommender systems challenge 2014, 2014, 8th ACM Conference on Recommender Systems, ACM, Foster City,CA, pp. 387-388.

[34] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, M. Sartin, Combining content-based and collaborative filters in an online newspaper, 1999, SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, CA.

[35] M. Balabanovi, Y. Shoham, 1997, Fab: content-based, collaborative recommendation, Commun. ACM, 40, 66-72.

[36] D. Geiger, M. Schader, Personalized task recommendation in crowdsourcing information systems — Current state of the art, 2014, Decis. Support Syst., 65, 3-16.

[37] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, 2003, J. Mach. Learn. Res., 3, 1157-1182.

[38] I.-Y. Song, Database design for real-world e-commerce systems, 2000, IEEE Data Engineering Bulletin, 23, 23-28.

[39] R. Kohavi, L. Mason, R. Parekh, Z. Zheng, Lessons and challenges from mining retail e-commerce data, 2004, Mach. Learn., 57, 83-113.

[40] S. Geuens, K. Coussement, K.W. De Bock, Evaluating collaborative filtering: Methods within a binary purchase setting, 2014, European Conference on Machine Learning (ECML), Nancy, France, pp. 81-90.

[41] M. Kellar, C. Watters, J. Duffy, M. Shepard, Effect of task on time spent reading as an implicit measure of interest, 2004, 67th Asis&T Annual Meeting, Medford: Information Today Inc, Providence, RI, pp. 168-175.

[42] C. Jin, L. Wang, Dimensionality dependent PAC-Bayes margin bound, in: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.) NIPS, Lake Tahoe, NV, 2012, pp. 1043-1051.

[43] Y. Tang, Deep learning using linear support vector machines, 2013, International Conference on Machine Learning, Atlanta, GA.

[44] Z. Wang, K. Crammer, S. Vucetic, Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale SVM training, 2012, J. Mach. Learn. Res., 13, 3103-3131.

[45] C. Cheng, F. Xia, T. Zhang, I. King, M.R. Lyu, Gradient boosting factorization machines, ACM Conference on Recommender Systems, 2014, ACM, Silicon Valley, CA, pp. 265-272.

[46] C.-M. Chen, M.-F. Tsai, J.-Y. Liu, Y.-H. Yang, Music recommendation based on multiple contextual similarity information, IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) , 2013, IEEE Computer Society, Atlanta, GA, pp. 65-72.

[47] S. Rendle, Z. Gantner, C. Freudenthaler, L. Schmidt-Thieme, Fast context-aware recommendations with factorization machines, 2011, 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, Beijing, China, pp. 635-644.

[48] T.V. Nguyen, A. Karatzoglou, L. Baltrunas, Gaussian process factorization machines for context-aware recommendations, 2014, 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, Gold Coast, Queensland, pp. 63-72.

[49] B. Loni, A. Said, M. Larson, A. Hanjalic, 'Free lunch' enhancement for collaborative filtering with factorization machines, 2014, 8th ACM Conference on Recommender Systems, ACM, Foster City, CA, pp. 281-284.

[50] R. Palovics, F. Ayala-Gomez, B. Csikota, B. Daroczy, L. Kocsis, D. Spadacene, A.A. Benczur, RecSys Challenge 2014: An ensemble of binary classifiers and matrix factorization, 2014 Recommender Systems Challenge, ACM, Foster City, CA, 2014, pp. 13-18.

[51] Z.C. Lipton, C. Elkan, B. Naryanaswamy, Optimal thresholding of classifiers to maximize F1 measure, in: T. Calders, F. Esposito, E. Hüllermeier, R. Meo (Eds.) Machine Learning and Knowledge Discovery in Databases, 2014, Springer Berlin Heidelberg, pp. 225-239.

[52] S. García, A. Fernández, J. Luengo, F. Herrera, Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, 2010, Inform. Sciences, 180 2044-2064.

[53] A. Rakotomamonjy, Variable selection using SVM based criteria, 2003, J. Mach. Learn. Res., 3, 1357-1370.

[54] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.) Advances in Neural Information Processing Systems, 2003, Curran Associates, Inc., pp. 431-439.

[55] A.E. Taylor, Statistical enhancement of support vector machines, 2009, Oregon State University.

[56] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: A survey, 2015, Decis. Support Syst., 74 12-32.

[57] J. Bauer, A. Nanopoulos, Recommender systems based on quantitative implicit customer feedback, 2014, Decis. Support Syst., 68, 77-88.

CHAPTER IV

THE EFFECT OF REVENUE MAXIMIZATION RECOMMENDATION SYSTEMS ON THE PURCHASE FUNNEL METRICS: A FIELD EXPERIMENT

CHAPTER IV

THE EFFECT OF REVENUE MAXIMIZATION RECOMMENDATION SYSTEMS ON THE PURCHASE FUNNEL METRICS: A FIELD EXPERIMENT

Abstract

This study addresses the measurement of added value of different types of recommendation systems and contributes to literature in three distinct ways. First, a framework for evaluating and comparing both traditional – and revenue maximization recommenders in terms of performance metrics throughout the entire purchase funnel is proposed. Rather than focusing on traditional metrics based on accuracy, the framework prescribes an assessment of the true impact of a recommendation system's algorithm configuration by focusing on multiple metrics throughout the purchase funnel. Additionally, the framework is used to assess the influence of three defining characteristics of popular algorithms for recommendation systems: (i) ability to personalize recommendations, (ii) hybridization strategy used to incorporate multiple data sources and (iii) revenue inclusion. Specifically, it is claimed that personalized recommendation systems outperform non-personalized recommenders in terms of click through -, view -, cart addition -, and conversion rate. Within personalized recommendation systems, a state-of-the-art feature combination technique outperforms an a posteriori combination of single data source recommendation systems. Additionally, the framework claims that revenue maximization recommenders positively affect the value per order because these systems include a revenue component. Even more, revenue maximization recommendation systems exhibit a greater effect on revenue compared to traditional recommenders as they influence both conversion and value per order. Second, the proposed framework is validated in a large-scaled field experiment executed in collaboration with a large European e-tailor. Finally, a business case shows that the best performing traditional recommendation system results in the highest number of orders and that the optimal revenue maximization recommender generates the highest revenue.

Keywords: E-commerce; Recommendation systems; Hybridization; Revenue maximization; Field experiment

1 Introduction

In e-commerce customers are typically exposed to an abundance of products. This overload of products and the limited human processing capabilities make it hard for customers to scan

every relevant product and make well-informed decisions. Recommendation systems are machine learning tools typically designed to help customers cope with this choice overload by recommending a limited set of relevant items to users [1-3]. In addition to serve as a mere decision aid, recommendation systems traditionally improve customer relationship by enhancing customer retention and increasing customer loyalty [4]. In accordance with these findings, recommendation systems' literature mainly focusses on algorithmic improvement, which resulted in the creation of many different algorithms.

In recent years two research subjects have gained in importance, i.e. hybrid - and revenue maximization recommendation systems. First, more complex algorithms have been developed and a shift towards hybrid recommendation systems is observed due to their higher accuracy and elimination of flaws associated with single data source recommenders [5]. Accordingly, this study focusses mainly on hybrid recommendation systems.

Second, both information systems and consumer psychology research demonstrates that recommendation systems influence customers' preferences [6, 7]. As a company is able to alter customers' preferences by proposing other products, it needs to find a trade-off between optimizing customer satisfaction, i.e. taking a customer focus, and maximizing revenue, i.e. taking a company focus. The majority of literature takes a customer focus, while companies often have a vested interest in recommending products with high revenue [8, 9]. In this context, several studies investigate revenue maximization recommendation systems [10-13]. These systems include a revenue component and have consequently a direct impact on the business results [10].

Most studies in recommendation systems' literature executes and evaluates recommenders on offline historical datasets. This evaluation procedure has as limitation that in most cases no business value is available. Consequently, evaluation is only possible in terms of statistical metrics, e.g. RSME, recall, precision, and F1 [14]. A limited number of studies conduct field experiments [15-17], allowing evaluation in terms of business metrics, e.g. click through rate, conversion rate, and revenue [15]. Field experiments executed in previous work are limited in algorithm complexity and current state-of-the-art hybridization methods and especially revenue maximization recommendation systems are never tested in a real-life setting. In contrast, recent research in several digital marketing domains evaluates communication and marketing effectiveness in terms of business metrics. Moreover, within the digital marketing literature the concept of purchase funnel is defined in a series of stages in which customers

move toward a purchase [18]. In this stream of research, business metrics are evaluated throughout different stages of the purchase funnel [18], which is never done in the recommendation systems literature.

This study proposes a framework identifying three effects of traditional - and revenue maximization recommendation systems on business metrics. First, it is argued that the recommendation system's algorithm configuration influences conversion business metrics throughout the entire purchase funnel for both traditional - and revenue maximization recommenders. Specifically, this study compares a non-personalized configuration to personalized ones, identified as the effect of personalization, and divides the personalized configurations further based on hybridization method, identified as the effect of hybridization algorithm. An a posteriori combination of collaboration filtering (CF) algorithms is compared to feature combination algorithm based on factorization machines (FM). These different configurations are evaluated in terms of click through, view, cart addition, and order stage of the purchase funnel. Second, the framework claims that the effect of revenue inclusion results in a positive effect on value per order. This effect is only observed for revenue maximization recommenders as traditional recommendation systems do not include a revenue component. Finally, revenue maximization recommendation systems exhibit a greater effect on revenue in the order stage compared to traditional recommenders as both a conversion effect and value per order effect drive revenue.

The proposed framework is validated in a large-scale email field experiment executed in collaboration with La Redoute, a large European e-commerce company. The execution of field experiment allows to evaluate an actual behavioral shift of various recommender approaches and increases the external validity [19]. Instead of evaluating performance in terms of statistical metrics, this study evaluates the performance of recommendation systems in terms of business metrics. Whereas e.g. Chen et al. (2008) and Panniello et al. (2016) only take order behavior into account, this study evaluates the effect of recommendation systems throughout the entire purchase funnel.

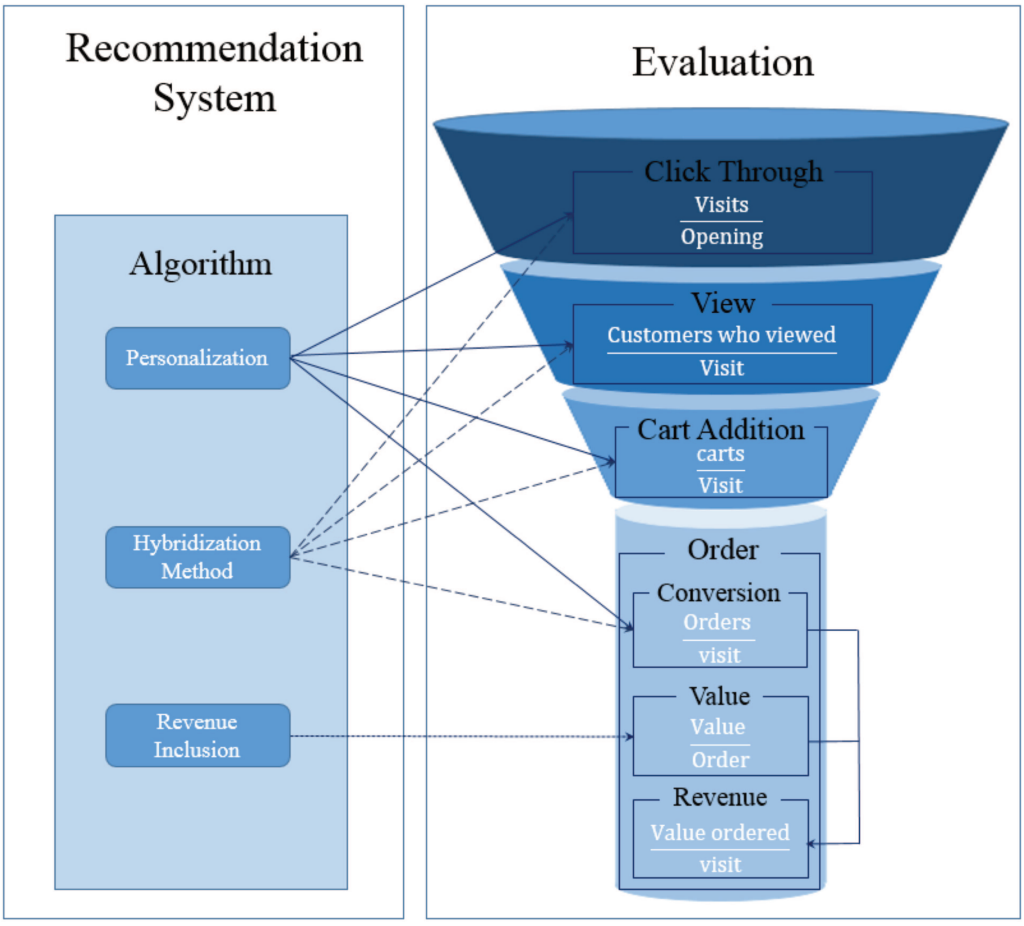
In the remainder of this study, section 2 proposes a theoretical framework and suggests research questions to analyze the effect of (revenue maximization) recommendation systems throughout the purchase funnel. Section 3 discusses the relevant literature related to hybrid recommendation systems, revenue maximization recommendations, the purchase funnel, and field experiments in recommendation systems. Section 4 describes the field experiment to

evaluate the different recommendation systems. The results of the analysis and answers to the research questions are formulated in section 5. Section 6 discusses the results and section 7 demonstrate the potential added value of (revenue maximization) recommendation systems in a business case. Finally, section 8 concludes the study and identifies some path for future research.

2 Framework and Research Questions

This section proposes a framework to identify three effects of traditional and revenue maximization recommendation systems on business metrics, as presented in Figure IV.1. First, the framework argues that algorithm configuration, of both traditional and revenue maximization recommendation systems, has an influence on click through rate, view rate, cart addition rate, and order conversion rate. As all these metrics represent conversion in a certain purchase funnel stage, we use the term conversion metrics to refer to this set of four business

Figure IV.1: Framework identifying the effect of traditional – and revenue maximization recommendation systems on business metrics throughout the purchase funnel.



metrics. Specifically, the framework distinguishes two algorithm configuration characteristics, i.e. personalization and hybridization method. The effect of personalization is tested by benchmarking a non-personalized method with personalized methods. The effect of hybridization method is evaluated by comparing two hybridization techniques for personalized recommendation systems, i.e. an a posteriori combination of CF algorithms and a feature combination algorithm based on FMs.

Second, the framework claims that revenue maximization recommendation systems have, next to the effect of algorithm configuration on conversion metrics, an effect on value. Specifically, it is argued that revenue inclusion positively affects value per order. This effect is only observed for revenue maximization recommenders as traditional recommendation systems do not include a revenue component.

Finally, the framework claims that revenue maximization recommendation systems exhibit a greater effect on the revenue metric in the order stage, i.e. value ordered per visit, compared

$$\frac{Order\ value}{\#\ Visits} = \frac{\#\ Orders}{\#\ Visits} * \frac{Value}{\#\ Orders}. \quad (1)$$

to traditional recommenders. This elevated revenue effect of revenue maximization recommenders is explained by the synergy between the conversion and the value per order effect. In the order funnel stage, a mathematical equality between revenue, conversion and value ($revenue = conversion * value$) exists. The business metric for revenue is decomposable in the business metrics for conversion and value:

In accordance with this mathematical equality, it is claimed that revenue maximization recommendation systems have a greater effect on revenue compared to traditional recommenders as they influence both conversion and value in the final stage of the purchase funnel.

To validate the proposed framework, six research questions are constructed:

RQ1a: Is there an effect of personalization on conversion metrics throughout the purchase funnel?

RQ1b: Is there an effect of hybridization method on conversion metrics throughout the purchase funnel?

RQ2: Is there an effect of revenue inclusion on value (value per order)?

RQ3a: Is there an effect of personalization on revenue (value per visit)?

RQ3b: Is there an effect of hybridization method on revenue (value per visit)?

RQ3c: Is there an effect of revenue inclusion on revenue (value per visit)?

3 Related Research

This study tests, evaluates, and compares traditional -, and revenue maximization hybrid recommendation systems throughout the purchase funnel in a field experiment. To be able to frame our study, this section discusses the related research about the hybridization methods used, revenue maximization algorithms, evaluation throughout the purchase funnel, and field experiments in a recommendation systems' setting.

3.1 Algorithms

The majority of recommendation systems' studies develop new algorithms for accurate prediction, resulting in a large body of literature. This section limits the discussion recommendation algorithms to a discussion of the most relevant work in traditional hybrid recommendation systems combing different data sources and revenue maximization recommendation systems.

3.1.1 Traditional Hybrid Recommendation Systems

An existing body of literature reviews hybrid algorithms combining different data sources. Previous studies have suggested a combination of behavioral data features, mainly explicit ratings, with product data or customer data to mitigate the drawbacks of collaborative filtering recommendation systems [5, 20, 21, e.g. 22, 23]. Other studies for example suggest a combination of product data with customer data [e.g. 20]. Some recent investigations also combine some specific behavioral, whether explicit or implicit, with customer and product information [24].

In addition to the hybridized data sources, the combination technique varies [5], reflecting three main categories:

1. Implementing different methods separately and combine their predictions a posteriori [e.g. 20, 25].
2. Incorporating some characteristics of recommendation model A into recommendation model B [e.g. 20, 26].
3. Constructing a general, unifying recommendation model that incorporates features of different models [e.g. 27, 28].

In this study two well-researched hybridization techniques are deployed: (1) a posteriori combining predictions, also known as weighting [27], and (2) constructing a unified recommendation model, known as feature combination [27].

Weighting. Many weighting schemes to a posteriori combine different individual recommendation systems are investigated in literature. Two examples are a linear combination of recommendation scores [25] and voting of different outcomes [20]. This studies estimates a support vector machine (SVM) with the recommendation scores of the individual CF recommenders as input to obtain a single recommendation [29].

Feature Combination. The feature combination technique deployed in this study is a FM. The FM technique [28] is based on SVMs and factorization algorithms and combines the advantages of both. Like SVMs, FMs are general predictors, so they work with any real valued feature vector. In contrast though, FMs can estimate interactions, even for problems with huge sparsity like recommender systems, where SVMs fail. The main reason for this interesting property is that feature interaction is calculated on the basis of factorized parameters, so it is possible to calculate the recommendation model in linear time. In contrast with other factorization recommendation algorithms like SVD++ [30], FM works with any real-valued feature vector, creating an opportunity to include different data types. The general model equation of an FM of degree 2 is formulated as follows:

$$Feature\ Combination\ FM = w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} \sum_{j=i=1}^{n} \langle v_i, v_j \rangle x_i x_j, \qquad (1)$$

where x_i represent the predictors that are features of the data sources; w_0 is the global bias; w_i are the parameters related to the first-degree effect of the n predictors; and $\langle v_i, v_j \rangle$ represents the dot product of v_i and v_j, such that these two terms are low dimensionality vector representations of rank k of the predictors x_i and x_j. The technique has already been shown to work well in situation with different data sources, especially in cases where behavioral data is combined with other information sources, e.g. product data and context [31-33].

3.1.2 Revenue Maximization Recommendation Systems

Literature shows that recommendation systems are able to influence customers' preferences [6, 7]. More specific, by recommending products a customer might be positively influenced towards these items, which leads to behavioral changes and has consequently a positive effect on the sales of the proposed products [34].

Revenue maximization recommendation systems [10-13] are a specific case of biased recommenders. These systems include a revenue component in their recommendation algorithm and therefore try to directly impact the business results instead of modelling customer preferences [10]. Table IV.1 gives an overview of revenue maximization studies together with the deployed algorithms, the applied revenue component, the empirical validation data set, and the evaluation metrics used.

Table IV.1 delivers three main insight. First, multiple algorithms are deployed to incorporate revenue. While most studies incorporate a revenue component in their loss function [11-13], Chen et al. (2008) estimate traditional recommenders (popular and user-based CF) and calculate revenue maximization recommendations by multiplying the resulting traditional recommendation scores with profit. Second, in terms of datasets to validate the results, it is

	Algorithms	**Validation data**	**Evaluation Metrics**
Chen et al. (2008)	Popular User-based CF Popular * profit User-based CF * profit	Publically available datasets	Accuracy Profit
Azaria et al. (2013)	Linear likelihood estimation	Lab experiment	Revenue Satisfaction
Lu et al. (2014)	Optimization problem solved with greedy algorithms	Publically available datasets	Revenue Efficiency
Wang et al. (2015)	Two Bayesian models	Publically available datasets	Accuracy Satisfaction Revenue
This Study	Popular **Weighting with SVM** **FM** **Weighting with SVM * revenue** **FM * revenue**	**Field Experiment**	Accuracy **six business metrics**

Note: Bolded items indicate novel considerations by the present study.

Table IV.1: Overview of studies investigating revenue maximization recommendation systems.

observed that most studies make use of publicly available datasets. In contrast Azaria et al. (2013) execute a controlled lab experiment in which a limited set of participants are invited in the lab and evaluate recommendations displayed in a digital interface. Third, most studies evaluate accuracy and revenue, and to a lesser extend satisfaction and efficiency.

This study leverages the combination strategy as described by Chen et al. (2008). Revenue maximization recommendation systems are created by multiplying traditional recommendation scores with a revenue component. This modular approach has the advantage that it is generalizable and could be deployed in combination with any set of traditional recommendation scores. Additionally, as demonstrated by Chen et al. (2008), the popular

algorithm is deployed as traditional recommendation algorithm. This algorithm is widely used and computationally inexpensive. Next to the poplar algorithm, two traditional hybrid recommendation systems, i.e. an a posteriori weighting using SVM and feature combination using FM, estimate recommendation scores. Additionally, the revenue variations are calculated by multiplying the two traditional recommendations with a revenue component.

Including a revenue component in recommendation systems is expected to increase revenue, but caution is needed. Favoring products with higher revenue might distort the original recommendation outcome which possibly lead to a suboptimal set of proposed products in terms of the customer's preferences. Specifically, if the revenue component is too dominant, more expensive products are proposed which leads to irrelevant product proposals. By proposing products diverting too much from the customer's preferences, the customer might react aversely to the recommendations. Consequently a customer might lose trust in the recommendation system [7-9, 15, 34]. McKnight et al. [35] define trust as the consumer's perception that a recommendation system has the ability, skills, and expertise to effectively and benevolently recommend products in the interest of customers. To resolve the distrust issue, Das et al. (2010) introduce a measure of trust. The inclusion of trust guarantees that the revenue maximization recommendations do not divert too much from the traditional recommendations to avoid distrust among customers. In this study trust is controlled by normalizing the revenue component to avoid it would become too dominant.

Finally, this study adds to the revenue maximization recommendation systems' literature by executing a large-scale field experiment allowing to evaluate results in a real-life situation. Section 3.3 discusses the related research and added value of field experiments in recommendation systems in more detail.

3.2 Purchase funnel

The main idea of the purchase funnel is that customers move toward a purchase in a series of stages [18]. In each of these stages, business metrics could be operationalized to evaluate the influence of marketing actions, like personalized product propositions based on recommendation systems, on the progression through the purchase funnel. In digital marketing a number of studies evaluate the online marketing activities throughout the purchase funnel [37]. Specifically, the effectiveness of display ads [18, 38, 39], search engine marketing [37, 39], and e-mail campaigns [18, 39] in different stages of the purchase funnel is already researched in depth.

Digital marketing literature formulates various classifications of the purchase funnel [37]. Lavidge and Steiner (1961) proposed to split up the purchase funnel in an awareness, knowledge, liking, preference, conviction, and purchase stage. More recent versions of the purchase funnel are for example proposed by Jansen and Schuster (2011) and Wiesel et al. (2011). The latter study divides the purchase funnel into a cognitive, affective, and conative stage.

While recommendation systems could also be evaluated in term of business metrics throughout the purchase funnel, the majority of studies evaluate algorithms in terms of statistical metrics, e.g. RSME, recall, precision, and F1 [14]. Only a few studies evaluate recommendation systems in terms of business metrics [15]. Even more, work evaluating recommenders throughout various stages of the purchase funnel is non-existing.

3.3 Field Experiments in Recommendation Systems

Controlled field experiments are the best way to measure a recommendation system's impact as customer behavior is influenced by the treatments, i.e. different recommendation system configurations. This is opposed to offline testing where the impact of recommenders is evaluated on historical data. Offline testing assess recommendation systems in terms of statistical metrics, e.g. RSME, recall, precision, and F1, without showing recommendations to users and a shift in behavior is consequently an assumption [41]. Despite the interesting properties of controlled field experiments, they also contain risks. First, the design of the tests needs to be flawless, because errors might directly affect the company's metrics. For example, it is important to sample users randomly, so that the comparisons between alternatives are fair. Second, recommendations need to be relevant, because irrelevant recommendations have a negative effect on the business metrics. Third, online testing is costly, especially in cases where a multitude of recommenders are tested. The creation of testing systems requires a lot of time and effort. Because of these pitfalls, Gunawardana and Shani (2009) propose to first test algorithms in an offline setting to afterwards deploy the most promising algorithms in field experiments. This procedure minimizes the chance of irrelevant recommendations and lowers the field experiments costs as less conditions need to be tested.

However, a well-defined procedure for field experiments is described by Gunawardana and Shani (2009), no field experiment is conducted. Even more, the overall number of studies executing a field experiment is limited in recommendation systems' literature [15]. Two studies focusing on satisfaction are conducted and find that personalized recommendation systems

outperform non-personalized recommendation systems [16, 17]. Additionally these studies show that more sophisticated algorithms perform better compared to basic algorithms (matrix factorization CF versus item-based CF and weighting of item-based CF and content-based versus item-based CF and content-based algorithm) in terms of satisfaction.

Panniello et al. (2016) execute an email field experiment to evaluate the effect of personalized recommendation systems on accuracy, diversification, trust, and purchase behavior. More specific, they compare a random non-personalized strategy, a content-based algorithm, and a context-aware algorithm. The authors find that personalized recommendation systems outperform random recommendations in terms of money spent, but results on quantity ordered and value ordered are less clear.

In line with Panniello et al. (2016), this study executes an email field experiment and complements the results in four distinct ways. First, Panniello et al. (2016) only evaluate a limited set of business metrics, while our study has a larger scope and creates a framework identifying the effect recommendation systems throughout the entire purchase funnel. Second, Panniello et al. (2016) include context in their recommendation systems, while our study includes a revenue component. Third, our study relies on more sophisticated recommendation algorithms. Finally, our study evaluates its results on a much larger customer base and uses six business metrics to measure the effect of recommendation systems on performance.

4 Field Experiment

This section discusses the setting of the field experiment together with the deployed recommendation algorithms, and the evaluation metrics.

4.1 Setting

La Redoute is a major European e-tailer specialized in apparel and home decoration. While their French branch has a client base of nine million customers, activation of these clients remains a challenge. Therefore, the company wants to investigate the effect of marketing communication personalization, achieved by deploying recommendation systems, on cross-sell and upsell.

This study executes a direct email field experiment to measure the effect of recommendation systems' properties, i.e. personalization, hybridization method, and revenue inclusion, on customer behavior in different stages of the purchase funnel. In total 6,195,735

emails containing nine recommendations out of a total set of 38,574 products are sent out in four different waves. The total population is constructed by the populations of each individual wave. Table IV.2 presents the number of emails sent in each of the waves and Appendix A displays an example email.

	Wave 1	Wave 2	Wave 3	Wave 4	Total
Emails Sent	1,286,937	1,636,794	1,470,053	1,801,851	6,195,735

Table IV.2: Number of emails sent per wave.

The population of each wave is randomly divided into five groups. Each of the groups receives an identical email only differing in terms of the nine recommended products. The recommendations are based on five different recommendation algorithms discussed in section 4.3.

4.2 Data

To be able to construct recommendation systems, input data is needed. In this study, the creation of the personalized recommendation systems uses four data sources as input, i.e. raw behavioral data, aggregated behavioral data on customer level, product data, and customer data. First, raw behavioral data comprises explicit rating, views, internal search, additions to cart, and orders on a customer-product combination level. Second, aggregated behavioral data consist of behavioral data aggregated on customer level. Length of relationship, value-based segmentation, and recency, frequency, monetary variables are good examples of aggregated behavioral features. Third, product data contains product information like brand and different categorizations. Finally, customer data comprises typical customer characteristics like age, gender, place of residence, and children's information.

4.3 Algorithms

In this section the deployed algorithms are discussed. A distinction is made between non-personalized and personalized algorithms. Furthermore, personalized algorithms are divided into traditional and revenue maximization recommenders [10]. In total five recommenders are proposed: one non-personalized algorithm, two traditional hybrid recommendation algorithms, and two revenue maximization algorithms.

Table IV.3 shows the five algorithms and an in depth discussion is given in the remainder of this section. The internal company recommender (ICR) is found on the top left of Table IV.3 with a white background. Two traditional recommendation systems, the hierarchical hybrid recommender (HHR) and the hybrid factorization machines (HFM) are found on the bottom left with a light gray background. Based on these traditional recommenders, two revenue maximization algorithms, featured in Table IV.3 on the bottom right with a dark gray background, are created. The RMR^{HHR} is a revenue maximization recommender in which purchase probability is based on the HHR algorithm. RMR^{HFM} is a revenue maximization recommendation system in which purchase probability is based on the HFM algorithm.

4.3.1 Non-Personalized Recommendation Algorithm

The company's current recommendation strategy for email campaigns is deploying an N most popular strategy [42]. This internal company recommender (*ICR*) is based on a non-personalized, non-computationally expensive algorithm proposing the most popular products to every targeted customer.

4.3.2 Personalized Traditional Recommendation Algorithms

This study incorporates two personalized, traditional, hybrid recommendation systems by deploying two hybridization algorithms, i.e. the hierarchical hybrid recommender (*HHR*) and the hybrid factorization machines (*HFM*) to combine raw behavioral -, aggregate behavioral -, product -, and customer data. These two hybrid recommendation systems are selected based on an offline test conducted in a previous study executed by the authors [29]. This test shows that the *HHR* and *HFM* recommenders, leveraging the four discussed data sources as input, perform best in terms of accuracy.

HHR Algorithm. The *HHR* algorithm consists of two hierarchical stages. In the first stage, individual memory-based algorithms estimate four recommendation models [43], i.e. a demographic -, a content-based-, a CF -, and an aggregated CF model. In the second stage,

			Revenue inclusion	
			No	**Yes**
Algorithm	Non-personalized	*Benchmark*	*ICR*	
	Personalized	*A posteriori weighting*	*HHR*	RMR^{HHR}
		Model-based Feature combination	*HFM*	RMR^{HFM}

Table IV.3: Overview of algorithms

individual recommendation scores for each user-item pair are used as input for an SVM combining the individual first stage results. To maintain generalizability and scalability, a linear SVM [44, 45] based on stochastic gradient decent optimization [46] is trained. The regularization parameter is set to 0.01, and L2 regularization is applied because of the differentiability [45].

HFM Algorithm. The *HFM* algorithm is an operationalization of FMs and is able to include the four different data sources in one unified model. The HFM's parameters are set similar to those of the SVM of the *HHR* algorithm to increase face validity. Specifically, a FM for classification with L2 regularization, a regularization parameter of 0.01 based on stochastic gradient decent optimization results, and 30 latent factors are retained in the recommendation model.

4.3.3 Personalized Revenue Maximization Recommendation Systems

This study incorporates a revenue component in the recommendation systems by multiplying purchase probability with normalized revenue [10]. Purchase probabilities are defined as (transformed) recommendation scores resulting from a traditional recommendation system. This study proposes to first normalize both recommendation scores and revenue component to obtain more comprehensive, scaled, and trustworthy revenue maximization

$$RMR_{u,i} = PP_{u,i} * NR_i. \quad (2)$$

recommendation scores:

In equation 2 $RMR_{u,i}$ represents the revenue maximization recommendation score for user-item pair *(u,i)*. *u* represents a user *u* within the total base of users *U*. *i* represent an item in the total set of items *I*. $PP_{u,i}$ refers to the purchase probability of user *u* for item *i*. NR_i refers to the normalized revenue of item *i*. Both $PP_{u,i}$ and NR_i are discussed in more detail below.

Purchase Probability. Recommendation scores of a traditional recommendation system are not always directly comparable and should therefore be normalized to obtain a purchase probability for each user-item pair. In this study purchase probabilities are calculated by rescaling the traditional recommendations scores (*HHR* and *HFM*) to values between 0 and 1. This rescaling serves two purposes: calculating a more intuitive value for purchase probability

$$PP_{u,i} = \frac{Rec_{u,i} - min\ (Rec)}{max(Rec) - min\ (Rec)}. \quad (3)$$

and making scores of recommendation systems originating from different algorithms comparable. The calculation of purchase probability is defined as follows:

In equation 3 $PP_{u,i}$ represents the purchase probability of user *u* for item *i*. *u* represents a user within the total population of users *U*. *i* represent an item in the total set of items *I*. *Rec* refers to the recommendation score of the traditional recommendation system. In this specific study the traditional recommendation scores are the results of the *HHR* or *HFM* recommendation systems, discussed in section 4.3.2. Consequently, *min(Rec)* and *max(Rec)* are the absolute minimum and maximum traditional recommendation scores over all user-item pairs.

Normalized Revenue. The revenue component, is normalized between 0 and 1 to obtain a normalized revenue measure. Instead of normalizing all product utilities at once, normalization is done by product category. The main reason for this normalization is trust [15, 36, 47]. First, if no normalization is done, revenue has a much bigger scale compared to purchase probability and would be too dominant in the revenue maximization recommendations, which might lead to distrust among customers. Second, revenue components are normalized by product category to avoid recommending only products from expensive product categories. The normalized revenue based on product revenue is calculated as follows:

$$NR_i = \frac{Revenue_i - min\ (Revenue_{PC})}{max(Revenue_{PC}) - min\ (Revenue_{PC})}. \quad (4)$$

In equation 4, NR_i represents the normalized revenue of item *i*. *i* represent an item in the total set of items *I*; *PC* indicates the product category of item *i*; $Revenue_i$ refers to the product revenue of item *i*; and min $(Revenue_{PC})$ and max$(Revenue_{PC})$ are the minimum and maximum item revenues in product category *PC* of which item is *i* is a member.

> Hypothetical example: Suppose we are interested in the revenue maximization recommendation scores (*RMR*) of two products, i.e. a pair of socks and a leather jacket for a specific user. The socks have a purchase probability (*PP*) of 99% and a revenue of €1. The leather jacket has a *PP* of 1% and a revenue of €200. If no revenue normalization is done, *RMR* would be calculated by multiplying *PP* with revenue. The socks would end up with a score of €0.99 and the jacket with a score of €2, while traditional recommendation scores, represented by *PP*, indicate that socks are a much more suitable recommendation, the jacket is recommended in this case.

If revenue normalization is applied, and we assume that the minimum revenue of socks is €0.30 and the maximum revenue is €2.5 and the minimum revenue of a jacket is €50 and the maximum revenue is €500, the normalized revenue (*NR*) of socks would be 0.32 and the normalized revenue of a jacket would be 0.33. If we now calculate the *RMR* for both products, socks would get a *RMR* of 0.316 and the leather jacket would get a *RMR* of 0.003. Based on these calculations, the socks are chosen over the jacket.

Going a bit further in this example, we could include a third product, a second pair of socks. Suppose these socks generate more revenue (€2). Assuming negative price elasticity, the *PP* (80%) of this pair of socks is lower than the *PP* of the first pair of socks. If we calculate the *RMR* for this second pair of socks, we have a score of 0.618. In this case we see that the second pair of socks is preferred over the first pair, while the first pair has the highest traditional recommendation score.

4.4 Evaluation

To evaluate the effectiveness of recommendation systems in the proposed field experiment, the behavior of customers related to the recommended products is measured in terms of business metrics throughout the company's purchase funnel [18]. This study operationalizes the cognitive stage by evaluating click through rate and view rate. Cart addition rate measures behavior in the affective stage. Finally, order behavior is a straightforward metric representing the conative stage. Note that the order behavior is measured by three metrics representing conversion, value, and revenue. Table IV.4 shows the different metrics evaluated throughout the purchase funnel. Notice that the evaluation done in this chapter is an extension of the evaluation done in Chapter 3. While Chapter 3 only evaluates recommendation systems in an offline setting on the F1 accuracy metric, this chapter includes business metric.

Funnel Stage		Business Metric
Click Through		$\frac{\#\text{ Visits}}{\#\text{ Openings}}$ = Click through rate (CTR)
View		$\frac{\#\text{ Customers who viewed}}{\#\text{ Visits}}$ = View rate (VR)
Cart Addition		$\frac{\#\text{ Carts}}{\#\text{ Visits}}$ = Cart addition rate (CAR)
Order	Conversion	$\frac{\#\text{ Orders}}{\#\text{ Visits}}$ = Conversion rate (CR)
	Value	$\frac{\text{Value}}{\#\text{ Orders}}$ = Value per order (VO)
	Revenue	$\frac{\text{Order value}}{\#\text{ Visits}}$ = Value per visit (VV)

Table IV.4: Overview of business metrics for each purchase funnel stage.

5 Results

In this section, the method of analysis is presented followed by a discussion of the results per purchase funnel stage.

5.1 Analysis

The statistical analysis to answer the research questions is done by means of general linear models (GLMs). The GLMs test the relationships between the six business metrics as the dependent variables and algorithm characteristics and revenue inclusion as the independent variables. Email wave is included in the equation as covariate. Each GLM has the appropriate

$$Metric_k = \beta_{0,k} + \beta_{1,k}FM + \beta_{2,k}HR + \beta_{3,k}RI + \beta_{4,k}Wave + \varepsilon_k. \quad (5)$$

link function. Specifically, a logit link function is used for GLMs with click through rate, view rate, cart addition rate, and conversion rate as target. An inverse Gaussian link function is used when value per order and value per visit are the dependent variables.

Equation 5 represents the general GLM. In this equation $Metric_k$ refers to a metric in the set of six business metrics proposed in Table 4. $\beta_{0,k}$ is the constant parameter for metric k. $\beta_{1,k}$, $\beta_{2,k}$, $\beta_{3,k}$, and $\beta_{4,k}$ are the parameter values for algorithm type FM (*FM*), algorithm type hierarchical recommender (*HR*), revenue inclusion (*RI*), and the email wave (*Wave*) for metric k. ε_k is the specific error term for metric k. Conclusions for the six research questions are drawn based the parameter values for hybridization methods $(\beta_{1,k}, \beta_{2,k})$ (RQ1a, RQ3a), a linear

hypothesis testing equality between parameters for the two hybridization methods $(\beta_{1,k} - \beta_{2,k=0} = 0)$ (RQ1b, RQ3b), and the parameter value for revenue inclusion($\beta_{3,k}$) (RQ2, RQ3c).

5.2 Click Through Stage

Parameters $\beta_{1,CTR}$ and $\beta_{2,CTR}$ show that both *FM* and *HR* are significantly outperforming *ICR*, the non-personalized reference algorithm. This result indicates that hybrid personalized algorithms outperform a non-personalized algorithm in terms of click through rate (RQ1a). The linear hypothesis testing the difference between $\beta_{1,CTR}$ and $\beta_{2,CTR}$ indicates that *FM* results in a higher click through rate compared to *HR*, but the difference is not significant (RQ1b).

Click Through Rate				
Coefficient		**Estimate**	**Z-value**	**p-value**
$\beta_{1,CTR}$	**RQ1a**	**0.047**	5.382	< 0.0001
$\beta_{2,CTR}$	**RQ1a**	**0.038**	4.519	< 0.0001
$\beta_{3,CTR}$		-0.005	-0.682	0.495
		Estimate	**F-value**	**p-value**
$\beta_{1CTR} - \beta_{2,CTR}$	**RQ1b**	0.009	1.74	0.187

Table IV.5: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on click through rate.

5.3 View Stage

Table IV.6 shows that parameters of *FM* ($\beta_{1,VR}$) and *HR* ($\beta_{2,VR}$) are significantly positive. These values show that hybrid personalized algorithms outperform a non-personalized algorithm in terms of view rate (RQ1a). The linear hypothesis testing the difference between $\beta_{1,VR}$ and $\beta_{1,VR}$ ($\beta_{1VR} - \beta_{2VR}$) indicates that *FM* results in a significantly higher view rate compared to *HR* (RQ1b).

View Rate				
Coefficient		**Estimate**	**Z-value**	**p-value**
$\beta_{1,VR}$	**RQ1a**	**0.532**	20.898	< 0.0001
$\beta_{2,VR}$	**RQ1a**	**0.479**	19.348	< 0.0001
$\beta_{3,VR}$		**0.057**	3.315	0.0009
		Estimate	**F-value**	**p-value**
$\beta_{1,VR} - \beta_{2,VR}$	**RQ1b**	**0.053**	9.231	0.002

Table IV.6: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion view rate.

5.4 Cart Addition Stage

Parameters $\beta_{1,CAR}$ and $\beta_{2,CAR}$ in Table IV.7 show that both *FM* and *HR* are significantly outperforming *ICR*. This result indicates that hybrid personalized algorithms outperform a non-personalized algorithm in terms of cart addition rate (RQ1a). The linear hypothesis testing the difference between $\beta_{1,CAR}$ and $\beta_{2,CAR}$ ($\beta_{1,CAR} - \beta_{2,CAR}$) indicates that *FM* results in a significantly higher cart addition rate compared to *HR* (RQ1b).

Cart Addition Rate				
Coefficient		**Estimate**	**Z-value**	**p-value**
$\beta_{1,CAR}$	**RQ1a**	**1.289**	15.376	< 0.0001
$\beta_{2,CAR}$	**RQ1a**	**1.181**	14.272	< 0.001
$\beta_{3,CAR}$		-0.003	-0.578	0.563
		Estimate	**F-value**	**p-value**
$\beta_{1,CAR} - \beta_{2,CAR}$	**RQ1b**	**0.108**	6,186	0,013

Table IV.7: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on cart addition rate.

5.5 Order Stage

The order stage is the final stage of the purchase funnel and the most interesting one. In this stage final conversion, value, and revenue are distinguished.

In terms of *conversion rate*, parameters $\beta_{1,CR}$ and $\beta_{2,CR}$ show that both *FM* and *HR* are significantly outperforming *ICR*. This result indicates that hybrid, personalized algorithms outperform a non-personalized algorithm in terms of conversion rate (RQ1a). The linear hypothesis testing the difference between $\beta_{1,CR}$ and $\beta_{2,CR}$ ($\beta_{1,CR} - \beta_{2,CR}$) indicates that *FM* results in a non-significantly higher conversion rate compared to *HR* (RQ1b).

Conversion Rate				
Coefficient		**Estimate**	**Z-value**	**p-value**
$\beta_{1,CR}$	**RQ1a**	**1.476**	5.689	< 0.0001
$\beta_{2,CR}$	**RQ1a**	**1.338**	5.175	< 0.0001
$\beta_{3,CR}$		0.055	0.452	0.651
		Estimate	**F-value**	**p-value**
$\beta_{1,CR} - \beta_{2,CR}$	**RQ1b**	0138	1.129	0.255

Table IV.8: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on conversion rate.

Regarding the *value effect*, a positive significant effect of revenue inclusion ($\beta_{3,VO}$) on value per order is detected (RQ2).

		Value		
Coefficient		**Estimate**	**t-value**	**p-value**
$\beta_{1,\mathrm{VO}}$		-10.974	-0.671	0.503
$\beta_{2,\mathrm{VO}}$		-25.094	-1.525	0.128
$\beta_{3,\mathrm{VO}}$	**RQ2**	**18.280**	2.365	0.019
		Estimate	**F-value**	**p-value**
$\beta_{1,\mathrm{VO}} - \beta_{2,VO}$		14.120	3.385	0.067

Table IV.9: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on value per order.

Finally, based on the figures in Table IV.10 can be concluded that the *revenue effect* in the order stage is significantly positively influenced by personalization $\beta_{1,VV}$ and $\beta_{2,VV}$ (RQ3a), hybridization method ($\beta_{1,VV} - \beta_{2,VV}$) (RQ3b), and revenue inclusion ($\beta_{3,VV}$) (RQ3c).

		Revenue		
Coefficient		**Estimate**	**t-value**	**p-value**
$\beta_{1,\mathrm{VV}}$	**RQ3a**	**0.078**	2.887	0.004
$\beta_{2,\mathrm{VV}}$	**RQ3a**	**0.043**	2.030	0.042
$\beta_{3,\mathrm{VV}}$	**RQ3c**	**0.044**	2.168	0.030
		Estimate	**F-value**	**p-value**
$\beta_{1,\mathrm{VV}} - \beta_{2,VV}$	**RQ3b**	**0.035**	4.51	0.034

Table IV.10: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on value per visit.

6 Discussion

This section discusses the results of the analyses done in section 5 and Table IV.11 summarizes the finding. Specifically, the three effects of traditional and revenue maximization recommendation systems on business metrics proposed by the framework in section 2 are discussed.

Second, analysis shows that revenue inclusion has a positive effect on the value metric in the order stage (RQ2). Finally, results show that revenue is positively influenced by personalization, hybridization method, the two algorithm characteristics driving conversion, and revenue inclusion, the driver of value (RQ3a-c). These results indicate that the revenue effect is driven by both the conversion effect and the value effect in the order stage.

Funnel Stage		RQ	Answer to research questions
Click Through		RQ1a	**Yes**
		RQ1b	ns
View		RQ1a	**Yes**
		RQ1b	**Yes**
Cart Addition		RQ1a	**Yes**
		RQ1b	**Yes**
Order	Conversion	RQ1a	**Yes**
		RQ1b	ns
	Value	RQ2	**Yes**
	Revenue	RQ3a	**Yes**
		RQ3b	**Yes**
		RQ3c	**Yes**

Note: '**Yes**' indicates that the research question is positively answered; 'ns' indicates that no significant result is found, but parameter values indicate a positive answer to the research question.

Table IV.11: Summary of GLM results.

Nevertheless revenue inclusion has a positive effect on value and revenue, being cautious is advised. If the revenue component becomes too dominant in recommendation systems, the recommendations might significantly divert from the original recommended products and customer preferences. Consequently distrust increases which negatively influences recommendation performance [15].

7 Business Case

Based on the observed visit rate, conversion rate, and value per visit in the four email waves, one could calculate the expected incremental orders and revenue of different personalized recommendation systems compared to the initial company benchmark. This is done by extrapolating the observed figures, assuming all emails contain product recommendations calculated by a specific algorithm. This exercise is done for the narrow set of nine recommended products and for the total product offering (38,574 products).

Table IV.12 shows the business case and indicates that *HFM*-based models generate the highest number of orders and the highest revenue. The traditional *HFM* results in the most orders within the recommendation set, resulting in an increment of 350% in orders compared to *ICR*. In terms of revenue, the revenue maximizing *HFM* (RMR^{HFM}) generates the highest revenue, which leads to an increment of 442% in revenue compared to *ICR* for the set of recommended products.

		ICR	*HHR*	RMR^{HHR}	*HFM*	RMR^{HFM}
Recommended Product Set	**% Increase Conversion Rate**		**233%**	**300%**	**350%**	**317%**
	Value per visit[1]	1	2.33	3.68	3.67	5.33
	% Increase Revenue		**139%**	**276%**	**269%**	**442%**
All products	**% Increase Conversion Rate**		**6.66%**	**3.08%**	**9.58%**	**4.22%**
	Value per visit[1]	1	1.030	1.032	1.072	1.127
	% Increase Revenue		**5.61%**	**5.38%**	**7.65%**	**14.62%**

[1] For confidentiality reasons, value per visit is expressed in relative figures compared with *ICR* as basis.

Table IV.12: Business case.

Table IV.12 also displays the total number of orders and the total revenue of visits generated by the emails when considering all products. However the majority of products were not featured in the email, an impact of recommendation algorithms is observed. Analogously with the results for the set of recommended products, *HFM*-based models generate the highest number of orders and the highest revenue. The traditional *HFM* results again in the most orders, resulting in an increment of 9.58% in orders compared to *ICR*. In terms of revenue, the revenue maximizing *HFM* (RMR^{HFM}) generates again the highest revenue, which leads to an incremental in revenue of 14.16% compared to *ICR*. It is observed that the relative increase is much more pronounced for the set of recommended products compared to the set of all products. This clearly indicates that the recommendation systems have more impact on the set of recommended products, as could be expected since these items are featured in the email.

8 Conclusions, Limitations and Future work

This study proposes and validates a framework that suggests three main observations. First, it is argued that recommendation systems' configurations have a positive effect on conversion business metrics throughout the purchase funnel. Specifically, personalization has a positive effect on click through rate, view rate, add to cart rate, and conversion rate. Second, a hybrid, state-of-the-art, model-based, feature combination recommendation system (*FM*) outperforms a simple a posteriori weighting of memory-based recommendation systems (*HR*) in terms of conversion metrics in all stage of the purchase funnel and in terms of revenue in the order stage. Third, revenue inclusion positively influences value and revenue in the order stage. These results indicate that it is worthwhile for a company to investigate both algorithm configuration to increase conversion and hence revenue. Additionally, it is useful to research revenue inclusion as this component increases value and hence revenue in the order stage of the purchase funnel.

The framework is validated by means of a real-life email field experiment executed at a large European e-tailor. This setup is unique in the recommendation systems' literature and can be used as a guidance to setup similar field experiments. Additionally does the execution of the field experiment result in a high external validity of the results.

In the business case, an extrapolation is made to concretize the results of the validated framework. This exercise shows that the recommendation systems based on FMs lead to the highest potential results in the final and most important stage of the purchase funnel for both the set of recommended products as well as for the complete product offering. Depending on the business metric to optimize, the traditional HFM (*HFM*) or the revenue maximization HFM (RMR^{HFM}) are preferable. *HFM* optimizes the conversion rate and results indicate an increase of 350% and 9.58% in number of orders for respectively the set of recommended products and total product set compared to the company benchmark (*ICR*). In terms of revenue, RMR^{HFM} shows the highest potential with an increase in revenue of 442% and 14.62% for respectively the set of recommendations and total product offering compared to *ICR*.

This study shows that revenue inclusion has a positive effect on the revenue of a recommendation system. However we also indicate that caution is needed. If the revenue component in a recommendation system becomes too dominant, recommendations will divert from the original preferences of the customer. This divergence might lead to distrust in a recommendation system resulting in lower direct revenue and to dissatisfaction and disloyalty among customers. To limit distrust, this study incorporates normalization of both the purchase probability and revenue component. To investigate this topic in more depth, it would be interesting to test the optimal weight of revenue component in future research.

Linked to this problem and in consultation with the company is decided not to incorporate a non-personalized algorithms (*ICR*) enhanced with a revenue component. As the *ICR* is already non-personalized, revenue inclusion could lead to the impression that only expensive products are proposed. This impression could result in distrust and dissatisfaction among the customer base. Independently from this decision, it could be interesting to include a revenue maximization non-personalized algorithm in future research.

This study only incorporates two traditional algorithms (*HHR* and *HFM*) and one revenue inclusion method. The set of algorithms and revenue inclusion methods could be extended to increase face validity. A logical selection of algorithms to include would for example be simple collaborative filtering or content-based systems.

References

[1] J.B. Schafer, J. Konstan, J. Riedl, Recommender systems in e-commerce, 1999, Proceedings of the 1st ACM conference on Electronic commerce, ACM, Denver, Colorado, USA, pp. 158-166.

[2] A. De Bruyn, J.C. Liechty, E.K.R.E. Huizingh, G.L. Lilien, Offering online recommendations with minimum customer input through conjoint-based decision aids, 2008, Marketing Science, 27, 443-460.

[3] G. Häubl, V. Trifts, Consumer decision making in online shopping environments: The effects of interactive decision aids, 2000, Marketing Science, 19, 4-21.

[4] A. Ansari, C.F. Mela, E-Customization, 2003, J. Marketing Res., 40, 131-145.

[5] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, 2005, IEEE Trans. Knowl. Data Eng., 17, 734-749.

[6] G. Häubl, K.B. Murray, Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents, 2003, Journal of Consumer Psychology, 13, 75-91.

[7] G. Adomavicius, J.C. Bockstedt, S.P. Curley, J. Zhang, Do recommender systems manipulate consumer preferences? A study of anchoring effects, 2013, Inform. Syst. Res., 24, 956-975.

[8] B. Xiao, I. Benbasat, Designing warning messages for detecting biased online product recommendations: An empirical investigation, 2015, Inform. Syst. Res., 26, 793-811.

[9] B. Xiao, I. Benbasat, Product-related deception in e-commerce: a theoretical perspective, 2011, MIS Q., 35, 169-196.

[10] L.-S. Chen, F.-H. Hsu, M.-C. Chen, Y.-C. Hsu, Developing recommender systems with the consideration of product profitability for sellers, 2008, Inform. Sciences, 178, 1032-1048.

[11] A. Azaria, A. Hassidim, S. Kraus, A. Eshkol, O. Weintraub, I. Netanely, Movie recommender system for profit maximization, 2013, Proceedings of the 7th ACM conference on Recommender systems, ACM, Hong Kong, China, pp. 121-128.

[12] W. Lu, S. Chen, K. Li, L.V.S. Lakshmanan, Show me the money: dynamic recommendations for revenue maximization, 2014, Proc. VLDB Endow., 7, 1785-1796.

[13] X. Wang, Y. Guo, C. Xu, Recommendation algorithms for optimizing hit rate, user satisfaction and website revenue, 2015, Proceedings of the 24th International Conference on Artificial Intelligence, AAAI Press, Buenos Aires, Argentina, pp. 1820-1826.

[14] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, 2004, ACM Trans. Inf. Syst., 22, 5-53.

[15] U. Panniello, M. Gorgoglione, A. Tuzhilin, Research note—In CARSs we trust: How context-aware recommendations affect customers' Trust and other business performance measures of recommender systems, 2016, Inform. Syst. Res., 27, 182-196.

[16] M.D. Ekstrand, D. Kluver, F.M. Harper, J.A. Konstan, Letting users choose recommender algorithms: An experimental study, 2015, Proceedings of the 9th ACM Conference on Recommender Systems, ACM, Vienna, Austria, pp. 11-18.

[17] M.A. Domingues, F. Gouyon, A.M. Jorge, J.P. Leal, J. Vinagre, L. Lemos, M. Sordo, Combining usage and content in an online recommendation system for music in the Long Tail, 2013, International Journal of Multimedia Information Retrieval, 2 3-13.

[18] T. Wiesel, K. Pauwels, J. Arts, Practice prize paper—Marketing's profit impact: Quantifying online and off-line funnel progression, 2011, Marketing Science, 30, 604-611.

[19] S.Y. Ho, D. Bodoff, The effects of web personalization on user attitude and behavior: an integration of the elaboration likelihood model and consumer search theory, 2014, MIS Q., 38, 497-520.

[20] M.J. Pazzani, A framework for collaborative, content-based and demographic filtering, 1999, Artif. Intell. Rev., 13, 393-408.

[21] M.G. Vozalis, K.G. Margaritis, Using SVD and demographic data for the enhancement of generalized collaborative filtering, 2007, Inform. Sciences, 177, 3017-3037.

[22] Y.-Y. Shih, D.-R. Liu, Product recommendation approaches: Collaborative filtering via customer lifetime value and customer demands, 2008, Expert Syst. Appl., 35, 350-360.

[23] A. Albadvi, M. Shahbazi, Integrating rating-based collaborative filtering with customer lifetime value: New product recommendation technique, 2010, Intell. Data Anal., 14, 143-155.

[24] A. Said, S. Dooms, B. Loni, D. Tikk, Recommender systems challenge 2014, 2014, 8th ACM Conference on Recommender Systems, ACM, Foster City,CA, pp. 387-388.

[25] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, M. Sartin, Combining content-based and collaborative filters in an online newspaper, 1999, SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, CA.

[26] M. Balabanovi, Y. Shoham, FAB: content-based, collaborative recommendation, 1997, Commun. ACM, 40, 66-72.

[27] R. Burke, Hybrid recommender systems: Survey and experiments, 2002, User Modeling and User-Adapted Interaction, 12, 331-370.

[28] S. Rendle, Factorization Machines, 2010, IEEE International Conference on Data Mining, Sydney, Australia.

[29] S. Geuens, A decision support system to evaluate recommendations systems combing multiple data sources and identify feature importance in e-commerce 2017, Personalization in e-commerce: A procedure to create and evaluate business relevant recommendation systems, Lille, pp. 63-87.

[30] Y. Koren, Factorization meets the neighborhood: A multifaceted collaborative filtering model, 2008, 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Las Vegas, NV, pp. 426-434.

[31] C. Cheng, F. Xia, T. Zhang, I. King, M.R. Lyu, Gradient boosting factorization machines, 2014, ACM Conference on Recommender Systems, ACM, Silicon Valley, CA, pp. 265-272.

[32] C.-M. Chen, M.-F. Tsai, J.-Y. Liu, Y.-H. Yang, Music recommendation based on multiple contextual similarity information, 2013, IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE Computer Society, Atlanta, GA, pp. 65-72.

[33] S. Rendle, Z. Gantner, C. Freudenthaler, L. Schmidt-Thieme, Fast context-aware recommendations with factorization machines, 2011, 34th international ACM SIGIR

conference on Research and development in Information Retrieval, ACM, Beijing, China, pp. 635-644.

[34] B. Pathak, R. Garfinkel, R. Gopal, R. Venkatesan, F. Yin, Empirical analysis of the impact of recommender systems on sales, 2010, Journal Management Information Systems, 27 159-188.

[35] D.H. McKnight, V. Choudhury, C. Kacmar, Developing and validating trust measures for e-commerce: An integrative typology, 2002, Inform. Syst. Res., 13, 334-359.

[36] A. Das, C. Mathieu, D. Ricketts, Maximizing profit using recommender systems, 2010, Conference WWW, Raleigh, NC, pp. 35-40.

[37] B.J. Jansen, S. Schuster, Bidding on the buying funnel for sponsored search and keyword advertising, 2011, Journal of Electronic Commerce Research, 12, 1-18.

[38] P.R. Hoban, R.E. Bucklin, Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment, 2015, J. Marketing Res., 52, 375-393.

[39] H. Li, P.K. Kannan, Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment, 2014, J. Marketing Res., 51, 40-56.

[40] R.J. Lavidge, G.A. Steiner, A Model for predictive measurements of advertising effectiveness, 1961, J. Marketing, 25, 59-62.

[41] A. Gunawardana, G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, 2009, J. Mach. Learn. Res., 10, 2935-2962.

[42] S. Prawesh, B. Padmanabhan, The “most popular news” recommender: Count amplification and manipulation resistance, 2014, Inform. Syst. Res., 25, 569-589.

[43] S. Lu, L. Xiao, M. Ding, A video-based automated recommender (VAR) system for garments, 2016, Marketing Science, 35, 484-510.

[44] C. Jin, L. Wang, Dimensionality dependent PAC-Bayes margin bound, 2012, in: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger, NIPS, Lake Tahoe, NV, pp. 1043-1051.

[45] Y. Tang, Deep learning using linear support vector machines, 2013, International Conference on Machine Learning, Atlanta, GA.

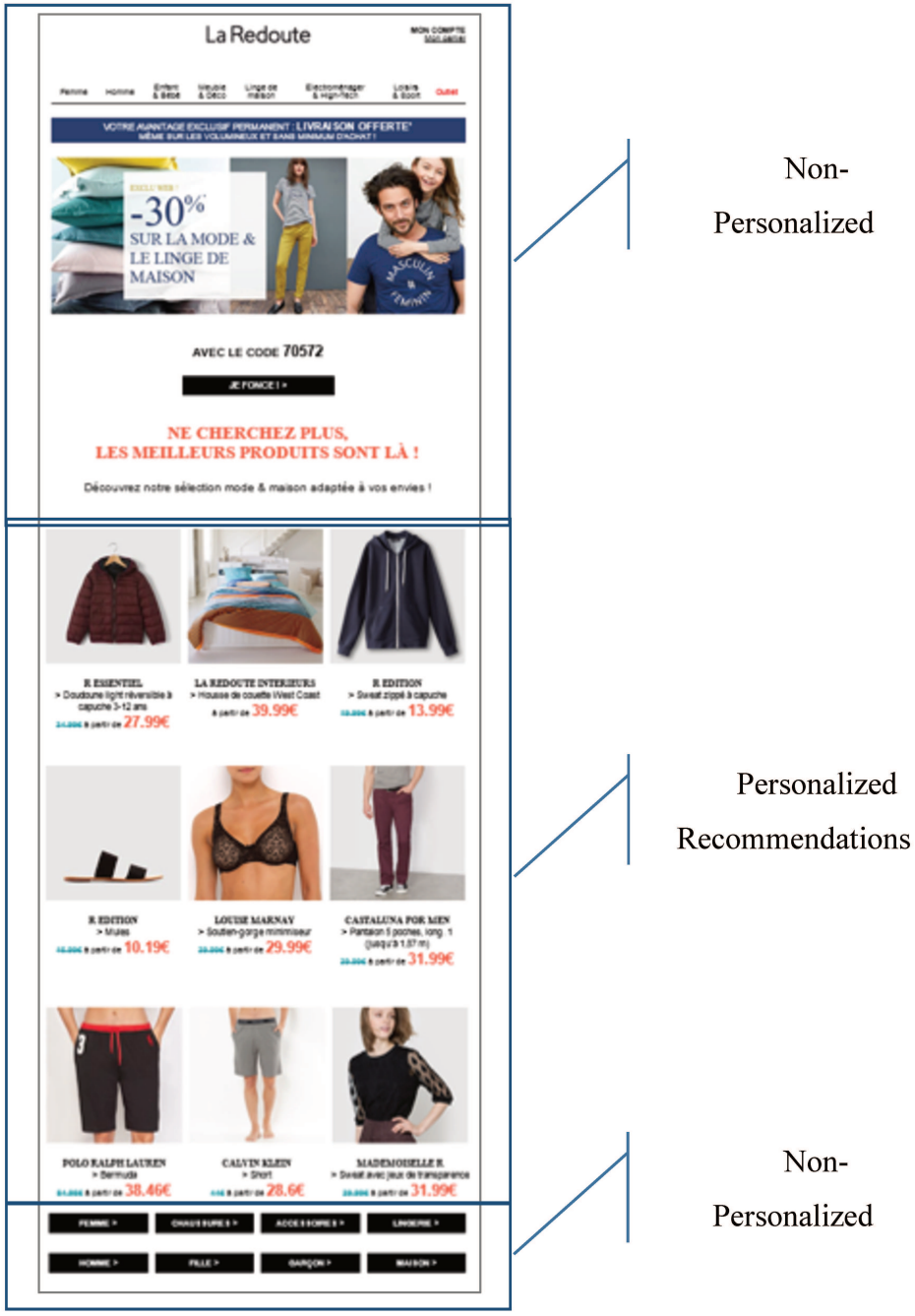
[46] Z. Wang, K. Crammer, S. Vucetic, Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale SVM training, 2012, J. Mach. Learn. Res., 13, 3103-3131.

[47] S.Y.X. Komiak, I. Benbasat, The effects of personalizaion and familiarity on trust and adoption of recommendation agents, 2006, MIS Q., 30 941-960.

Appendix

This appendix contains an example of personalized emails sent. Within each mailing, the header and footer of the emails were similar, only the product selection was personalized for each receiver.

Figure IV.A.1: Example Email.



CHAPTER V

GENERAL CONCLUSION & FURTHER WORK

CHAPTER V

GENERAL CONCLUSION & DIRECTIONS FOR FURTHER WORK

1 General Conclusion

Chapters II – IV construct the body of this dissertation and present results on their own. Additionally, the individual chapters contribute to one general objective, i.e. the creation of a phased design to setup field experiments in an e-commerce recommendation systems' context. Concretely, Chapter II investigates a single algorithm, CF, on a single data source, purchase data, into detail to determine the optimal configuration in an offline setting on historical data. These results are used in Chapter III to open up the scope to hybrid recommendation systems deploying multiple data sources as input. Finally, Chapter IV adopts the best performing recommendation systems in the offline tests on historical data of Chapter III. These top performing systems are transformed into revenue maximization recommendation systems and a field experiment allowing to draw conclusions in terms of business metrics throughout the purchase funnel is executed. The remainder of this section discusses the most important conclusions for each chapter.

Chapter II uses the binary purchase matrix as input for different CF algorithms. For many companies explicit data is not readily available in their databases, but they do collect a large amount of transactional data like purchases. As purchase data can be collected at relatively low cost and it is directly related to firm performance, Chapter II leverages purchase data as input data source. Concretely, a framework that guides marketers in building better recommendation systems advises on how to find a suitable recommendation algorithm in terms of accuracy, diversity, and computation time. To do so, this study analyzes the performance of different CF algorithm configurations. In this regard, we use synthetic data sets with different binary purchase input characteristics as well as two real-life validation sets.

Results show that the accuracy and diversity of the generated recommendations depend on the data reduction technique, the CF method, and the similarity measure. Computation time is influenced only by the data reduction technique, mainly in the sense that LPCA and NMF are based on multiplicative updating algorithms, whereas SVD and CA are calculated in one step. Second, different input characteristics can lead to other optimal algorithms. For accuracy, the optimal model is stable (CA/Item/Cos, Corr), but various characteristics lead to different optimal configurations for diversity and computation time. In addition, the optimal model configuration for each data set is influenced by input characteristics. For example, accuracy

depends on sparsity, while diversity is influenced by sparsity, the purchase distribution, and the item-user ratio. Computation time is only influenced by the purchase distribution and the item-user ratio.

Chapter III broadens the scope of this dissertation by combining different data sources in hybrid recommendation systems for e-commerce. This chapter constructs a framework for optimizing the hybridization process of data sources and introduces feature importance into recommendation literature. The validation of the framework on eight historical data sets from a European e-commerce company produces five distinct findings. First, RBD and PD are the most predictive sources; companies should focus their efforts to create recommendation systems primarily on these two data sources. Second, combining data sources adds value, and more data sources lead to higher predictive performance. Third, despite the higher predictive performance of recommendation models with four data sources it is not always possible to investigate them all. If a company lacks the ability and time to investigate all four sources, it can concentrate its efforts. It should do so in the following order: RBD, PD, CD, and then ABD. Fourth, this study suggests using feature combination based on FM for the optimal combination of all four data sources (hybrid factorization machine (*HFM*)). This technique outperforms an a posteriori weighting of different single data source recommendation models (*HHR*). Fifth and finally, the accuracy of a recommendation system is very important, but beyond having a highly predictive recommendation model, it is also insightful to open the black box to determine which data sources and features contribute to recommendation success. According to the current study findings, RBD contributes most to the model, followed by PD and CD respectively. Finally ABD is the least important data source.

In terms of the importance of individual features, implicit RBD features are very important, indicating that log data from the e-commerce site is vital. Explicit ratings are less important, mainly because this information is only available in smaller amounts. If a business model does not thrive on ratings (like e.g., Netflix, LastFM), explicit ratings are less important to consider. Furthermore, PD is important information to gather, especially product division and brand data. Although somewhat less important, individual CD features can add value to recommendation systems. Finally, ABD features have relatively little importance and can be less emphasized, if the time and resources available to create recommendation systems are limited.

Chapter IV leverages the results of Chapter III and takes the two best performing recommendation systems (*HHR* and *HFM*) out of the lab. The resulting traditional hybrid

recommendation scores are multiplied with a revenue component to create revenue maximization recommendation systems (RMR^{HHR} and RMR^{HFM}).

Based on the different recommendation systems, a framework identifying three effects of (revenue maximization) recommendation systems on business metrics throughout the purchase funnel is created. First, it is argued that recommendation systems' configurations have an effect on conversion business metrics throughout the purchase funnel. Concretely, personalization has a positive effect on click through rate, view rate, add to cart rate, and conversion rate. Second, a hybrid, state-of-the-art, model-based, feature combination recommendation system (*FM*) outperforms a simple a posteriori weighting of memory-based recommendation systems (*HR*) in terms of conversion metrics in all stage of the purchase funnel and in terms of revenue in the order stage. Third, revenue inclusion positively influences value per order. Consequently, revenue maximization recommendation systems outperform traditional recommendation systems in terms of revenue in the order stage as the revenue effect is driven by both the conversion and value per order effect. A large-scale email field experiment executed at La Redoute shows that the proposed framework is valid.

A business case demonstrates that the factorization machines based models have the highest potential in term of conversion and revenue. First, the traditional *HFM* results in an increase in the number of orders of 350% and 9.58% for respectively the set of recommended products and the total product offering compared to the *ICR* model. Second, the RMR^{HFM} obtains the highest incremental revenue of 442% and 14.16% for respectively the recommendation set and the total product offering compared to the company benchmark. The validated framework and business case demonstrates that state-of-the-art hybrid recommendation systems improve conversion and revenue. Consequently, a company could benefit from adopting the proposed recommendation strategies or investigating other suitable sophisticated recommendation algorithms to improve recommendation performance. Additionally, results show the usefulness of investigating revenue inclusion as it increases value per order and thus revenue.

2 Limitations and Directions for Future Research

In this dissertation is deliberately chosen not to use publicly available data sets based on two main reasons. First, most publicly available data sets do not contain all four data sources, i.e. product-, customer-, behavioral-, and aggregated behavioral data, that we have at our disposal via the collaboration with La Redoute. Additionally, most available data sets do not contain a revenue component making it impossible to investigate revenue maximization

recommendation systems. Second, public data sets contain only historical data and do not allow for field experimentation. Despite these restrictions, replicating the results, of mainly Chapter III, on publicly available data sets would be an interesting topic for further research. In a machine learning context this benchmarking could lead to higher external validity and acceptance within the community.

In line with previous comment additional benchmark algorithms, like matrix factorization CF, could be deployed throughout this dissertation. Especially in Chapter IV stronger benchmarks would increase the value of the results. This additional benchmarking would increase face validity and again acceptance within the machine learning community. However, I argue that this benchmarking is not done because this dissertation focusses on the data science aspect and adds to literature by giving business insights rather than contributing by merely benchmarking algorithms. The designed frameworks are generic, consequently it is easy to leverage them in other implementations deploying other algorithms which could be done in further research.

In Chapter II the focus lies on implicit ratings/feedback, more specific purchases. Additionally, because we only consider a purchase yes or no, this data type looks at first sight like binary data as we only have two values. In contrast, several studies suggest that this data cannot be treated as binary, because an absence of purchase is not necessarily a negative signal. If a user purchases an item, an implicit signal for preference is observed, but not purchasing an item does not necessarily imply an implicit dislike. Not purchasing can have multiple reasons. The absence of an action can for example be the result of a real dislike, but can also be a consequence of the user being unaware of the existence of the product. Another reason could be that the user likes the item, but is currently not looking for this type of product or the product is too expensive, etc. Consequently, only a purchase implicitly indicates preference, while no purchase does not necessarily indicates a dislike. Therefore purchase could be coded as 1 and no purchases as missing. This type of data is called unary data. In this study, we consider purchase data as binary [eg. 3, 4], while no interpretation of unary data is done.

Because of the unary character of the purchase data used in Chapter II, traditional classification accuracy metrics, like recall, precision, and F1, are not optimal to use for two main reasons. First the unary character of data does not allow to give a clear-cut evaluation of no purchases. An absence of purchase might have multiple reasons and therefore an interpretation as dislike might not be correct. Second, the purchase data is very sparse, resulting

in only a few positive hits to evaluate the recommendation systems. In Chapter II and Chapter III we use a ranking based classification variation of the F1 metrics, to be able to cope with the unary data problems. Nevertheless this strategy resolves some of the issues, specific ranking based accuracy metrics are developed to overcome the concerns related to sparse and unary data [5, 6]. The deployment of ranking metrics complements the use of explicit binary classification. Most popular metrics are the mean reciprocal rank (MRR), mean average precision (MAP), NDCG, and AUC [5]. The incorporation of ranking based measures in Chapter II and Chapter III is a specific path for future research.

Next to accuracy, diversity is measured in Chapter II. In literature two types of diversity are distinguished, i.e. individual diversity and aggregate diversity [7]. Individual diversity represents the diversity in the set of recommended items of a single user. In contrast, aggregate diversity denotes the overall diversity, i.e. the diversity of all recommended items in all the recommendation lists to every user. The deployed measure of diversity in Chapter II is the Intra-List Similarity (ILS), a measure for individual diversity. Nevertheless Chapter II computes an average ILS over all users, ILS remains a measure of individual diversity. This signifies that Chapter II only considers individual diversity, without accounting for aggregated diversity. In future research a measure for aggregated diversity could be integrated to cover both diversity types in the analyses.

Additionally, computation time is evaluated in Chapter II, but a throughout discussion of scalability is not include. Future work could include an analysis measuring the effect of the characteristics, mainly sparsity and size, of the input matrix and algorithm configuration to assess scalability. This subject is important because one of the major disadvantages of CF is its limited scalability. Nevertheless, I argue that modern big data technologies allow distributed computing to help solving the scalability issue of CF.

Next to limitations and future work considerations related to specific evaluation metrics, this dissertation evaluates several different evaluation aspects. In Chapter II, accuracy (F1), computation time, and diversity (ILS) are evaluated. Chapter III only focusses on accuracy (F1) and Chapter IV evaluates business metrics. In future research, it would be interesting to evaluate all the results in terms of computation time, diversity, as done in Chapter II and additionally novelty, scalability, serendipity, and trust could be analyzed [1]. As business metrics are only quantifiable in field experiments, analyzing these metrics in Chapter II and

Chapter III would be less useful as these chapters are evaluated in an offline setting in which recommendation systems have no impact on actual customer behavior.

As mentioned in the list of possible additional evaluation metrics, trust is an important issue currently investigated in literature [e.g. 2]. Nowadays, recommendation systems are more frequently used in different settings, resulting in more awareness of these personalization tools. This awareness leads to more prudence of customers reflected by a critical assessment of recommendation systems. In this respect, I believe it is important to create trustworthy recommendation systems. A customer should experience a recommendation system as a guidance tool instead of advertisement or spam. Therefore it is important to take this trust factor into account in further research.

This dissertation focusses on traditional data sources, i.e. customer-, product- and behavioral data in a recommendation systems' setting. Nevertheless these data sources have important predictive power, other data sources exists and might contribute to recommendation performance. Context is for example a promising additional data source [2]. Customers' needs are not static and dependent on situations and contexts. Take for example La Redoute's activities in the apparel industry. Deciding on which sweater to buy/wear depends on e.g. weather, seasonality, and location. These context factors could be important predictors to address in recommendation systems. A special case of context is the purchase cycle. A customer looking for a sweater a week ago is not necessarily looking for a sweater today. To account for this phenomenon, real-time data could help to improve recommendations. In this perspective investigation of real-time and context-aware recommendation systems is an interesting path for future research to improve recommendations.

Whereas a decade ago recommendation systems were only a machine learning topic, in recent years other fields of study like customer psychology, marketing, and data science gained interest in the subject. In my opinion this broadening of the scope is a positive evolution. Nowadays recommendation literature, like this dissertation, goes beyond the creation sophisticated algorithms, by investigating topics like trust, context, user perception, satisfaction, and profitability. The investigations of these topics make recommendation systems more accessible for the business and increases their interest in the subject. This interest results in more investments boosting both commercial and academic development of recommendation systems. This synergy is important to build practically relevant recommendation systems. Therefore in my opinion academic research should also create recommendation systems that

are relevant for companies. This goal could for example be achieved by investigating real-time hybrid recommendation systems. Real-time is currently a very important topic in data science and companies. Leveraging this trend could boost the visibility of academic recommendation systems' research.

Additionally, I believe nowadays big data and streaming technologies allow to create state-of-the art recommendation systems with a shorter throughput time which makes them feasible for practical implementation. These technologies should be leveraged in academic research. Even more, I believe that the use of big data and streaming technologies should be investigated and discussed in academic research. Investigation of these subjects increases the impact of scientific research on real world applications.

References

[1] J.L. Herlocker, J.A. Konstan, K. Terveen, J.T. Riedl, Evaluating collaborative filtering recommender systems, 2004, Acm Transactions on Information Systems, 22, 5-53.

[2] U. Panniello, M. Gorgoglione, A. Tuzhilin, Research note—In CARSs we trust: How context-aware recommendations affect customers' Trust and other business performance measures of recommender systems, 2016, Inform. Syst. Res., 27, 182-196.

[3] B.M. Sarwar, G. Karypis, J.A. Konstan, J.T. Riedl, Analysis of recommendation algorithms for e-commerce, 2000, 2nd ACM Conference on Electronic Commerce, ACM, Minneapolis, Minnesota, pp. 158-167.

[4] M. Deshpande, G. Karypis, Item-based top-N recommendation algorithms, 2004, ACM Trans. Inf. Syst., 22 143-177.

[5] C. C. Aggarwal, Recommender systems: The Textbook, 2016, Springer Cham Heidelberg, New York

[6] D. Jannach, L. Lerche, and M. Jugovac, Adaptation and Evaluation of Recommendations for Short-term Shopping Goals, 2015, In *Proceedings of the 9th ACM Conference on Recommender Systems* (RecSys '15). ACM, New York, NY, USA, 211-218.

[7] G. Adomavicius, Y. Kwon, Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques, 2012, IEEE Trans. on Knowl. and Data Eng., 24 896-911.

CONCLUSIONS GENERALES

Les chapitres II et IV constituent le corpus de cette thèse et présentent chacun leurs propres résultats. De plus, chaque chapitre contribue à l'objectif général, qui est la création d'un canevas progressif pour la définition d'expériences de terrain dans le contexte de systèmes de recommandation pour le commerce en ligne. Concrètement, le chapitre II se penche en détail sur un algorithme, CF, qui opère sur une seule base de données, les données d'achat, dans le but de déterminer sa configuration optimale dans un environnement hors-ligne en travaillant sur des données historiques. Ces résultats sont utilisés dans le chapitre III pour élargir le champ d'investigation aux systèmes de recommandations hybrides qui traitent des sources de données multiples. Enfin, le chapitre IV adopte les systèmes de recommandations ayant présenté les meilleures performances sur des données historiques au cours des tests hors-ligne effectués au chapitre III. Ces systèmes à haute performance sont transformés en systèmes de recommandation avec maximisation du chiffre d'affaires ; une expérience de terrain est réalisée pour permettre de tirer des conclusions en termes d'indicateurs à travers l'ensemble du funnel d'achats. Le reste de cette section discute des conclusions les plus importantes de chaque chapitre.

Le *chapitre II* utilise la matrice d'achats binaire comme base pour les différents algorithmes CF. Même si de nombreuses entreprises ne possèdent pas de données explicites immédiatement disponibles dans leurs bases de données, elles collectent une grande quantité de données transactionnelles telles que les données d'achat. Puisque les données d'achat peuvent être récoltées à un cout relativement bas et sont directement liées à la performance de l'entreprise, le chapitre II tire parti de ces données d'achat en tant que source de données d'entrée. Concrètement, un cadre qui aide les vendeurs à construire de meilleurs systèmes de recommandations leur donne des conseils sur la manière de trouver un algorithme de recommandation adapté en termes de précision, de diversité et de temps de calcul. Pour ce faire, cette étude analyse la performance de différentes configurations d'algorithme CF. À cette fin, nous utilisons des jeux de données synthétiques avec différentes caractéristiques de données d'achat binaires ainsi que deux jeux de données tirés de la vie réelle pour validation.

Nos résultats montrent que la précision et la diversité des recommandations générées dépendent de la technique de réduction des données, de la méthode CF et de la mesure de similarité. Le temps de calcul est influencé uniquement par la technique de réduction des

données, principalement dans le sens que LPCA et NMF sont basés sur des algorithmes multiplicatifs automatiquement réactualisés, alors que SVD et CA sont calculés en une étape. Deuxièmement, différentes caractéristiques d'entrée peuvent nous conduire à d'autres algorithmes optimaux. En ce qui concerne la précision, le modèle optimal est stable (CA/Item/Cos, Corr), mais différentes caractéristiques induisent différentes configurations optimales pour la diversité et le temps de calcul. De plus, la configuration du modèle optimal pour chaque jeu de données est influencée par les caractéristiques d'entrée. Par exemple, la précision dépend de la rareté, tandis que la diversité est influencée par la rareté, la répartition des achats et le taux objet/utilisateur. Le temps de calcul n'est quant à lui influencé que par la répartition des achats et le taux objet/utilisateur.

Le *chapitre III* élargit le champ d'investigation de cette thèse en combinant différentes sources de données dans des systèmes de recommandation hybrides pour le commerce en ligne. Ce chapitre établit un cadre pour l'optimisation du processus d'hybridation des sources de données et introduit l'importance des caractéristiques dans la littérature sur les systèmes de recommandation. La validation du cadre au moyen de huit jeux de données historiques provenant d'une entreprise de commerce en ligne européenne produit cinq résultats. Premièrement, les données d'achat et les données comportementales brutes sont les sources les plus prédictives ; les entreprises devraient concentrer leurs efforts sur la création de systèmes de recommandation basés avant tout sur ces deux sources de données. Deuxièmement, le fait de combiner les sources de données ajoute de la valeur : plus de sources de données donnent une meilleure performance prédictive. Troisièmement, malgré la performance prédictive plus élevée des modèles de recommandation basés sur quatre sources de données, il n'est pas toujours possible d'analyser l'ensemble de ces jeux de données. Si une entreprise n'a pas assez de temps ou de capacités pour analyser l'ensemble de ces quatre sources, elle peut décider de concentrer ses efforts sur certaines d'entre elles seulement. Dans ce cas, elle devrait le faire dans l'ordre de priorité suivant : données comportementales brutes, données d'achat, données client, données comportementales agrégées. Quatrièmement, cette étude suggère l'utilisation d'une combinaison de caractéristiques basée sur une machine à factorisation pour une combinaison optimale de l'ensemble de ces quatre sources de données (machine à factorisation hybride, *HFM*). Cette technique s'avère plus performante que la pondération *a posteriori* de différents modèles de recommandation tirés d'une seule source de données (*HHR*). Cinquièmement et pour terminer, même si la précision d'un système de recommandation est

très importante, il est aussi extrêmement intéressant, au-delà d'un modèle de recommandation hautement prédictif, d'ouvrir la boite noire pour déterminer quelles sources de données et caractéristiques contribuent le plus à de bonnes recommandations. Selon les résultats de notre étude, les données comportementales brutes sont celles qui contribuent le plus au modèle, suivies des données d'achat et des données client ; les données comportementales agrégées étant la source de données la moins importante.

En termes de l'importance des caractéristiques individuelles, les caractéristiques implicites tirées des données de comportement brutes sont très importantes, ce qui suggère que les données de journal du site de commerce en ligne jouent un rôle crucial. Les évaluations explicites sont moins importantes, surtout parce que cette information n'est disponible qu'en plus petites quantités. Pour les modèles d'entreprise dans lesquels les évaluations ne jouent qu'un faible rôle (comme Netflix ou LastFM), les évaluations explicites sont moins importantes à considérer. De plus, les données d'achat sont des informations importantes à récolter, surtout en ce qui concerne la division des produits et les données de marque. Bien qu'un peu moins importantes, les caractéristiques individuelles tirées des données client peuvent ajouter une valeur aux systèmes de recommandations. Enfin, les données comportementales agrégées ont relativement peu d'importance, ce qui fait qu'on peut y consacrer moins d'attention si le temps et les ressources disponibles pour créer les systèmes de recommandation sont limités.

Le *chapitre IV* part des résultats du chapitre III pour faire sortir du laboratoire les deux systèmes de recommandation les plus performants (*HHR* et *HFM*). Les scores de recommandation hybride traditionnelle ainsi obtenus sont multipliés par une composante « chiffre d'affaires » pour créer des systèmes de recommandation avec maximisation du chiffre d'affaires (RMR^{HHR} et RMR^{HFM}).

À partir de ces différents systèmes de recommandation, nous avons créé un cadre pour identifier les trois effets des systèmes de recommandation (avec maximisation du chiffre d'affaires) sur les indicateurs d'entreprise tout au long du funnel d'achats. Premièrement, il est suggéré que les configurations des systèmes de recommandation ont un effet sur les indicateurs de conversion tout au long du funnel d'achats. Concrètement, la personnalisation a un effet positif sur le taux de clics, le taux de visites, le taux d'ajout au chariot et le taux de conversion. Deuxièmement, un système de recommandation hybride haut de gamme basé sur un modèle et combinant différentes caractéristiques (*FM*) offre de meilleures performances en termes

d'indicateurs de conversion à toutes les étapes du funnel d'achat et en termes de chiffre d'affaires à l'étape de la commande qu'une simple pondération *a posteriori* de systèmes de recommandation basés sur la mémoire (*HR*). Troisièmement, le fait d'inclure le facteur « chiffre d'affaires » influence de manière positive la valeur par commande. De ce fait, les systèmes de recommandation avec maximisation du chiffre d'affaires donnent de meilleures performances que les systèmes de recommandation traditionnels en termes de chiffre d'affaires à l'étape de la commande, puisque l'effet « chiffre d'affaires » est accru par la conversion et par l'effet « valeur par commande ». Une expérience de terrain par courrier électronique à grande échelle exécutée en collaboration avec La Redoute a démontré que le cadre proposé est valable.

Notre étude de cas a démontré que les modèles obtenus à partir des machines à factorisation ont le plus haut potentiel en termes de conversion et de revenus. Premièrement, comparé au modèle de l'*ICR*, les *HFM* traditionnelles donnent une hausse du nombre de commandes de 350 % pour la gamme de produits recommandés et de 9,58 % pour l'offre de produits globale. Deuxièmement, comparé aux données de l'entreprise, la RMR^{HFM} obtient la meilleure hausse de recettes, de 442 % et de 14,16 % respectivement pour les produits recommandés et pour l'offre de produits globale. Le cadre validé et l'étude de cas démontrent que les systèmes de recommandation hybrides de pointe améliorent la conversion et le chiffre d'affaires. Par conséquent, une entreprise pourrait bénéficier de l'adoption des stratégies de recommandation proposées ou rechercher d'autres algorithmes de recommandation sophistiqués et adaptés pour améliorer la performance des recommandations. De plus, nos résultats montrent l'utilité de la recherche portant sur l'inclusion du facteur « chiffre d'affaires » puisque celle-ci augmente la valeur par commande et donc, le chiffre d'affaires.

LIST OF TABLES

Table I.1: Input data, algorithms, and validation data used in the different chapters. 24

Table II.1: Previous studies incorporating CF algorithms in a binary purchase setting 38

Table II.2: Overview of effects of CF algorithm variations as function of the evaluation metric 44

Table II.3: Overview of effects of CF algorithm variations and input data characteristics as a function of the evaluation metric 46

Table II.4: Influence of sparsity on best performing models: accuracy and diversity 48

Table II.5: Real-Life Data sets and their Characteristics 49

Table II.6: Overview of Effects of CF Algorithm Configurations 50

Table II.7: Comparison of Accuracy, Diversity, and Computation Times for Different Algorithm Configurations 52

Table II.8: Effects of input characteristics on three evaluation metrics 53

Table II.A.1: Confusion matrix of item vectors for items *i* and *j* 59

Table II.A.2: μ –values as function of *n* and *spar* for *m* = 1,000 63

Table II.B.1: AN(C)OVA results for RQ1 65

Table II.B.2: Accuracy, diversity, and computation time results of the pairwise t-test of data reduction techniques, indicating t_{df} and *p*-values 65

Table II.B.3: Accuracy, diversity, and computation time results of the pairwise t-test of CF methods, indicating t_{df} and *p*-values 66

Table II.B.4: Accuracy, diversity, and computation time results of the pairwise t-test of similarity measures, indicating t_{df} and *p*-values 66

Table II.C.1: Best performing models for accuracy as a function of item–user ratio, sparsity, and purchase distribution 67

Table II.C.2: Best performing models for of diversity as a function of item–user ratio, sparsity, and purchase distribution 67

Table II.C.3: Best performing models for computation time as a function of item–user ratio, sparsity, and purchase distribution 68

Table II.D.1: ANOVA results for accuracy sensitivity of the CA/Item/Corr model to sparsity, purchase distribution, and item/user ratio 68

Table II.D.2: ANOVA results for diversity sensitivity of the best performing models to sparsity, purchase distribution, and item/user ratio 68

Table II.D.3: ANOVA results for computation time sensitivity of the best performing models to sparsity, purchase distribution, and item/user ratio 69

Table II.D.4: Means, t-values (df), and *p*-values related to the sensitivity of diversity to purchase distribution 69

Table II.D.5: Means, t-values (df), and *p*-values related to the sensitivity of diversity to item–user ratio 70

Table II.D.6: Means, t-values (df), and *p*-values related to the sensitivity of computation time to item–user ratio 70

Table III.1: Advantages and Shortcomings of the three recommendation systems. 76

Table III.2: Number of visitors and products in different product categories. 83

Table III.3: Data source importance scores per product category. 93

Table IV.1: Overview of studies investigating revenue maximization recommendation systems. 110

Table IV.2: Number of emails sent per wave. 114

Table IV.3: Overview of algorithms 115

Table IV.4: Overview of business metrics for each purchase funnel stage. 119

Table IV.5: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on click through rate. 120

Table IV.6: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion view rate. 120

Table IV.7: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on cart addition rate. 121

Table IV.8: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on conversion rate. 121

Table IV.9: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on value per order. 122

Table IV.10: Parameter estimates and statistics of the GLM estimating the effects of personalization, hybridization method, and utility inclusion on value per visit. 122

Table IV.11: Summary of GLM results. 123

Table IV.12: Business case. 123

LIST OF FIGURES

Figure II.1: Overview of the experimental design 39

Figure II.2: F1 Measure (a), ILS (b), and Computation Time (c) as function of selection size 44

Figure II. 3: Accuracy (a) and diversity (b) for different sparsity levels as a function of selection size for the CA/Item/Corr algorithm 48

Figure II.4: Diversity (a) and computation time (b) as a function of item–user ratios for the None/User/Jaccard and CA, NMF, SVD/Item/Corr algorithms 49

Figure II.5: Algorithm variations as a function of accuracy, diversity, and computation time 51

Figure II.A.1: Example of uniform (a) and linear (b) purchase distributions for three alternative sparsity levels (m=20; n=1,000) 62

Figure II.A.2: Example of an exponential purchase distribution for three alternative sparsity levels (m=20; n=1,000) 63

Figure III.1: Empirical DSS. 82

Figure III.2:Data structure. 84

Figure III.3: Data collection timeline 85

Figure III.4: Post hoc test results for different single data source models. 90

Figure III.5: Post hoc test results for algorithms with different numbers of data sources. 91

Figure III. 6: Post hoc test results to identify the best second data source to combine with RBD. 92

Figure III.7: Friedman test results to identify the best third data source to combine with RBD and PD 92

Figure III.8: Friedman test results comparing feature combination (FM) and a posteriori weighting for recommendation models with four data sources. 92

Figure III. 9: Aggregated data source importance scores. 93

Figure III.10: Feature importance scores. 94

Figure IV.1: Framework identifying the effect of traditional – and revenue maximization recommendation systems on business metrics throughout the purchase funnel. 106

Figure IV.A.1: Example Email. 129